



**Εθνικό Μετσόβιο Πολυτεχνείο**  
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών  
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

---

**Αξιοποίηση Μοντέλων Διάχυσης για την Παραγωγή  
Τρισδιάστατων Κινούμενων Σχεδίων Σώματος Καθοδηγούμενη  
από Ομιλία και Συναίσθημα**

---

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Δημουλά Ιωάννη

**Επιβλέπων:** Αθανάσιος Βουλόδημος  
Επίκουρος Καθηγητής Ε.Μ.Π.

**Συνεπιβλέπουσα:** Παρασκευή Τζούβελη  
Μέλος Ε.ΔΙ.Π. Ε.Μ.Π.

Αθήνα, Οκτώβριος 2025







## Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών  
Εργαστήριο Συστημάτων Τεχνητής Νοημοσύνης και Μάθησης

---

### Αξιοποίηση Μοντέλων Διάχυσης για την Παραγωγή Τρισδιάστατων Κινούμενων Σχεδίων Σώματος Καθοδηγούμενη από Ομιλία και Συναίσθημα

---

#### ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Δημουλά Ιωάννη

Επιβλέπων: Αθανάσιος Βουλόδημος  
Επίκουρος Καθηγητής Ε.Μ.Π.

Συνεπιβλέπουσα: Παρασκευή Τζούβελη  
Μέλος Ε.ΔΙ.Π. Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 30<sup>η</sup> Οκτώβρη 2025.

.....  
Αθανάσιος Βουλόδημος  
Επίκουρος Καθηγητής  
Ε.Μ.Π.

.....  
Ανδρέας-Γεώργιος  
Σταφυλοπάτης  
Καθηγητής  
Ε.Μ.Π.

.....  
Γεώργιος Στάμου  
Καθηγητής  
Ε.Μ.Π.

Αθήνα, Οκτώβριος 2025

.....

**Δημουλάς Ιωάννης**

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

© Δημουλάς Ιωάννης, 2025. Με επιφύλαξη παντός δικαιώματος. All rights reserved.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

## Περίληψη

---

Η παραγωγή τρισδιάστατων κινούμενων σχεδίων του ανθρώπινου σώματος συνιστά έναν ερευνητικό χώρο που καθίσταται ολοένα και πιο επίκαιρος σε έναν κόσμο αυξανόμενης συνδεσιμότητας και εικονικής αλληλεπίδρασης. Ειδικά η εξάρτησή της από την ομιλία και το συνοδεύον συναίσθημα παρέχει της κατάλληλες σημασιολογικές και ρυθμικές πληροφορίες για την δημιουργία πιο ρεαλιστικών και διαδραστικών έργων. Η αξιοποίηση των μοντέλων διάχυσης για τον σκοπό αυτό έχει επιδείξει, αναμφίβολα, εντυπωσιακά αποτελέσματα, οδηγώντας στην ανάπτυξη εξελιγμένων εργαλείων. Ωστόσο, παρά την εν λόγω πρόοδο, ένα από τα βασικά προβλήματα που παραμένουν, αποτελεί το “φαινόμενο της απόκοσμης κοιλάδας”, όπου η σχεδόν ρεαλιστική, αλλά όχι πλήρως φυσική κίνηση, προκαλεί αισθήματα δυσaréσκειας ή ανησυχίας στον θεατή.

Για την αντιμετώπιση του συγκεκριμένου προβλήματος, προτείνουμε μια εναλλακτική προσέγγιση βασισμένη στην χρήση χαρακτήρων καρτούν—και δη χαρακτήρων anime. Υποθέτουμε ότι οι έντονα στυλιζαρισμένες φιγούρες δύναται να επιτρέψουν στα μοντέλα να εστιάσουν περισσότερο στην εκφραστικότητα και συναισθηματική συνέπεια, παρά στην φωτορεαλιστική απόδοση. Στο πλαίσιο αυτό, δημιουργήσαμε το πρώτο *speech-to-motion* σύνολο δεδομένων από σύντομα βίντεο κλιπ anime που απαρτίζεται από τον ακατέργαστο ήχο και την ακολουθία της κίνησης.

Το νέο αυτό σύνολο δεδομένων αξιοποιήθηκε σε δύο κύρια πειράματα τα οποία έγκεινται στην εξ’ αρχής εκπαίδευση, αφενός, και fine-tuning, αφετέρου, του βαθέος μοντέλου AMUSE. Σκοπός μας αποτελεί η ενσωμάτωση βασικών κινησιολογικών μοτίβων που εφαρμόζουν επαγγελματίες animators στην παραγωγική διαδικασία. Μέσω ποσοτικών και ποιοτικών αξιολογήσεων των αποτελεσμάτων μας διαπιστώθηκε ότι η χρήση δεδομένων anime παρέχει πράγματι την δυνατότητα βελτίωσης τόσο της χρονικής ευθυγράμμισης ομιλίας-κίνησης, όσο και της συνολικής φυσικότητας του παραγόμενου τρισδιάστατου animation.

Η παρούσα διπλωματική εργασία συνεισφέρει τόσο θεωρητικά, όσο και πρακτικά στην έρευνα της πολυτροπικής μάθησης, προτείνοντας μια νέα προσέγγιση στην δημιουργία 3D κινούμενων σχεδίων του ανθρώπινου σώματος, με εφαρμογές στην εικονική πραγματικότητα, ψηφιακούς χαρακτήρες και την βιομηχανία ψυχαγωγίας.

## Λέξεις-Κλειδιά

Μοντέλα Διάχυσης, 3D Animation Σώματος, Anime, Ομιλία, Συναίσθημα.



# Abstract

---

The production of 3D human body animation constitutes a research area that becomes progressively more relevant in a world of increasing connectivity and virtual interaction. Especially, its conditioning on speech and its accompanying emotion provides the appropriate semantic and rhythmic cues for the creation of more realistic and interactive motion sequences. The leveraging of diffusion models to this end has, undoubtedly, demonstrated impressive results, leading to the development of sophisticated generative tools. Nevertheless, despite the progress, one of the main problems that persists is that of the so-called “Uncanny Valley Effect”, where movements that are almost realistic, yet not entirely natural, may evoke emotions of discomfort or unease in the viewer.

To address this issue, we propose an alternative approach based on the usage of cartoon characters—predominantly anime ones. We hypothesize that such highly stylized figures may allow models to focus more effectively on expressiveness and emotional consistency, rather than photorealistic depiction. In this context, we constructed the first speech-to-motion dataset, composed from short anime video clips, that includes both the raw speech audio and the corresponding motion sequence.

We then employed this new dataset in two main experiments, involving the training from scratch and fine-tuning of the deep generative model *AMUSE*. Our goal is to incorporate fundamental motion patterns used by professional animators into the inference process. Through quantitative and qualitative evaluation, we found that the use of anime-based data indeed provides the potential to improve the temporal alignment between speech and motion, as well as the overall naturalness of the generated 3D animations.

This thesis contributes both theoretically and practically to the field of multimodal learning, proposing a novel approach for the creation of 3D human body animations with potential applications in virtual reality, digital avatars, and the entertainment industry.

## Keywords

Diffusion Models, 3D Body Animation, Anime, Speech, Emotion.



## Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κύριο Αθανάσιο Βουλόδημο, για την εμπιστοσύνη και την ευκαιρία που μου προσέφερε να εργαστώ στο Εργαστήριο Τεχνητής Νοημοσύνης και Συστημάτων Μάθησης και να εξερευνήσω ένα τόσο ενδιαφέρον πεδίο που συνδυάζει δημιουργικότητα και τεχνολογία. Παράλληλα, θα ήθελα να ευχαριστήσω την κυρία Παρασκευή Τζούβελη για την θερμή και εγκάρδια υποστήριξη που μου παρείχε στην διάρκεια της εκπόνησης της διπλωματικής μου εργασίας. Ασφαλώς, με ειλικρινή ευγνωμοσύνη, απευθύνω τις ευχαριστίες μου στον Ηλία Μήτσουρα για την άριστη συνεργασία που είχαμε, την υποστήριξή του, και την προθυμία του να βοηθήσει σε κάθε δυσκολία που αντιμετώπιζα, με όποιον τρόπο ήταν δυνατό για τον ίδιο. Τέλος, δεν θα μπορούσα να παραλείψω όλους τους συγγενείς και φίλους μου, και ιδιαίτερα τους γονείς μου και τα αδέρφια μου, που υπήρξαν πραγματικά συνοδοιπόροι μου σε αυτό μου το ταξίδι, και που δίχως την συμπαράστασή τους, δεν θα βρισκόμουν στο σημείο που βρίσκομαι τώρα, ως άνθρωπος και ως ερευνητής.

Η παρούσα διπλωματική εργασία σηματοδοτεί το πέρας μιας μακράς και επίπονης πορείας για την ολοκλήρωση των προπτυχιακών μου σπουδών και την απόκτηση του τίτλου του Διπλωματούχου Ηλεκτρολόγου Μηχανικού και Μηχανικού Υπολογιστών του Ε.Μ.Π. Οι γνώσεις και οι αναμνήσεις που απέκτησα κατά την διάρκειά της είναι αναντικατάστατες. Ωστόσο, η εργασία αυτή δεν αποτελεί το τέλος, αλλά εφαλτήριο για νέες αναζητήσεις και και πιο ώριμους στόχους.





# Περιεχόμενα

---

|          |                                                    |           |
|----------|----------------------------------------------------|-----------|
| <b>1</b> | <b>Εισαγωγή</b>                                    | <b>1</b>  |
| 1.1      | Εισαγωγή . . . . .                                 | 1         |
| 1.2      | Διάρθρωση Διπλωματικής Εργασίας . . . . .          | 2         |
| <b>2</b> | <b>Θεωρητικό Υπόβαθρο</b>                          | <b>5</b>  |
| 2.1      | Denoising Diffusion Probabilistic Models . . . . . | 6         |
| 2.2      | Score-Based Generative Models. . . . .             | 13        |
| 2.2.1    | Score Matching . . . . .                           | 15        |
| 2.2.2    | Sliced Score Matching . . . . .                    | 18        |
| 2.2.3    | Denoising Score Matching . . . . .                 | 20        |
| 2.2.4    | Langevin Dynamics . . . . .                        | 21        |
| 2.3      | Stochastic Differential Equations . . . . .        | 22        |
| 2.4      | Denoising Diffusion Implicit Models . . . . .      | 27        |
| <b>3</b> | <b>Σχετική Έρευνα</b>                              | <b>33</b> |
| 3.1      | Rule-Based Προσεγγίσεις . . . . .                  | 34        |
| 3.2      | Data-Based Προσεγγίσεις . . . . .                  | 35        |
| <b>4</b> | <b>Περιγραφή του Προβλήματος</b>                   | <b>37</b> |
| 4.1      | Αντίληψη του Προβλήματος . . . . .                 | 37        |
| 4.1.1    | Το Φαινόμενο της Απόκοσμης Κοιλιάδας . . . . .     | 37        |
| 4.1.2    | Οι Δώδεκα Αρχές των Κινουμένων Σχεδίων . . . . .   | 39        |

|       |                                                     |           |
|-------|-----------------------------------------------------|-----------|
| 4.2   | Ερευνητικοί Στόχοι . . . . .                        | 40        |
| 4.2.1 | Διατύπωση του Προβλήματος . . . . .                 | 41        |
| 5     | <b>Κατασκευή Συνόλου Δεδομένων Εκπαίδευσης</b>      | <b>45</b> |
| 5.1   | Δομή και Περιεχόμενα Συνόλου Δεδομένων . . . . .    | 45        |
| 5.2   | Μεθοδολογία . . . . .                               | 49        |
| 5.3   | Επιλογή Αρχικού Συνόλου Anime . . . . .             | 50        |
| 5.4   | Φιλτράρισμα Βίντεο Κλιπς . . . . .                  | 52        |
| 5.5   | Εκτίμηση Τρισδιάστατου Πλέγματος . . . . .          | 54        |
| 5.5.1 | Δισδιάστατη Εκτίμηση Στάσης . . . . .               | 54        |
| 5.5.2 | Ανύψωση στην Τρίτη Διάσταση . . . . .               | 56        |
| 5.5.3 | Εξαγωγή Επιπρόσθετων Keypoints . . . . .            | 58        |
| 5.5.4 | Υπολογισμός 6D Αναπαραστάσεων Περιστροφής . . . . . | 61        |
| 5.6   | Λεπτομέρειες Εκτέλεσης και Αποτελέσματα . . . . .   | 69        |
| 6     | <b>Εκπαίδευση</b>                                   | <b>73</b> |
| 6.1   | Η Αρχιτεκτονική του <i>AMUSE</i> . . . . .          | 74        |
| 6.1.1 | Μπλοκ Αποσύνθεσης Ήχου . . . . .                    | 75        |
| 6.1.2 | Μπλοκ Παραγωγής Κινησιολογίας . . . . .             | 78        |
| 6.2   | Προεπεξεργασία Δεδομένων . . . . .                  | 82        |
| 6.3   | Εκπαίδευση από την Αρχή . . . . .                   | 84        |
| 6.3.1 | Λεπτομέρειες Εκπαίδευσης . . . . .                  | 84        |
| 6.3.2 | Αποτελέσματα . . . . .                              | 84        |
| 6.3.3 | Συμπεράσματα . . . . .                              | 84        |
| 6.4   | Εκπαίδευση Fine-Tuning . . . . .                    | 86        |
| 6.4.1 | Λεπτομέρειες Εκπαίδευσης . . . . .                  | 86        |
| 6.4.2 | Αποτελέσματα . . . . .                              | 87        |
| 6.4.3 | Συμπεράσματα . . . . .                              | 87        |

---

|                                              |            |
|----------------------------------------------|------------|
| 6.5 Ποιοτική Σύγκριση . . . . .              | 89         |
| 6.6 Ποσοτική Σύγκριση . . . . .              | 90         |
| <b>7 Ανακεφαλαίωση και Μελλοντική Έρευνα</b> | <b>95</b>  |
| <b>A Επιπλέον Δείγματα Συνόλου Δεδομένων</b> | <b>97</b>  |
| <b>B Επιπλέον Παραδείγματα Σύνθεσης</b>      | <b>101</b> |
| <b>Βιβλιογραφία</b>                          | <b>113</b> |



## Κατάλογος Σχημάτων

---

|     |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           |    |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.1 | Διαισθητική απεικόνιση της διαδικασίας εκπαίδευσης ενός DDPM [10]                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | 6  |
| 2.2 | Απεικονίσεις της έννοιας του Stein Score Function. <b>Αριστερά</b> διακρίνεται το score της πραγματικής συνάρτησης πυκνότητας πιθανότητας $\nabla_{\mathbf{x}} \log q(\mathbf{x})$ , και <b>δεξιά</b> το εκτιμημένο score $s_{\theta}(\mathbf{x})$ . Η συνάρτηση πυκνότητας πιθανότητας των δεδομένων $q(\mathbf{x})$ κωδικοποιείται μέσω του χρωματικού χάρτη: περιοχές σκουρότερου χρώματος υποδηλώνουν περιοχές υψηλότερης πυκνότητας. Τα κόκκινα διακεκομμένα πλαίσια επισημαίνουν περιοχές όπου $s_{\theta}(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log q(\mathbf{x})$ [7]. . . . . | 15 |
| 2.3 | Η αρχή λειτουργίας των score-based παραγωγικών μοντέλων μέσω στοχαστικών διαφορικών εξισώσεων (score SDEs) [23]. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                  | 22 |
| 2.4 | Γραφική αναπαράσταση μιας μαρκοβιανής δειγματοληπτικής διαδικασίας, όπως των DDPMs (αριστερά), και μιας μη-μαρκοβιανής δειγματοληπτικής διαδικασίας, όπως των DDIMs (δεξιά) [25]. . . . .                                                                                                                                                                                                                                                                                                                                                                                                 | 28 |
| 4.1 | Παραχθέντα δείγματα του μοντέλου <i>DiffSHEG</i> [27]. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            | 37 |
| 4.2 | Το αρχικό διάγραμμα του <i>Masahiro Mori</i> [43], που απεικονίζει το φαινόμενο της απόκοσμης κοιλάδας (uncanny valley). Όσο αυξάνεται ο ρεαλισμός ενός αντικειμένου, αυξάνεται και η συμπάθεια του παρατηρητή προς αυτό. Ωστόσο, η ανοδική τάση διακόπτεται καθώς το αντικείμενο προσεγγίζει σε ρεαλισμό τον άνθρωπο· οπότε και η συμπάθεια ελαττώνεται απότομα. Επιπλέον, αυτό το φαινόμενο εντείνεται περαιτέρω στην περίπτωση των κινούμενων αντικειμένων, όπως υποδεικνύει η διακεκομμένη γραμμή. . . . .                                                                            | 38 |
| 4.3 | Οι δώδεκα αρχές της κινούμενης εικόνας. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | 40 |
| 5.1 | Η διαδικασία ανακατασκευής του τρισδιάστατου πλέγματος του ανθρώπινου σώματος. Κατ' αρχάς, εκτελείται <i>2D pose estimation</i> , ακολουθούμενο από <i>3D lifting</i> . Τέλος γίνεται <i>fitting</i> του τρισδιάστατου σκελετού στο μοντέλο <i>SMPL-X</i> [54]. . . . .                                                                                                                                                                                                                                                                                                                   | 51 |
| 5.2 | Κατηγορίες (Genres) και Θέματα (Themes) παρόντα στο σύνολο δεδομένων <i>Anim-400K</i> [85]. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 52 |
| 5.3 | Η διαδικασία φιλτραρίσματος. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      | 55 |
| 5.4 | Οι σκελετοί και η αντίστοιχη απαρίθμηση των αρθρώσεων κατά <i>COCO</i> [95] (αριστερά) και κατά <i>Human3.6M</i> [96] (δεξιά). . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                    | 57 |
| 5.5 | Οι 21 αρθρώσεις του χεριού που παράγει το <i>MediaPipe Hands</i> [101]. . . . .                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                           | 60 |

|     |                                                                                                                                                                                                                                                                                                                                                                                                             |     |
|-----|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 5.6 | Ασάφεια Στρέψης: Διαφορετικές στάσεις δύναται να οριστούν με τις ίδιες σχετικές θέσεις αρθρώσεων [107]. . . . .                                                                                                                                                                                                                                                                                             | 66  |
| 5.7 | Σχηματική αναπαράσταση μιας διάταξης δύο αρθρώσεων κι ενός άκρου. Η αριστερή σφαίρα συνιστά την άρθρωση-γονέα και η δεξιά την άρθρωση-παιδί. Η περιστροφή της γονικής άρθρωσης αναλύεται σε περιστροφές ταλάντωσης ( <i>swing</i> ) και στρέψης ( <i>twist</i> ) [107]. . . . .                                                                                                                             | 66  |
| 5.8 | Η διαδικασία εκτίμησης του τρισδιάστατου πλέγματος. . . . .                                                                                                                                                                                                                                                                                                                                                 | 70  |
| 5.9 | Ενδεικτικά δείγματα (μεμονωμένα frames) του συνόλου δεδομένων μας. Για κάθε ζεύγος, η αριστερή εικόνα είναι το αρχικό frame και η δεξιά το τρισδιάστατο πλέγμα που εκτιμήθηκε με την διαδικασία της ενότητας 5.5. . . . .                                                                                                                                                                                   | 71  |
| 6.1 | Σχηματική αναπαράσταση της αρχιτεκτονικής του <i>AMUSE</i> [1]. . . . .                                                                                                                                                                                                                                                                                                                                     | 74  |
| 6.2 | Σχηματική αναπαράσταση της αρχιτεκτονικής του Μπλοκ Αποσύνθεσης Ήχου του <i>AMUSE</i> [1]. . . . .                                                                                                                                                                                                                                                                                                          | 78  |
| 6.3 | Σχηματική αναπαράσταση του Motion Prior Μπλοκ [1]. . . . .                                                                                                                                                                                                                                                                                                                                                  | 79  |
| 6.4 | Σχηματική αναπαράσταση της αρχής λειτουργίας του <i>conditional latent diffusion model</i> του <i>AMUSE</i> [1]. Κατά την προς τα εμπρός διαδικασία διάχυσης (δεξιά προς τα αριστερά), λαμβάνεται η θορυβώδης λανθάνουσα αναπαράσταση της κίνησης, ενώ κατά την αντίστροφη διαδικασία αποθορυβοποίησης (αριστερά προς τα δεξιά) εξάγεται το εξαρτώμενο αποθορυβοποιημένο λανθάνον διάνυσμα κίνησης. . . . . | 80  |
| 6.5 | Σχηματική αναπαράσταση της διαδικασίας παραγωγής νέων δειγμάτων του <i>AMUSE</i> [1]. . . . .                                                                                                                                                                                                                                                                                                               | 82  |
| 6.6 | Δείγματα που παρήχθησαν από το <i>from scratch</i> εκπαιδευμένο μοντέλο. Κάθε σειρά αντιπροσωπεύει ένα διαφορετικό δείγμα και απεικονίζει μια ακολουθία ενδεικτικών frames συνοδευόμενη από μια λεζάντα του περιεχομένου της ομιλίας. . . . .                                                                                                                                                               | 85  |
| 6.7 | Δείγματα που παρήχθησαν από το <i>fine-tuned</i> μοντέλο. . . . .                                                                                                                                                                                                                                                                                                                                           | 88  |
| 6.8 | Άμεση σύγκριση των αποτελεσμάτων του fine-tuned (πάνω) και προεκπαιδευμένου (κάτω) μοντέλου υπό την ίδια είσοδο. . . . .                                                                                                                                                                                                                                                                                    | 90  |
| 6.9 | Σχηματική αναπαράσταση της αρχιτεκτονικής του <i>Autoencoder Εξαγωγής Χαρακτηριστικών</i> , βασισμένη στους <i>Yoon et al.</i> [34]. Το BN συμβολίζει <i>Batch Normalization</i> , το FC ένα <i>fully connected layer</i> , και το ConvTr ένα επίπεδο <i>transpose convolution</i> . . . . .                                                                                                                | 92  |
| A.1 | Επιπλέον δείγματα του συνόλου δεδομένων εκπαίδευσης. . . . .                                                                                                                                                                                                                                                                                                                                                | 97  |
| A.2 | Επιπλέον δείγματα του συνόλου δεδομένων εκπαίδευσης (συνέχεια). . . . .                                                                                                                                                                                                                                                                                                                                     | 98  |
| A.3 | Επιπλέον δείγματα του συνόλου δεδομένων εκπαίδευσης (συνέχεια). . . . .                                                                                                                                                                                                                                                                                                                                     | 99  |
| A.4 | Επιπλέον δείγματα του συνόλου δεδομένων εκπαίδευσης (συνέχεια). . . . .                                                                                                                                                                                                                                                                                                                                     | 100 |
| B.1 | Επιπλέον παραδείγματα σύνθεσης. . . . .                                                                                                                                                                                                                                                                                                                                                                     | 101 |
| B.2 | Επιπλέον παραδείγματα σύνθεσης (συνέχεια). . . . .                                                                                                                                                                                                                                                                                                                                                          | 102 |
| B.3 | Επιπλέον παραδείγματα σύνθεσης (συνέχεια). . . . .                                                                                                                                                                                                                                                                                                                                                          | 103 |
| B.4 | Επιπλέον παραδείγματα σύνθεσης (συνέχεια). . . . .                                                                                                                                                                                                                                                                                                                                                          | 104 |

## Κατάλογος Πινάκων

---

|     |                                                                                                                                                                                                                                                                                                                                                              |    |
|-----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 5.1 | Σύγκριση των χαρακτηριστικών διαφορών speech-to-motion συνόλων δεδομένων . . . . .                                                                                                                                                                                                                                                                           | 49 |
| 5.2 | Οι 55 Αρθρώσεις του <i>SMPL-X</i> : 21 για την στάση του σώματος· 30 για τα δάκτυλα των χεριών· 3 για το πρόσωπο· και μία ( <i>pelvis</i> ) για τον καθολικό προσανατολισμό του μοντέλου. . . . .                                                                                                                                                            | 59 |
| 6.1 | Τιμές των επιμέρους ορών απώλειας του βέλτιστου μοντέλου εκπαιδευμένου από την αρχή. . . . .                                                                                                                                                                                                                                                                 | 84 |
| 6.2 | Τιμές των επιμέρους ορών απώλειας του βέλτιστου fine-tuned μοντέλου. . . . .                                                                                                                                                                                                                                                                                 | 87 |
| 6.3 | Υπερπαραμέτροι για το προεκπαιδευμένο μοντέλο των <i>Chhatre et al.</i> [1], το από την αρχή εκπαιδευμένο μοντέλο, και το fine-tuned μοντέλο. . . . .                                                                                                                                                                                                        | 87 |
| 6.4 | Τιμές των επιμέρους ορών απώλειας του μοντέλου των <i>Chhatre et al.</i> [1] που εκπαιδεύτηκε για 6000 εποχές στο στο σύνολο δεδομένων <i>BEAT</i> [45]. . . . .                                                                                                                                                                                             | 89 |
| 6.5 | Αποτελέσματα της ποσοτικής σύγκρισης του προεκπαιδευμένου, from-scratch, και fine-tuned μοντέλου. Για τις μετρικές Div και BA, όσο μεγαλύτερη είναι η τιμή τους, τόσο καλύτερα τα αποτελέσματα, ενώ για το FGD ισχύει το αντίστροφο. Με πράσινο χρώμα έχουμε σημειώσει το εκάστοτε μοντέλο με την καλύτερη επίδοση, ενώ με μπλε το δεύτερο καλύτερο. . . . . | 93 |





# Κεφάλαιο 1

## Εισαγωγή

---

### 1.1 Εισαγωγή

Η παραγωγή τρισδιάστατων κινουμένων σχεδίων του ανθρώπινου σώματος αποτελεί έναν ερευνητικό τομέα ο οποίος γνωρίζει ραγδαία ανάπτυξη. Αυτό καθίσταται ιδιαίτερα αληθές τα τελευταία χρόνια, καθώς διατρέχουμε μια εποχή η οποία έχει χαρακτηριστεί, επανειλημμένως, ως ο “Αιών της Πληροφορίας”. Ουδείς δύναται να αμφισβητήσει ότι με την πάροδο του χρόνου, ο κόσμος γίνεται ολοένα και πιο συνδεδεμένος, διαδραστικός, αλλά και εξοικειωμένος με νέες τεχνολογίες. Σε αυτό, ακριβώς, το συγκείμενο έγκειται η ανάγκη αποδοτικής και αυτόματης παραγωγής 3D body animations, αφού απαντώνται σε μια ευρεία πληθώρα διαφορετικών εφαρμογών, συμπεριλαμβανομένων, μεταξύ άλλων, της τηλεπαρουσίας σε συστήματα εικονικής και επαυξημένης πραγματικότητας, τον σχεδιασμό χαρακτήρων σε ταινίες ή βιντεοπαιχνίδια, και της “ενσάρκωσης” διαδραστικών ψηφιακών πρακτόρων.

Ωστόσο, το συγκεκριμένο πρόβλημα δεν είναι διόλου απλοϊκό. Η κίνηση του ανθρώπου είναι εξαιρετικά περίπλοκη και εξαρτάται από έναν μεγάλο αριθμό διαφορετικών παραγόντων. Ενδεικτικά, η εκάστοτε δραστηριότητα στην οποία εστιάζει, το περιεχόμενο της ομιλίας του, η συναισθηματική του και σωματική του κατάσταση, καθώς και το προσωπικό του στυλ είναι μονάχα ορισμένοι από τους εν λόγω παράγοντες που επηρεάζουν την γλώσσα του σώματος του κάθε ανθρώπου. Καθίσταται, συνεπώς, σαφές ότι η παραγωγή ανθρώπινης κινησιολογίας αποκλειστικά με δεδομένο κάποιο κείμενο δεν είναι αρκετή. Η ομιλία, και το υποκείμενο συναίσθημα που το συνοδεύει, αν και ασφαλώς δεν αποτυπώνουν το σύνολο των απαραίτητων παραμέτρων, εντούτοις, δύναται να παράσχουν σημαντικές ρυθμικές και σημασιολογικές πληροφορίες που να καταστήσουν τις παραγόμενες κινήσεις πολύ πιο εκφραστικές, ρεαλιστικές, αλλά και προσίτες στον θεατή.

Τα τελευταία χρόνια, η πρόοδος των μεθόδων βαθιάς μάθησης έχει επιτρέψει την ανάπτυξη προχωρημένων αλγορίθμων που επιχειρούν να προσομοιώσουν την ανθρώπινη κίνηση με ρεαλιστικό και εκφραστικό τρόπο. Μεταξύ αυτών, τα μοντέλα διάχυσης (*diffusion models*) έχουν αναδειχθεί ως μια από τις πιο ισχυρές και ευέλικτες κατηγο-

ρίες γενετικών μοντέλων, ξεπερνώντας προηγούμενες προσεγγίσεις όπως τα GANs και τα VAEs σε ποικιλία εφαρμογών, από τη σύνθεση εικόνων και βίντεο, έως και τη δημιουργία ήχου και, κριτικά για την προκειμένη περίπτωση, κίνησης. Στο πλαίσιο, αυτό μια ευρεία πληθώρα διαφορετικών μοντέλων, που επιστρατεύουν διαφορετικές τεχνικές για την εξαρτώμενη από ήχο παραγωγή κινουμένων σχεδίων, έχει ήδη καθιερωθεί στην σχετική βιβλιογραφία.

Ωστόσο, η εκπαίδευση των εν λόγω μοντέλων σε σύνολα, που απαρτίζονται από δεδομένα πραγματικών ανθρώπων, οδηγεί, πολύ συχνά, σε αποτελέσματα που καλλιεργούν κάποιο περίεργο συναίσθημα δυσφορίας στον θεατή. Το φαινόμενο αυτό, γνωστό ως “το φαινόμενο της απόκοσμης κοιλάδας”, προκαλείται στην περίπτωση κάποιου αντικειμένου που είναι αρκετά ρεαλιστικό και ανθρωπόμορφο, αλλά όχι ακριβώς. Υποθέτουμε ότι η εκπαίδευση ενός τέτοιου μοντέλου σε έντονα στυλιζαρισμένες φιγούρες έχει δυνατότητα να αναιρέσει αυτό το πρόβλημα.

Στα πλαίσια, συνεπώς, της παρούσης διπλωματικής εργασίας, θα συνθέσουμε το πρώτο, από ό,τι γνωρίζουμε, *speech-to-motion* σύνολο δεδομένων εκπαίδευσης από σύντομα βίντεο κλιπς καρτούν—και πιο συγκεκριμένα, από βίντεο κλιπς Anime, μίας μορφής κινούμενης τέχνης με προέλευση από την Ιαπωνία που έχουν, όμως, εξαπλωθεί ραγδαία και στον δυτικό κόσμο. Εν συνεχεία, θα χρησιμοποιήσουμε το καινούριο αυτό *dataset* για να διεξάγουμε μια σειρά πειραμάτων που έγκεινται στην εκπαίδευση από την αρχή και τύπου *fine-tuning* του *State-of-the-Art* μοντέλου *AMUSE* [1], που κάνει χρήση διάχυσης για την κατευθυνόμενη από ομιλία και συναίσθημα παραγωγή τρισδιάστατων σχεδίων του ανθρώπινου σώματος. Μέσα από αυτά τα πειράματα αποσκοπούμε να εξερευνήσουμε την πιθανότητα ενσωμάτωσης βασικών κινησιολογικών μοτίβων, που χρησιμοποιούνται από επαγγελματίες animators, στην παραγωγική διαδικασία, ώστε να μετριαστεί το προαναφερθέν φαινόμενο της “απόκοσμης κοιλάδας”.

## 1.2 Διάρθρωση Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία διαρθρώνεται σε 7 κεφάλαια, εκ των οποίων το παρόν Κεφάλαιο 1 συνιστά μια πρώτη γενική εισαγωγή στο εξεταζόμενο πρόβλημα και μια σύντομη περιγραφή της προσέγγισης που θα ακολουθήσουμε για την αντιμετώπισή του. Τα υπόλοιπα κεφάλαια δύνανται, σε γενικές γραμμές, να διακριθούν σε δύο μέρη που έγκεινται στην καθιέρωση του απαραίτητου θεωρητικού υποβάθρου και εξερεύνηση της υπάρχουσας σχετικής έρευνας για την βαθύτερη κατανόηση του προβλήματος· και στην περιγραφή της ακριβούς υλοποίησης της προτεινόμενης μεθόδου, των πειραμάτων που διεξαγάγαμε, των αποτελεσμάτων αυτών, και, τελικά, των εξαγόμενων συμπερασμάτων. Πιο αναλυτικά, σε κάθε κεφάλαιο εξετάζονται τα εξής:

- **Κεφάλαιο 2:** Στο κεφάλαιο αυτό παρουσιάζεται το θεωρητικό υπόβαθρο των μοντέλων διάχυσης. Παρουσιάζονται οι διαφορετικές μορφές με τις οποίες αυτά εμφανίζονται στην βιβλιογραφία και περιγράφονται αναλυτικώς οι βασικές αρχές λειτουργίας τους.
- **Κεφάλαιο 3:** Στο κεφάλαιο αυτό γίνεται αναφορά στην υφιστάμενη βιβλιογραφία που πραγματεύεται την παραγωγή τρισδιάστατων κινουμένων σχεδίων του

ανθρώπινου σώματος. Εξηγούνται οι διαφορετικές κατηγορίες στις οποίες μπορούν να υπαχθούν οι διαφορετικές προσεγγίσεις, και περιγράφονται οι συνεισφορές καθιερωμένων μεθόδων.

- **Κεφάλαιο 4:** Στο κεφάλαιο αυτό γίνεται η επίσημη διατύπωση του προβλήματος, καθώς αυτό προκύπτει από την υπάρχουσα βιβλιογραφία, και περιγράφονται ρητώς οι υποθέσεις που καλούμαστε να εξετάσουμε και οι τελικοί στόχοι της παρούσης διπλωματικής εργασίας.
- **Κεφάλαιο 5:** Στο κεφάλαιο αυτό πραγματοποιείται αναλυτική περιγραφή της διαδικασίας παραγωγής του πρώτου *speech-to-motion* συνόλου δεδομένων εκπαίδευσης από σύντομα βίντεο κλιπς anime.
- **Κεφάλαιο 6:** Στο κεφάλαιο αυτό παρουσιάζονται οι πειραματικές μας διατάξεις, τα αποτελέσματα αυτών, και τα συμπεράσματά μας, που αφορούν την εκπαίδευση ενός βαθέος μοντέλου διάχυσης με την χρήση του *speech-to-motion* συνόλου δεδομένων μας.
- **Κεφάλαιο 7:** Στο κεφάλαιο αυτό γίνεται μια σύντομη ανακεφαλαίωση των συνεισφορών της παρούσης διπλωματικής εργασίας, καθώς και μιας σύντομης αναφοράς στη μελλοντική έρευνα και σε τρόπους επέκτασης του προτεινόμενου πλαισίου.



## Κεφάλαιο 2

### Θεωρητικό Υπόβαθρο

---

Τα βαθιά παραγωγικά μοντέλα (Deep Generative Models) κατέχουν εξέχουσα θέση στον χώρο της μηχανικής μάθησης. *Παραγωγικά Αντιπαραθετικά Δίκτυα* (Generative Adversarial Networks, GANs) [2, 3], *Αυτοπαλίνδρομα Μοντέλα* (Autoregressive Models) [4], *Ροές Κανονικοποίησης* (Normalizing Flows) [5] και *Μεταβλητοί Αυτοκωδικοποιητές* (Variational Autoencoders, VAEs) [6] έχουν αναδείξει, τα τελευταία χρόνια, δείγματα υψηλής ποιότητας σε μία ευρεία ποικιλία εφαρμογών, συμπεριλαμβανομένων της σύνθεσης εικόνων αυξημένης πιστότητας και της παραγωγής αποσπασμάτων ανθρώπινης ομιλίας, μουσικής, και άλλου είδους δεδομένων [7, 8, 9].

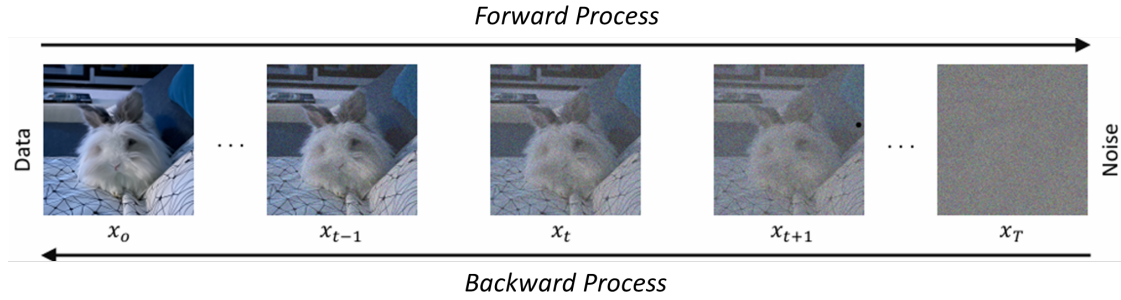
Τα *μοντέλα διάχυσης* (Diffusion Models) αποτελούν μια επιπλέον, πιθανοτική, κατηγορία των προαναφερθέντων παραγωγικών μοντέλων που, πρόσφατα, έχουν καθιερωθεί ως η κυρίαρχη προσέγγιση στον χώρο της συνθετικής γεννήτριας δεδομένων καταλαμβάνοντας μεγάλο μέρος της σχετικής έρευνας. Έως την συγγραφή της παρούσης εργασίας, τα diffusion models έχουν εφαρμοστεί και διακριθεί σε ένα ευρύ φάσμα παραγωγικών διαδικασιών, όπως στην δημιουργία, διευκρίνιση, συμπλήρωση και επεξεργασία εικόνων και βίντεο. Την ίδια στιγμή, η λανθάνουσα αναπαράσταση δεδομένων που μαθαίνεται από τα μοντέλα διάχυσης έχει αποδειχθεί χρήσιμη και σε διακριτικές διαδικασίες, όπως στην τμηματοποίηση και ταξινόμηση εικόνων, και στον εντοπισμό ανωμαλιών, μεταξύ άλλων. Τα παραπάνω επιβεβαιώνουν την ευρεία εφαρμοσιμότητα και ευελιξία των μοντέλων διάχυσης και δικαιολογούν την περίοπτη θέση που κατέχουν στην πρόσφατη βιβλιογραφία [10, 11].

Σε αυτό το κεφάλαιο εξετάζονται οι θεωρητικές βάσεις των μοντέλων διάχυσης, η αρχή λειτουργίας των οποίων βασίζεται σε μια διαδικασία εισαγωγής θορύβου και της επακόλουθης αντίστροφης διαδικασίας αποθορυβοποίησης που τους δίδει την ικανότητα μάθησης της υποκείμενης κατανομής δεδομένων. Κατ' ουσίαν, η διαδικασία εκπαίδευσης αποτελείται από δύο φάσεις: μια διαδικασία προώθησης της διάχυσης, στην οποία τα δεδομένα καταστρέφονται σταδιακά με την εισαγωγή θορύβου, και μια διαδικασία οπισθοδρόμησης, στην οποία ο θόρυβος αφαιρείται διαδοχικά. Η τρέχουσα έρευνα επικεντρώνεται ως επί το πλείστον σε τρεις διακριτές κατηγορίες των μοντέλων διάχυσης: τα *Denoising Diffusion Probabilistic Models* (DDPMs), τα *Score-Based Generative Models* (SGMs), και τα *Stochastic Differential Equations* (score SDEs) [10, 11,

12], τα οποία και αναλύονται διεξοδικά παρακάτω.

## 2.1 Denoising Diffusion Probabilistic Models

Τα *Denoising Diffusion Probabilistic Models* (DDPMs) [8, 13] υλοποιούν τις δύο προαναφερθείσες φάσεις της εκπαιδευτικής διαδικασίας μέσω της χρήσης δύο Μαρκοβιανών αλυσίδων: μία προς τα εμπρός αλυσίδα που διαταράσσει τα δεδομένα εγχέοντάς τα προοδευτικά με θόρυβο—συνήθως Γκαουσιανό—κατά την διάρκεια πολλαπλών διακριτών χρονικών σταδίων, και μία, αντίστροφη, προς τα πίσω αλυσίδα που επιχειρεί να μετατρέψει τον θόρυβο εκ νέου σε δεδομένα. Η τελευταία επιτυγχάνει το σκοπό της μέσω της εκμάθησης πυρήνων μετάβασης που παραμετροποιούνται από βαθιά συνελκτικά νευρωνικά δίκτυα, συνήθως βασισμένα σε κάποια αρχιτεκτονική τύπου U-Net [14].



Σχήμα 2.1: Διαισθητική απεικόνιση της διαδικασίας εκπαίδευσης ενός DDPM [10]

Πιο αναλυτικά, ας υποθέσουμε  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$  μια κατανομή δεδομένων. Η προς τα εμπρός διαδικασία Markov ξεκινάει με αφετηρία τα αρχικά δεδομένα  $\mathbf{x}_0$  και παράγει μια ακολουθία τυχαίων μεταβλητών  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$  σε  $T$  βήματα διάχυσης. Το πέρασμα από το ένα βήμα στο επόμενο γίνεται μέσα από έναν πυρήνα μετάβασης (transition kernel)  $q(\mathbf{x}_t | \mathbf{x}_{t-1})$ , όπου  $t \in \{1, 2, \dots, T\}$ . Χρησιμοποιώντας τον κανόνα της αλυσίδας και την μαρκοβιανή ιδιότητα, δύναται να παραγοντοποιήσουμε την από κοινού κατανομή των  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$  δεδομένης της κατανομής  $\mathbf{x}_0$  ως εξής:

$$q(\mathbf{x}_{1:T} | \mathbf{x}_0) = \prod_{t=1}^T q(\mathbf{x}_t | \mathbf{x}_{t-1}) \quad (2.1)$$

Στα DDPMs, ο πυρήνας μετάβασης  $q(\mathbf{x}_t | \mathbf{x}_{t-1})$  κατασκευάζεται, συνήθως, με τέτοιο τρόπο που να μετατρέπει σταδιακά την αρχική κατανομή δεδομένων  $q(\mathbf{x}_0)$  σε μία εκ των προτέρων γνωστή κατανομή. Όπως αναφέρθηκε και προηγουμένως, μια συνήθης επιλογή για αυτόν τον σκοπό είναι ο Γκαουσιανός πυρήνας που δίδεται από την κάτωθι σχέση:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_t = \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \boldsymbol{\Sigma}_t = \beta_t \mathbf{I}) \quad (2.2)$$

όπου  $\beta_t \in (0, 1)$  είναι μια υπερπαραμέτρος που επιλέγεται κατά τον σχεδιασμό του μοντέλου και που αναπαριστά το σχήμα διακύμανσης θορύβου· ελέγχει, δηλαδή, την ποσότητα θορύβου που προστίθεται σε κάθε βήμα. Επιπλέον, στην εξίσωση (2.2), το  $\mathbf{I}$  συμβολίζει τον μοναδιαίο πίνακα, και το  $\mathcal{N}(\mathbf{x}; \mu, \sigma)$  είναι η κανονική κατανομή με μέση τιμή  $\mu$  και συνδιακύμανση  $\sigma$ .

Η (2.2), χρησιμοποιώντας την τεχνική της επαναπαραμετροποίησης, γράφεται:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (2.3)$$

Αν αντικαταστήσουμε, πλέον, στην (2.3), αναδρομικά, την τυχαία μεταβλητή  $\mathbf{x}_{t-1}$  με την προηγούμενή της στην μαρκοβιανή αλυσίδα έως ότου φθάσουμε στην  $\mathbf{x}_0$  μπορούμε, κατ' ουσίαν, να εκφράσουμε ολόκληρη την προς τα εμπρός διαδικασία διάχυσης συναρτήσει της αρχικής κατανομής δεδομένων:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon \quad (2.4)$$

$$= \mathcal{N}(\mathbf{x}_t; \mu_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0, \Sigma_t = (1 - \bar{\alpha}_t) \mathbf{I}) \quad (2.5)$$

όπου  $\alpha_t := 1 - \beta_t$  και  $\bar{\alpha}_t := \prod_{i=1}^t \alpha_i$ . Αξίζει, σε αυτό το σημείο, να παρατηρηθεί ότι για  $\bar{\alpha}_T \approx 0$ , η τυχαία μεταβλητή  $\mathbf{x}_T$  γίνεται σχεδόν Γκαουσιανή ως κατανομή:

$$q(\mathbf{x}_T) = \int q(\mathbf{x}_T | \mathbf{x}_0) q(\mathbf{x}_0) d\mathbf{x}_0 \approx \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I}) \quad (2.6)$$

Καθ' αυτόν τον τρόπο ολοκληρώνεται ο στόχος της προς τα εμπρός διαδικασίας εκπαίδευσης των μοντέλων διάχυσης η οποία, διαισθητικά, αποσκοπεί στην σταδιακή έγχυση των δεδομένων με θόρυβο έως την πλήρη καταστροφή κάθε πληροφορίας που φέρουν. Ακολουθεί η ανάστροφη διαδικασία η οποία θα επιχειρήσει να επαναφέρει την χαμμένη πληροφορία από τον καθαρό θόρυβο.

Πιο συγκεκριμένα, κατά την *προς τα πίσω διαδικασία*, θόρυβος αφαιρείται σταδιακά από το διάνυσμα θορύβου που παρήχθη από την εκ των προτέρων γνωστή κατανομή. Αυτό γίνεται μέσω μιας εκπαιδευσίμης αλυσίδας Markov που κινείται κατά αντίστροφη χρονική κατεύθυνση με προηγούμενως. Οπότε, η αντίστροφη αλυσίδα παραμετροποιείται, τώρα, από μία εκ των προτέρων γνωστή κατανομή,  $p(\mathbf{x}_T) \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ , και τον πυρήνα μετάβασης προς μάθηση,  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$ . Σε αντιστοιχία με τον προς τα εμπρός πυρήνα, ο εκπαιδευσίμος αντίστροφος πυρήνας μετάβασης λαμβάνει την ακόλουθη μορφή:

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)) \quad (2.7)$$

όπου η μέση τιμή  $\mu_\theta$  και η συνδιακύμανση  $\Sigma_\theta$  αποτελούν τις παραμέτρους που πρέπει να εκτιμηθούν μέσω της εκπαίδευσης βαθέων συνελκτικών νευρωνικών δικτύων. Γενικά, ο δείκτης  $\theta$  θα υποδηλώνει στο εξής παραμέτρους μοντέλου.

Κάνοντας κανείς χρήση της εν λόγω αντίστροφης μαρκοβιανής αλυσίδας, δύναται να παραγάγει ένα δείγμα δεδομένων,  $\mathbf{x}_0$ , δειγματοληπτώντας, αρχικά, ένα διάνυσμα

θορύβου,  $\mathbf{x}_T \sim p(\mathbf{x}_T)$ , και, κατόπιν, δειγματοληπτώντας, από τον εκπαιδευσιμο πυρήνα μετάβασης, διάνυσμα  $\mathbf{x}_{t-1} \sim p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$ , αναδρομικά, έως ότου  $t = 1$ . Σαφώς, απαραίτητη προϋπόθεση για την επιτυχία αυτής της δειγματοληπτικής διαδικασίας, αποτελεί η εκπαίδευση της αντίστροφης αλυσίδας έτσι ώστε να προσεγγίζει την πραγματική χρονική αναστροφή της προς τα εμπρός. Με άλλα λόγια, οφείλουμε να προσαρμόσουμε την παράμετρο  $\theta$  με τρόπο που η από κοινού κατανομή της αντίστροφης αλυσίδας:

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \quad (2.8)$$

να προσεγγίζει όσο το δυνατόν περισσότερο αυτήν της προς τα εμπρός αλυσίδας:

$$q(\mathbf{x}_{0:T}) := q(\mathbf{x}_0) \prod_{t=1}^T q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \quad (2.9)$$

Σε ένα ιδανικό σενάριο, αυτό θα υλοποιούταν με στόχο την *μέγιστη πιθανοφάνεια* (maximum likelihood)· την μεγιστοποίηση, δηλαδή, της πιθανότητας που αποδίδεται από το μοντέλο  $p_\theta(\mathbf{x}_0)$  σε κάθε δείγμα εκπαίδευσης  $\mathbf{x}_0$ . Αυτή δίδεται από την παρακάτω σχέση:

$$p_\theta(\mathbf{x}_0) = \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \quad (2.10)$$

Παρατηρούμε, εντούτοις, ότι το  $p_\theta(\mathbf{x}_0)$ , όπως δίδεται από την εξίσωση (2.10), είναι δυσεπίλυτο στην πράξη, λόγω της υψηλής διαστατικότητας της ολοκλήρωσης. Για να το υπολογίσουμε, οφείλουμε να αθροίσουμε όλες τις πιθανές αντίστροφες τροχές, ήτοι όλες τις πιθανές διαδρομές μέσω των οποίων μπορεί να ανακατασκευαστεί το δείγμα  $\mathbf{x}_0$  με αφετηρία τον θόρυβο. Η λύση στο εν λόγω πρόβλημα δόθηκε από τους *Sohl-Dickstein et al.* [13] που αντιπρότειναν την ελαχιστοποίηση του *variational lower-bound* της *αρνητικής λογαριθμικής πιθανοφάνειας* (negative log-likelihood). Λογαριθμίζοντας, λοιπόν, τα δύο μέλη της εξίσωσης (2.10), και, κατόπιν, πολλαπλασιάζοντάς τα επί του  $-1$ , προκύπτει:

$$-\log p_\theta(\mathbf{x}_0) = -\log \int p_\theta(\mathbf{x}_{0:T}) d\mathbf{x}_{1:T} \quad (2.11)$$

$$= -\log \int \frac{p_\theta(\mathbf{x}_{0:T}) q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} d\mathbf{x}_{1:T} \quad (2.12)$$

$$= -\log \mathbb{E}_{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} \left[ \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)} \right] \quad (2.13)$$

όπου το  $\mathbb{E}$  δηλώνει την μέση τιμή. Σε αυτό το σημείο πρέπει να εισαγάγουμε την ανισότητα του Jensen.



**Θεώρημα 2.1: Ανισότητα του Jensen**

Έστωσαν  $(\Omega, \mathcal{F}, \mathbb{P})$  χώρος πιθανότητας,  $I \subseteq \mathbb{R}$  διάστημα,  $X : \Omega \rightarrow I$  ολοκληρώσιμη τυχαία μεταβλητή, και  $f : I \rightarrow \mathbb{R}$  κυρτή συνάρτηση τέτοια ώστε η  $Y = f(X)$  να είναι επίσης ολοκληρώσιμη. Τότε, ισχύει:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Επειδή, λοιπόν, η αρνητική λογαριθμική συνάρτηση είναι κυρτή, μπορούμε να εφαρμόσουμε την ανισότητα του Jensen στην (2.13):

$$-\log p_\theta(\mathbf{x}_0) = -\log \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \leq \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ -\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \right] \quad (2.14)$$

Αυτός ο τρόπος μας επιτρέπει, ουσιαστικά, να μεταφέρουμε τον στόχο μας στην ελαχιστοποίηση του ορίου έναντι της ίδιας της αρνητικής λογαριθμικής πιθανοφάνειας.

Για αυτόν τον σκοπό, ας ασχοληθούμε, κατ' αρχάς, με την τιμή εντός της μέσης τιμής της  $q$ . Υπενθυμίζουμε ότι, εξ' ορισμού, τόσο η προς τα εμπρός, όσο και η αντίστροφη διαδικασία αποτελούν μαρκοβιανές αλυσίδες. Εκμεταλλευόμενοι, λοιπόν, της ιδιότητας του Markov, μπορούμε να αντικαταστήσουμε τις πιθανότητες της (2.14) με την βοήθεια των (2.8) και (2.9) ως εξής:

$$-\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} = -\log \frac{p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\prod_{t=1}^T q(\mathbf{x}_t|\mathbf{x}_{t-1})} \quad (2.15)$$

$$= -\log p(\mathbf{x}_T) + \sum_{t=1}^T \log \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \quad (2.16)$$

$$= -\log p(\mathbf{x}_T) + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)} + \sum_{t=2}^T \log \frac{q(\mathbf{x}_t|\mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \quad (2.17)$$

Ο όρος  $q(\mathbf{x}_t|\mathbf{x}_{t-1})$  της (2.17) μπορεί να γραφτεί εκ νέου βάσει του νόμου του Bayes ως  $\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t) q(\mathbf{x}_t)}{q(\mathbf{x}_{t-1})}$ . Ωστόσο η πραγματική εκ των υστέρων κατανομή της προς τα εμπρός διαδικασίας  $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$  είναι δυσεπίλυτη. Αυτό αλλάζει εάν, επιπρόσθετα, την εξαρτήσουμε στην αρχική κατανομή  $\mathbf{x}_0$ . Μία τέτοια κίνηση δεν θα ήταν προβληματική αφού λόγω της ιδιότητας του Markov ισχύει ότι  $q(\mathbf{x}_t|\mathbf{x}_{t-1}) = q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0)$ . Τελικά, βάσει των παραπάνω, ο ζητούμενος όρος της (2.17) προκύπτει από την εφαρμογή του νόμου του Bayes στην εξαρτημένη από την  $\mathbf{x}_0$  εκ των προτερών κατανομή:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{x}_0) = \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) q(\mathbf{x}_t, \mathbf{x}_0)}{q(\mathbf{x}_{t-1}, \mathbf{x}_0)} \quad (2.18)$$

Βρισκόμαστε, πλέον, σε θέση να αντικαταστήσουμε την (2.18) στην (2.15) προκει-

μένου να πάρουμε την μορφή που λαμβάνει το variational lower-bound:

$$\begin{aligned}
\mathcal{L}_{vlb} &:= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ -\log p(\mathbf{x}_T) + \log \frac{q(\mathbf{x}_1 | \mathbf{x}_0)}{p_\theta(\mathbf{x}_0 | \mathbf{x}_1)} + \sum_{t=2}^T \log \frac{q(\mathbf{x}_t | \mathbf{x}_{t-1})}{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ -\log p(\mathbf{x}_T) + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{p_\theta(\mathbf{x}_0|\mathbf{x}_1)} + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \cdot \frac{q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t)} \right] \\
&= \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[ \log \frac{q(\mathbf{x}_T | \mathbf{x}_0)}{p(\mathbf{x}_T)} + \sum_{t=2}^T \log \frac{q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)}{p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)} - \log p_\theta(\mathbf{x}_0 | \mathbf{x}_1) \right] \\
&= \mathbb{D}_{KL} [q(\mathbf{x}_T | \mathbf{x}_0) \parallel p(\mathbf{x}_T)] + \sum_{t=2}^T \mathbb{D}_{KL} [q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)] \\
&\quad - \log p_\theta(\mathbf{x}_0 | \mathbf{x}_1)
\end{aligned} \tag{2.19}$$

όπου  $\mathbb{D}_{KL}$  η απόκλιση Kullback-Leibler μεταξύ δύο κατανομών πιθανότητας.

Η εξίσωση (2.19) αποτελεί την *συνάρτηση απώλειας* (loss function)  $\mathcal{L}_{vlb}$  ενός DDPM. Ωστόσο, μπορεί να απλοποιηθεί περεταίρω. Αναλύοντας πιο προσεκτικά τους επιμέρους όρους, μπορούμε να διακρίνουμε ότι ο όρος  $\mathbb{D}_{KL} [q(\mathbf{x}_T | \mathbf{x}_0) \parallel p(\mathbf{x}_T)]$  είναι σταθερός και δεν περιλαμβάνει καμία παράμετρο ικανή να βελτιστοποιηθεί. Ως εκ τούτου, μπορεί να παραληφθεί. Επιπλέον, οι *Ho et al.* [8], διαπίστωσαν εμπειρικά ότι η παράλειψη του όρου  $-\log p_\theta(\mathbf{x}_0 | \mathbf{x}_1)$  οδηγεί εν γένει σε καλύτερα αποτελέσματα. Τελικά, η αντικειμενική συνάρτηση παίρνει την ακόλουθη μορφή:

$$\mathcal{L}_{simple} := \sum_{t=2}^T \mathbb{D}_{KL} [q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)] \tag{2.20}$$

$$= \mathbb{E}_{t \sim \mathcal{U}[1, T]} \left[ \mathbb{D}_{KL} [q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)] \right] \tag{2.21}$$

Όσον αφορά, πιο συγκεκριμένα, τον όρο  $q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$  αυτός γράφεται, διαφορετικά, βάσει του νόμου του Bayes:

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \frac{q(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0) q(\mathbf{x}_{t-1}, \mathbf{x}_0)}{q(\mathbf{x}_t, \mathbf{x}_0)} \tag{2.22}$$

Λαμβάνοντας, δε, υπ' όψιν, ότι η προς τα εμπρός αλυσίδα είναι μαρκοβιανή, μπορούμε να γράψουμε εκ νέου την (2.22) κάνοντας χρήση των (2.3) και (2.4):

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \frac{\mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_{t-1}, (1 - \alpha_t)\mathbf{I}) \cdot \mathcal{N}(\mathbf{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0, (1 - \bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I})} \tag{2.23}$$

Παρατηρούμε ότι όλοι οι επί μέρους όροι είναι γκαουσιανοί. Ως εκ τούτου, μπορούμε να εργαστούμε με τις εκθετικές μορφές τους. Υπενθυμίζουμε ότι:

$$\mathcal{N}(\mathbf{x}; \mu, \sigma^2 \mathbf{I}) \propto \exp \left( -\frac{\|\mathbf{x} - \mu\|^2}{2\sigma^2} \right) \tag{2.24}$$

Άρα:

$$\begin{aligned} q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) &\propto \exp \left[ - \left( \frac{(\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}_{t-1})^2}{2(1 - \alpha_t)} + \frac{(\mathbf{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}} \mathbf{x}_0)^2}{2(1 - \bar{\alpha}_{t-1})} - \frac{(\mathbf{x}_t - \sqrt{\bar{\alpha}_t} \mathbf{x}_0)^2}{2(1 - \bar{\alpha}_t)} \right) \right] \\ &\propto \mathcal{N} \left( \mathbf{x}_{t-1}; \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\mathbf{x}_0}{1 - \bar{\alpha}_t}, \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{I} \right) \end{aligned} \quad (2.25)$$

Από την (2.25), συμπεραίνουμε ότι σε κάθε χρονικό βήμα, το δείγμα  $\mathbf{x}_{t-1} \sim q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$  ακολουθεί κανονική κατανομή με μέση τιμή και συνδιακύμανση που διακρίνονται στις (2.26) και (2.27), αντιστοίχως:

$$\mu_q(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t)\mathbf{x}_0}{1 - \bar{\alpha}_t} \quad (2.26)$$

$$\Sigma_q(t) := \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{I} \quad (2.27)$$

Σε αυτό το σημείο, μπορούμε να μεταφέρουμε την προσοχή μας στον πυρήνα μετάβασης της προς τα πίσω αλυσίδας  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$ . Όπως είχε εξηγηθεί νωρίτερα, ο εν λόγω πυρήνας οφείλει να προσεγγίζει όσο το δυνατόν περισσότερο την πραγματική εκ των υστέρων κατανομή  $q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0)$ . Για τον λόγο αυτό, ακολουθεί κι εκείνος την γκαουσιανή κατανομή που περιγράφηκε στην εξίσωση (2.7)· ήτοι  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t))$ . Δεδομένου ότι η συνδιακύμανση  $\Sigma_q$  αποδείχθηκε σταθερή, μπορούμε να θέσουμε:

$$\Sigma_\theta(\mathbf{x}_t, t) = \Sigma_q(t) = \frac{(1 - \alpha_t)(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t} \mathbf{I} \quad (2.28)$$

Άρα, πλέον, αρκεί να βρούμε εκείνη την μέση τιμή  $\mu_\theta(\mathbf{x}_t, t)$  για την οποία ελαχιστοποιείται η συνάρτηση απώλειας  $\mathcal{L}_{simple}$  (2.21), ή, διαφορετικά, η απόκλιση Kullback-Leibler  $\mathbb{D}_{KL}[q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)]$ . Η απόκλιση Kullback-Leibler δύο πολυμεταβλητών γκαουσιανών κατανομών δίδεται από το κάτωθι θεώρημα:

### Θεώρημα 2.2: Απόκλιση Kullback-Leibler Κανονικών Κατανομών

Έστω  $\mathbf{x}$  ένα  $d \times 1$  τυχαίο διάνυσμα, και έστω δυο πολυμεταβλητές κανονικές κατανομές  $P$  και  $Q$  που καθορίζουν την κατανομή πιθανότητας του  $\mathbf{x}$  ως:

$$P : \mathbf{x} \sim \mathcal{N}(\mathbf{x}; \mu_1, \Sigma_1)$$

$$Q : \mathbf{x} \sim \mathcal{N}(\mathbf{x}; \mu_2, \Sigma_2)$$

Τότε η απόκλιση Kullback-Leibler των  $P$  και  $Q$  δίδεται από την σχέση:

$$\mathbb{D}_{KL}[P \parallel Q] = \frac{1}{2} \left[ (\mu_2 - \mu_1)^\top \Sigma_2^{-1} (\mu_2 - \mu_1) + \text{tr}(\Sigma_2^{-1} \Sigma_1) - \ln \frac{|\Sigma_1|}{|\Sigma_2|} - d \right]$$

Στην προκειμένη περίπτωση, τα δύο μητρώα συνδιακύμανσης είναι ίσα μεταξύ τους. Ως εκ τούτου, η KL απόκλιση λαμβάνει την μορφή:

$$\mathbb{D}_{KL} [q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)] = \mathbb{E}_q \left[ \frac{1}{2} (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)^\top \Sigma_q^{-1} (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q) \right] \quad (2.29)$$

$$= \mathbb{E}_q \left[ \frac{1}{2\sigma^2} \|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q\|_2^2 \right] \quad (2.30)$$

όπου τα  $\boldsymbol{\mu}_q$  και  $\boldsymbol{\mu}_\theta$  αποτελούν συντομογραφίες των  $\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0)$  και  $\boldsymbol{\mu}_\theta(\mathbf{x}_t, t)$  αντίστοιχα, και όπου το  $\sigma$  είναι η τυπική απόκλιση των δύο κατανομών, δηλαδή  $\Sigma_q = \Sigma_\theta = \sigma^2 \mathbf{I}$ . Όσον αφορά, τώρα, την μέση τιμή  $\boldsymbol{\mu}_q$ , παρατηρούμε από την (2.26) ότι εξαρτάται από το αρχικό διάνυσμα  $\mathbf{x}_0$ . Αυτό, εντούτοις, προκύπτει κατόπιν μετασχηματισμού της (2.4) ως εξής:

$$\mathbf{x}_0 = \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon}{\sqrt{\bar{\alpha}_t}} \quad (2.31)$$

Οπότε, αντικαθιστώντας την (2.31) στην (2.26):

$$\boldsymbol{\mu}_q(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})\mathbf{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1 - \alpha_t) \frac{\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t} \epsilon}{\sqrt{\bar{\alpha}_t}}}{1 - \bar{\alpha}_t} \quad (2.32)$$

$$= \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\alpha_t}} \epsilon \quad (2.33)$$

Αναμενόμενα, η μέση τιμή του πυρήνα αποθορυβοποίησης  $\boldsymbol{\mu}_\theta$  οφείλει να προσεγγίζει την (2.33). Λαμβάνοντας, όμως, υπ' όψιν ότι το  $\mathbf{x}_t$  αποτελεί, κατ' ουσίαν, την είσοδο στον πυρήνα, οι *Ho et al.* [8] πρότειναν την παρακάτω επαναπαραμετροποίηση:

$$\boldsymbol{\mu}_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\alpha_t}} \hat{\epsilon}_\theta(\mathbf{x}_t, t) \quad (2.34)$$

Βρισκόμαστε, πλέον, σε θέση να αντικαταστήσουμε τις (2.33) και (2.34) στην (2.30), και, κατόπιν, στην (2.21). Δηλαδή:

$$\begin{aligned} \mathcal{L}_{simple} &= \mathbb{E}_{t \sim \mathcal{U}[1, T]} \left[ \mathbb{D}_{KL} [q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)] \right] \\ &= \mathbb{E}_{t \sim \mathcal{U}[1, T], \mathbf{x}_t \sim q(\mathbf{x}_t), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[ \frac{\left\| \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\alpha_t}} \hat{\epsilon}_\theta(\mathbf{x}_t, t) - \frac{1}{\sqrt{\alpha_t}} \mathbf{x}_t + \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t} \sqrt{\alpha_t}} \epsilon \right\|_2^2}{2\sigma^2(t)} \right] \\ &= \mathbb{E}_{t \sim \mathcal{U}[1, T], \mathbf{x}_t \sim q(\mathbf{x}_t), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[ \frac{1}{2\sigma^2(t)} \frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t) \alpha_t} \left\| \epsilon - \hat{\epsilon}_\theta(\mathbf{x}_t, t) \right\|_2^2 \right] \\ &= \mathbb{E}_{t \sim \mathcal{U}[1, T], \mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[ \frac{1}{2\sigma^2(t)} \frac{(1 - \alpha_t)^2}{(1 - \bar{\alpha}_t) \alpha_t} \left\| \epsilon - \hat{\epsilon}_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|_2^2 \right] \end{aligned} \quad (2.35)$$

Ο όρος αυτός δεν είναι, ουσιαστικά, παρά μια χρονικά σταθμισμένη απόσταση μεταξύ του πραγματικού θορύβου της εικόνας  $\mathbf{x}_t$  και της εκτίμησης του νευρωνικού δικτύου. Οι *Ho et al.* [8] προτείνουν μία επιπρόσθετη τελευταία απλοποίηση. Διαπίστωσαν εμπειρικά ότι η εκπαίδευση του μοντέλου διάχυσης γίνεται πιο αποδοτική όταν παραλείπονται οι όροι στάθμισης  $\frac{1}{2\sigma^2(t)} \frac{(1-\alpha_t)^2}{(1-\bar{\alpha}_t)\alpha_t}$ .

Άρα, τελικά, η συνάρτηση απώλειας ενός DDPM δίδεται από την εξίσωση (2.36), βάσει της οποίας εξάγουμε τους αλγορίθμους εκπαίδευσης και δειγματοληψίας που παραθέτονται παρακάτω [8, 9, 10, 11, 12, 13, 15].

$$\mathcal{L}_{simple} = \mathbb{E}_{t \sim \mathcal{U}[1, T], \mathbf{x}_0 \sim q(\mathbf{x}_0), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[ \left\| \epsilon - \hat{\epsilon}_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t) \right\|_2^2 \right] \quad (2.36)$$

---

### Αλγόριθμος 2.1 Εκπαίδευση DDPM

---

```

1: repeat
2:    $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ 
3:    $t \sim \mathcal{U}[1, T]$ 
4:    $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5:   Take gradient step on  $\nabla_\theta \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t)\|^2$ 
6: until converged

```

---



---

### Αλγόριθμος 2.2 Δειγματοληψία DDPM

---

```

1:  $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
2: for  $t = T$  to 1 do
3:   if  $t > 1$  then
4:      $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
5:   else
6:      $\mathbf{z} = \mathbf{0}$ 
7:   end if
8:    $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left( \mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \hat{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_q(t) \mathbf{z}$ 
9: end for

```

---

## 2.2 Score-Based Generative Models

Τα προαναφερθέντα DDPMs ανήκουν στην κατηγορία των likelihood-based παραγωγικών μεθόδων λόγω του τρόπου εξαγωγής της αντικειμενικής τους συνάρτησης, όπως αυτή περιγράφηκε στην ενότητα 2.1. Τα likelihood-based μοντέλα έχουν σημειώσει, αναμφίβολα, εντυπωσιακά αποτελέσματα τα τελευταία χρόνια· παρουσιάζουν, ωστόσο, ορισμένους εγγενείς περιορισμούς. Στην προκειμένη περίπτωση οι εν λόγω περιορισμοί έγκεινται στην ανάγκη χρήσης υποκατάστατων συναρτήσεων απωλειών (του variational lower-bound). Η ανικανότητα εκτίμησης του maximum likelihood ενδέχεται

να καταστήσει τόσο την εκπαίδευση όσο και την δειγματοληψία ανακριβείς, περίπλοκες και αργές. Η ανάγκη επίλυσης αυτών των περιορισμών οδήγησε στην ανάπτυξη των *score-based* μεθόδων, στις οποίες υπάγονται τα SGMs.

Η ειδοποιός διαφορά των *Score-Based Generative Models* (SGMs) [7, 16] συγκριτικά με τα DDPMs έγκειται στην προσέγγιση που ακολουθούν για την εκμάθηση της υποκείμενης κατανομής δεδομένων ενός δείγματος. Αντί, λοιπόν, να μοντελοποιούν ρητά την αντίστροφη διαδικασία διάχυσης, τα SGMs εκμεταλλεύονται την έννοια του Stein Score Function.

Το *Stein Score*, *Score Function*, ή ακόμη πιο απλοϊκά, *Score* [17] μίας κατανομής δεδομένων  $\mathbf{x} \sim p(\mathbf{x})$  ορίζεται ως η κλίση του λογαρίθμου της συνάρτησης πυκνότητας πιθανότητας της  $\mathbf{x}$ . Μαθηματικώς:

$$f_{\text{score}}(\mathbf{x}) := \nabla_{\mathbf{x}} \log p(\mathbf{x}) \quad (2.37)$$

Στην ουσία, το stein score πρόκειται για ένα διανυσματικό πεδίο που δείχνει προς τις κατευθύνσεις πάνω στις οποίες η συνάρτηση πυκνότητας πιθανότητας παρουσιάζει τον μεγαλύτερο ρυθμό αύξησης.

Η αρχή λειτουργίας ενός SGM βασίζεται στην προοδευτική έγχυση γκαουσιανού θορύβου αυξανόμενης έντασης στα δεδομένα, δημιουργώντας, καθ' αυτόν τον τρόπο, μια ακολουθία θορυβωδών κατανομών. Κατόπιν, ένα βαθύ νευρωνικό δίκτυο ονόματι *Noise Conditional Score Network* (NCSN) [7] εκπαιδεύεται να εκτιμάει την συνάρτηση score των εν λόγω θορυβωδών κατανομών εξαρτημένα από το εκάστοτε εφαρμοζόμενο επίπεδο θορύβου.

Πιο αναλυτικά, ας υποθέσουμε  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$  την κατανομή δεδομένων ενός δείγματος εκπαίδευσης. Η *εκπαιδευτική διαδικασία* παράγει, κατ' αρχάς, μια φθίνουσα θετική γεωμετρική ακολουθία τυπικών αποκλίσεων  $\{\sigma_t\}_{t=1}^T$  που υπακούει στην σχέση:

$$\frac{\sigma_1}{\sigma_2} = \frac{\sigma_2}{\sigma_3} = \dots = \frac{\sigma_t}{\sigma_{t+1}} = \dots = \frac{\sigma_{T-1}}{\sigma_T} > 1 \quad (2.38)$$

Η  $\{\sigma_t\}_{t=1}^T$  αποτελεί την ακολουθία επιπέδου θορύβου, αφού μετά την παραγωγή της, χρησιμοποιείται για την μετατροπή της αρχικής κατανομής  $\mathbf{x}_0$  σε  $\mathbf{x}_t$  εγχύοντας την πρώτη με γκαουσιανό θόρυβο ως εξής:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \mathbf{x}_0, \sigma_t^2 \mathbf{I}) \quad (2.39)$$

Σε αυτό το σημείο αξίζει να σχολιάσουμε ότι η επιλογή της ακολουθίας επιπέδου θορύβου γίνεται με τρόπο τέτοιο που η  $\sigma_T$  να είναι αρκετά μικρή έτσι ώστε να ελαχιστοποιεί την επίδραση στα αρχικά δεδομένα, και η  $\sigma_1$  να είναι, τουναντίον, αρκετά μεγάλη ώστε να μετριάξει τις δυσκολίες που σχετίζονται με την *score-based* παραγωγική μοντελοποίηση. Αυτές οι δυσκολίες συνδέονται κυρίως με την “*Υπόθεση της Πολλαπλότητας*” (The Manifold Hypothesis), η οποία αναφέρεται στο ότι, συχνά, τα δεδομένα στον πραγματικό κόσμο βρίσκονται πάνω σε πολλαπλότητες χαμηλής διάστασης, ενσωματωμένες μέσα σε έναν χώρο υψηλής διάστασης—τον περιβάλλοντα χώρο. Αυτή η πραγματικότητα μπορεί, ενδεχομένως, να οδηγήσει σε ανακριβή εκτίμηση της συνάρτησης score.

Επιστρέφοντας στην εκπαιδευτική διαδικασία, παρατηρούμε ότι η περιγραφείσα

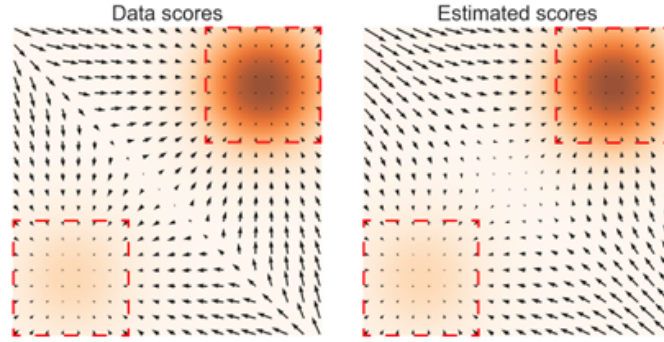
μέθοδος παράγει μια νέα ακολουθία θορυβωδών κατανομών δεδομένων  $\{q(\mathbf{x}_t)\}_{t=1}^T$ , όπου:

$$q(\mathbf{x}_t) := \int q(\mathbf{x}_t) q(\mathbf{x}_0) d\mathbf{x}_0 \quad (2.40)$$

Τελικός στόχος είναι το NCSN να εκτιμήσει από κοινού τα score όλων των διαταραγμένων κατανομών δεδομένων. Δηλαδή:

$$\forall t \in \{1, \dots, T\} : s_\theta(\mathbf{x}_0, \sigma_t) \approx \nabla_{\mathbf{x}_t} \log q(\mathbf{x}_t) \quad (2.41)$$

όπου το  $s_\theta(\mathbf{x}_0, \sigma_t)$  συμβολίζει το βαθύ νευρωνικό δίκτυο. Τα SGMs μπορούν να είναι αρκετά ευέλικτα ως προς την ακριβή μέθοδο υλοποίησης της εκπαιδευτικής διαδικασίας. Προσφέρεται, δηλαδή, μια πληθώρα καθιερωμένων τεχνικών για την εκτίμηση των score, συμπεριλαμβανομένων του Score Matching, Sliced Score Matching, και Denoising Score Matching, που περιγράφονται πιο αναλυτικά παρακάτω.



**Σχήμα 2.2:** Απεικονίσεις της έννοιας του Stein Score Function. **Αριστερά** διακρίνεται το score της πραγματικής συνάρτησης πυκνότητας πιθανότητας  $\nabla_{\mathbf{x}} \log q(\mathbf{x})$ , και **δεξιά** το εκτιμημένο score  $s_\theta(\mathbf{x})$ . Η συνάρτηση πυκνότητας πιθανότητας των δεδομένων  $q(\mathbf{x})$  κωδικοποιείται μέσω του χρωματικού χάρτη: περιοχές σκουρότερου χρώματος υποδηλώνουν περιοχές υψηλότερης πυκνότητας. Τα κόκκινα διακεκομμένα πλαίσια επισημαίνουν περιοχές όπου  $s_\theta(\mathbf{x}) \approx \nabla_{\mathbf{x}} \log q(\mathbf{x})$  [7].

### 2.2.1 Score Matching

Το *Score Matching* [17] είναι ιδιαίτερα κατάλληλο για την εκμάθηση μη κανονικοποιημένων στατιστικών μοντέλων, όπως των ενεργειακών μοντέλων. Βασίζεται στην ελαχιστοποίηση της απόστασης των συναρτήσεων score των κατανομών των δεδομένων και του μοντέλου.

Έστω, λοιπόν, μια ακολουθία ανεξάρτητων και ομοιόμορφα κατανεμημένων δειγμάτων  $\{\mathbf{x}_i\}_{i=1}^N \subset \mathbb{R}^d$  από μία κατανομή δεδομένων  $q_d(\mathbf{x})$ . Όπως προείπαμε, στόχος αποτελεί η εκμάθηση μίας μη κανονικοποιημένης συνάρτησης πυκνότητας πιθανότητας,  $\tilde{p}_\theta(\mathbf{x})$ . Επιπλέον, συμβολίζουμε την συνάρτηση κανονικοποίησης του μοντέλου ως  $Z_\theta$  την οποία και θεωρούμε υφιστάμενη αλλά δυσεπίλυτη. Αν  $p_\theta(\mathbf{x})$  είναι η κανονικοποι-



ημένη συνάρτηση πυκνότητας πιθανότητας που καθορίζεται από το μοντέλο, τότε:

$$p_\theta(\mathbf{x}) = \frac{\tilde{p}_\theta(\mathbf{x})}{Z_\theta}, \quad (2.42)$$

$$Z_\theta = \int \tilde{p}_\theta(\mathbf{x}) d\mathbf{x} \quad (2.43)$$

Επιπλέον, παραθέτουμε τις κάτωθι συναρτήσεις score:

$$\mathbf{s}_\theta(\mathbf{x}) := \nabla_{\mathbf{x}} \log p_\theta(\mathbf{x}) \quad (2.44)$$

$$\mathbf{s}_q(\mathbf{x}) := \nabla_{\mathbf{x}} \log q(\mathbf{x}) \quad (2.45)$$

Η εκμάθηση μη κανονικοποιημένων μοντέλων δια της μεθόδους της εκτίμησης μέγιστης πιθανοφάνειας μπορεί να γίνει εξαιρετικά δύσκολη λόγω της δυσεπίλυτης συνάρτησης κανονικοποίησης  $Z_\theta$ . Η τεχνική του Score Matching αποφεύγει το εν λόγω πρόβλημα ελαχιστοποιώντας την απόκλιση Fisher μεταξύ των κατανομών  $q$  και  $p_\theta$ . Αυτή δίδεται από την παρακάτω σχέση:

$$\mathcal{J}(\theta) := \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} [\|\mathbf{s}_\theta(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\|_2^2] \quad (2.46)$$

Στο μεταξύ η (2.44), κάνοντας χρήση της (2.42) μπορεί να γραφτεί ως:

$$\mathbf{s}_\theta(\mathbf{x}) = \nabla_{\mathbf{x}} \log \left( \frac{\tilde{p}_\theta(\mathbf{x})}{Z_\theta} \right) \quad (2.47)$$

$$= \nabla_{\mathbf{x}} \log \tilde{p}_\theta(\mathbf{x}) - \nabla_{\mathbf{x}} \log Z_\theta \quad (2.48)$$

$$= \nabla_{\mathbf{x}} \log \tilde{p}_\theta(\mathbf{x}) \quad (2.49)$$

όπου η ισότητα των (2.48) και (2.49) ισχύει δεδομένου ότι η συνάρτηση κανονικοποίησης  $Z_\theta$  είναι ανεξάρτητη του  $\mathbf{x}$ .

Παρατηρούμε, λοιπόν, από την (2.49), ότι η  $\mathbf{s}_\theta(\mathbf{x})$  είναι ανεξάρτητη της  $Z_\theta$ . Αυτό συνεπάγεται ότι το ίδιο συμβαίνει και για την απόκλιση Fisher. Παρόλ' αυτά, η εξίσωση (2.46) δεν μπορεί να χρησιμοποιηθεί ακόμη, καθώς δεν γνωρίζουμε την συνάρτηση score της κατανομής του δείγματος,  $\mathbf{s}_q(\mathbf{x})$ . Στο πλαίσιο αυτό, οι Hyvärinen et al. [17] έδειξαν ότι η ζητούμενη απόκλιση Fisher μπορεί να γραφτεί ισοδυνάμως ως εξής:

$$\mathcal{J}(\theta) := \frac{1}{2} \mathbb{E}_q [\|\mathbf{s}_\theta(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\|_2^2] \quad (2.50)$$

$$= \frac{1}{2} \mathbb{E}_q [\|\mathbf{s}_\theta(\mathbf{x})\|_2^2] - \mathbb{E}_q [\mathbf{s}_\theta(\mathbf{x}) \cdot \mathbf{s}_q(\mathbf{x})] + \frac{1}{2} \mathbb{E}_q [\|\mathbf{s}_q(\mathbf{x})\|_2^2] \quad (2.51)$$

Όμως, ο τελευταίος όρος  $\frac{1}{2} \mathbb{E}_q [\|\mathbf{s}_q(\mathbf{x})\|_2^2]$  δεν εξαρτάται από την παράμετρο  $\theta$  και, για τον λόγο αυτό, συχνά παραλείπεται όταν ελαχιστοποιούμε ως προς  $\theta$ , όπως εν προκειμένω. Συνεπώς, για τις ανάγκες της εκπαίδευσης, γράφουμε:

$$\mathcal{J}(\theta) = \frac{1}{2} \mathbb{E}_q [\|\mathbf{s}_\theta(\mathbf{x})\|_2^2] - \mathbb{E}_q [\mathbf{s}_\theta(\mathbf{x}) \cdot \mathbf{s}_q(\mathbf{x})] + \text{const.} \quad (2.52)$$

Χρησιμοποιούμε, τώρα, την (2.45) που, αξιοποιώντας τον κανόνα παραγώγισης λογα-



ρίθμου, γράφουμε εκ νέου ως:

$$\mathbf{s}_q(\mathbf{x}) := \nabla_{\mathbf{x}} \log q(\mathbf{x}) = \frac{\nabla_{\mathbf{x}} q(\mathbf{x})}{q(\mathbf{x})} \quad (2.53)$$

Συνεπώς, ο δεύτερος όρος της (2.52) δίδεται από την σχέση:

$$\mathbb{E}_q [\mathbf{s}_\theta(\mathbf{x}) \cdot \mathbf{s}_q(\mathbf{x})] = \int \mathbf{s}_\theta(\mathbf{x}) \cdot \mathbf{s}_q(\mathbf{x}) \cdot q(\mathbf{x}) d\mathbf{x} \quad (2.54)$$

$$= \int \mathbf{s}_\theta(\mathbf{x}) \cdot \frac{\nabla_{\mathbf{x}} q(\mathbf{x})}{q(\mathbf{x})} \cdot q(\mathbf{x}) d\mathbf{x} \quad (2.55)$$

$$= \int \mathbf{s}_\theta(\mathbf{x}) \cdot \nabla_{\mathbf{x}} q(\mathbf{x}) d\mathbf{x} \quad (2.56)$$

Εφαρμόζουμε στην (2.56) ολοκλήρωση κατά μέλη, οπότε:

$$\mathbb{E}_q [\mathbf{s}_\theta(\mathbf{x}) \cdot \mathbf{s}_q(\mathbf{x})] = \int \nabla_{\mathbf{x}} \cdot (q(\mathbf{x}) \mathbf{s}_\theta(\mathbf{x})) d\mathbf{x} - \int q(\mathbf{x}) \nabla_{\mathbf{x}} \cdot \mathbf{s}_\theta(\mathbf{x}) d\mathbf{x} \quad (2.57)$$

Ας σταθούμε, τώρα, σε κάθε έναν όρο της (2.57) ξεχωριστά. Παρατηρούμε ο δεύτερος όρος εξ' ορισμού με την παρακάτω αναμενόμενη τιμή:

$$\int q(\mathbf{x}) \nabla_{\mathbf{x}} \cdot \mathbf{s}_\theta(\mathbf{x}) d\mathbf{x} =: \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} [\nabla_{\mathbf{x}} \cdot \mathbf{s}_\theta(\mathbf{x})] \quad (2.58)$$

Ο πρώτος, δε, όρος παρουσιάζει μεγαλύτερο ενδιαφέρον. Μπορούμε να χρησιμοποιήσουμε το *Θεώρημα του Gauss*.

### Θεώρημα 2.1: Θεώρημα Απόκλισης (Gauss)

Έστω  $\Omega \subset \mathbb{R}^d$  μία συμπαγής περιοχή με ικανοποιητικά ομαλό σύνορο  $\partial\Omega$ , και έστω  $\mathbf{F} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  ένα συνεχώς διαφορίσιμο διάνυσμα πεδίο. Τότε ισχύει η ακόλουθη σχέση:

$$\int_{\Omega} \nabla \cdot \mathbf{F}(\mathbf{x}) d\mathbf{x} = \oint_{\partial\Omega} \mathbf{F}(\mathbf{x}) \cdot d\mathbf{S},$$

όπου  $d\mathbf{S}$  δηλώνει το στοιχειώδες διανυσματικό στοιχείο επιφανείας στο  $\partial\Omega$ , με κατεύθυνση προς τα έξω.

Χρησιμοποιώντας, λοιπόν, το άνωθεν θεώρημα, δύναται να γραφτεί ο πρώτος όρος της (2.58) ως εξής:

$$\int_{\Omega} \nabla_{\mathbf{x}} \cdot (q(\mathbf{x}) \mathbf{s}_\theta(\mathbf{x})) d\mathbf{x} = \oint_{\partial\Omega} q(\mathbf{x}) \mathbf{s}_\theta(\mathbf{x}) d\mathbf{S} \quad (2.59)$$

Οι *Hynvärinen et al.* [17] θεωρούν ότι στην πράξη  $\Omega = \mathbb{R}^d$ , οπότε το  $\partial\Omega$  τείνει σε σφαίρα με ακτίνα  $\|\mathbf{x}\| \rightarrow \infty$ . Υποθέτουν, επιπρόσθετα, ότι η συνάρτηση πυκνότητας πιθανότητας

$q(\mathbf{x})$  παρουσιάζει την ιδιότητα:

$$\lim_{\|\mathbf{x}\| \rightarrow \infty} q(\mathbf{x}) = 0 \quad (2.60)$$

και ότι το διάνυσμα score του μοντέλου  $\mathbf{s}_\theta(\mathbf{x})$  δεν αυξάνει ταχύτερα από πολυωνυμικό ρυθμό. Υπό αυτές τις ήπιες υποθέσεις, το γινόμενο  $q(\mathbf{x}) \mathbf{s}_\theta(\mathbf{x})$  φθίνει ταχύτερα από  $\frac{1}{\|\mathbf{x}\|^d}$ , με αποτέλεσμα το επικαμπύλιο ολοκλήρωμα να μηδενίζεται:

$$\oint_{\partial\Omega} q(\mathbf{x}) \mathbf{s}_\theta(\mathbf{x}) dS = 0 \quad (2.61)$$

Αντικαθιστώντας, επομένως, τις (2.58) και (2.61) στην (2.57) καταλήγουμε με:

$$\mathbb{E}_q [\mathbf{s}_\theta(\mathbf{x}) \cdot \mathbf{s}_q(\mathbf{x})] = -\mathbb{E}_q [\nabla_{\mathbf{x}} \cdot \mathbf{s}_\theta(\mathbf{x})] \quad (2.62)$$

Επιστρέφοντας στην (2.52), η απόκλιση Fisher δίδεται από:

$$\mathcal{J}_{sm}(\theta) = \frac{1}{2} \mathbb{E}_q [\|\mathbf{s}_\theta(\mathbf{x})\|_2^2] + \mathbb{E}_q [\nabla_{\mathbf{x}} \cdot \mathbf{s}_\theta(\mathbf{x})] \quad (2.63)$$

Τελικά, αν λάβουμε υπ' όψιν ότι:

$$\nabla_{\mathbf{x}} \cdot \mathbf{s}_\theta(\mathbf{x}) = \text{tr}(\nabla_{\mathbf{x}} \mathbf{s}_\theta(\mathbf{x})) \quad (2.64)$$

μας δίδεται η δυνατότητα να γράψουμε συμπαγώς την ζητούμενη απόκλιση με την παρακάτω σχέση:

$$\mathcal{J}_{sm}(\theta) = \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} \left[ \text{tr}(\nabla_{\mathbf{x}} \mathbf{s}_\theta(\mathbf{x})) + \frac{1}{2} \|\mathbf{s}_\theta(\mathbf{x})\|_2^2 \right] \quad (2.65)$$

Άμεσα από την (2.65) προκύπτει ο αμερόληπτος εκτιμητής που χρησιμοποιείται για την εκπαίδευση της  $\tilde{p}_\theta(\mathbf{x})$ :

$$\hat{\mathcal{J}}_{sm}(\theta; \{\mathbf{x}_i\}_{i=1}^N) := \frac{1}{N} \sum_{i=1}^N \left[ \text{tr}(\nabla_{\mathbf{x}} \mathbf{s}_\theta(\mathbf{x}_i)) + \frac{1}{2} \|\mathbf{s}_\theta(\mathbf{x}_i)\|_2^2 \right] \quad (2.66)$$

### 2.2.2 Sliced Score Matching

Η μέθοδος του Score Matching παρουσιάζει ένα μη αμελητέο ελάττωμα: ο υπολογισμός των διαγωνίων στοιχείων του Ιακωβιανού μητρώου του score του μοντέλου,  $\text{tr}(\nabla_{\mathbf{x}} \mathbf{s}_\theta(\mathbf{x}))$ , στην εξίσωση (2.65) είναι αρκετά ακριβός, ειδικά σε περιπτώσεις δεδομένων υψηλών διαστάσεων. Προς επίλυση αυτού του προβλήματος, αναπτύχθηκε μια εναλλακτική μέθοδος, η *Sliced Score Matching* [18].

Η βασική ιδέα του Sliced Score Matching είναι, αντί να αντιστοιχίζει τα πλήρη διανύσματα των score του μοντέλου,  $\mathbf{s}_\theta(\mathbf{x})$ , και των δεδομένων,  $\mathbf{s}_q(\mathbf{x})$ , στο  $\mathbb{R}^d$ , να τα προβάλλει σε τυχαίες κατευθύνσεις  $\mathbf{v}$  και να αντιστοιχίζει τις προβολές. Ως εκ τούτου, ορίζεται η κάτωθι αντικειμενική συνάρτηση αντί της απόκλισης Fisher:

$$\mathcal{J}_{ssm}(\theta; p_v) := \frac{1}{2} \mathbb{E}_{\mathbf{v} \sim p_v(\mathbf{v})} \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} \left[ \left( \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) - \mathbf{v}^\top \mathbf{s}_q(\mathbf{x}) \right)^2 \right] \quad (2.67)$$

όπου η  $p_v$  είναι μια κατανομή των διανυσμάτων προβολών, ανεξάρτητη της  $q$ , και για την οποία απαιτούμε  $\mathbb{E}_{p_v} [\mathbf{v}^\top \mathbf{v}] > 0$  και  $\mathbb{E}_{p_v} [\|\mathbf{v}\|_2^2] < \infty$ . Συννηθισμένες επιλογές περιλαμβάνουν μια πολυμεταβλητή κανονική κατανομή  $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ , μια πολυμεταβλητή κατανομή Rademacher (ήτοι η ομοιόμορφη κατανομή στο  $\{\pm 1\}^d$ ), ή μια ομοιόμορφη κατανομή στην υπερσφαίρα  $\mathbb{S}^{d-1}$ .

Στην (2.67), παρατηρούμε ότι αυτή εξαρτάται από το άγνωστο  $\mathbf{s}_q(\mathbf{x})$ . Για την απλοποίησή του, χρησιμοποιείται παρόμοια τεχνική με αυτήν που χρησιμοποιήθηκε και προηγουμένως:

$$\mathcal{J}_{ssm}(\theta; p_v) = \frac{1}{2} \mathbb{E}_{p_v, q} \left[ \left( \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) - \mathbf{v}^\top \mathbf{s}_q(\mathbf{x}) \right)^2 \right] \quad (2.68)$$

$$\begin{aligned} &= \frac{1}{2} \mathbb{E}_{p_v, q} \left[ \left( \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) \right)^2 \right] - \mathbb{E}_{p_v, q} \left[ \left( \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) \right) \left( \mathbf{v}^\top \mathbf{s}_q(\mathbf{x}) \right) \right] \\ &\quad + \frac{1}{2} \mathbb{E}_{p_v, q} \left[ \left( \mathbf{v}^\top \mathbf{s}_q(\mathbf{x}) \right)^2 \right] \end{aligned} \quad (2.69)$$

Όπως και στο κλασσικό Score Matching, έτσι και εδώ ο τρίτος όρος είναι ανεξάρτητος της  $\theta$  και αγνοείται. Ο δεύτερος όρος γράφεται:

$$\mathbb{E}_{p_v, q} \left[ \left( \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) \right) \left( \mathbf{v}^\top \mathbf{s}_q(\mathbf{x}) \right) \right] = \mathbb{E}_{p_v} \left[ \int \left( \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) \right) \left( \mathbf{v}^\top \mathbf{s}_q(\mathbf{x}) \right) q(\mathbf{x}) d\mathbf{x} \right] \quad (2.70)$$

$$= \mathbb{E}_{p_v} \left[ \int \left( \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) \right) \left( \mathbf{v}^\top \nabla_{\mathbf{x}} \log q(\mathbf{x}) \right) q(\mathbf{x}) d\mathbf{x} \right] \quad (2.71)$$

$$= \mathbb{E}_{p_v} \left[ \int \left( \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) \right) \left( \mathbf{v}^\top \frac{\nabla_{\mathbf{x}} q(\mathbf{x})}{q(\mathbf{x})} \right) q(\mathbf{x}) d\mathbf{x} \right] \quad (2.72)$$

$$= \mathbb{E}_{p_v} \left[ \int \left( \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) \right) \left( \mathbf{v}^\top \nabla_{\mathbf{x}} q(\mathbf{x}) \right) d\mathbf{x} \right] \quad (2.73)$$

$$= \mathbb{E}_{p_v} \left[ \sum_{i=1}^d \int \left( \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) \right) \left( v_i \frac{\partial q(\mathbf{x})}{\partial x_i} \right) d\mathbf{x} \right] \quad (2.74)$$

$$\begin{aligned} &= \mathbb{E}_{p_v} \left[ \sum_{i=1}^d \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) v_i q(\mathbf{x}) \right]_{-\infty}^{+\infty} \\ &\quad - \mathbb{E}_{p_v} \left[ \sum_{i=1}^d \int v_i q(\mathbf{x}) \frac{\partial \left( \mathbf{v}^\top \mathbf{s}_\theta(\mathbf{x}) \right)}{\partial x_i} d\mathbf{x} \right] \end{aligned} \quad (2.75)$$

$$= 0 - \mathbb{E}_{p_v} \left[ \int \mathbf{v}^\top \nabla_{\mathbf{x}} \mathbf{s}_\theta(\mathbf{x}) \mathbf{v} \cdot q(\mathbf{x}) d\mathbf{x} \right] \quad (2.76)$$

$$= -\mathbb{E}_{p_v, q} \left[ \mathbf{v}^\top \nabla_{\mathbf{x}} \mathbf{s}_\theta(\mathbf{x}) \mathbf{v} \right] \quad (2.77)$$

Άρα η εξίσωση (2.67) γράφεται συλλογικά:

$$\mathcal{J}_{ssm}(\theta; p_v) = \mathbb{E}_{\mathbf{v} \sim p_v(\mathbf{x})} \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} \left[ \mathbf{v}^\top \nabla_{\mathbf{x}} \mathbf{s}_\theta(\mathbf{x}) \mathbf{v} + \frac{1}{2} \|\mathbf{s}_\theta(\mathbf{x})\|_2^2 \right] \quad (2.78)$$

από την οποία λαμβάνεται ο κάτωθι αμερόληπτος εκτιμητής;

$$\hat{\mathcal{J}}_{ssm}(\theta; \{\mathbf{x}_i\}_{i=1}^N, \{\mathbf{v}_{ij}\}_{1 \leq i \leq N, 1 \leq j \leq M}) := \frac{1}{N} \frac{1}{M} \sum_{i=1}^N \sum_{j=1}^M \left[ \mathbf{v}_{ij}^\top \nabla_{\mathbf{x}} \mathbf{s}_\theta(\mathbf{x}_i) \mathbf{v}_{ij} + \frac{1}{2} \|\mathbf{s}_\theta(\mathbf{x}_i)\|_2^2 \right] \quad (2.79)$$

### 2.2.3 Denoising Score Matching

Το *Denoising Score Matching* [19] αποτελεί μια ακόμη μέθοδο που δημιουργήθηκε με σκοπό την επίλυση του περιορισμού του κλασσικού Score Matching, αλλά, αυτή τη φορά, αποφεύγει εξ' ολοκλήρου τον προβληματικό όρο  $\text{tr}(\nabla_{\mathbf{x}} \mathbf{s}_\theta(\mathbf{x}))$  της (2.65). Η βασική ιδέα έγκειται στην έγχυση των δεδομένων  $\mathbf{x} \sim q(\mathbf{x})$  με γκαουσιανό θόρυβο. Αυτό μας επιτρέπει να μεταστρέψουμε την προσοχή μας από την συνάρτηση πυκνότητας πιθανότητας  $q$  στην κατανομή του θορυβώδους δείγματος:

$$q_\sigma(\tilde{\mathbf{x}} | \mathbf{x}) = \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \sigma^2 \mathbf{I}) \quad (2.80)$$

Νέος στόχος εκπαίδευσης του παραμετρικού score function  $\mathbf{s}_\theta(\tilde{\mathbf{x}})$  είναι να προσεγγίζει το αληθινό score της θορυβώδους κατανομής, η οποία δίδεται από την κάτωθι σχέση:

$$q_\sigma(\tilde{\mathbf{x}}) := \int q_\sigma(\tilde{\mathbf{x}} | \mathbf{x}) q(\mathbf{x}) d\mathbf{x} \quad (2.81)$$

Σε αυτήν την περίπτωση, η αντικειμενική συνάρτηση ορίζεται ως εξής:

$$\mathcal{J}_{dsm}(\theta; q_\sigma) = \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} \mathbb{E}_{\tilde{\mathbf{x}} \sim q_\sigma(\tilde{\mathbf{x}} | \mathbf{x})} \left[ \|\mathbf{s}_\theta(\tilde{\mathbf{x}}) - \nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}} | \mathbf{x})\|_2^2 \right] \quad (2.82)$$

Όμως, επειδή η  $q_\sigma(\tilde{\mathbf{x}} | \mathbf{x})$  είναι γκαουσιανή, ισχύει:

$$\nabla_{\tilde{\mathbf{x}}} \log q_\sigma(\tilde{\mathbf{x}} | \mathbf{x}) = -\frac{1}{\sigma^2} (\tilde{\mathbf{x}} - \mathbf{x}) \quad (2.83)$$

Συνεπώς, η (2.82) δύναται να απλοποιηθεί περεταίρω:

$$\mathcal{J}_{dsm}(\theta; q_\sigma) = \frac{1}{2} \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} \mathbb{E}_{\tilde{\mathbf{x}} \sim \mathcal{N}(\tilde{\mathbf{x}}; \mathbf{x}, \sigma^2 \mathbf{I})} \left[ \|\mathbf{s}_\theta(\tilde{\mathbf{x}}) + \frac{1}{\sigma^2} (\tilde{\mathbf{x}} - \mathbf{x})\|_2^2 \right] \quad (2.84)$$

Ομοίως με τις άλλες δύο προαναφερθείσες μεθόδους, η (2.85) δίδει τον αντικειμενικό εκτιμητή του Denoising Score Matching:

$$\hat{\mathcal{J}}_{dsm}(\theta; \{\mathbf{x}_i\}_{i=1}^N, \{\tilde{\mathbf{x}}_i\}_{i=1}^N) := \frac{1}{2N} \sum_{i=1}^N \left[ \|\mathbf{s}_\theta(\tilde{\mathbf{x}}_i) + \frac{1}{\sigma^2} (\tilde{\mathbf{x}}_i - \mathbf{x}_i)\|_2^2 \right] \quad (2.85)$$

Αυτή η τρίτη στρατηγική εκτίμησης του score είναι και αυτή που έχει συνδεθεί ως επί το πλείστον με την εκπαίδευση ενός NCSN. Υπενθυμίζουμε ότι ο στόχος, που δίδεται από την σχέση (2.41), είναι η από κοινού εκτίμηση των score όλων των διαταραγμένων

κατανομών δεδομένων. Αυτό, βάσει της (2.84) παίρνει την επακόλουθη μορφή:

$$\mathcal{L}(\theta; \{\sigma_t\}_{t=1}^T) := \frac{1}{T} \sum_{t=1}^T \lambda(\sigma_t) \mathcal{J}_{dsm}(\theta; q_{\sigma_t}) \quad (2.86)$$

$$= \frac{1}{T} \sum_{t=1}^T \frac{1}{2} \lambda(\sigma_t) \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0)} \mathbb{E}_{\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_t; \mathbf{x}, \sigma_t^2 \mathbf{I})} \left[ \|\mathbf{s}_\theta(\mathbf{x}_t, \sigma_t) + \frac{1}{\sigma_t^2} (\mathbf{x}_t - \mathbf{x}_0)\|_2^2 \right] \quad (2.87)$$

όπου  $\lambda(\sigma_t)$  είναι μια θετική συνάρτηση στάθμισης που εξαρτάται από το επίπεδο θορύβου  $\sigma_t$ . Οι Song και Ermon [7, 16] συχνά θέτουν  $\lambda(\sigma_t) = \sigma_t^2$  ως αντιστάθμιση στον παράγοντα  $\frac{1}{\sigma_t^2}$  που συναντάται στην αντικείμενική συνάρτηση.

#### 2.2.4 Langevin Dynamics

Κατόπιν εκπαίδευσης του NCSN, μπορεί να ακολουθήσει η διαδικασία της δειγματοληψίας. Όπως και στην εκπαιδευτική διαδικασία, έτσι και στην δειγματοληπτική δύναται να χρησιμοποιηθεί μια πληθώρα score-based προσεγγίσεων συμπεριλαμβανομένων στοχαστικών και σύνθητων διαφορικών εξισώσεων. Μία από τις πλέον κλασικές τέτοιες μεθόδους αποτελεί αυτή των Langevin Dynamics.

Η μέθοδος *Langevin Dynamics* πρόκειται, κατ' ουσίαν, για μία Markov Chain Monte Carlo, η οποία μπορεί να παραγάγει δείγματα από μία εκ των προτέρων γνωστή πυκνότητα πιθανότητας  $p(\mathbf{x})$  κάνοντας χρήση της συνάρτησης score της,  $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ . Έστω ένα σταθερό βήμα  $\alpha > 0$  και ένα αρχικό δείγμα  $\mathbf{x}_0 \sim p(\mathbf{x}_0)$  όπου  $p$  μια εκ των προτέρων κατανομή. Τότε η μέθοδος Langevin Dynamics αναδρομικά υπολογίζει για  $t = 1, \dots, T$ :

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \frac{\alpha}{2} \nabla_{\mathbf{x}} \log p(\mathbf{x}_{t-1}) + \sqrt{\alpha} \mathbf{z}_t \quad (2.88)$$

όπου  $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Ο πρώτος όρος της (2.88) αντιστοιχεί σε μια βαθμιδωτή άνοδο κατά μήκος του εκπαιδευμένου πεδίου score, ενώ ο δεύτερος όρος έγχυσης γκαουσιανού θορύβου προστίθεται για την εξασφάλιση της στοχαστικής εξερεύνησης της κατανομής και την αποτροπή κατάρρευσης σε κάποια ντετερμινιστική στάσιμη κατάσταση. Αποδεικνύεται ότι καθώς  $\alpha \rightarrow 0$  και  $T \rightarrow \infty$ , το τελικό παραχθέν δείγμα  $\mathbf{x}_T$  συγκλίνει στην πραγματική κατανομή δεδομένων  $p(\mathbf{x})$  [20].

Ωστόσο η άνωθεν κλασική μέθοδος του Langevin Monte Carlo δεν είναι αποτελεσματική σε περιπτώσεις που υπάρχουν περιοχές χαμηλής πυκνότητας στην κατανομή των δεδομένων· περιορισμός που, όπως σχολιάστηκε νωρίτερα, συναντάται σε μεγάλο μέρος των δεδομένων στον πραγματικό κόσμο (η Υπόθεση της Πολλαπλότητας). Για τον λόγο αυτό, οι Song και Ermon [7] πρότειναν μια παραλλαγή που ονόμασαν *Annealed Langevin Dynamics* (ALD), εμπνευσμένοι από την προσομοιωμένη ανόπτηση [21] και την ανόπτηση δειγματοληψίας σημασίας [22]. Επισήμως, η ALD ξεκινάει με την αρχικοποίηση ενός δείγματος  $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Κατόπιν, εκτελείται ο αλγόριθμος Langevin για την λήψη του δείγματος  $\mathbf{x}_1$  με την χρήση βήματος  $\alpha_1$  και την εκπαιδευμένη συνάρτηση score  $\mathbf{s}_\theta(\mathbf{x}_0, \sigma_1)$ . Ο ίδιος αλγόριθμος επαναλαμβάνεται, στην συνέχεια, για την λήψη του δείγματος  $\mathbf{x}_2$ , του  $\mathbf{x}_3$ , και ούτω καθεξής, έως ότου ληφθεί το τελικό

δείγμα  $\mathbf{x}_T$  το οποίο προσεγγίζει την πραγματική κατανομή των δεδομένων για  $\sigma_T \approx 0$ . Μαθηματικώς, για κάθε  $t = 1, \dots, T$  υπολογίζεται το:

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \frac{\alpha_t}{2} \mathbf{s}_\theta(\mathbf{x}_{t-1}, \sigma_t) + \sqrt{\alpha_t} \mathbf{z}_t \quad (2.89)$$

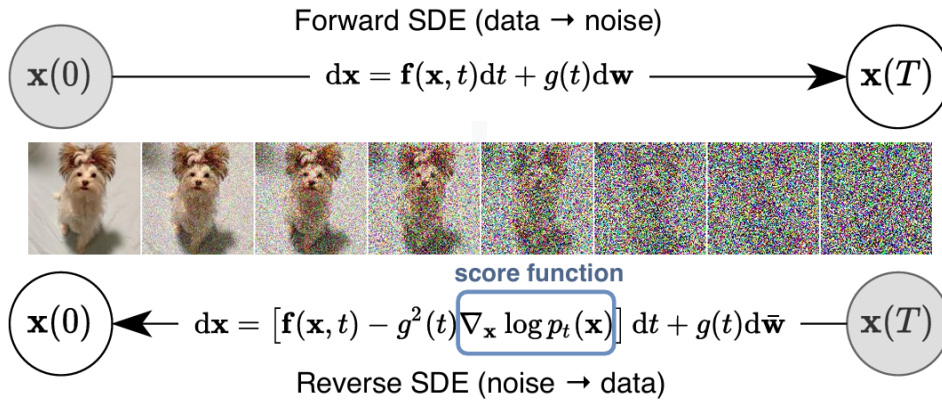
Το βήμα  $\alpha_t$ , σε κάθε επανάληψη, λαμβάνει τιμή σύμφωνα με την σχέση:

$$\alpha_t = \epsilon \cdot \frac{\sigma_t^2}{\sigma_T^2} \quad (2.90)$$

όπου  $\epsilon > 0$  μια υπερπαράμετρος κλιμάκωσης.

### 2.3 Stochastic Differential Equations

Στις δύο προηγούμενες ενότητες, εξετάστηκαν δύο διαφορετικές δημοφιλείς μορφές με τις οποίες μπορούν να εμφανισθούν τα μοντέλα διάχυσης—τα DDPMs και τα NCSNs. Και οι δύο περιπτώσεις περιλαμβάνουν την εκμάθηση της εγγενούς κατανομής δεδομένων ενός δείγματος μέσω μιας διαδικασίας σταδιακής εισαγωγής θορύβου σε ένα διακριτό, πεπερασμένο αριθμό χρονικών βημάτων. Προκειμένου να ενισχύσουν τις δυνατότητές τους, οι *Song et al.* [23] πρότειναν, το 2021, ένα πλαίσιο που ενοποιεί τις έννοιες των DDPMs και NCSNs, και τις γενικεύει στην περίπτωση των άπειρων χρονικών βημάτων ή επιπέδων θορύβου, κάνοντας χρήση στοχαστικών διαφορικών εξισώσεων. Ουσιαστικά, τόσο η προς τα εμπρός διαδικασία διάχυσης, όσο και η προς τα πίσω, μοντελοποιούνται ως χρονικά συνεχείς, αντιπροσωπεύοντας καθ' αυτόν τον τρόπο, λύσεις στοχαστικών διαφορικών εξισώσεων. Ονομάζουμε αυτήν τη διατύπωση των μοντέλων διάχυσης *score SDEs* καθώς αξιοποιεί στοχαστικές διαφορικές εξισώσεις (stochastic differential equations, SDEs) για τις διαδικασίες διαταραχής θορύβου, αποιορυβοποίησης, και δειγματοληψίας, ενώ η εκπαίδευση περατώνεται με την υλοποίηση κάποιας score matching τεχνικής.



**Σχήμα 2.3:** Η αρχή λειτουργίας των score-based παραγωγικών μοντέλων μέσω στοχαστικών διαφορικών εξισώσεων (score SDEs) [23].

Όπως και στις προηγούμενες ενότητες, ας ξεκινήσουμε την ανάλυσή μας από την μοντελοποίηση της *προς τα εμπρός διαδικασίας*. Έστω  $\mathbf{x}(0) \in \mathbb{R}^d$  ένα δείγμα δεδομένων

πυκνότητας πιθανότητας  $q(\mathbf{x}(0))$ . Η προς τα εμπρός SDE κατασκευάζει μια οικογένεια περιθωριακών, χρονικά εξαρτώμενων κατανομών  $\mathbf{x}(t) \sim q(\mathbf{x}(t))$  για συνεχές  $t \in [0, T]$  με αφετηρία την  $\mathbf{x}(0)$  μέσω της χρήσης μιας Itô SDE που περιγράφεται στην κάτωθι σχέση. Επίσης, για λόγους ευκολίας, από τούδε και στο εξής θα συμβολίζουμε το δείγμα δεδομένων την χρονική στιγμή  $t$ ,  $\mathbf{x}(t)$ , με  $\mathbf{x}_t$ :

$$d\mathbf{x}_t = f(\mathbf{x}_t, t)dt + g(t)d\mathbf{w}_t \quad (2.91)$$

Στην (2.91) η συνάρτηση  $f : \mathbb{R}^d \times [0, T] \rightarrow \mathbb{R}^d$  είναι ένα διάνυσμα που ονομάζεται συντελεστής μετατόπισης του  $\mathbf{x}_t$  (drift coefficient) και που μοντελοποιεί το ντετερμινιστικό μέρος της SDE, καθορίζοντας τον βαθμό στον οποίο αναμένεται να αλλάξει η διαδικασία  $d\mathbf{x}_t$  με την πάροδο του χρόνου. Η, δε, συνάρτηση  $g : [0, T] \rightarrow \mathbb{R}$  πρόκειται για τον βαθμωτό συντελεστή διάχυσης του  $\mathbf{x}_t$  (diffusion coefficient), που, με την σειρά της, αντιπροσωπεύει το στοχαστικό μέρος της SDE καθορίζοντας το εύρος της διαδικασίας θορυβοποίησης σε βάθος χρόνου. Τέλος, το  $\mathbf{w}_t$  είναι η τυπική διαδικασία Wiener (γνωστή και ως κίνηση Brown). Έτσι, ο όρος  $g(t)d\mathbf{w}_t$  της (2.91) απεικονίζει το σύνολο της διαδικασίας θορύβου.

Συναντώνται δύο κύριες οικογένειες των προς τα εμπρός SDEs: οι *Variance-Preserving (VP) SDEs* και οι *Variance-Exploding (VE) SDEs*. Συνδέονται, μάλιστα, άμεσα με τις μορφές των μοντέλων διάχυσης διακριτού χρόνου: των DDPMs και των SMLD (Score Matching με Langevin Dynamics), αντίστοιχα. Αποτελούν, στην ουσία, την έκφρασή τους σε συνεχή χρόνο.

Όσον αφορά τα VP SDEs ως συνεχή έκφραση των DDPMs, θα πρέπει πρώτα να υπενθυμίσουμε την εξίσωση (2.3) που περιγράφει τον πυρήνα μετάβασης  $q(\mathbf{x}_t | \mathbf{x}_{t-1})$  ως μία διακριτή μαρκοβιανή αλυσίδα με την εξής μορφή:

$$\mathbf{x}_t = \sqrt{1 - \beta_t} \mathbf{x}_{t-1} + \sqrt{\beta_t} \mathbf{z}_{t-1}, \quad \mathbf{z}_{t-1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (2.92)$$

Τώρα θα περάσουμε στο όριο συνεχούς χρόνου· θα θεωρήσουμε, δηλαδή, συνολικό πλήθος χρονικών βημάτων  $T \rightarrow \infty$ , και θα εισαγάγουμε ένα συνεχές χρόνο  $t \in [0, 1]$ . Σε αυτήν την περίπτωση μπορούμε να θεωρήσουμε κάθε υπερπαραμέτρο  $\beta_t = \beta(t) \Delta t$ , όπου  $\Delta t = \frac{1}{T}$  το στοιχειώδες χρονικό βήμα. Από την (2.92), θεωρώντας, πλέον, το διακριτό χρονικό βήμα  $t-1$  ως τη συνεχή χρονική στιγμή  $t$  και, αντιστοίχως, το διακριτό χρονικό βήμα  $t$  ως τη συνεχή χρονική στιγμή  $t + \Delta t$ , προκύπτει:

$$\mathbf{x}(t + \Delta t) = (\sqrt{1 - \beta(t + \Delta t)\Delta t}) \mathbf{x}(t) + \sqrt{\beta(t + \Delta t)\Delta t} \mathbf{z}(t) \quad (2.93)$$

Σε αυτό το σημείο μπορεί να χρησιμοποιηθεί το ανάπτυγμα Taylor ως εξής:

$$\sqrt{1 - \beta(t)} \approx 1 - \frac{1}{2}\beta(t) - \frac{1}{8}\beta^2(t) - \frac{1}{16}\beta^3(t) + \dots \approx 1 - \frac{1}{2}\beta(t) \quad (2.94)$$

Στην (2.94) κρατάμε μόνο τους δύο πρώτους όρους δεδομένου ότι οι υπόλοιποι γίνονται πολύ μικροί. Εξάλλου, υπενθυμίζουμε ότι  $\beta(t) \in (0, 1)$ . Άρα:

$$\mathbf{x}(t + \Delta t) \approx \left(1 - \frac{1}{2}\beta(t + \Delta t)\Delta t\right) \mathbf{x}(t) + \sqrt{\beta(t + \Delta t)\Delta t} \mathbf{z}(t) \quad (2.95)$$

$$\Leftrightarrow \mathbf{x}(t + \Delta t) \approx \mathbf{x}(t) - \frac{1}{2}\beta(t + \Delta t)\Delta t \mathbf{x}(t) + \sqrt{\beta(t + \Delta t)\Delta t} \mathbf{z}(t) \quad (2.96)$$



$$\Leftrightarrow \mathbf{x}(t + \Delta t) - \mathbf{x}(t) \approx -\frac{1}{2}\beta(t)\Delta t \mathbf{x}(t) + \sqrt{\beta(t)} \sqrt{\Delta t} \mathbf{z}(t) \quad (2.97)$$

Όμως στο όριο του συνεχούς χρόνου, καθώς  $\Delta t \rightarrow 0$ , τότε, επιπλέον,  $\mathbf{x}(t + \Delta t) - \mathbf{x}(t) \rightarrow d\mathbf{x}_t$ , ενώ το  $\mathbf{z}(t) \sqrt{\Delta t} \rightarrow d\mathbf{w}_t$ , το στοιχειώδες Brownian βήμα. Τελικά, η (2.97) συγκλίνει στο:

$$d\mathbf{x}_t = -\frac{1}{2}\beta(t) \mathbf{x}_t dt + \sqrt{\beta(t)} d\mathbf{w}_t \quad (2.98)$$

Αποδείχθηκε, συνεπώς, ότι η μοντελοποίηση ενός DDPM στον συνεχή χρόνο οδηγεί σε μια Itô στοχαστική διαφορική εξίσωση με συντελεστή μετατόπισης  $f(\mathbf{x}_t, t) = -\frac{1}{2}\beta(t) \mathbf{x}_t$  και συντελεστή διάχυσης  $g(t) = \sqrt{\beta(t)}$ . Αυτή η διαδικασία ικανοποιεί:

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon \quad (2.99)$$

με:

$$\epsilon = \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad \alpha_t = \exp\left(-\frac{1}{2} \int_0^t \beta(t) dt\right), \quad \sigma_t^2 = 1 - \alpha_t^2 \quad (2.100)$$

Επειδή  $\alpha_t^2 + \sigma_t^2 = 1$ , η διασπορά της  $\mathbf{x}_t$  που δίδεται στην (2.101) παραμένει περίπου σταθερή, δικαιολογώντας την ονομασία Variance-Preserving SDE:

$$\text{Var}[\mathbf{x}_t] = \alpha_t^2 \text{Var}[\mathbf{x}_0] + \sigma_t^2 \quad (2.101)$$

Από την άλλη πλευρά, για την εξαγωγή των VE SDEs ως συνεχή έκφραση των SMLD θα ακολουθηθεί μια παρόμοια διαδικασία. Στην προκειμένη περίπτωση ο πυρήνας μετάβασης δόθηκε στην (2.39) και κατόπιν εφαρμογής της τεχνικής της επαναπαραμετροποίησης γίνεται:

$$\mathbf{x}_t = \mathbf{x}_0 + \sigma_t^2 \mathbf{z}_t, \quad \mathbf{z}_t = \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (2.102)$$

Για να μοντελοποιήσουμε την SMLD ως μία διακριτή μαρκοβιανή αλυσίδα θα πρέπει να εκφράσουμε τον πυρήνα  $q(\mathbf{x}_t | \mathbf{x}_{t-1})$ . Προς αυτό το σκοπό, μπορούμε να εκφράσουμε τον γκαουσιανό θόρυβο  $\mathbf{z}_t$  συναρτήσει του  $\mathbf{z}_{t-1}$  κι ενός ανεξάρτητου γκαουσιανού θορύβου  $\mathbf{z}'_t$ :

$$\mathbf{z}_t = \frac{\sigma_{t-1}}{\sigma_t} \mathbf{z}_{t-1} + \sqrt{1 - \frac{\sigma_{t-1}^2}{\sigma_t^2}} \mathbf{z}'_t \quad (2.103)$$

Αντικαθιστώντας στην (2.102) προκύπτει:

$$\mathbf{x}_t = \mathbf{x}_0 + \sigma_{t-1} \mathbf{z}_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2} \mathbf{z}'_{t-1} \quad (2.104)$$

$$= \mathbf{x}_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2} \mathbf{z}'_{t-1} \quad (2.105)$$

Μετονομάζοντας το  $\mathbf{z}'_{t-1} \rightarrow \mathbf{z}_{t-1}$  παίρνουμε τον ζητούμενο πυρήνα μετάβασης:

$$\mathbf{x}_t = \mathbf{x}_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2} \mathbf{z}_{t-1} \quad (2.106)$$



Μπορούμε, πλέον, να περάσουμε στο όριο συνεχούς χρόνου· ορίζουμε, δηλαδή, εκ νέου πλήθος χρονικών βημάτων  $T \rightarrow \infty$ , συνεχές χρόνο  $t \in [0, 1]$ , και στοιχειώδες  $\Delta t = \frac{1}{T}$ . Ξαναγράφουμε την (2.106) ως εξής:

$$\mathbf{x}(t + \Delta t) = \mathbf{x}(t) + \sqrt{\sigma^2(t + \Delta t) - \sigma^2(t)} \mathbf{z}(t) \quad (2.107)$$

$$\approx \mathbf{x}(t) + \sqrt{\frac{d[\sigma^2(t)]}{dt} \Delta t} \mathbf{z}(t) \quad (2.108)$$

Ομοίως με προηγουμένως, καθώς  $\Delta t \rightarrow 0$ , ισχύει  $\mathbf{x}(t + \Delta t) - \mathbf{x}(t) \rightarrow d\mathbf{x}_t$  και  $\mathbf{z}(t) \sqrt{\Delta t} \rightarrow d\mathbf{w}_t$ . Συνεπώς η (2.105) μετατρέπεται σε:

$$d\mathbf{x}_t = \sqrt{\frac{d[\sigma_t^2]}{dt}} d\mathbf{w}_t \quad (2.109)$$

Αποδείχθηκε ότι η μοντελοποίηση ενός SMLD στον συνεχή χρόνο είναι μια Itô στοχαστική διαφορική εξίσωση με συντελεστή μετατόπισης  $f(\mathbf{x}_t, t) = \mathbf{0}$ , και συντελεστής διάχυσης  $g(t) = \sqrt{\frac{d[\sigma_t^2]}{dt}}$ . Στην προκειμένη περίπτωση, η λύση είναι:

$$\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \epsilon, \quad \epsilon = \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (2.110)$$

Η διασπορά της  $\mathbf{x}_t$  είναι:

$$\text{Var}[\mathbf{x}_t] = \text{Var}[\mathbf{x}_0] + \sigma_t^2 \quad (2.111)$$

Εφόσον η  $\sigma_t$  αυξάνεται (και συνήθως εκθετικά), η συνολική διασπορά “εκρύννεται”, εξού και ο χαρακτηρισμός Variance-Exploding SDE.

Έχοντας, σε αυτό το σημείο, ορίσει την προς τα εμπρός SDE και βρει τις παραμέτρους της για τις δύο κύριες οικογένειες με τις οποίες εμφανίζεται, επόμενο βήμα αποτελεί η εύρεση της αντίστροφης SDE. Ο *Anderson* [24] έδειξε ότι η SDE αντίστροφου χρόνου δίδεται από την:

$$d\mathbf{x}_t = [f(\mathbf{x}_t, t) - g^2(t) \nabla_{\mathbf{x}} \log q(\mathbf{x}_t)] dt + g(t) d\bar{\mathbf{w}}_t \quad (2.112)$$

όπου το  $d\bar{\mathbf{w}}_t$  είναι το στοιχειώδες Brownian βήμα, αλλά στην αντίθετη χρονική κατεύθυνση· ρέει, δηλαδή, από το  $T$  στο 0, ενώ το  $dt$  συμβολίζει το απειροελάχιστο αρνητικό χρονικό βήμα. Οι τροχιές λύσεων της αντίστροφης SDE μοιράζονται τις ίδιες περιθωριακές πυκνότητες  $q(\mathbf{x}_t)$  με αυτές της προς τα εμπρός SDE, με την διαφορά ότι πλέον εξελίσσονται στην αντίστροφη χρονική κατεύθυνση.

Παρατηρούμε ότι ο άγνωστος όρος που εμφανίστηκε στον συντελεστή μετατόπισης της αντίστροφης SDE δεν είναι παρά η χρονικά εξαρτώμενη συνάρτηση score της συνάρτησης πυκνότητας πιθανότητας του  $\mathbf{x}_t$ ,  $\nabla_{\mathbf{x}} \log q(\mathbf{x}_t)$ . Η βασική ιδέα ενός score SDE έγκειται στην εκπαίδευση εκτίμησης της εν λόγω συνάρτησης score, γεγονός που θα επέτρεπε την προσομοίωση της αντίστροφης SDE για την μετατροπή θορύβου σε δεδομένα. Παραμετροποιούμε ένα βαθύ νευρωνικό δίκτυο  $\mathbf{s}_\theta(\mathbf{x}_t, t)$  για να προσεγγίζει το score  $\mathbf{s}_q(\mathbf{x}_t, t) = \nabla_{\mathbf{x}} \log q(\mathbf{x}_t)$ . Εκτελούμε, ουσιαστικά, Score Matching· ήτοι η εκπαιδευτική διαδικασία των score SDEs ταυτίζεται με αυτήν των SGMs. Ως πρακτική επιλογή

επιλέγεται η αντικειμενική συνάρτηση (2.82) του *Denoising Score Matching*, ανηγμένη στον συνεχή χρόνο. Δηλαδή:

$$\mathcal{L}(\theta) := \mathbb{E}_{t \sim \mathcal{U}[0, T]} \left\{ \lambda(t) \mathbb{E}_{\mathbf{x}_0 \sim q(\mathbf{x}_0)} \mathbb{E}_{\mathbf{x}_t \sim q(\mathbf{x}_t | \mathbf{x}_0)} \left[ \frac{1}{2} \left\| \mathbf{s}_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}} \log q(\mathbf{x}_t | \mathbf{x}_0) \right\|_2^2 \right] \right\} \quad (2.113)$$

Ειδικά για την περίπτωση γκαουσιανής κατανομής  $\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \mathbf{z}_t$  τότε στην (2.113) μπορούμε να αντικαταστήσουμε:

$$\nabla_{\mathbf{x}} \log q(\mathbf{x}_t | \mathbf{x}_0) = -\frac{\mathbf{x}_t - \alpha_t \mathbf{x}_0}{\sigma_t^2} = -\frac{\mathbf{z}_t}{\sigma_t} \quad (2.114)$$

Κατόπιν εκπαίδευσης του χρονοεξαρτώμενου score-based εκτιμητή  $\mathbf{s}_\theta(\mathbf{x}_t, t)$ , μπορούμε να τον αντικαταστήσουμε αντί του αγνώστου  $\nabla_{\mathbf{x}} \log q(\mathbf{x}_t)$  στην SDE αντίστροφου χρόνου που δίδεται από την (2.112). Η εκπαιδευμένη score-based στοχαστική διαφορική εξίσωση δύναται να χρησιμοποιηθεί για την διαδικασία της δειγματοληψίας:

$$d\mathbf{x}_t = [f(\mathbf{x}_t, t) - g^2(t) \mathbf{s}_\theta(\mathbf{x}_t, t)] dt + g(t) d\bar{\mathbf{w}}_t \quad (2.115)$$

Η διαφορική εξίσωση (2.115) αποτελεί την γεννήτρια διαδικασία νέων δειγμάτων. Αρχικοποιώντας για  $t = T$  με  $\mathbf{x}_T \sim p(\mathbf{x}_T) \approx \mathcal{N}(\mathbf{0}, \mathbf{I})$  και ολοκληρώνοντας στην αντίστροφη χρονική κατεύθυνση έως ότου  $t = 0$ , μπορούμε να λάβουμε δείγματα κατά προσέγγιση κατανεμημένα σύμφωνα με την αρχική κατανομή δεδομένων  $q(\mathbf{x}_0)$ .

Στην πράξη, αυτή η SDE πρέπει να λυθεί αριθμητικά. Έστω μια γνησίως μονότονα φθίνουσα χρονική ακολουθία  $T = t_N > t_{N-1} > \dots > t_0 = 0$  με βήμα μεγέθους  $\Delta t = t_{i+1} - t_i > 0$ . Κάνοντας χρήση της τεχνικής διακριτοποίησης *Euler-Maruyama* (EM) για στοχαστικές διαφορικές εξισώσεις, δίδεται η προσέγγιση διακριτού χρόνου:

$$\mathbf{x}_{i-1} = \mathbf{x}_i + [f(\mathbf{x}_i, t_i) - g^2(t_i) \mathbf{s}_\theta(\mathbf{x}_i, t_i)] \Delta t + g(t_i) \sqrt{\Delta t} \mathbf{z}_i, \quad \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (2.116)$$

Το νέο δείγμα λαμβάνεται αντικαθιστώντας αναδρομικά στην (2.116) μέχρι  $i = 1$ . Σε αυτό το σημείο, αξία έχει να σταθούμε λίγο παραπάνω στην ειδική περίπτωση της VP SDE, όπου  $f(\mathbf{x}_i, t_i) = -\frac{1}{2} \beta(t_i) \mathbf{x}_i$  και  $g(t_i) = \sqrt{\beta(t_i)}$ . Η παραπάνω διακριτοποίηση καθίσταται ισοδύναμη με το αντίστροφο βήμα διάχυσης ενός DDPM. Πράγματι, αντικαθιστώντας την αναλυτική έκφραση του ακριβούς score της γκαουσιανής διαταραχής  $\mathbf{s}_q(\mathbf{x}_t) = -\frac{1}{\sigma_t^2} \epsilon$  (2.114), και τον όρο  $\beta(t_i) \Delta t$  με την διακριτή υπερπαραμέτρο  $\beta_i$ , ανακτούμε τον κλασσικό κανόνα ενημέρωσης του DDPM, όπου το μοντέλο προβλέπει τον θόρυβο  $\epsilon_\theta(\mathbf{x}_t, t)$ , αντί της συνάρτησης score. Καθ' αυτόν τον τρόπο, η στοχαστική αντίστροφη SDE προσφέρει ένα συνεχές θεωρητικό υπόβαθρο για το διακριτό DDPM.

Αν και η διακριτοποίηση Euler-Maruyama ορίζει μια έγκυρη γενετική τεχνική, η ακρίβειά της ελλοπαίνεται δραστικά σε περιπτώσεις μεγάλων χρονικών βημάτων ή αυστηρών προγραμμάτων θορύβου. Για την βελτίωση της ποιότητας των δειγμάτων, οι Song et al. [23], εισήγαγαν την έννοια των *Δειγματοληπτών Πρόβλεψης-Διόρθωσης* (Predictor-Corrector Samplers, PC). Ο ρόλος τους είναι να εμπλουτίζουν κάθε βήμα ολοκλήρωσης της SDE (2.116) (ο “Προβλέπτης”) με ένα στοχαστικό βήμα βελτίωσης (ο “Διορθωτής”). Στο πλαίσιο αυτό, ο προβλέπτης αντιστοιχεί σε κάποιον λύτη της αντίστροφης SDE, όπως η προαναφερθείσα τεχνική Euler-Maruyama. Ο διορθωτής, από

την άλλη πλευρά, εφαρμόζει μια σύντομη αλληλουχία ενημερώσεων κάποιας μεθόδου Markov Chain Monte Carlo (MCMC) για μια σταθερή χρονική στιγμή  $t_i$ , που σκοπό έχει να μετακινήσει στοχαστικά τα δείγματα προς περιοχές υψηλότερης πυκνότητας πιθανότητας, υπό την εκάστοτε εκτίμηση του score. Τον ρόλο του διορθωτή θα μπορούσε να διαδραματίσει οιαδήποτε μέθοδος MCMC. Έτσι, για παράδειγμα, στην περίπτωση του Annealed Langevin Dynamics, ο διορθωτής εκτελεί  $K$  βήματα:

$$\mathbf{x}_{t_i} \leftarrow \mathbf{x}_{t_i} + \frac{\alpha_{t_i}}{2} \mathbf{s}_\theta(\mathbf{x}_{t_i}, \sigma_{t_i}) + \sqrt{\alpha_{t_i}} \mathbf{z}_{t_i}, \quad \mathbf{z}_{t_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (2.117)$$

με μικρό βήμα  $\alpha_{t_i} > 0$ . Η μέθοδος αφήνει την επιθυμητή κατανομή  $p_{t_i}$  προσεγγιστικά αναλλοίωτη όταν η εκτίμηση  $\mathbf{s}_\theta(\mathbf{x}_{t_i}, \sigma_{t_i})$  είναι ακριβής, αντισταθμίζοντας καθ' αυτόν τον τρόπο την μεροληψία διακριτοποίησης του βήματος πρόβλεψης. Κατ' ουσίαν, ο συνδυασμός των μηχανισμών πρόβλεψης και διόρθωσης, εναλλάσσει προβλεπτικά βήματα που μετακινούν τα δείγματα στον χρόνο, με διορθωτικά βήματα, που τα αποθορυβώζουν τοπικά σε κάθε χρονικό επίπεδο.

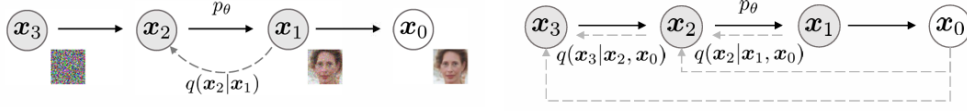
## 2.4 Denoising Diffusion Implicit Models

Σε αυτήν την ενότητα θα γίνει μια σύντομη αναφορά σε μια ειδική κατηγορία μοντέλων διάχυσης—τα *Denoising Diffusion Implicit Models* [25]—τα οποία αποτελούν, κατ' ουσίαν, μια επέκταση των προαναφερθέντων Denoising Diffusion Probabilistic Models (DDPMs). Σχεδιάστηκαν με σκοπό να ενισχύσουν την απόδοση της δημιουργίας νέων δεδομένων εισάγοντας μια μη-μαρκοβιανή δειγματοληπτική διαδικασία, διατηρώντας ταυτόχρονα τις υψηλής ποιότητας γενετικές δυνατότητες του αρχικού πλαισίου διάχυσης.

Τα βαθιά μοντέλα που έχουν ήδη εξεταστεί—τα DDPMs και τα NCSNs—έχουν, τα τελευταία χρόνια, παραγάγει εντυπωσιακά και ρεαλιστικά δείγματα αμφισβητώντας την μακροχρόνια πρωτοκαθεδρία των GANs στον χώρο της γενετικής μοντελοποίησης. Μολοταύτα, εμφανίζουν ένα κρίσιμο μειονέκτημα: είναι εξαιρετικά κοστοβόρα τόσο σε χρόνο όσο και σε υπολογιστικούς πόρους, απαιτώντας πολλές επαναλήψεις για να παραγάγουν ένα μεμονωμένο δείγμα. Ο λόγος έγκειται στο γεγονός ότι αυτές οι τεχνικές υιοθετούν μια δειγματοληπτική διαδικασία που προσομοιάζει την αντίστροφη διαδικασία διάχυσης, η οποία, ωστόσο, ενδέχεται να απαριθμεί χιλιάδες χρονικά βήματα.

Τα DDIMs υπήρξαν το αποτέλεσμα μιας προσπάθειας να μετριάσει ο εν λόγω περιορισμός και, ως εκ τούτου, συνδέονται στενά με τα DDPMs. Για την ακρίβεια, εκπαιδεύονται κάνοντας χρήση της ίδιας αντικειμενικής συνάρτησης, που δίδεται στην (2.36). Η κατ' εξοχήν παρατήρηση, που οδήγησε τελικά στην ανάπτυξή τους, είναι ότι η τελευταία εξαρτάται αποκλειστικά από τις περιθωριακές κατανομές  $q(\mathbf{x}_t | \mathbf{x}_0)$  και όχι άμεσα από την από κοινού κατανομή  $q(\mathbf{x}_{1:T} | \mathbf{x}_0)$ . Δεδομένου ότι υφίστανται πολλαπλές από κοινού κατανομές με τα ίδια περιθώρια, είναι δυνατόν να εξερευνηθούν εναλλακτικές τροχιές που είναι μη-μαρκοβιανές στην φύση τους, όπως διακρίνεται στην εικόνα 2.4.

Ας υποθέσουμε, λοιπόν,  $\mathcal{Q}$  μια οικογένεια κατανομών δεικτοδοτούμενες από  $\sigma \in$



Σχήμα 2.4: Γραφική αναπαράσταση μιας μαρκοβιανής δειγματοληπτικής διαδικασίας, όπως των DDPMs (αριστερά), και μιας μη-μαρκοβιανής δειγματοληπτικής διαδικασίας, όπως των DDIMs (δεξιά) [25].

$\mathbb{R}_{\geq 0}^T$  ένα πραγματικό, μη αρνητικό διάνυσμα διάστασης  $T$ :

$$q_{\sigma}(\mathbf{x}_{1:T} | \mathbf{x}_0) := q_{\sigma}(\mathbf{x}_T | \mathbf{x}_0) \prod_{t=2}^T q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \quad (2.118)$$

για την οποία ισχύουν:

$$\begin{cases} q_{\sigma}(\mathbf{x}_T | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_T; \sqrt{\alpha_T} \mathbf{x}_0, (1 - \alpha_T) \mathbf{I}) \\ q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{t-1}; \sqrt{\alpha_{t-1}} \mathbf{x}_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \frac{\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}_0}{\sqrt{1 - \alpha_t}}, \sigma_t^2 \mathbf{I}\right), \forall t > 1 \end{cases} \quad (2.119)$$

όπου  $\sigma_t$  μια υπερπαραμέτρος. Η επιλογή της μέσης τιμής στην δεύτερη εξίσωση της (2.119) επιλέγεται έτσι ώστε να εγγυάται για κάθε  $t$  ότι:

$$q_{\sigma}(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_0, (1 - \alpha_t) \mathbf{I}) \quad (2.120)$$

Αυτό είναι επιθυμητό έτσι ώστε να ορίζεται μια από κοινού καντανομή με επιθυμητά περιθώρια.

Η απόδειξη για την παραπάνω πρόταση είναι η εξής: Ας υποθέσουμε ότι για κάθε  $t < T$ , το  $q_{\sigma}(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_0, (1 - \alpha_t) \mathbf{I})$  ισχύει εάν

$$q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \sqrt{\alpha_{t-1}} \mathbf{x}_0, (1 - \alpha_{t-1}) \mathbf{I}) \quad (2.121)$$

Τότε μπορούμε να αποδείξουμε την πρόταση με ένα επαγωγικό επιχείρημα για  $t$  από  $T$  έως 1, δεδομένου ότι η βασική περίπτωση που  $t = T$  ισχύει ήδη. Κατ' αρχάς έχουμε ότι:

$$q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_0) := \int_{\mathbf{x}_t} q_{\sigma}(\mathbf{x}_t | \mathbf{x}_0) q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) d\mathbf{x}_t \quad (2.122)$$

και, επιπλέον:

$$\begin{cases} q_{\sigma}(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_0, (1 - \alpha_t) \mathbf{I}) \\ q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{t-1}; \sqrt{\alpha_{t-1}} \mathbf{x}_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \frac{\mathbf{x}_t - \sqrt{\alpha_t} \mathbf{x}_0}{\sqrt{1 - \alpha_t}}, \sigma_t^2 \mathbf{I}\right) \end{cases} \quad (2.123)$$

Από τον Bishop [26] γνωρίζουμε ότι η κατανομή  $q_{\sigma}(\mathbf{x}_{t-1} | \mathbf{x}_0)$  είναι γκαουσιανή και

συμβολίζεται με  $\mathcal{N}(\mathbf{x}_{t-1}; \mu_{t-1}, \Sigma_{t-1})$ , όπου:

$$\begin{cases} \mu_{t-1} = \sqrt{\alpha_{t-1}} \mathbf{x}_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \frac{\mathbf{x}_0 - \sqrt{\alpha_t} \mathbf{x}_0}{\sqrt{1 - \alpha_t}} = \sqrt{\alpha_{t-1}} \mathbf{x}_0 \\ \Sigma_{t-1} = \sigma_t^2 \mathbf{I} + \frac{1 - \alpha_{t-1} - \sigma_t^2}{1 - \alpha_t} (1 - \alpha_t) \mathbf{I} = (1 - \alpha_{t-1}) \mathbf{I} \end{cases} \quad (2.124)$$

Δηλαδή  $q_\sigma(\mathbf{x}_{t-1} | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \sqrt{\alpha_{t-1}} \mathbf{x}_0, (1 - \alpha_{t-1}) \mathbf{I})$  που μας επιτρέπει να χρησιμοποιήσουμε το επαγωγικό επιχείρημα.

Συνεχίζοντας την ανάλυσή μας, μπορούμε να εξαγάγουμε την προς τα εμπρός διαδικασία κάνοντας χρήση του νόμου του Bayes, δηλαδή:

$$q_\sigma(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0) = \frac{q_\sigma(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) q_\sigma(\mathbf{x}_t | \mathbf{x}_0)}{q_\sigma(\mathbf{x}_{t-1} | \mathbf{x}_0)} \quad (2.125)$$

η οποία είναι επίσης γκαουσιανή. Σε αντίθεση, όμως, με τον πυρήνα μετάβασης των DDPMs, που δίδεται από την (2.3), εδώ, η forward διαδικασία δεν είναι πλέον μαρκοβιανή, αφού κάθε κατανομή  $\mathbf{x}_t$  μπορεί να εξαρτάται τόσο από την  $\mathbf{x}_{t-1}$  όσο και από την  $\mathbf{x}_0$ . Το μέγεθος της υπερπαραμέτρου  $\sigma_t$  είναι αυτό που καθορίζει την στοχαστικότητα της προς τα εμπρός διαδικασίας. Μάλιστα, στην οριακή περίπτωση που  $\sigma_t \rightarrow 0$ , διαδικασία γίνεται πλήρως ντετερμινιστική: εφόσον μπορούμε να παρατηρήσουμε τα  $\mathbf{x}_0$  και  $\mathbf{x}_t$  για κάποιο  $t$ , τότε το  $\mathbf{x}_{t-1}$  γίνεται γνωστό και σταθερό.

Κατόπιν μοντελοποίησης της προς τα εμπρός διαδικασίας, ακολουθεί ο ορισμός μια εκπαιδεύσιμης διαδικασίας παραγωγής  $p_\theta(\mathbf{x}_{0:T})$ , στην οποία κάθε  $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t)$  κάνει χρήση της  $q_\sigma(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$ . Με απλούστερα λόγια, δεδομένου ενός θορυβώδους δείγματος  $\mathbf{x}_t$ , γίνεται κατ' αρχάς μια πρόβλεψη του  $\mathbf{x}_0$ . Στην συνέχεια, κατασκευάζεται το δείγμα  $\mathbf{x}_{t-1}$  με την χρήση της ορισμένης κατανομής  $q_\sigma(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$ . Η περιγραφείσα διαδικασία επαναλαμβάνεται αναδρομικά έως ότου κατασκευαστεί το δείγμα  $\mathbf{x}_0$ . Όσον αφορά την εκτίμηση της  $\mathbf{x}_0$  από το εκάστοτε δείγμα  $\mathbf{x}_0$ , αυτή μπορεί να γίνει μέσω της εξίσωσης (2.5), όπως διακρίνεται παρακάτω:

$$f_\theta(\mathbf{x}_t) := \frac{\mathbf{x}_t - \sqrt{1 - \alpha_t} \epsilon_\theta(\mathbf{x}_t)}{\sqrt{\alpha_t}} \quad (2.126)$$

Δύναται, πλέον, να οριστεί η παραγωγική διαδικασία δοθείσης μιας σταθερής πρότερης κατανομής  $p_\theta(\mathbf{x}_T) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ :

$$p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t) = \begin{cases} \mathcal{N}(f_\theta(\mathbf{x}_1), \sigma_1^2 \mathbf{I}) & , \text{αν } t = 1 \\ q_\sigma(\mathbf{x}_{t-1} | \mathbf{x}_t, f_\theta(\mathbf{x}_t)) & , \text{διαφορετικά} \end{cases} \quad (2.127)$$

όπου ο όρος  $q_\sigma(\mathbf{x}_{t-1} | \mathbf{x}_t, f_\theta(\mathbf{x}_t))$  δίδεται από την εξίσωση (2.125) αντικαθιστώντας το  $\mathbf{x}_0$  με την πρόβλεψή του,  $f_\theta(\mathbf{x}_t)$ . Επιπρόσθετα, ένας γκαουσιανός θόρυβος συνδιακύμανσης  $\sigma_1^2 \mathbf{I}$  προστίθεται όταν  $t = 1$  προκειμένου να διασφαλιστεί ότι η γενετική διαδικασία υποστηρίζεται παντού.

Βάσει, τώρα, της (2.127), η παραγωγή ενός δείγματος  $\mathbf{x}_{t-1}$  συναρτήσει του  $\mathbf{x}_t$  γίνεται

ως εξής:

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}} \left( \frac{\mathbf{x}_t - \sqrt{1 - \alpha_t} \epsilon_\theta(\mathbf{x}_t)}{\sqrt{\alpha_t}} \right) + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \epsilon_\theta(\mathbf{x}_t) + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad (2.128)$$

Στην (2.128), ο πρώτος όρος,  $\sqrt{\alpha_{t-1}} \left( \frac{\mathbf{x}_t - \sqrt{1 - \alpha_t} \epsilon_\theta(\mathbf{x}_t)}{\sqrt{\alpha_t}} \right)$  αποτελεί το προβλέψαν  $\mathbf{x}_0$ , ο δεύτερος όρος,  $\sqrt{1 - \alpha_{t-1} - \sigma_t^2} \epsilon_\theta(\mathbf{x}_t)$  δείχνει την κατεύθυνση προς το  $\mathbf{x}_t$ , ενώ, τέλος, ο τρίτος όρος,  $\sigma_t \epsilon$  είναι ένα τυχαίος εισαγώμενος θόρυβος. Είναι σημαντικό να σημειωθεί, ότι κάθε διαφορετική τιμή του  $\sigma_t$  οδηγεί σε διαφορετική τροχιά δειγματοληψίας, παρόλο που χρησιμοποιείται το ίδιο μοντέλο  $\epsilon_\theta$ . Ιδιαίτερο ενδιαφέρον παρουσιάζει ότι για την επιλογή  $\sigma_t = \sqrt{\frac{1 - \alpha_{t-1}}{1 - \alpha_t}} \sqrt{1 - \frac{\alpha_t}{\alpha_{t-1}}}$  για κάθε  $t$ , τότε η προς τα εμπρός διαδικασία γίνεται μαρκοβιανή, και, ως εκ τούτου, η παραγωγική διαδικασία ταυτίζεται με αυτή των DDPMs. Από την άλλη πλευρά, όπως αναφέρθηκε προηγουμένως, όταν  $\sigma_t = 0$ ,  $\forall t$ , η forward διαδικασία γίνεται πλήρως ντετερμινιστική, εξαιρουμένως, φυσικά, για  $t = 1$ . Οι Song et al. [25] ονόμασαν το μοντέλο που προκύπτει από αυτήν την τελευταία ειδική περίπτωση, *Denoising Diffusion Implicit Model* (DDIM), καθώς πρόκειται για ένα απόλυτο (“implicit”, ντετερμινιστικό) πιθανοτικό μοντέλο, το οποίο, εντούτοις, εκπαιδεύεται βάσει της ίδιας αντικειμενικής συνάρτησης με αυτήν των DDPMs.

Έως αυτό το σημείο, η παραγωγική διαδικασία εξακολουθεί να προσεγγίζει την αντίστροφη διαδικασία διάχυσης που ενδέχεται να απαριθμεί χιλιάδες χρονικά βήματα. Για την ελάττωσή τους, θεωρούμε μια forward διαδικασία η οποία δεν ορίζεται για ολόκληρη την ακολουθία παραμέτρων  $\mathbf{x}_{1:T}$ . αντ’ αυτού, ορίζεται για ένα υποσύνολό της  $\{\mathbf{x}_{\tau_1}, \dots, \mathbf{x}_{\tau_S}\}$ , όπου  $\tau$  μια αύξουσα υπακολουθία του  $\{1, \dots, T\}$  μήκους  $S$ . Κάνοντας χρήση της άνωθι υπακολουθίας, η εξίσωση (2.118) δύναται να παραγοντοποιηθεί ως εξής:

$$q_{\sigma, \tau}(\mathbf{x}_{1:T} | \mathbf{x}_0) := q_{\sigma, \tau}(\mathbf{x}_{\tau_S} | \mathbf{x}_0) \prod_{i=1}^S q_{\sigma, \tau}(\mathbf{x}_{\tau_{i-1}} | \mathbf{x}_{\tau_i}, \mathbf{x}_0) \prod_{t \in \bar{\tau}} q_{\sigma, \tau}(\mathbf{x}_t | \mathbf{x}_0) \quad (2.129)$$

όπου το  $\bar{\tau}$  συμβολίζει το συμπλήρωμα της  $\tau$ . Επιπροσθέτως, κατ’ αντιστοιχία με τις εξισώσεις της (2.123), ορίζονται:

$$\begin{cases} q_{\sigma, \tau}(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t} \mathbf{x}_0, (1 - \alpha_t) \mathbf{I}), \quad \forall t \in \bar{\tau} \cup \{T\} \\ q_{\sigma, \tau}(\mathbf{x}_{\tau_{i-1}} | \mathbf{x}_{\tau_i}, \mathbf{x}_0) = \mathcal{N}\left(\mathbf{x}_{\tau_{i-1}}; \sqrt{\alpha_{\tau_{i-1}}} \mathbf{x}_0 + \sqrt{1 - \alpha_{\tau_{i-1}} - \sigma_{\tau_i}^2} \cdot \frac{\mathbf{x}_{\tau_i} - \sqrt{\alpha_{\tau_i}} \mathbf{x}_0}{\sqrt{1 - \alpha_{\tau_i}}}, \sigma_{\tau_i}^2 \mathbf{I}\right), \quad \forall i \in [S] \end{cases} \quad (2.130)$$

όπου, πάλι, η επιλογή των συντελεστών γίνεται προκειμένου να οδηγούν στις επιθυμητές περιθώριες κατανομές, σύμφωνα με την σχέση:

$$q_{\sigma, \tau}(\mathbf{x}_{\tau_i} | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{\tau_i}; \sqrt{\alpha_{\tau_i}} \mathbf{x}_0, (1 - \alpha_{\tau_i}) \mathbf{I}), \quad \forall i \in [S] \quad (2.131)$$

Βάσει όλων των παραπάνω, μπορούμε να κατασκευάσουμε την παραγωγική διαδι-

κασία ως εξής:

$$p_\theta(\mathbf{x}_{0:T}) := p_\theta(\mathbf{x}_T) \prod_{i=1}^S p_\theta(\mathbf{x}_{t_{i-1}} | \mathbf{x}_{t_i}) \prod_{t \in \bar{\tau}} p_\theta(\mathbf{x}_t | \mathbf{x}_0) \quad (2.132)$$

όπου οι δεσμευμένες κατανομές δίδονται από τις σχέσεις:

$$\begin{cases} p_\theta(\mathbf{x}_{t_{i-1}} | \mathbf{x}_{t_i}) = q_{\sigma, \tau}(\mathbf{x}_{t_{i-1}} | \mathbf{x}_{t_i}, f_\theta(\mathbf{x}_{t_{i-1}})) \\ p_\theta(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(f_\theta(\mathbf{x}_t), \sigma_t^2 \mathbf{I}) \end{cases} \quad (2.133)$$

Στην εξίσωση (2.132), ο πρώτος όρος,  $p_\theta(\mathbf{x}_T) \prod_{i=1}^S p_\theta(\mathbf{x}_{t_{i-1}} | \mathbf{x}_{t_i})$ , αντιπροσωπεύει την παραγωγική διαδικασία και χρησιμοποιείται μόνος του κατά την διάρκειά της. Ο, δε, δεύτερος όρος,  $\prod_{t \in \bar{\tau}} p_\theta(\mathbf{x}_t | \mathbf{x}_0)$ , αντιπροσωπεύει την αντικειμενική συνάρτηση, η οποία, τελικά, γίνεται:

$$\mathcal{J} := \mathbb{E}_{\mathbf{x}_{0:T} \sim q_{\sigma, \tau}(\mathbf{x}_{0:T})} [\log q_{\sigma, \tau}(\mathbf{x}_{1:T} | \mathbf{x}_0) - \log p_\theta(\mathbf{x}_{0:T})] \quad (2.134)$$

$$\begin{aligned} &= \mathbb{E}_{\mathbf{x}_{0:T} \sim q_{\sigma, \tau}(\mathbf{x}_{0:T})} \left[ \sum_{t \in \bar{\tau}} \mathbb{D}_{KL}[q_{\sigma, \tau}(\mathbf{x}_t | \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_0 | \mathbf{x}_t)] \right. \\ &\quad \left. + \sum_{i=1}^L \mathbb{D}_{KL}[q_{\sigma, \tau}(\mathbf{x}_{t_{i-1}} | \mathbf{x}_{t_i}, \mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t_{i-1}} | \mathbf{x}_{t_i})] \right] \quad (2.135) \end{aligned}$$

Εν τέλει, αποδεικνύεται ότι η αντικειμενική συνάρτηση της (2.135) μπορεί να μετασχηματιστεί σε μια εξίσου αποδοτική αντικειμενικής συνάρτησης μορφής (2.36). Αυτό, πρακτικά, συνεπάγεται ότι ένα μοντέλο μπορεί να εκπαιδευτεί με έναν τυχαίο αριθμό χρονικών βημάτων κατά την προς τα εμπρός διαδικασία, και, στην συνέχεια, η παραγωγή νέων δειγμάτων να πραγματοποιηθεί κάνοντας χρήση μόνο ενός υποσυνόλου αυτών των βημάτων, επιταχύνοντας σημαντικά την δειγματοληψία.





## Κεφάλαιο 3

### Σχετική Έρευνα

---

Μη-λεκτικές πληροφορίες, όπως οι εκφράσεις του προσώπου, η γλώσσα του σώματος και οι χειρονομίες διαδραματίζουν καθοριστικό ρόλο στην αποδοτικότητα της ανθρώπινης επικοινωνίας. Ως εκ τούτου, έχουν αναδειχθεί σε σημαντικό παράγοντα κατά τον σχεδιασμό υπολογιστικών συστημάτων που αποσκοπούν στην αλληλεπίδραση με τους ανθρώπους με φυσικό τρόπο [1, 27]. Σε αυτό το Κεφάλαιο, θα ασχοληθούμε με την εξερεύνηση ορισμένων εξ' αυτών των καθιερωμένων συστημάτων, καθώς και τον τρόπο με τον οποίο επιτυγχάνουν την ενσωμάτωση των εν λόγω μη-λεκτικών πληροφοριών στην διαδικασία παραγωγής τρισδιάστατων κινουμένων σχεδίων του ανθρώπινου σώματος.

Πρώιμες έρευνες επί του θέματος εστίαζαν στην πρόβλεψη ή την παραγωγή ανθρώπινης κινησιολογίας δίχως να υλοποιούν κάποιου είδους πολυτροπικό έλεγχο. Τα τελευταία, ωστόσο, χρόνια, αυτή η πραγματικότητα αναιρείται, μέσω της εξερεύνησης της δυνατότητας εξάρτησης της παραγωγικής διαδικασίας από εναλλακτικές τροπικότητες, συμπεριλαμβανομένων κειμένου, ομιλίας, ετικετών δραστηριότητας—μέχρι και μουσικής [1]. Σε αυτά τα πλαίσια, θα εστιάσουμε την προσοχή μας σε μεθόδους κατευθυνόμενης από την ομιλία παραγωγής κίνησης, δεδομένου ότι αυτές είναι πιο σχετικές με το θέμα της παρούσης διπλωματικής εργασίας.

Σε γενικές γραμμές, τα τρισδιάστατα κινούμενα σχέδια σώματος δύναται να διακριθούν, ως προς την μέθοδο παραγωγής τους, σε δύο κατ' εξοχήν κατηγορίες: τα *rule-based* και τα *data-driven*. τα τελευταία εκ των οποίων μπορούν να χωριστούν, επιπλέον, σε *στατιστικές* και *learning-based* προσεγγίσεις. Τα *rule-based* συστήματα κάνουν χρήση μεθοδικά σχεδιασμένων ευρετικών (*heuristics*) για να επιλέξουν την εκάστοτε κατάλληλότερη κίνηση. Τουναντίον, τα *data-based* μαθαίνουν να επιλέγουν την αντίστοιχη κίνηση για κάθε περίπτωση από δεδομένα. Πιο συγκεκριμένα, τα *στατιστικά* συστήματα, συνήθως, προϋπολογίζουν πιθανότητες ή εκχωρούν μια εκ των προτέρων κατανομή σε δεδομένα κίνησης, και εν συνεχεία δειγματοληπτούν δείγματα από την εν λόγω κατανομή. Από την άλλη πλευρά, τα *learning-based* συστήματα κάνουν τις λίγότερες υποθέσεις για τις κατανομές των δεδομένων κινήσεων· αντ' αυτού, βελτιστοποιούν παραμέτρους μίας περίπλοκης μη-γραμμικής συνάρτησης, συνήθως ενός βαθύς νευρωνικού δικτύου, προκειμένου να αντιστοιχήσουν την κατάλληλη κίνηση στην εκάστοτε είσοδο [28]. Ας εντυπώσουμε περεταίρω στις εν λόγω κατηγορίες και τις *speech-driven*

προσεγγίσεις που υπάγονται σε καθεμία εξ' αυτών.

### 3.1 Rule-Based Προσεγγίσεις

Όσον αφορά τα *rule-based* συστήματα, αυτά εμφανίζονται, ως επί το πλείστον, με τη μορφή ενσαρκωμένων συνομιλιακών πρακτόρων (*Embodied Conversational Agents, ECA*), που συνιστούν ένα από τα σημαντικότερα επιτεύγματα στον τομέα της αλληλεπίδρασης του ανθρώπου με την μηχανή. Ένας ECA δεν περιορίζεται στην λεκτική επικοινωνία, αλλά σχεδιάζεται με σκοπό να εσωματώνει πολυτροπικά στοιχεία επικοινωνίας, όπως εκφράσεις του προσώπου, κινήσεις των χεριών, και στάση του σώματος, με απώτερο σκοπό να μιμηθεί βασικές πτυχές της ανθρώπινης επικοινωνίας και να επιτύχει, καθ' αυτόν τον τρόπο, μια πιο φυσική, πειστική και συναισθηματικά εκφραστική επικοινωνία [29].

Η εισαγωγή της *Behavior Markup Language (BML)* [30] αποτέλεσε ένα καθοριστικό βήμα προς την δημιουργία ενός κοινού πλαισίου περιγραφής της συμπεριφοράς για τους ECAs. Η BML σχεδιάστηκε έτσι ώστε να επιτρέπει την τυποποιημένη περιγραφή πολυτροπικών συμπεριφορών, ανεξάρτητα από την εκάστοτε πλατφόρμα ή μηχανισμό απόδοσης κίνησης. Με άλλα λόγια, λειτουργεί ως ένα ενδιάμεσο επίπεδο επικοινωνίας ανάμεσα στο σύστημα που αποφασίζει τι πρέπει να εκφράσει ο πράκτορας (*behavior planner*) και το σύστημα που υλοποιεί την ίδια την κίνηση (*behavior realizer*). Η γλώσσα παρέχει ένα σύνολο εντολών και χρονικών περιορισμών που επιτρέπουν τον ακριβή συγχρονισμό ανάμεσα σε διαφορετικά εκφραστικά κανάλια.

Κάνοντας χρήση της BLM, δύναται κανείς να κατασκευάσει *rule-based* συστήματα ανθρωποειδών βασισμένα σε προκαθορισμένες συμπεριφορές. Χαρακτηριστικό παράδειγμα αποτελεί το *BEAT (Behavior Expression Animation Toolkit)* [31], ένα σύνολο *rule-based* εργαλείων που αναπτύχθηκε με σκοπό την αυτόματη προσθήκη μη-λεκτικών συμπεριφορών πάνω σε μία προϋπάρχουσα, προ-ανιμαρισμένη φιγούρα. Αναλύει την ομιλία εισόδου, οπότε και εντοπίζει την σημασιολογική της δομή και επιλέγει αναλόγως τις κατάλληλες εκφράσεις. Το *BEAT* αποτέλεσε ένα εκ των πρώτων συστημάτων που επέδειξαν την σημασία της συντονισμένης πολυτροπικής έκφρασης σε ψηφιακούς πράκτορες.

Στο πλαίσιο αυτό, οι *Poggi et al.* [29] ανέπτυξαν τον πράκτορα *GRETA*. Η αρχιτεκτονική του βασίζεται στην λογική του πλαισίου *SAIBA (Situation-Affect-Intention-Behavior-Animation)*, οπότε και η συμπεριφορά του προκύπτει κατόπιν επεξεργασίας των προθέσεων, των συναισθημάτων, και της εκάστοτε κατάστασης της αλληλεπίδρασης, οδηγώντας, εν τέλει, στην δημιουργία BLM εντολών που μεταφράζονται σε κίνηση και έκφραση.

Την ίδια στιγμή, οι *Thiebaux et al.* [32] δημιούργησαν έναν ECA επιστρατεύοντας τεχνικές διαδικαστικού animation και *keyframe interpolation*—ένα γνωστό και καθιερωμένο *task* που αφορά την παραγωγή των ενδιάμεσων απόντων καρέ, δεδομένων δύο διαδοχικών keyframes. Η μέθοδος αυτή επιτρέπει στον πράκτορα να δημιουργεί φυσικές κινήσεις, σε πραγματικό χρόνο, βασιζόμενος σε προκαθορισμένα “σημεία-κλειδιά” της κίνησης, δίχως την ανάγκη συνεχούς αναπαράστασης ή αποθήκευσης ολόκληρων ακολουθιών κίνησης. Καθ' αυτόν τον τρόπο, η δημιουργία εκφραστικών κινήσεων γίνεται πιο αποδοτική, ενώ διατηρείται η δυνατότητα ελέγχου μέσω κανόνων.

Οι Marsella *et al.* [33], από την άλλη, πρότειναν έναν πιο γενικευμένο *rule-based* πράκτορα, σχεδιασμένο να δημιουργεί εκφράσεις, χειρονομίες, ακόμη και να κατευθύνει το βλέμμα του, απευθείας από το περιεχόμενο της ομιλίας. Το σύστημά τους επιχειρεί να μοντελοποιήσει τη σχέση μεταξύ προφορικού λόγου και μη λεκτικής συμπεριφοράς, παράγοντας συνεπείς και εκφραστικές αντιδράσεις. Μέσω ενός συνόλου προκαθορισμένων κανόνων, ο πράκτορας μπορεί να προσαρμόζει το βλέμμα, τις χειρονομίες ή τη στάση του σώματος, ώστε να αντικατοπτρίζουν τη συναισθηματική κατάσταση ή τη σημασιολογική έμφαση του λόγου.

Όλες αυτές οι προαναφερθείσες προσεγγίσεις έχουν ως κοινό παρονομαστή το γεγονός ότι βασίζονται σε μη-εκπαιδεύσιμες τεχνικές, οι οποίες απαιτούν εκτεταμένη χειροκίνητη μοντελοποίηση και προσεκτικό καθορισμό των κανόνων συμπεριφοράς. Αν και τα *rule-based* συστήματα προσφέρουν υψηλό επίπεδο ελέγχου και προβλεψιμότητας, η προσαρμογή τους σε νέες καταστάσεις ή σενάρια αλληλεπίδρασης παραμένει ιδιαίτερα χρονοβόρα και δύσκολη. Παρ' όλα αυτά, αποτελούν τη βάση πάνω στην οποία αναπτύχθηκαν οι μεταγενέστερες, πιο ευέλικτες *data-driven* προσεγγίσεις που αξιοποιούν δεδομένα και μηχανική μάθηση για τη σύνθεση εκφραστικής συμπεριφοράς.

### 3.2 Data-Based Προσεγγίσεις

Τα τελευταία χρόνια, οι μέθοδοι βασισμένες σε δεδομένα έχουν σταδιακά αντικαταστήσει τα παραδοσιακά *rule-based* συστήματα, εισάγοντας μεγαλύτερη ευελιξία, προσαρμοστικότητα και δυνατότητα γενίκευσης—ικανότητας, η έλλειψη των οποίων, καθώς είδαμε στην προηγούμενη ενότητα, ταλάνιζε τα τελευταία. Πλέον, αντί να στηρίζονται σε ρητούς κανόνες, τα *data-driven* μοντέλα μαθαίνουν αυτόματα τις σχέσεις μεταξύ ομιλίας, κειμένου και κινήσεων μέσα από εκτενή σύνολα δεδομένων, ελαττώνοντας την ανάγκη χειροκίνητης μοντελοποίησης και καθιστώντας δυνατή την δημιουργία πιο φυσικών, πιο ρεαλιστικών, πιο ευέλικτων κινήσεων.

Μία από τις πρώτες σημαντικές προσεγγίσεις αποτελεί εκείνη των Yoon *et al.* [34], οι οποίοι συνδύασαν τρεις διαφορετικές μορφές πληροφορίας—το κείμενο, τον ήχο και τις κινήσεις του άνω μέρους του σώματος—προκειμένου να εκπαιδεύσουν ένα *avatar* να ακολουθεί συντονισμένες κινήσεις. Το μοντέλο μαθαίνει τη συσχέτιση μεταξύ των λεκτικών και ακουστικών χαρακτηριστικών της ομιλίας και των αντίστοιχων κινήσεων, επιτυγχάνοντας μια αρκετά ρεαλιστική απόδοση. Ωστόσο, ο έλεγχος του στυλ παραμένει περιορισμένος, καθώς η προσαρμογή γίνεται μόνο στο επίπεδο μεμονωμένων ομιλητών μέσω δειγματοληψίας από τον λανθάνοντα χώρο τους, χωρίς να υπάρχει ρητός μηχανισμός καθοδήγησης.

Στο ίδιο μήκος κύματος κυμαίνονται και οι Liang *et al.* [35], οι οποίοι πρότειναν ένα σύστημα που αξιοποιεί συνδυαστικά την σημασιολογική πληροφορία του κειμένου και τα ακουστικά χαρακτηριστικά, με σκοπό να διαχωρίσει και να συνθέσει δύο διαφορετικούς τύπους χειρονομιών—τις “beat gestures”, που ακολουθούν τον ρυθμό και τη δυναμική της ομιλίας, και τις “semantic gestures”, που μεταφέρουν νοηματικό περιεχόμενο. Η διάκριση αυτή αποδείχθηκε καθοριστική για τη βελτίωση της χρονικής ευθυγράμμισης και της φυσικότητας της κίνησης.

Οι Ahuja *et al.* [36] εισήγαγαν ένα πιο σύνθετο πλαίσιο βασισμένο σε *adversarial*

*learning* και τη χρήση *multi-scale transformers*, προκειμένου να μάθουν τη λανθάνουσα σχέση μεταξύ λόγου και κίνησης. Απώτερος στόχος του εν λόγω συστήματος αποτελεί το να ευθυγραμμίσει υπο-λεκτικές (*sub-word*) ακουστικές ενδείξεις με χειρονομίες σε διαφορετικές χρονικές κλίμακες, επιτυγχάνοντας καλύτερη ισορροπία μεταξύ ακρίβειας και ποικιλίας της παραγόμενης κίνησης.

Αντίστοιχα, οι *Kucherenko et al.* [37] πρότειναν ένα *autoregressive* μοντέλο που ενσωματώνει σημασιολογικά χαρακτηριστικά βασισμένα στο *BERT* [38], σε συνδυασμό με ακουστικά χαρακτηριστικά, για τη δημιουργία κινήσεων που συνδέονται τόσο με το περιεχόμενο της ομιλίας όσο και με τη φωνητική της ροή. Η *autoregressive* φύση του επιτρέπει στο μοντέλο να λαμβάνει υπ' όψιν τόσο το παρελθόν όσο και το μέλλον της κίνησης, διασφαλίζοντας, καθ' αυτόν τον τρόπο, ομαλή συνέχεια και χρονική συνέπεια στην παραγωγή. Εξίσου εμπνευσμένοι από τα αυτοπαλίνδρομα παραγωγικά μοντέλα (*autoregressive generative models*), οι *Korzun et al.* [39] εισήγαγαν την χρήση ενός *recurrent context encoder* για την δημιουργία ρευστών και φυσικών κινήσεων, στο πλαίσιο του *GENEA Challenge 2020* [40].

Η ενοποίηση των διαφορετικών μορφών πληροφορίας σε ένα ενιαίο πλαίσιο επιτυγχάνεται από το *Diff-TTSG* [41], το οποίο προβλέπει ταυτόχρονα τόσο την ομιλία όσο και τις κινήσεις. Με τον τρόπο αυτό, οι δύο μορφές επικοινωνίας μαθαίνονται από κοινού, οδηγώντας σε πιο συνεκτικές και φυσικές αλληλεπιδράσεις. Το *DiffGAN* [42], από την άλλη, εστιάζει περισσότερο στη μεταφορά χειρονομιών μεταξύ διαφορετικών ομιλητών (*gesture retargeting*), επιτρέποντας τη διατήρηση του στυλ ή των κινήσεων ενός ομιλητή σε περιβάλλοντα με περιορισμένα δεδομένα (*low-resource*).

Κατά τα πιο πρόσφατα χρόνια, η εισαγωγή των μοντέλων διάχυσης έχουν μονοπωλήσει το ενδιαφέρον της σχετικής έρευνας, δεδομένου ότι έχουν αποδειχθεί εξαιρετικά αποτελεσματικά στην μοντελοποίηση πολύπλοκων πολυτροπικών σχέσεων. Ένα χαρακτηριστικό παράδειγμα αποτελεί το βαθύ μοντέλο *AMUSE* των *Chhatre et al.* [1]. Αυτό συνδυάζει δύο ξεχωριστά εκπαιδευμένα υποσυστήματα: ένα *Audio Disentanglement Module*, που κωδικοποιεί την ομιλία σε λανθάνουσες αναπαραστάσεις περιεχομένου, συναισθήματος και ύφους, και ένα *latent diffusion model*, το οποίο, ξεκινώντας από τυχαίο θόρυβο, παράγει ρεαλιστικές ακολουθίες ανθρώπινης κίνησης. Το *AMUSE* επιτρέπει έτσι τη διαχωρισμένη αλλά ταυτόχρονα ελεγχόμενη παραγωγή κινήσεων που ανταποκρίνονται τόσο στον ρυθμό όσο και στη συναισθηματική χροιά της ομιλίας.

Αντίστοιχα, το *DiffSHEG* των *Chen et al.* [27] εισάγει ένα ενοποιημένο πλαίσιο που συνδυάζει την παραγωγή χειρονομιών και εκφράσεων προσώπου, αξιοποιώντας ένα *expression-gesture denoising network* βασισμένο στην αρχή λειτουργίας της διάχυσης. Η αρχιτεκτονική του περιλαμβάνει έναν *audio encoder* και έναν *Transformer-based Unidirectional Expression-Gesture Generator*, που επιτρέπει ροή πληροφορίας από την έκφραση προς τη χειρονομία, εξασφαλίζοντας έτσι συνέπεια και εκφραστική αρμονία. Επιπρόσθετα, η προτεινόμενη μέθοδος *Fast Out Painting-based Partial Autoregressive Sampling* (*FOPPAS*) επιτρέπει τη γρήγορη παραγωγή ακολουθιών σε πραγματικό χρόνο χωρίς εξάρτηση από προηγούμενα καρέ. Αυτό καθιστά το *DiffSHEG* ένα από τα πιο προηγμένα και ολοκληρωμένα *diffusion-based* συστήματα έως σήμερα.

## Κεφάλαιο 4

### Περιγραφή του Προβλήματος

---

#### 4.1 Αντίληψη του Προβλήματος

Όπως διαπιστώθηκε στο προηγούμενο κεφάλαιο, η έρευνα επί της παραγωγής τρισδιάστατων animations τα τελευταία χρόνια είναι ιδιαίτερος εκτενής. Οι πρόσφατες εξελίξεις, ειδικά στον τομέα της παραγωγής χειρονομιών και γενικότερα της κινησιολογίας του σώματος, έχουν αναδείξει εντυπωσιακά αποτελέσματα, επιδεικνύοντας αυξανόμενα πιο ποικιλόμορφες κινήσεις, επιτρεπτές από την ανθρώπινη φυσιολογία, μεγαλύτερο συγχρονισμό μεταξύ της ομιλίας και της κίνησης των χειριών κι εν γένει βελτίωση σε διάφορους παράγοντες που καθιστούν ένα 3D animation ρεαλιστικό.

##### 4.1.1 Το Φαινόμενο της Απόκοσμης Κοιλιάδας

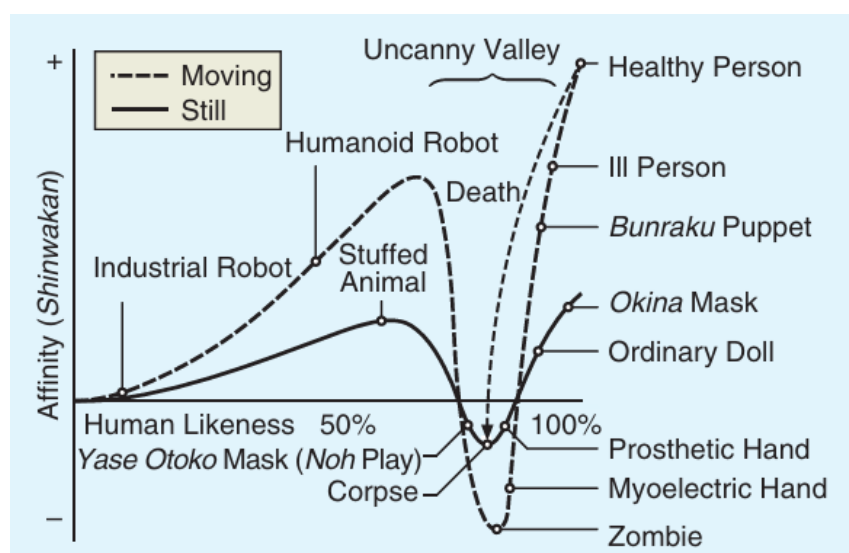
Ωστόσο, παρά την τεχνική πολυπλοκότητα που έχει επιτευχθεί, πολλά από τα παραγόμενα animations προκαλούν μια λεπτή αλλά διακριτή αίσθηση ανησυχίας ή δυσφορίας στους θεατές. Για παράδειγμα η εικόνα 4.1 δημιουργήθηκε με την χρήση του μοντέλου *DiffSHEG* [27]. Τα αποτελέσματα είναι αδιαφιλονίκητα εντυπωσιακά, και οι χαρακτήρες που απεικονίζονται διαπνέονται από υψηλό βαθμό ρεαλισμού. Μολαταύτα, ορισμένοι από αυτούς ενδέχεται να εμπνέουν κάποιο απόκοσμο συναίσθημα σε ορισμένους ανθρώπους. Αυτή η αντίδραση είναι φυσιολογική και συνδέεται με ένα ψυχολογικό φαινόμενο γνωστό ως “*Uncanny Valley*” (απόκοσμη κοιλάδα). Ο όρος, που επινοήθηκε από τον *Masahiro Mori* το 1979 [43], χρησιμοποιείται για να περιγράψει την παρατήρηση ότι τεχνητά αντικείμενα που μοιάζουν με ανθρώπους, αλλά κάτι φαντάζει λάθος πάνω τους, μπορούν να προκαλέσουν αρνητικά συναισθήματα σε έναν παρατηρητή.



Σχήμα 4.1: Παραχθέντα δείγματα του μοντέλου *DiffSHEG* [27].



Αυτή ακριβώς η έννοια είναι υπεύθυνη για τον περιστασιακό χαρακτηρισμό ανθρωποειδών ρομπότ, κουκλών ή κινούμενων σχεδίων δημιουργημένων στον ηλεκτρονικό υπολογιστή ως “παράξενων” ή “ανατριχιαστικών” [44]. Κρίσιμο, μάλιστα, για την περίπτωση μας είναι το γεγονός ότι ο Mori παρατήρησε, επίσης, πως το φαινόμενο της απόκοσμης κοιλάδας εντείνεται όταν τα αντικείμενα με τα οποία ερχόμαστε ενώπιον είναι κινούμενα, όπως διακρίνεται στο διάγραμμα της εικόνας 4.2.



Σχήμα 4.2: Το αρχικό διάγραμμα του Masahiro Mori [43], που απεικονίζει το φαινόμενο της απόκοσμης κοιλάδας (uncanny valley). Όσο αυξάνεται ο ρεαλισμός ενός αντικειμένου, αυξάνεται και η συμπάθεια του παρατηρητή προς αυτό. Ωστόσο, η ανοδική τάση διακόπτεται καθώς το αντικείμενο προσεγγίζει σε ρεαλισμό τον άνθρωπο· οπότε και η συμπάθεια ελαττώνεται απότομα. Επιπλέον, αυτό το φαινόμενο εντείνεται περαιτέρω στην περίπτωση των κινούμενων αντικειμένων, όπως υποδεικνύει η διακεκομμένη γραμμή.

Εγείρεται, συνεπώς, η εύλογη απορία γιατί τα προαναφερθέντα μοντέλα, παρά την πολυπλοκότητά τους, όλα προκαλούν το φαινόμενο της απόκοσμης κοιλάδας. Για να απαντήσουμε στο εν λόγω ερώτημα, αρκεί να επιθεωρήσουμε τα δεδομένα στα οποία έχουν εκπαιδευτεί.

Ένα από τα πλέον δημοφιλή datasets για αυτόν τον σκοπό αποτελεί το BEAT από τους Liu et al. [45]. Με 76 ώρες συνολικής διάρκειας δεδομένων, αποτελεί, έως την συγγραφή της παρούσης διπλωματικής εργασίας, το μεγαλύτερο motion-capture, speech-to-motion dataset, ενσωματώνοντας ταυτόχρονα κείμενο, ταυτότητες των ομιλητών, συναισθήματα, και σημασιολογία. Ακολούθησαν τα βήματα προηγούμενων εργασιών, όπως των Takeuchi et al. [46] και Trinity [47], που επίσης χρησιμοποίησαν τεχνολογία motion-capture για να παράσχουν τρισδιάστατα keypoints του σώματος. Για την αποτύπωση, δε, των εκφράσεων το προσώπου, έχουν επιστρατευτεί τα τελευταία χρόνια μια ευρεία ποικιλία τεχνικών, συμπεριλαμβανομένων πολυκαμερών (multi-cameras) από το Multiface [48], 3D-scanner από το BIWI [49], και 4D-scan από το VOCASET [50], εκ των οποίων όλα ανέδειξαν υψηλής ποιότητας τρισδιάστατα πλέγματα (3D mesh) του προσώπου. Παράλληλα, η σχετικά ακριβή και εργατικά εντατική φύση των άνωθεν μεθόδων, έχει οδηγήσει άλλους ερευνητές να διερευνήσουν εναλλακτικές

μεθόδους. Η χρήση διαφορετικών learning-based τεχνικών, που δημιουργούν ψευδο-δεδομένα αναφοράς (*pseudo-ground truth*, p-GT) για την κίνηση του σώματος από *in-the-wild* βίντεο, έχει ανοίξει τον δρόμο για πιο εκτενή και ποικίλα datasets. Οι *Ginosar et al.* [51] καινοτόμησαν δημιουργώντας ένα γιγαντιαίο dataset συνολικής διάρκειας 144 ωρών, το *Speech2Gesture*, που περιελάμβανε βίντεο παρουσιαστών τηλεοπτικών εκπομπών, λεκτόρων, και τηλεαγγελιστών, των οποίων είχαν εξαγάγει τα δισδιάστατα *pose estimations* των χεριών και σωμάτων τους. Αργότερα, οι *Habibie et al.* [52] χρησιμοποίησαν τα βίντεο του *Speech2Gesture* για να προτείνουν την πρώτη προσέγγιση αυτόματης εξαγωγής δεδομένων 3D παραμέτρων του προσώπου, 3D στάση κεφαλής, 3D χεριών και 3D χειρονομιών του άνω μέρους του σώματος. Τέλος, οι *Yi et al.* [53] χρησιμοποίησαν μια fine-tuned έκδοση του *SMPLify-X* [54] για να μετατρέψουν τα βίντεο των *Habibie et al.* σε ένα νέο dataset ολιστικών τρισδιάστατων πλεγμάτων, το *SHOW*, συνολικής διάρκειας 27 ωρών.

Τα datasets που μόλις παρουσιάστηκαν εν συντομία, αν και διαφέρουν αρκετά ως προς την μεθοδολογία τους, μοιράζονται ως κοινό χαρακτηριστικό την παροχή δεδομένων υψηλής ποιότητας. Καθένα εξ' αυτών επιμελήθηκε με προσοχή χρησιμοποιώντας μια ευρεία ποικιλία διαφορετικών τεχνικών—από χειροκίνητη σημειογραφία έως προηγμένα αυτοματοποιημένα συστήματα—προκειμένου να διασφαλιστεί η αξιοπιστία και η εφαρμοσιμότητά τους σε πολλαπλούς τομείς. Ωστόσο, παρά την εκ πρώτης όψεως έλλειψη ομοιοτήτων στις διαδικασίες σύνθεσής τους, παρουσιάζουν έναν κρίσιμο κοινό παρονομαστή: βασίζονται σε βίντεο πραγματικών ανθρώπων. Όταν μοντέλα για την δημιουργία τρισδιάστατων κινουμένων σχεδίων, όπως αυτά που αναφέρθηκαν στο Κεφάλαιο 3, εκπαιδεύονται σε τέτοια σύνολα δεδομένων, στοχεύουν στην ελαχιστοποίηση μετρικών, όπως των *αποστάσεων Fréchet* των κινήσεων, των χειρονομιών και των εκφράσεων ή της *ανάκλησης σημασιολογικά σχετικών κινήσεων* (Semantic-Relevant Gesture Recall, SRGR). Με άλλα λόγια, επιχειρούν να φθάσουν στο ground truth. Περιορισμοί, όμως, όπως πεπερασμένα και συχνά θορυβώδη δεδομένα εκπαίδευσης, μοντελοποιήσεις περιορισμένης εκφραστικής ικανότητας, αλλά και η χαοτική και απρόβλεπτη φύση της ανθρώπινης κινησιολογίας, συνεπάγονται ότι, ακόμη και με τέλεια εκπαίδευση, πάντα θα υπάρχει ένα υπόλοιπο σφάλμα σε αυτό τους το εγχείρημα να προσεγγίσουν τον άνθρωπο, καθιστώντας εξαιρετικά πιθανό το ενδεχόμενο να προκαλέσουν το φαινόμενο της απόκοσμης κοιλάδας.

#### 4.1.2 Οι Δώδεκα Αρχές των Κινουμένων Σχεδίων

Η λύση για το παραπάνω πρόβλημα μπορεί να βρεθεί εξετάζοντας την ίδια την ιστορία της κινούμενης εικόνας. Παρά το γεγονός ότι το φαινόμενο της απόκοσμης κοιλάδας δεν θα περιγραφόταν επίσημα για τουλάχιστον δύο δεκαετίες, είχε ήδη εκδηλωθεί ως σημείο ανησυχίας για τους πρώτους δημιουργούς animation που επιδίωκαν να ισορροπήσουν ρεαλισμό και ελκυστικότητα στο έργο τους. Εξάλλου, τεχνικές animation, όπως η ροτοσκοπία και το motion-capture, έχουν κατηγορηθεί στο παρελθόν ακριβώς για αυτό το λόγο. Αυτή η συνειδητοποίηση οδήγησε, τελικά, στην ανάπτυξη των λεγόμενων “*Δώδεκα Αρχών των Κινουμένων Σχεδίων*”—ενός συνόλου θεμελιωδών κατευθυντήριων γραμμών που προήλθαν από τις πρακτικές που χρησιμοποιούσαν οι δημιουργοί κινουμένων σχεδίων στα *Walt Disney Studios* κατά την διάρκεια των δεκαετιών του

1930 και 1940, που συχνά χαρακτηρίζεται ως η “Χρυσή Εποχή της Κινούμενης Εικόνας”. Οι εν λόγω αρχές, που κωδικοποιήθηκαν από τους *Frank Thomas* και *Ollie Johnston*—δύο από τους “εννέα γέρους” της Disney—στο βιβλίο τους “*The Illusion of Life: Disney Animation*” (1981) [55], αν και δεν είναι όλες σχετικές με την δική μας περίπτωση, εν γένει δημιουργήθηκαν με σκοπό να διορθώσουν μια κοινή παρανόηση: τα κινούμενα σχέδια δεν αποτελούν μια τέλεια απεικόνιση της πραγματικής ζωής, αλλά μάλλον μια καρικατούρα της. Δεν χρειάζεται να ακολουθούν τους φυσικούς νόμους· τουναντίον, πρέπει να τους υπερβαίνουν. Καθ’ αυτόν τον τρόπο, το υποκείμενο μήνυμα μεταδίδεται πιο καθαρά στους θεατές, ενώ αποτυχία σε αυτό ενέχει το κίνδυνο να προκαλέσει το φαινόμενο της απόκοσμης κοιλάδας [56].



Σχήμα 4.3: Οι δώδεκα αρχές της κινούμενης εικόνας.

## 4.2 Ερευνητικοί Στόχοι

Στόχος της παρούσης διπλωματικής εργασίας είναι η παραγωγή τρισδιάστατων κινούμενων σχεδίων του ανθρώπινου σώματος με την χρήση μοντέλων διάχυσης. Αντί, λοιπόν, να επιδιώκουμε την υπέρ-ρεαλιστική απεικόνιση της ανθρώπινης κίνησης, που όπως αναφέρθηκε ενέχει τον κίνδυνο πρόκλησης του φαινομένου της απόκοσμης κοιλάδας, θα μπορούσαμε να ακολουθήσουμε το παράδειγμα των επαγγελματιών animators που έχουν τελειοποιήσει την εν λόγω τέχνη ανά τις δεκαετίες. Αντί, λοιπόν, να εκπαιδεύσουμε κάποιο μοντέλο διάχυσης σε δεδομένα εξαχθέντα από την κινησιολογία πραγματικών ανθρώπων, όπως στην υπάρχουσα βιβλιογραφία, η βασική μας ιδέα έγκειται στην εκπαίδευση ενός diffusion model με ένα speech-to-motion dataset από cartoon clips.



### 4.2.1 Διατύπωση του Προβλήματος

Όπως καθίσταται σαφές, πρωταρχικό μας μέλημα αποτελεί η εύρεση ενός τέτοιου συνόλου δεδομένων. Αυτό αποδείχθηκε πολύ δυσκολότερο εγχείρημα από ό,τι θεωρούσαμε αρχικά.

Το πρώτο πρόβλημα που αντιμετωπίσαμε ήταν η απουσία κάποιου έτοιμου speech-to-motion dataset από cartoon clips στην βιβλιογραφία. Περιορισμοί όπως πνευματικά δικαιώματα, και η παρουσία ανομοιογενών και υπερβολικών χαρακτήρων με έντονες, γρήγορα εναλλασσόμενες και αφύσικες κινήσεις έχουν αποδειχθεί ικανοί να αποτρέψουν πολλούς ερευνητές από το να αναλάβουν ένα τέτοιο έργο. Ακριβώς σε αυτό το σημείο έγκειται ο πρώτος στόχος μας για τα πλαίσια της παρούσης εργασίας: η δημιουργία του πρώτου, από ό,τι γνωρίζουμε, speech-to-motion συνόλου δεδομένων απαρτιζόμενο από σύντομα βίντεο καρτούν.

Η πρώτη μας προσέγγιση συμπεριελάμβανε την συλλογή και επεξεργασία σύντομων βίντεο clips προερχόμενων κυρίως από δυτικότροπα έργα κινουμένων σχεδίων. Μεγάλα στούντιο παραγωγής του Hollywood και της Ευρώπης, όπως η *Disney*, η *Pixar*, η *DreamWorks*, η *Illumination*, και πολλές άλλες, έχουν παραγάγει διαχρονικά ορισμένα από τα πιο επιτυχημένα και τεχνικά αρτιότερα έργα animations παγκοσμίως. Οι δημιουργίες τους χαρακτηρίζονται από εξαιρετικά ρεαλιστικά μοντέλα κίνησης, πλούσια εκφραστικότητα, ακριβή μοτίβα χειρονομιών, αλλά και υψηλό συγχρονισμό μεταξύ του λόγου και των κινήσεων τόσο των χεριών όσο κι εν γένει του υπόλοιπου σώματος. Δυστυχώς για την εργασία μας, τα εν λόγω έργα προστατεύονται από πολύ αυστηρούς νόμους περί προστασίας της πνευματικής ιδιοκτησίας. Αυτό καθιστά εξαιρετικά δύσκολη την πρόσβαση όχι μόνο σε βίντεο κλιπς των εν λόγω έργων, αλλά και, επακόλουθα, στην ανάπτυξη learning-based μοντέλων για την αυτοματοποιημένη εξαγωγή κινησιολογικών παραμέτρων. Η απόκτηση σχετικών βίντεο που θεωρητικά ανήκουν στις κατηγορίες των *Public Domain* ή *Creative Commons*, κατόπιν εξονυχιστικών αναζητήσεων σε πλατφόρμες, όπως το *YouTube*, και ανοιχτές διαδικτυακές βιβλιοθήκες, σαν το <https://archive.org/>, είναι, μεν, εφικτή· νομικά αμφίβολη, δε.

Μία εναλλακτική προσέγγιση που εξετάστηκε αποτέλεσε και η συλλογή παλαιών ταινιών κινουμένων σχεδίων που έχουν, πλέον, περάσει στο *Public Domain*. Το νομικό πλαίσιο για αυτήν την διαδικασία είναι πολυσύνθετο και διαφέρει από κώδικα σε κώδικα. Για παράδειγμα, στην Ελλάδα και την Ευρωπαϊκή Ένωση τα δικαιώματα λήγουν μετά το πέρας των 70 ετών από την πρώτη δημοσίευση του έργου [57]. Αντιστοίχως, στις Ηνωμένες Πολιτείες Αμερικής, αυτό συμβαίνει μετά από 95 χρόνια για έργα δημοσιευμένα πριν το 1978 [58, 59]. Οι λεπτομέρειες είναι ασφαλώς πολύ πιο περίπλοκες από την απλοϊκή εικόνα που μόλις παρουσιάστηκε, αλλά ορισμένα κλασικά έργα των δεκαετιών του 1910 και 1920 έχουν ήδη περάσει στον δημόσιο τομέα. Σύντομες ταινίες, όπως το αρχικό “*Popeye*” και το “*Steamboat Willy*” μπορούν να συλλεχθούν και χρησιμοποιηθούν απευθείας, δίχως κάποια παραβίαση ενδεχόμενων πνευματικών δικαιωμάτων. Εντούτοις, και αυτή η προσέγγιση σύντομα εγκαταλείφθηκε λόγω ορισμένων απροσπέλαστων δυσκολιών. Κατ’ αρχάς, το καλλιτεχνικό στυλ εκείνης της εποχής ήταν ριζικά διαφορετικό από το σύγχρονο. Πολλές από τις τεχνικές animation που διέπουν τα σημερινά καρτούν, καθιστώντας τα ομαλά και ρεαλιστικά, δεν είχαν ανακαλυφθεί και αποκρυσταλλωθεί ακόμη. Ελλόχευε σοβαρός κίνδυνος τα δεδομένα που

θα παραγάγαμε να ήταν θορυβώδη, ασυνεχή και, γενικότερα, χαμηλότερης ποιότητας. Επιπροσθέτως, όπως ήδη σχολιάστηκε, υφίσταται εμφανής έλλειψη learning-based εργαλείων για την εξαγωγή των χαρακτηριστικών της κίνησης στυλιζαρισμένων χαρακτήρων. Αυτό, ενδεχομένως, να συνιστά και το μεγαλύτερο τροχοπέδη για την έρευνά μας και θα αποτελέσει ένα επανεμφανιζόμενο μοτίβο για το υπόλοιπο της εργασίας. Ακόμη και όταν πειραματιστήκαμε με την χρήση μοντέλων εκπαιδευμένων σε πραγματικά άτομα, που ούτως ή άλλως θα υπολειτουργούσαν σε καρτούν, στην προκειμένη περίπτωση, όπου οι χαρακτήρες βρίσκονται ακόμη πιο μακριά σε ομοιότητα από τον άνθρωπο, τα αποτελέσματα ήταν απαγορευτικά.

Οι συνθήκες για την παραγωγή του δικού μας speech-to-motion dataset από βίντεο κλιπς δυτικών ταινιών καρτούν του Hollywood και της Ευρώπης είναι, σαφώς, αρκετά δύσκολες. Το αυστηρό νομοθετικό πλαίσιο για την προστασία των πνευματικών δικαιωμάτων καθιστά την νόμιμη συλλογή δεδομένων δύσκολη έως αδύνατη· γεγονός που με την σειρά του παρακωλύει την έρευνα και ανάπτυξη learning-based εργαλείων, που επίσης είναι απαραίτητα για την σύνθεση του dataset. Ευτυχώς για εμάς, η ίδια εικόνα δεν παρατηρείται στην περίπτωση της Ιαπωνίας και του καλλιτεχνικού στυλ γνωστού ως “Anime”.

Το *Anime*, και η έντυπη έκφρασή του σε μορφή κόμικ, το *Manga*, συνιστούν ένα ιδιαίτερο υποείδος animation που αναπτύχθηκε στην Ιαπωνία ήδη από τις αρχές του 20ού αιώνα, αν και οι απαρχές τους μπορούν να ανιχνευτούν βαθύτερα, μέσα στις μακράς καλλιτεχνικές παραδόσεις της ιαπωνικής κουλτούρας και τα εικονογραφημένα χειρόγραφα βουδιστών μοναχών της φεουδαρχικής εποχής. Παρόλο που, λόγω διαφόρων γεωπολιτικών και πολιτιστικών συμβάντων που έχουν σημαδέψει την ιαπωνική ιστορία, τα Anime έχουν επηρεαστεί σε μεγάλο βαθμό από τα δυτικά καρτούν, η αισθητική τους διαφοροποιείται σημαντικά από αυτά, καθώς βασίζονται σε έντονη εκφραστικότητα, παραμορφωμένες αλλά ομαλές κινήσεις, και χαρακτηριστική χρονική ρυθμικότητα. Σαν είδος, γνώρισε εκρηκτική ανάπτυξη περί την δεκαετία του 1970, ειδικά εντός της ίδιας της Ιαπωνίας, ενώ η εισαγωγή της τηλεόρασης ως μέσον προβολής του, στις αρχές της δεκαετίας του 1980, αποτέλεσε καταλύτης για την ευρεία κατανάλωσή τους και στη Δύση. Ως τα τέλη της δεκαετίας του 1990, το ονομαζόμενο “*Manga Boom*” ήταν γεγονός, με τα Anime να λαμβάνουν διαστάσεις παγκοσμίου φαινομένου. Σήμερα, το Anime έχει ξεπεράσει το αρχικό του στίγμα ως “cult” και “εξειδικευμένη αγορά”, και αντιπροσωπεύει ένα σημαντικό μέρος της εκδοτικής βιομηχανίας [60, 61]. Και ακριβώς σε αυτό το σημείο έγκειται το μεγάλο πλεονέκτημά τους στα πλαίσια της δικής μας έρευνας. Λόγω της εκτεταμένης και μαζικής παραγωγής τους σε συνδυασμό με την ιδιαίτερα ενεργή κοινότητα, πλήθος από σειρές και βίντεο βρίσκονται διαθέσιμα σε διαδικτυακές πλατφόρμες υπό καθεστώς πιο χαλαρής ή ελεύθερης διανομής. Ταυτοχρόνως, υφίσταται μεγαλύτερο κίνητρο για ερευνητές, που, ενδεχομένως, να αποτελούν, συνάμα, μέλη της εν λόγω κοινότητας, να αναπτύξουν learning-based εργαλεία—απαραίτητα για την εξέλιξη της εργασίας μας. Τα παραπάνω καθιστούν τα Anime μια φυσική επιλογή πηγής δεδομένων για την σύνθεση του δικού μας dataset.

Συμπερασματικά, ο πρώτος ερευνητικός μας στόχος αποτελεί η ανάπτυξη του πρώτου, από όσο γνωρίζουμε, speech-to-motion dataset με annotations εξαχθέντα από σύντομα βίντεο κλιπς Anime. Λεπτομέρειες για τα περιεχόμενα του εν λόγω dataset, καθώς και η μεθοδολογία που θα ακολουθηθεί για τη δημιουργία του, περιγράφονται ανα-

λυτικά στο Κεφάλαιο 5. Κατόπιν ολοκλήρωσης αυτού του έργου, επόμενο ερευνητικό στόχο συνιστά η χρήση του νέου συνόλου δεδομένων για το fine-tuning ενός μοντέλου διάχυσης, του *AMUSE* [1], που αναφέρθηκε και στο Κεφάλαιο 3. Πάλι, οι λεπτομέρειες της διαδικασίας μπορούν να βρεθούν στο Κεφάλαιο 6.



## Κεφάλαιο 5

### Κατασκευή Συνόλου Δεδομένων Εκπαίδευσης

---

Τα υπάρχοντα speech-to-motion σύνολα δεδομένων μπορούν, σε γενικές γραμμές, να διακριθούν σε δύο κατηγορίες: στα in-house και τα in-the-wild [53]. Για τη δημιουργία των in-house συνόλων [45, 46, 47, 48, 49, 50], δύναται να χρησιμοποιηθεί μια πληθώρα διαφορετικών τεχνικών, συμπεριλαμβανομένων, ενδεικτικά, συστημάτων πολυκαμερών, τρισδιάστατων και τετραδιάστατων σαρωτών, και, πιο πρόσφατα, τεχνολογία motion-capture. Δεδομένου ότι, στην προκειμένη περίπτωση, η παραγωγή δεδομένων ελέγχεται και ρυθμίζεται αυστηρά, η ακρίβεια και ποιότητα των in-house συνόλων είναι σχεδόν πάντα εγγυημένη. Ωστόσο αυτά τα συστήματα είναι συχνά πολύ ακριβά στην απόκτηση, διατήρηση και λειτουργία τους, και απαιτούν πολλές εργατοώρες. Ως αποτέλεσμα οδηγούν τις περισσότερες φορές σε σχετικά περιορισμένα datasets. Για την εκμάθηση πλουσιότερων και πιο ποικίλων στυλ ομιλίας και συναισθημάτων, οι Yoon *et al.* [34, 62] πρότειναν την χρήση των in-the-wild συνόλων δεδομένων [51, 52, 53]. Τα δεδομένα, πλέον, είναι pseudo-ground truth, που παράγονται κατόπιν επεξεργασίας κάποιων αρχικά συλλεχθέντων δειγμάτων από διάφορες πηγές, και εμφανίζονται με την μορφή δισδιάστατων ή τρισδιάστατων keypoints, τρισδιάστατων πλεγμάτων, ή κάποιον συνδυασμό τους. Ο έλεγχος για την ποιότητα των παραγόμενων δεδομένων είναι, ασφαλώς, αρκετά μικρότερος, αλλά οι αυτοματοποιημένες ή ημι-αυτοματοποιημένες μέθοδοι που χρησιμοποιούνται ανοίγουν τον δρόμο για πιο ποικίλα και εκτενή datasets. Προκειμένου, επομένως, να εκμεταλλευτούμε την επαγγελματική δουλειά των animators, απαραίτητη για την αυξημένη ποιότητα του συνόλου δεδομένων μας, θα ακολουθήσουμε την δεύτερη προσέγγιση των in-the-wild.

#### 5.1 Δομή και Περιεχόμενα Συνόλου Δεδομένων

Καθίσταται σαφές ότι θα πρέπει να επινοήσουμε μια μεθοδολογία για την μετατροπή των συλλεχθέντων Anime βίντεο κλιπς στα annotation που θα χρειαστούμε. Για να το κάνουμε, ωστόσο, αυτό, θα πρέπει πρώτα να κατασταλάξουμε στο τι ακριβώς δεδομένα θα περιλαμβάνει το σύνολό μας. Αυτό καθορίζεται τόσο από τις ανάγκες του AMUSE που επιλέχθηκε για fine-tuning, όσο και από τα διαθέσιμα εργαλεία. Υπενθυμίζουμε ότι,

παρά την σχετικά θετική τάση που παρατηρείται ειδικά στον χώρο των Anime, η βιβλιογραφία για learning-based μεθόδους επεξεργασίας στυλιζαρισμένων χαρακτήρων εξακολουθεί να υστερεί αισθητά.

Ας υποθέσουμε, λοιπόν, ένα αρχικό σύντομο βίντεο κλιπ Anime. Κατ' αρχάς, όντας ένα speech-to-motion dataset, έπεται πως θα πρέπει να περιλαμβάνει δύο βασικά στοιχεία: *ομιλία* και *κινησιολογία*. Το πρώτο, ακολουθώντας το παράδειγμα των προκατόχων μας, λαμβάνει την μορφή ενός WAV αρχείου με τον ακατέργαστο ήχο του αρχικού βίντεο. Η ομιλία θα πρέπει να έχει τα χαρακτηριστικά μονολόγου ενός μοναδικού χαρακτήρα. Ο λόγος για αυτήν την ανάγκη είναι το γεγονός ότι όλα τα datasets που αναφέρθηκαν στην ενότητα 4.1.1 επίσης ακολουθούν αυτήν την σχεδιαστική απόφαση. Το *Speech2Gesture* [51], και κατ' επέκτασιν, όλα τα in-the-wild datasets που προέκυψαν από περαιτέρω επεξεργασία αυτού [52, 53], κάνουν χρήση βίντεο μονολόγων διαφόρων τηλεπερσόνων, όπως παρουσιαστών, λεκτόρων και τηλεαγγελιστών. Το ίδιο ισχύει και για τα αναφερθέντα in-house σύνολα. Για παράδειγμα, το *BEAT* [45] επιστρατεύει 30 ηθοποιούς που είτε μονολογούν για χρονική διάρκεια ενός λεπτού περίπου, είτε συνομιλούν σε δεκάλεπτες συνεδρίες με τον σκηνοθέτη, κατά την διάρκεια των οποίων, όμως, μόνον η φωνή του εκάστοτε ηθοποιού καταγράφεται. Το μοντέλο διάχυσης που επιλέχθηκε, το *AMUSE* [1] χρησιμοποιεί το σύνολο δεδομένων *BEAT* για την εκπαίδευσή του. Ως εκ τούτου, για να αποφύγουμε ενδεχόμενη θορυβώδη επανεκπαίδευσή του, οφείλουμε κι εμείς να περιοριστούμε σε βίντεο-μονολόγους.

Ο δεύτερος πυλώνας του συνόλου μας—η *κινησιολογία*—παρουσιάζει εξίσου μεγάλο ενδιαφέρον. Διαφορετικοί τρόποι έχουν χρησιμοποιηθεί για την περιγραφή της κίνησης. Ορισμένα σύνολα προσφέρουν τρισδιάστατα keypoints ανά frame, άλλα τρισδιάστατα keypoints, και κάποια άλλα χρησιμοποιούν τρισδιάστατα πλέγματα. Σε μερικές περιπτώσεις, μάλιστα, γίνεται χρήση ενός συνδυασμού των παραπάνω. Οι λεπτομέρειες επί τούτου παρουσιάζονται στον πίνακα 5.1. Ωστόσο, το *AMUSE* απαιτεί την χρήση τρισδιάστατων πλεγμάτων (3D mesh) ανά frame, τα οποία ορίζονται με την συμπλήρωση των κατάλληλων παραμέτρων του *SMPL-X* [54], ενός τρισδιάστατου μοντέλου του ανθρώπινου σώματος. Αποτέλεσμα της συνένωσης των χαρακτηριστικών των επί μέρους μοντέλων *SMPL* [63] για το σώμα, *MANO* [64] για τα χέρια, και *FLAME* [65] για τις εκφράσεις του προσώπου, το *SMPL-X* ορίζεται ως μια συνάρτηση  $M(\beta, \theta, \psi)$  που παράγει ένα τρισδιάστατο πλέγμα. Παραμετροποιείται από το σχήμα του σώματος  $\beta \in \mathbb{R}^{300}$ , την στάση του σώματος  $\theta \in \mathbb{R}^{55 \times 3}$  για 3D περιστροφές γωνίας-άξονα γύρω από τις 55 αρθρώσεις που υποστηρίζει το μοντέλο, και την έκφραση του προσώπου  $\psi \in \mathbb{R}^{100}$ . Δεδομένων των παραμέτρων στάσης και οποιασδήποτε παραμέτρου σχήματος, μπορούμε να λάβουμε τις κορυφές του πλέγματος του σώματος,  $V$ , χρησιμοποιώντας το διαφορίσιμο επίπεδο του *SMPL-X*. Ωστόσο, οι συγγραφείς του *AMUSE*, επέλεξαν να υιοθετήσουν την μέθοδο της 6D συνεχούς αναπαράστασης περιστροφής, σύμφωνα με τους *Zhou et al.* [66] που διαπίστωσαν εμπειρικά ότι αποδίδει καλύτερα κατά την εκμάθηση βαθένων νευρωνικών δικτύων, έναντι των 3D περιστροφών γωνίας-άξονα και των 4D τετραδονίων. Ως εκ τούτου, η παράμετρος της στάσης του σώματος γίνεται, πλέον,  $\theta \in \mathbb{R}^{55 \times 6}$ . Επιπλέον, παραλείπουν τις παραμέτρους των εκφράσεων του προσώπου, θέτοντας  $\psi = 0$ , καθώς και τις οκτώ αρθρώσεις του κάτω μέρους του σώματος, δεδομένου ότι τόσο ο δικός τους, όσο και ο δικός μας ερευνητικός στόχος επικεντρώνεται στην σύνθεση χειρονομιών τους σώματος και όχι στην μετατόπισή του. Για



τον ίδιο λόγο, και σε συνδυασμό με την έλλειψη σχετικής βιβλιογραφίας, αποφασίσαμε να θέσουμε, επιπρόσθετα,  $\beta = 0$ . Εξάλλου, η έντονα στυλιζαρισμένη φύση των Anime δεν θα μεταφραζόταν ομαλά σε ρεαλιστικές παραμέτρους σχήματος σώματος. Τελικά, στο σύνολο δεδομένων μας, η κινησιολογία του χαρακτήρα ενός βίντεο κλιπ N frames, θα αντιπροσωπεύεται από ένα NPZ αρχείο που θα περιλαμβάνει τα εξής διανύσματα: την στάση του σώματος, διάστασης [N, 55, 6]· την σχετική μετατόπιση του σώματος από την αρχή των αξόνων, διάστασης [N, 3]· το εκμηδενισμένο σχήμα του σώματος, διάστασης [N, 300]· και το φύλο του απεικονιζόμενου χαρακτήρα που τίθεται πάντα σε “neutral” (ουδέτερο), αφού, πάλι, στόχος μας αποτελεί η παραγωγή χειρονομιών.

Αυτά τα δύο αρχεία είναι αρκετά για την εκπαίδευση του AMUSE και ως εκ τούτου αυτά θα περιλαμβάνονται στο σύνολο δεδομένων εκπαίδευσης. Θα απαρτίζεται, κοινώς, από φακέλους ονομασμένους κατά αύξοντα ακέραιο αριθμό, καθένας εκ των οποίων θα περιλαμβάνει τα annotations ενός σύντομου βίντεο κλιπ Anime· θα περιλαμβάνει, δηλαδή, τα αρχεία WAV και NPZ που περιεγράφηκαν παραπάνω.

Αξίζει, σε αυτό το σημείο, να σχολιαστεί ότι κατά την διάρκεια της έρευνάς μας, εξετάστηκε ο ενδεχόμενος εμπλουτισμός του συνόλου δεδομένων μας με επιπρόσθετα annotations και μεταδεδομένα για κάθε βίντεο. Εξάλλου, πρόκειται για το πρώτο speech-to-motion dataset από Anime. Μπορεί να μην απαιτούνταν για την χρήση που σκοπεύαμε εμείς να κάνουμε, αλλά αυτό δεν αναιρεί την χρησιμότητα ενός πιο ολοκληρωμένου dataset για μελλοντικές εφαρμογές. Για το υπόλοιπο της παρούσης ενότητας θα πραγματοποιηθεί μια συνοπτική παρουσίαση επιπλέον χαρακτηριστικών που διερευνήθηκαν και οι λόγοι για τους οποίους αποφασίστηκε η παράλειψή τους, εισάγοντας συνάμα μια χρήσιμη αφετηρία για ενδεχόμενη μελλοντική προσθήκη τους.

Είναι κοινά αποδεκτό ότι η γλώσσα του σώματος είναι άμεσα συνυφασμένη με την προσωπικότητα του εκάστοτε ατόμου. Δεν παρουσιάζουν όλοι οι άνθρωποι του πλανήτη τις ίδιες αντιδράσεις όταν βρίσκονται αντιμέτωποι με τα ίδια ερεθίσματα. Τουναντίον, το σύνολο χειρονομιών, από το οποίο αντλεί ο κάθε άνθρωπος στην καθημερινότητά του, είναι ένα μοναδικό αποτέλεσμα του χαρακτήρα του, αλλά και του οικογενειακού, φιλικού και πολιτισμικού του υποβάθρου. Ορισμένα καθιερωμένα σύνολα, όπως το BEAT [45] και το SHOW [53] αντιπροσωπεύουν αυτήν την πραγματικότητα μέσω της προσθήκης της ταυτότητας του ομιλητή. Στην δική μας περίπτωση, κάτι τέτοιο θα μπορούσε ενδεχομένως να εκδηλωθεί ως η προσθήκη του ονόματος του χαρακτήρα anime που μονολογεί στο εκάστοτε βίντεο. Στο Διαδίκτυο διατίθεται μια πληθώρα δημόσια διαθέσιμων learning-based εργαλείων που επιστρατεύουν διαφορετικές μεθόδους για την αναγνώριση anime χαρακτήρων. Ενδεικτικά, το YesChat AI [67] πρόκειται για ένα εξειδικευμένο GPT, υποστηριζόμενο από το ChatGPT-4o, σχεδιασμένο για την αναγνώριση και παροχή εκτενών πληροφοριών χαρακτήρων από anime και video games. Παράλληλα, οι συγγραφείς του AniWho [68] εξερεύνησαν την απόδοση γνωστών και ευρέως καθιερωμένων βαθμών συνελκτικών νευρωνικών δικτύων, συμπεριλαμβανομένων, των InceptionV3 [69], InceptionResNetV2 [70], MobileNetV2 [71] και EfficientNet [72] στο έργο της ταξινόμησης προσώπων anime χαρακτήρων. Στο πλαίσιο της μελέτης τους, αξιοποίησαν τεχνικές μεταφοράς μάθησης (transfer learning), και μάθησης λίγων δειγμάτων (few-shot learning) επιτυγχάνοντας υψηλά ποσοστά ακριβείας, με το EfficientNet-B7 να σημειώνει την καλύτερη συνολική επίδοση. Εναλλακτικά, μια πιο έμμεση προσέγγιση είναι μέσω κάποιου learning-based tagger, όπως του WD ConvNext

*Tagger* v3 [73] που αξιοποιεί *ViT* για την παραγωγή περιγραφικών ετικετών στην μορφή αυτών που συναντώνται να συνοδεύουν τις εικόνες του *Danbooru* [74], ενός εκ των μεγαλύτερων ανοιχτών βιβλιοθηκών για fan-art στα Anime. Εμπειρικά, παρατηρήσαμε ότι στις περισσότερες των περιπτώσεων, μία από τις παραγόμενες ετικέτες για είσοδο κάποιο frame, είναι και το όνομα του απεικονιζόμενου χαρακτήρα. Αυτή η μέθοδος, ασφαλώς, δεν είναι εξίσου αξιόπιστη με τις προηγούμενες που αναφέρθηκαν· ωστόσο, είναι χρήσιμη στην περίπτωση που επιθυμούμε να προσθέσουμε επιπλέον πληροφορίες όπως το φύλο του απεικονιζόμενου χαρακτήρα, τα ρούχα που φοράει, το φόντο, και ούτω καθεξής. Συμπεραίνεται ότι το πρόβλημα της ταξινόμησης χαρακτήρων μπορεί εύκολα να αντιμετωπιστεί με την υπάρχουσα έρευνα. Αποφασίσαμε, όμως, να μην το συμπεριλάβουμε στο σύνολο δεδομένων μας. Στα άλλα datasets η προσθήκη ταυτοτήτων παρουσιάζει κάποιο νόημα λόγω του μικρού πλήθους διαφορετικών ομιλητών. Χαρακτηριστικά, στο *BEAT* συμμετέχουν 30 ηθοποιοί, ενώ το *SHOW* προσφέρει βίντεο κλιπς μόνο τεσσάρων τηλεπαρουσιαστών. Στην δική μας περίπτωση, ο αριθμός των διαφορετικών χαρακτήρων που απεικονίζονται είναι, ενδεχομένως, τόσο εκτεταμένος, με πολλούς από αυτούς να εμφανίζονται μονάχα μία φορά, που η προσθήκη των ονομάτων τους στον ρόλο ταυτότητας δεν προσφέρει κάποιο στρατηγικό πλεονέκτημα. Εναλλακτικά, τέθηκε η ιδέα του ονόματος της σειράς, στην οποία ανήκει το εκάστοτε βίντεο κλιπ, ή του στούντιο παραγωγής στον ρόλο της ταυτότητας. Πάλι, όμως, το προσωπικό στυλ του κάθε χαρακτήρα μπορεί να διαφέρει σε εξαιρετικά μεγάλο βαθμό, καθιστώντας την εν λόγω προσέγγιση επίσης μη-πρακτική.

Ένας άλλος παράγοντας που επηρεάζει εξίσου την εκδήλωση χειρονομιών στον άνθρωπο είναι η συναισθηματική του κατάσταση. Αυτός είναι ο λόγος που οδήγησε τους συγγραφείς του *BEAT* να κατηγοριοποιήσουν επιπλέον τα βίντεό τους βάσει οκτώ συναισθημάτων. Η συναισθηματική ταξινόμηση θεωρείται ένα σχετικά δύσκολο πρόβλημα, που εμπλέκει πολλαπλούς τομείς έρευνας, όπως επεξεργασία φυσικής γλώσσας, δεδομένα πολυμέσων, ανακάλυψη σημασιολογικής γνώσης, και εξόρυξη δεδομένων από κοινωνικά δίκτυα. Έχουν υπάρξει, στο παρελθόν, αρκετές απόπειρες από καθιερωμένους ψυχολόγους να χαρτογραφήσουν τον περίπλοκο κόσμο του ανθρώπινου αισθήματος. Παραδείγματος χάριν, ο *Plutchik* [75] διαχώρισε τα συναισθήματα σε 24 δυάδες ή τριάδες, ενώ ο *Russell* [76] τα ανέλυσε ως προς δύο άξονες: την θετικότητα (valence) και την διέγερση (arousal). Εντούτοις, η πιο δημοφιλής προσέγγιση αποτελεί αυτή του *Ekman* [77], ο οποίος αντιμετώπισε κάθε συναισθήμα ως μια διακριτή κατηγορία και τα ταξινόμησε σε επτά διαφορετικούς τύπους: θυμός, αηδία, φόβος, χαρά, λύπη, έκπληξη, και ουδετερότητα. Αυτή η τελευταία προσέγγιση, ή κάποια παραλλαγή της, ταιριάζει καλύτερα στις ανάγκες σύνθεσης συνόλων δεδομένων, αφού δημιουργεί αυστηρώς καθορισμένες και διακριτές κατηγορίες. Για το έργο της ταξινόμησης του συναισθήματος, δύναται να αναλυθούν οι εκφράσεις του προσώπου, η τονικότητα της φωνής, χειρονομίες, και φυσιολογικά σήματα. Μάλιστα, διαφορετικά μέσα είναι κατάλληλότερα για τον προσδιορισμό διαφορετικών συναισθημάτων. Για παράδειγμα, ηχητικά στοιχεία είναι περισσότερο ενδεικτικά της χαράς και του ενθουσιασμού, ενώ συναισθήματα όπως ο θυμός, η λύπη και η αηδία, γίνονται ευκολότερα αντιληπτά μέσω οπτικών ερεθισμάτων, και δη μέσω των εκφράσεων του προσώπου [78]. Το ευτυχές στην προκειμένη περίπτωση συνιστά το γεγονός ότι ένα σύντομο βίντεο κλιπ Anime παρέχει πληροφορίες για όλους τους άνωθι παράγοντες. Το δυστυχές, από την άλλη,



είναι ότι δεν υφίσταται επαρκής έρευνα περί της αξιοποίησης μη λεκτικών στοιχείων σε έντονα στυλιζαρισμένους χαρακτήρες [79, 80]. Μοντέλα που κατηγοριοποιούν το συναίσθημα με είσοδο το κείμενο ή το ηχητικό δείγμα υπάρχουν διαθέσιμα και θα μπορούσαν να χρησιμοποιηθούν, αλλά εμπειρικά παρατηρήσαμε ότι συχνά υπολειτουργούν. Εξάλλου, ο *Mehrabian* [81] παρατήρησε ότι μόνον τον 7% και 38% της συναισθηματικής πληροφορίας μεταδίδεται λεκτικά και φωνητικά, αντίστοιχα, με το υπόλοιπο 55% να αποδίδεται στην γλώσσα του σώματος. Την ίδια στιγμή, πειράματα με μοντέλα εκπαιδευμένα σε δεδομένα πραγματικών ανθρώπων, επίσης, δεν επέφεραν ενθαρρυντικά αποτελέσματα. Ως εκ τούτου, αποφασίσαμε να εγκαταλείψουμε το συναίσθημα ως μέρος του dataset μας. Ούτως ή άλλως, το *AMUSE* θα λαμβάνει υπ' όψιν συναισθηματικές πληροφορίες εξάγοντάς τες από το *WAV* αρχείο κατά τρόπο που θα επεξηγηθεί στο Κεφάλαιο 6.

| Dataset               | Head    | Hand         | Body         | Duration | Annotations  |
|-----------------------|---------|--------------|--------------|----------|--------------|
| Multiface [48]        | 3D mesh | -            | -            | -        | Multi-camera |
| BIWI [49]             | 3D mesh | -            | -            | -        | 3D-scanner   |
| VOCASET [50]          | 3D mesh | -            | -            | 0.5h     | 4D-scan      |
| Takeutchi et al. [46] | -       | -            | 3D keypoints | 5h       | MoCap        |
| Trinity [47]          | -       | -            | 3D keypoints | 4h       | MoCap        |
| BEAT [45]             | 3D mesh | 3D keypoints | 3D keypoints | 76h      | MoCap        |
| Yoon et al. [34, 62]  | -       | -            | 3D keypoints | 52h      | p-GT         |
| S2G [51]              | -       | 2D keypoints | 2D keypoints | 144h     | p-GT         |
| Habibie et al. [52]   | 3D mesh | 3D keypoints | 3D keypoints | 33h      | p-GT         |
| SHOW [53]             | 3D mesh | 3D mesh      | 3D mesh      | 27h      | p-GT         |
| <b>Το δικό μας</b>    | 3D mesh | 3D mesh      | 3D mesh      | 6hr      | p-GT         |

Πίνακας 5.1: Σύγκριση των χαρακτηριστικών διαφόρων speech-to-motion συνόλων δεδομένων

## 5.2 Μεθοδολογία

Έχοντας, πλέον, εντοπίσει επακριβώς τα δεδομένα τα οποία θα πρέπει να συμπεριλάβουμε στο σύνολο δεδομένων μας, επόμενο βήμα είναι η κατάστρωση της μεθοδολογίας για την εξαγωγή τους.

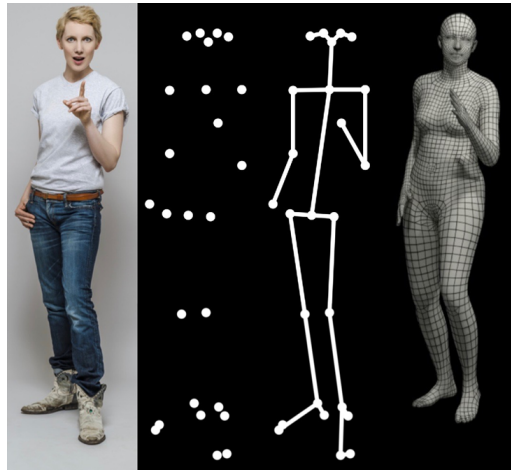
Ας υποθέσουμε, λοιπόν, ένα αρχικό dataset με σύντομα ακατέργαστα βίντεο κλιπ *Anime*. Οι λεπτομέρειες περί διαλογής του εν λόγω dataset βρίσκονται στην ενότητα 5.3. Πρώτη μας μέριμνα συνιστά η εκπόνηση μιας διαδικασίας φιλτραρίσματος· ήτοι, της διάκρισης εκείνων των βίντεο κλιπς που κρίνουμε χρήσιμα και η απόρριψη των υπολοίπων. Όπως εξηγήθηκε, εν συντομία, στην προηγούμενη ενότητα, είναι αναγκαίο αυτά να πληρούν συγκεκριμένες προϋποθέσεις. Τα υφιστάμενα speech-to-motion datasets παρέχουν κινησιολογικά annotations για μόνον ένα άτομο την φορά το οποίο μονολογεί. Συνεπώς, και τα σχετικά μοντέλα κατασκευάζονται λαμβάνοντας υπ' όψιν αυτές τις συνθήκες. Κριτικά για εμάς, το *AMUSE* δεν αποτελεί εξαίρεση. Ως εκ τούτου, για καθένα βίντεο κλιπ θα πρέπει να ελέγξουμε κατά πόσο περιλαμβάνουν μόνον έναν χαρακτήρα ανά frame του οποίου η φωνή είναι η μοναδική που ακούγεται. Επιπλέον, δεδομένου ότι αποσκοπούμε να παραμετροποιήσουμε την κινησιολογία του

άνω μέρους του σώματός του, οφείλουμε να εγγυηθούμε ότι είναι ορατός τουλάχιστον από την μέση και πάνω. Τέλος, θα πρέπει να ελέγξουμε ότι η γλώσσα της ομιλίας είναι η αγγλική, και ότι τα βίντεο κλιπς δεν είναι εξαιρετικά σύντομα. Για το τελευταίο, κρίνεται σημαντικό να αναφέρουμε ότι το *AMUSE* είναι εκπαιδευμένο σε βίντεο των 300 frames. Η αρχική μας διαίσθηση, επομένως, ήταν, κι εμείς με την σειρά μας, να θέσουμε το κατώτερο όριο των frames ενός βίντεο κλιπ στα 300. Ωστόσο, παρατηρήσαμε ότι κάτι τέτοιο οδηγεί σε πολύ περιορισμένο αριθμό βίντεο που περνούν την διαδικασία φιλτραρίσματος. Για τον λόγο αυτό, πήραμε την απόφαση να ελαττώσουμε το κατώτερο όριο στα 100 frames, οπότε και κατά την διάρκεια της εκπαίδευσης ακολουθείται μια διαδικασία padding. Περισσότερες λεπτομέρειες για τον ακριβή τρόπο και τα εργαλεία που χρησιμοποιήθηκαν για την περάτωση της άνωθι διαδικασίας φιλτραρίσματος συναντώνται στην ενότητα 5.4.

Αφού έχουμε ολοκληρώσει την διαδικασία του φιλτραρίσματος, απορρίπτοντας τα βίντεο που κρίνουμε ότι δεν πληρούν τις αναγκαίες προϋποθέσεις, μπορούμε να προχωρήσουμε στην εξαγωγή των απαραίτητων δεδομένων. Η διαδικασία δημιουργίας του WAV αρχείου για την ομιλία είναι αρκετά απλοϊκή και ευθεία. Μάλιστα, θα πραγματοποιηθεί ήδη κατά την διαδικασία του φιλτραρίσματος, καθώς θα χρειαστεί για τον έλεγχο ορισμένων προαπαιτούμενων, όπως της γλώσσας (βλ. ενότητα 5.4). Τουναντίον, η διαδικασία εξαγωγής των χαρακτηριστικών του ανθρώπινου σώματος, καθώς και της τρισδιάστατης ανακατασκευής του είναι μακράν πιο περίπλοκη και ενδιαφέρουσα, περιλαμβάνοντας μια σειρά από επί μέρους βήματα. Κατ' αρχάς, οφείλουμε να προεπεξεργαστούμε ελαφρώς το αρχικό βίντεο διαχωρίζοντάς το σε frames. Εργαζόμενοι τώρα σε κάθε ένα frame ξεχωριστά, εντοπίζουμε τον απεικονιζόμενο χαρακτήρα και εκτελούμε pose estimation για την εξαγωγή των δισδιάστατων keypoints. Ακολουθεί η ανύψωσή τους στην τρίτη διάσταση κάνοντας χρήση κάποιου βαθέως νευρωνικού δικτύου εκπαιδευμένου για *3D lifting* από μονοφθαλμική προοπτική. Τέλος, για κάθε άρθρωση του παραγόμενου σκελετού, υπολογίζεται ντετερμινιστικά η ζητούμενη 6D αναπαράσταση περιστροφής, που θα αποτελέσουν τις παραμέτρους  $\theta$  του *SMPL-X* [54].

### 5.3 Επιλογή Αρχικού Συνόλου Anime

Προτού εφαρμόσουμε την μεθοδολογία της 5.2, θα πρέπει να βρούμε ένα σύνολο ακατέργαστων βίντεο κλιπ Anime, που θα λάβουν τον ρόλο των αρχικών μας δεδομένων. Μία πρώτη σκέψη ήταν η χειροκίνητη ή, μάλλον, ημιαυτόματη συλλογή τους. Διαλέγοντας μία εκ των πολλαπλών δημοσίως διαθέσιμων λιστών ονομάτων σειρών και ταινιών που ανήκουν στο καλλιτεχνικό στυλ του Anime, θα μπορούσαμε να επιστρατεύσουμε κάποιον *Web Scraper*· ένα πρόγραμμα, δηλαδή, σαν το *AnimDL* [82], που σαρώνει διάφορες ιστοσελίδες του Διαδικτύου για την λήψη περιεχομένου. Κατόπιν, μέσω της χρήσης βιβλιοθηκών, όπως του *OpenCV* [83], θα ήταν δυνατό να χωρίσουμε τα ληφθέντα αρχεία σε έγκυρες σκηνές σύμφωνα με τις προϋποθέσεις μας. Ωστόσο, γρήγορα, απορρίψαμε την εν λόγω ιδέα. Κατ' αρχάς, αν και το ίδιο το *AnimDL* δεν είναι παράνομο από μόνο του, η χρήση που θα θέλαμε εμείς να κάνουμε συχνά είναι, λόγω του γεγονότος ότι οι ιστοσελίδες τις οποίες επισκέπτεται για την λήψη του περιεχομένου συνιστούν, κατ' ουσίαν, υπηρεσίες *Streaming* που επιβάλλουν αυστηρούς



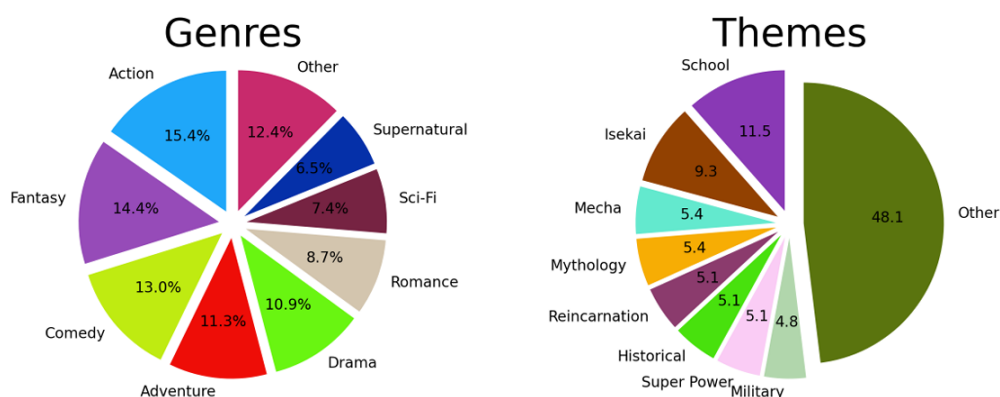
**Σχήμα 5.1:** Η διαδικασία ανακατασκευής του τρισδιάστατου πλέγματος του ανθρώπινου σώματος. Κατ' αρχάς, εκτελείται *2D pose estimation*, ακολουθούμενο από *3D lifting*. Τέλος γίνεται *fitting* του τρισδιάστατου σκελετού στο μοντέλο *SMPL-X* [54].

όρους. Σε κάθε περίπτωση, βρίσκουμε την συγκεκριμένη μέθοδο ηθικά αμφίβολη. Αυτό, σε συνδυασμό με την δυσκολία χρήσης του εργαλείου εξαιτίας των μέτρων προστασίας των προαναφερθέντων υπηρεσιών streaming έναντι των bots και την επιπλέον εργασία που θα απαιτούνταν για τον διαχωρισμό των σκηνών, ειδικά αν λάβουμε υπ' όψιν την ύπαρξη σχετικών δημοσίως διαθέσιμων datasets, κατέστησε την εν λόγω τεχνική μη πρακτική.

Στο σημείο αυτό, παρατηρήσαμε ότι μεγάλο μέρος της έρευνας επί της ανάπτυξης μοντέλων που υλοποιούν κάποιο task πάνω σε Anime, έχει εκπαιδευτεί είτε στο *Danbooru* [74] είτε σε λοιπά σύνολα βασισμένα στο *Danbooru*. Υπενθυμίζουμε αυτό συνιστά μία—αν όχι την—μεγαλύτερη ανοιχτή βιβλιοθήκη για Anime fan-art στο Διαδίκτυο. Αυτό, αν και θετικό εκ πρώτης όψευς λόγω του όγκου δεδομένων, εντούτοις, συνδέεται με μια σειρά από αξιοσημείωτα προβλήματα. Το πιο βασικό, ενδεχομένως, εξ' αυτών είναι ότι, εξαιτίας των πιο χαλαρών υφιστάμενων κανονισμών, μια πολύ μεγάλη μερίδα των δεδομένων χαρακτηρίζεται ως ακατάλληλου περιεχομένου. Βέβαια, ανά τα χρόνια, έχουν δημοσιευτεί αρκετές εκδόσεις που εφαρμόζουν κάποιο φίλτρο για την απόρριψη των ακατάλληλων ή αμφιβόλων δεδομένων, όπως, για παράδειγμα, το [84]. Παρόλα αυτά, η γενικότερη ερασιτεχνικότητα που χαρακτηρίζει πολλά από έργα του dataset, όντας fan-art, σε συνδυασμό με το γεγονός ότι αφορά, ως επί το πλείστον, εικόνες αντί για βίντεο, μας οδήγησε στην απόφαση να αναζητήσουμε διαφορετική πηγή δεδομένων.

Τελικά, καταλήξαμε στο *Anim-400K* από τους *Cai et al.* [85]. Αν και ο αρχικός σκοπός της δημιουργίας του ήταν η αντιμετώπιση του προβλήματος της αυτόματης παραγωγής υπότιτλων, απαρτίζεται εν μέρει από σύντομα βίντεο κλιπ Anime που το καθιστά ιδανικό για την έρευνά μας. Δημιουργήθηκε από την συγκέντρωση δημόσια διαθέσιμων βίντεο Anime προερχόμενων από δημοφιλείς ιστότοπους προβολής τους. Κατά την στιγμή της λήψης τους, κανένα από τα συλλεχθέντα βίντεο δεν βρίσκονταν πίσω από *paywall*, ούτε απαιτούνταν η οιαδήποτε μορφή σύνδεσης για την πρόσβασή τους. Οι συγγραφείς συνέλεξαν ακατέργαστα επεισόδια σε ανάλυση  $1920 \times 1080$ , και

ήχο συχνότητας 48kHz. Ο διαχωρισμός τους σε επιμέρους κλιπς επιτυγχάνεται μέσω μιας περίπλοκης επαναληπτικής διαδικασίας που περιλαμβάνει, αρχικά, την παραγωγή ASR μεταγραφών ομιλίας και, στην συνέχεια, την αναδρομική συγχώνευση χρονικά επικαλυπτόμενων μεταγραφών με μέγιστη απόκλιση 125ms. Αποτέλεσμα ήταν η δημιουργία μιας συλλογής 437 χιλιάδων βίντεο κλιπς Anime από μια ευρεία ποικιλία κατηγοριών και θεμάτων, όπως φαίνεται στα διαγράμματα της εικόνας 5.2. Αυτό είναι υψίστης σημασίας για το έργο μας, καθώς ευρύτερη ποικιλία διαφορετικών συγκεκριμένων συνεπάγεται και μεγαλύτερη ποικιλία εμφανιζόμενων χειρονομιών. Επιπροσθέτως, κατόπιν εφαρμογής μιας μεθόδου *speaker diarization* (γνωστό task καθορισμού του πλήθους διακριτών ομιλητών που συναντώνται σε ένα ηχητικό δείγμα), οι Cai *et al.* διαπίστωσαν ότι από τα 437 χιλιάδες βίντεο, τα 323 χιλιάδες κρίθηκαν κλιπς ενός μονάχα ομιλητή. Σαφώς, αυτό δεν συνεπάγεται ότι ο εν λόγω ομιλητής είναι ο μοναδικός παρόν στο βίντεο, καθώς απαιτούμε επιπλέον εμείς, αλλά, σίγουρα, το αναφερόμενο πλήθος είναι ενθαρρυντικό.



Σχήμα 5.2: Κατηγορίες (Genres) και Θέματα (Themes) παρόντα στο σύνολο δεδομένων *Anim-400K* [85].

## 5.4 Φιλτράρισμα Βίντεο Κλιπς

Όπως περιεγράφηκε στην ενότητα 5.2, πρώτο βήμα αποτελεί το φιλτράρισμα των βίντεο κλιπ του *Anim-400K* [85], διατηρώντας μονάχα αυτά που κρίνουμε χρήσιμα. Υπενθυμίζουμε ότι κατά την διάρκεια αυτής της διαδικασίας, θα πρέπει να ελέγξουμε το μέγεθος του εκάστοτε βίντεο κλιπ ώστε να περιλαμβάνει τουλάχιστον 100 frames· να υπάρχει ένας μονάχα ομιλητής και η γλώσσα ομιλίας του να είναι η αγγλική· και, τέλος, ο συγκεκριμένος ομιλητής να είναι ο μοναδικός παρόν σε όλα τα frames του dataset και να είναι ορατός τουλάχιστον από την μέση και πάνω. Πρόκειται για μια πολύπλοκη διαδικασία, καθώς για τον έλεγχο των παραπάνω απαιτήσεων θα πρέπει να αξιοποιηθούν λεκτικά, ηχητικά και οπτικά χαρακτηριστικά του κάθε βίντεο.

Το πρώτο απαιτούμενο που θα ελεγχθεί είναι η γλώσσα. Για τον σκοπό αυτό θα πρέπει, κατ' αρχάς, να εξαγάγουμε τον ακατέργαστο ήχο από το βίντεο κλιπ. Κάπως έτσι, δημιουργείται και το πρώτο αρχείο που απαιτούμε να περιλαμβάνει το σύνολο

δεδομένων μας—το WAV αρχείο. Η εξαγωγή γίνεται με το *FFmpeg* [86], ένα ευέλικτο και ευρέως χρησιμοποιούμενο εργαλείο επεξεργασίας βίντεο. Ο ήχος δειγματοληπτείται στα 16kHz λόγω των απαιτήσεων του *AMUSE* [1]. Εφόσον η εξαγωγή είναι επιτυχής, το παραχθέν αρχείο ήχου οδηγείται στο *turbo* μοντέλο *Whisper* της *OpenAI* [87]. Πρόκειται για ένα *sequence-to-sequence* μοντέλο, που δημοσιεύτηκε το 2022, και που χρησιμοποιείται κατά κόρον για μεταγραφή (audio transcription). Εντούτοις, προσφέρει την δυνατότητα να επιστρέψει ένα JSON αρχείο με πολύτιμα μεταδεδομένα· ένα εξ' αυτών αποτελεί η γλώσσα της ομιλίας. Αυτό μας επιτρέπει να διατηρήσουμε τα αγγλικά βίντεο του *Anim-400K* και να απορρίψουμε τα ιαπωνικά.

Εάν ένα βίντεο κλιπ κριθεί ότι είναι στην αγγλική γλώσσα, μπορούμε να προχωρήσουμε στην εξαγωγή των επί μέρους frames του. Αυτό πραγματοποιείται εκμεταλλεύοντας την βιβλιοθήκη *OpenCV* [83]. Σε αυτό το σημείο ελέγχουμε εάν το συνολικό τους πλήθος είναι μεγαλύτερο των 100. Σε αντίθετη περίπτωση απορρίπτονται.

Στην συνέχεια ακολουθείται μια διαδικασία εύρεσης των σκηνών. Οι συγγραφείς του *Anim-400K* διαχώρισαν τα συλλεχθέντα επεισόδιά τους σε σκηνές σύμφωνα με την μέθοδο που περιεγράφηκε στην 5.3, αλλά ο διαχωρισμός είναι άτσαλος πολλές φορές και δεν πραγματοποιήθηκε με τις ίδιες ανάγκες με τις δικές μας υπ' όψιν. “Αδέσποτα” frames στην αρχή ή στο τέλος του βίντεο κλιπ εξαιτίας της προϋπάρχουσας κωδικοποίησης του *MP4* αρχείου, αλλά και γενικότερα βίντεο με πολλαπλές εναλλαγές χαρακτήρων και υποβάθρων είναι συχνό φαινόμενο. Στην δική μας περίπτωση που τα annotations της κινησιολογίας καταγράφονται για κάθε frame, είναι σημαντικό τα κλιπς να είναι “κομμένα” όσο πιο “καθαρά” γίνεται, χωρίς υπολειπόμενα frames από διπλανές σκηνές. Για την εύρεση των frames που συναντάται αλλαγή σκηνής, χρησιμοποιήσαμε τα εργαλεία της *OpenCV* με τα οποία υπολογίσαμε, για κάθε frame, ένα score διαφοράς αυτού με το προηγούμενό του. Εάν το score διαφοράς πέσει κάτω από ένα συγκεκριμένο κατώφλι, τότε υποθέτουμε αλλαγή σκηνής. Πειραματικά, το εν λόγω κατώφλι ρυθμίστηκε στο 60%. Εδώ, θα μπορούσαμε, απλώς, να απορρίψουμε τα βίντεο κλιπ που εμφανίζουν κάποια αλλαγή σκηνής. Αυτό, ωστόσο, θα συνεπαγόταν την διαγραφή πολλών υποψήφιων επιμέρους βίντεο κλιπς. Για τον λόγο αυτό, σε περίπτωση που εντοπιστούν σκηνές ανώ των 100 frames, θεωρούνται από μόνα τους νέα βίντεο: υπολογίζονται τα ακριβή τους *timestamps* διαιρώντας τους δείκτες των εναρκτήριων και τελικών τους frames δια του *FPS* (που εξάγεται με την βοήθεια του *FFmpeg*) και περικόπτεται το αρχικό WAV αρχείο (επίσης με την βοήθεια του *FFmpeg*). Έχοντας, πλέον, εγγυηθεί, κατά το δυνατόν, ότι τα βίντεο-δεδομένα μας είναι περικομμένα με ακρίβεια, τα νέα WAV αρχεία—εφόσον υπάρχουν—περνάνε εκ νέου από το μοντέλο *Whisper* για μεταγραφή. Αυτό συμβαίνει για να ελεγχθεί ότι τα νέα βίντεο παρουσιάζουν πράγματι ομιλία και δεν εντοπίστηκε κάποια εκτενής σκηνή όπου οι χαρακτήρες παραμένουν σιωπηλοί.

Σε αυτό το σημείο, το μόνο που απομένει από την διαδικασία φιλτραρίσματος είναι ο οπτικός έλεγχος της παρουσίας των χαρακτήρων. Για αυτό τον σκοπό, αποδείχθηκαν εξαιρετικά χρήσιμες οι συναρτήσεις της βιβλιοθήκης *dghs-imgutils* [88], και δη αυτές που αφορούν την ανίχνευση μερών του σώματος ενός χαρακτήρα anime. Οι συγκεκριμένες συναρτήσεις αποτελούν ορισμένες από τα ελάχιστα δημοσίως διαθέσιμα learning-based εργαλεία που εξειδικεύονται σε αυτόν τον σκοπό. Η συγκεκριμένη βιβλιοθήκη παρέχει εργαλεία για την ανίχνευση κεφαλής, ανθρώπου και άνω μέρους του



σώματος ενός στυλιζαρισμένου ανθρώπου anime. Παράλληλα, επιστρέφουν τα αντίστοιχα οριοθετημένα πλαίσια. Για κάθε frame, συνεπώς, ενός βίντεο κλιπ, καλούμαστε να ανιχνεύσουμε τους απεικονιζόμενους ανθρώπους. Εφόσον το πλήθος αυτών ισούται με την μονάδα, ανιχνεύουμε την κεφαλή και το άνω μέρος το σώματός τους. Παρατηρήσαμε, ωστόσο, ότι αν, φερ' ειπείν, ένα frame απεικονίζει μονάχα ένα πρόσωπο, τότε και τα δύο τελευταία μοντέλα θα επιστρέψουν κάποια επιτυχή ανίχνευση. Για να εγγυηθούμε, λοιπόν, ότι αυτά τα βίντεο θα απορρίπτονται, θέτουμε έναν επιπρόσθετο έλεγχο: ότι το εμβαδόν του οριοθετημένου πλαισίου της κεφαλής οφείλει να είναι μικρότερο από το 40% του αντίστοιχου εμβαδού του άνω μέρους του σώματος.

Η διαδικασία φιλτραρίσματος που μόλις περιεγράφηκε είναι αρκετά συνεπής και αυστηρή, εγγυώντας ότι τα βίντεο κλιπς που θα περάσουν διαθέτουν τις ιδανικές συνθήκες για την μετέπειτα εφαρμογή του αλγορίθμου εκτίμησης του τρισδιάστατου πλέγματος. Ωστόσο, μόνο, ένα μικρό ποσοστό—της τάξεως του 5%— ικανοποιεί τις αναγκαίες προϋποθέσεις μας. Η διαδικασία φιλτραρίσματος διακρίνεται σχηματική αναπαράσταση στην εικόνα 5.3.

## 5.5 Εκτίμηση Τρισδιάστατου Πλέγματος

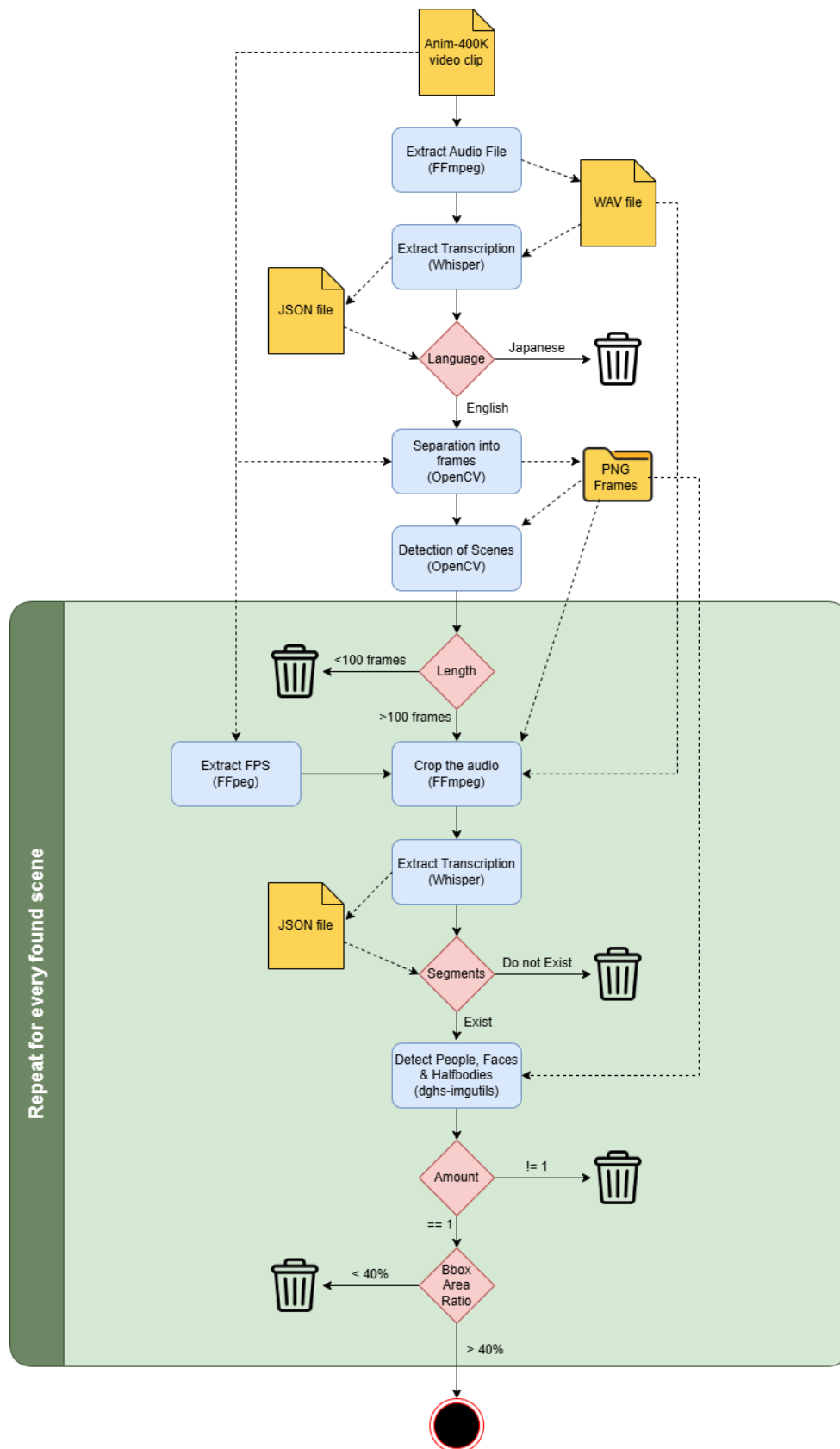
Σε αυτήν την ενότητα θα περιγραφεί αναλυτικά η μεθοδολογία και τα εργαλεία που θα χρειαστούμε για την διαδικασία εκτίμησης του πλέγματος στα βίντεο τα οποία πέρασαν την διαδικασία φιλτραρίσματος ικανοποιώντας τις αναγκαίες μας προϋποθέσεις. Όπως, περιεγράφηκε εν συντομία στην 5.2, η διαδικασία περιλαμβάνει χονδρικά τα εξής βήματα: *δισδιάσταση εκτίμηση στάσης · ανύψωση στην τρίτη διάσταση · και υπολογισμός των 6D αναπαραστάσεων περιστροφής*.

### 5.5.1 Δισδιάστατη Εκτίμηση Στάσης

Το αρχικό και, ομολογουμένως, ένα εκ των πιο δυσεπίλυτων βημάτων αποτελεί η εκτέλεση της δισδιάστατης εκτίμησης στάσης (2D pose estimation). Δοθέντος, δηλαδή, ενός frame θα πρέπει να υπολογίσουμε τα δισδιάστατα keypoints του χαρακτήρα της εικόνας.

Στην πραγματικότητα, το συγκεκριμένο πρόβλημα είναι χρόνια καθιερωμένο και η ερευνητική κοινότητα έχει αναπτύξει μια ευρεία ποικιλία διαφορετικών λύσεων [89]. Εξέχοντα μεταξύ τους, τα *OpenPose* [90], *MediaPipe* [91], *AlphaPose* [92], και *Detectron2* [93] παρουσιάζουν αδιαμφισβήτητα εντυπωσιακά αποτελέσματα. Δυστυχώς, τα συγκεκριμένα μοντέλα εξειδικεύονται στην εκτίμηση στάσης φωτογραφιών και βίντεο πραγματικών ανθρώπων. Η αναζήτηση κάποιας fine-tuned έκδοσής τους σε σύνολα δεδομένων από Anime ή Manga, επίσης, δεν τελεσφόρησε. Σε αυτό το σημείο ελέγχσαμε, πειραματικά ορισμένα από αυτά [90, 93], για να αξιολογήσουμε την επίδοσή τους στους έντονα στυλιζαρισμένους χαρακτήρες με τους οποίους ασχολούμαστε. Όπως ήταν αναμενόμενο, στην μεγάλη πλειονότητα των δειγμάτων δεν κατάφεραν καν να εντοπίσουν τον απεικονιζόμενο χαρακτήρα στην εικόνα, που αποτελεί το πρώτο βήμα για την εκτέλεση ενός pose estimation.

Κατέστη, γρήγορα, σαφές ότι για να ολοκληρώσουμε αυτό το βήμα της διαδικασίας θα πρέπει να αναζητήσουμε κάποιο εργαλείο που εξειδικεύεται στην εκτίμηση στάσης



Σχήμα 5.3: Η διαδικασία φιλτραρίσματος.

στα Anime και Manga και να εγκαταλείψουμε τα προαναφερθέντα *state-of-the-art* μοντέλα. Κι αυτό, εντούτοις, αποδείχτηκε πιο δύσκολο από ό,τι αναμενόταν. Τουλάχιστον έως την συγγραφή της παρούσης εργασίας, η σχετική βιβλιογραφία υστερεί αισθητά. Εν τέλει, η έρευνά μας μας οδήγησε στο *Bizarre-Pose-Estimator* [94] που επιστρατεύει *transfer learning*. Συγκεκριμένα, περνάει την εικόνα-είσοδο από δύο προεκπαιδευμένα δίκτυα: το *Mask R-CNN*, που εξάγει *task-specific* χαρακτηριστικά σχετικά με την εκτίμηση στάσης και το *ResNet Tagger* που εξάγει *domain-specific* χαρακτηριστικά προσαρμοσμένα στο είδος των δεδομένων—τα Anime. Τα δύο σύνολα χαρακτηριστικών μετασχηματίζονται μέσω μικρών εκπαιδευσιμων converter δικτύων, ώστε να αποκτήσουν κοινή αναπαράσταση, και έπειτα συνενώνονται, αφού, πρώτα, ευθυγραμμιστούν χωρικά. Το ενιαίο αυτό διάλυμα χαρακτηριστικών τροφοδοτεί το εκπαιδευσιμο *keypoint head*, το οποίο παράγει *heatmaps* για τις θέσεις των keypoints. Η εκπαίδευση γίνεται με *Binary Cross-Entropy (BCE) loss*, ώστε τα προβλεπόμενα *heatmaps* να προσεγγίζουν τα *ground-truth heatmaps* της εκτίμησης στάσης.

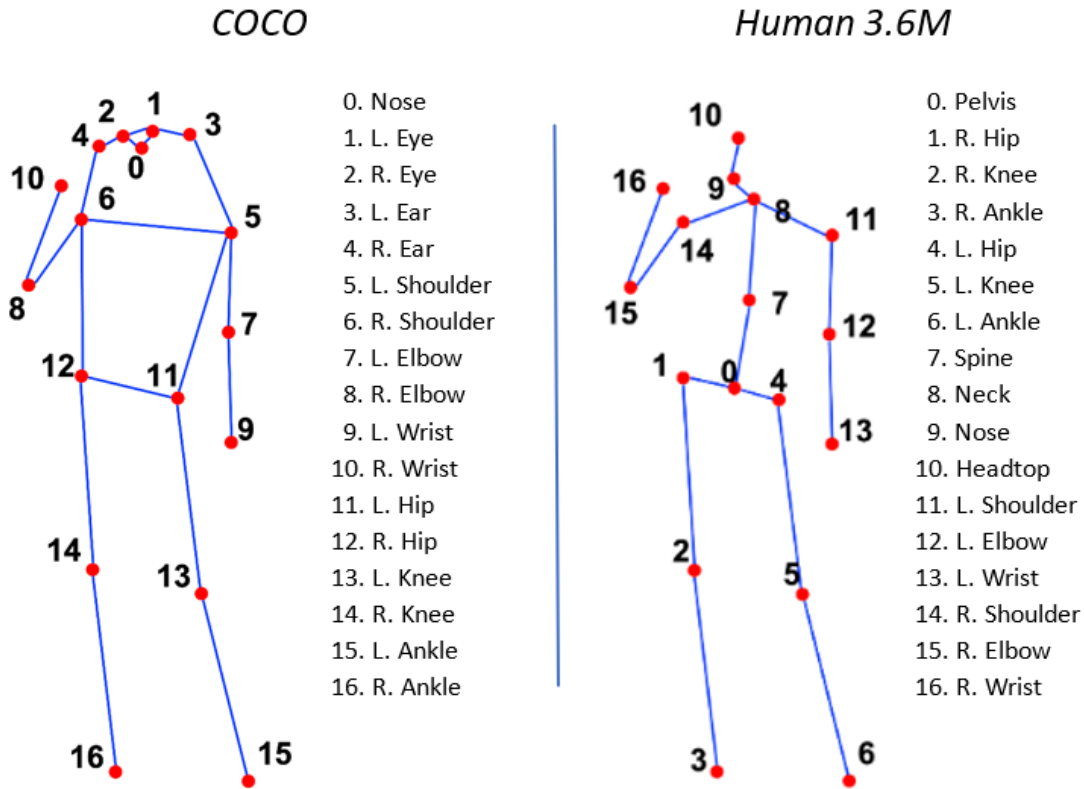
Το *Bizarre-Pose-Estimator* παράγει 17 δισδιάστατα keypoints σε μορφή σκελετού COCO [95], τα οποία διακρίνονται στην εικόνα 5.4. Επιπλέον, όπως αναφέρθηκε προηγουμένως η είσοδος του μοντέλου είναι εικόνες, όχι βίντεο. Για αυτόν τον λόγο, θα πρέπει εκτελέσουμε αναδρομικά τον αλγόριθμο για όλα τα frames του εκάστοτε βίντεο κλιπ. Αυτή η στρατηγική συνοδεύεται από δύο κρίσιμα προβλήματα: Κατ' αρχάς, προκαλούνται απαγορευτικά μεγάλες καθυστερήσεις κατά την διάρκεια προσπέλασης όλων των frames. Κατόπιν, η παραγωγική διαδικασία του μοντέλου είναι, προφανώς, αποσυνδεδεμένη για διαφορετικές εισόδους. Ως εκ τούτου, για ένα frame, το ίδιο keypoint δύναται να βρίσκεται σε ελαφρώς διαφορετικό σημείο από το αμέσως προηγούμενό του, παρά το γεγονός ότι ο απεικονιζόμενος χαρακτήρας είναι πρακτικά ακίνητος. Αυτό δημιουργεί ένα τρεμάμενο εφέ στο τελικό αποτέλεσμα. Για την επίλυση και των δύο αυτών προβλημάτων υιοθετήσαμε μια παρόμοια τεχνική για τον εντοπισμό των αλλαγών σκηνών κατά την διαδικασία φιλτραρίσματος. Συνεπώς, με την βοήθεια των συναρτήσεων της *OpenCV* [83], για κάθε frame υπολογίζουμε την διαφορά του από εκείνο για το οποίο εκτιμήθηκε τελευταία η στάση, και σε περίπτωση που αυτή πέσει κάτω από το κατώφλι του 95% τότε επανεκτιμάται· διαφορετικά, αντιγράφουμε την εκτιμημένη στάση του προηγούμενου frame. Αυτή, βέβαια, η μέθοδος δημιουργεί το καινούριο πρόβλημα ότι οι μεταβάσεις από frame σε frame είναι συχνά απότομες. Ωστόσο, αυτό θα αντιμετωπισθεί στην πορεία.

### 5.5.2 Ανύψωση στην Τρίτη Διάσταση

Έχοντας, πλέον, εξασφαλίσει την επιτυχή εξαγωγή των δισδιάστατων keypoints στα δεδομένα βίντεο κλιπ Anime, επόμενο λογικό βήμα, καθώς προείπαμε, αποτελεί η ανύψωσή τους στην τρίτη διάσταση (*3D lifting*). Το θετικό, από αυτό το σημείο κι εξής, είναι δεν χρειάζεται να περιορίσουμε τα εργαλεία μας σε αυτά τα οποία είναι συγκεκριμένα εκπαιδευμένα σε έντονα στυλιζαρισμένους χαρακτήρες Anime και Manga. Κι αυτό, διότι το δυσκολότερο κομμάτι αποτελεί, κατ' ουσίαν, η ίδια η ανίχνευσή τους και η εκτίμηση της στάσης τους. Αφού κατέχουμε, όμως, ήδη τα 2D keypoints, μπορούμε να χρησιμοποιήσουμε αυτά ως είσοδο χωρίς ιδιαίτερα προβλήματα.

Αναζητούμε μοντέλα για την τρισδιάστατη ανύψωση ανθρώπινων στάσεων από





Σχήμα 5.4: Οι σκελετοί και η αντίστοιχη απαρίθμηση των αρθρώσεων κατά *COCO* [95] (αριστερά) και κατά *Human3.6M* [96] (δεξιά).

μονοφθαλμικά βίντεο. Αυτό παραμένει ένα ιδιαίτερα δύσκολο πρόβλημα δοθέντων της αμφισημίας βάθους και των συχνών αυτοκαλύψεων. Ορισμένες προσεγγίσεις που βασίζονται σε βαθιά συνελκτικά δίκτυα [97, 98] και δίκτυα γράφων [99] έχουν σημειώσει αδιαμφισβήτητα εντυπωσιακά αποτελέσματα· εντούτοις υποαποδίδουν στην εκμετάλλευση πληροφοριών συγκεκλιμένων από όλα τα frames. Η εισαγωγή, ωστόσο, των ViTs δύναται να αλλάξει αυτήν την πραγματικότητα. Για τον λόγο αυτό, διαλέξαμε τον *MHFormer* των *Li et al.* [100], που βασίζεται στην έννοια της “Πολλαπλής Υπόθεσης” (Multi-Hypothesis Transformer). Το μοντέλο κάνει χρήση πολλαπλών Transformers για να αντιμετωπίσει την προαναφερθείσα ασάφεια που προκύπτει κατά την εκτίμηση 3D στάσεων από 2D εικόνες. Πιο συγκεκριμένα, για να το επιτύχει αυτό, αντί να παράγει μόνο μία, δημιουργεί πολλαπλές πιθανές τρισδιάστατες στάσεις, μέσω ενός μηχανισμού *temporal attention* που ενσωματώνει πληροφορία από διαδοχικά frames. Η αρχιτεκτονική του βασίζεται σε stacked transformer blocks που μαθαίνουν τόσο τις χωρικές όσο και τις χρονικές σχέσεις μεταξύ των αρθρώσεων του σώματος.

Ένα σημαντικό πρόβλημα με τον *MHFormer* το οποίου καλούμαστε να αντιμετωπίσουμε είναι η ασυμβατότητα του σκελετικού προτύπου στο οποίο έχει εκπαιδευτεί με αυτό που χρησιμοποιείται από τον *Bizarre-Pose-Estimator*. Ενώ ο τελευταίος βασίζεται σε keypoints τύπου *COCO* [95], ο πρώτος κάνει χρήση του συνόλου δεδομένων *Human3.6M* [96]. Οι αρθρώσεις του εν λόγω προτύπου διακρίνονται, επίσης, στην εικόνα 5.4. Για να χρησιμοποιήσουμε, λοιπόν, το μοντέλο ανύψωσης, θα πρέπει πρώτα, να

μετατρέψουμε τα εξαχθέντα keypoints από το ένα πρότυπο στο άλλο. Δυστυχώς, δεν υφίσταται, από όσο γνωρίζουμε κάποιο δημοσίως διαθέσιμο εργαλείο που να εκτελεί την μετατροπή αυτή. Ως εκ τούτου, θα πρέπει να συμβιβαστούμε με την προσέγγιση των απόντων keypoints μέσω απλών μαθηματικών πράξεων. Έτσι, για παράδειγμα, θέτουμε την λεκάνη (άρθρωση 0) στο γεωμετρικό μέσον των εκατέρωθεν ισχίων, και την σπονδυλική στήλη (άρθρωση 7) στο μέσον της λεκάνης και του θώρακος, ο οποίος πρόκειται για μια βοηθητική άρθρωση που ορίζεται, με την σειρά του, ως το γεωμετρικό μέσον των δύο ώμων. Ο λαιμός (άρθρωση 8) τίθεται στο ένα τέταρτο της απόστασης από τον θώρακα στην μύτη, ενώ το πάνω μέρος της κεφαλής (άρθρωση 10) υπολογίζεται, με την βοήθεια της συνάρτησης της *dghs-imgutils* [88] για την ανίχνευση προσώπου, ως το μέσον της άνω πλευράς του οριοθετημένου πλαισίου.

Τέλος, η δυνατότητα του *MHFormer* να λαμβάνει υπ' όψιν συγκεκριμένες λεπτομέρειες από τα γειτονικά frames λύνει το πρόβλημα των απότομων μεταβάσεων που προέκυψαν λόγω της μεθόδου που ακολουθήσαμε κατά την 2D pose estimation. Πειραματικά, παρατηρήσαμε ότι κατά την τρισδιάστατη ανύψωση, το μοντέλο εξομαλύνει τις εν λόγω μεταβάσεις δίχως να θυσιάζεται η ακρίβεια των κινήσεων. Αξίζει, βέβαια, να σχολιάσουμε ότι το *MHFormer* έχει κατασκευαστεί κατά τρόπο που να δέχεται keypoints από βίντεο των συγκεκριμένα 351 frames. Εμείς, κατέχουμε βίντεο κλιπς των 100 frames και άνω. Για τον λόγο αυτό, εάν μια ακολουθία καρέ είναι μικρότερη των 351, συμπληρώνουμε την αρχή και το τέλος του βίντεο με επαναληπτικά καρέ έως ότου φθάσουμε το ζητούμενο πλήθος· διαφορετικά εφαρμόζουμε μια μέθοδο *Sliding Window*.

### 5.5.3 Εξαγωγή Επιπρόσθετων Keypoints

Έως αυτό το σημείο της διαδικασίας με τελικό στόχο την τρισδιάστατη εκτίμηση του πλέγματος, έχουμε κατορθώσει να συγκεντρώσει τις τρισδιάστατες συντεταγμένες των 17 αρθρώσεων του *Human3.6M*. Μολαταύτα, το ανθρώπινο μοντέλο *SMPL-X* [54] απαιτεί 55 αρθρώσεις ως μέρος των  $\theta$  παραμέτρων του. Αυτές οι 55 αρθρώσεις αναγράφονται στον πίνακα 5.2. Παρατηρούμε ότι, εξ' αυτών, μόνον οι 21 αρθρώσεις αφορούν την στάση του σώματος. Οι υπόλοιπες συμπληρώνονται από 30 αρθρώσεις για τα δάκτυλα των δύο χεριών και 3 για το πρόσωπο (μία για το πιγούνι και δύο για τα μάτια), καθώς και από μία επιπρόσθετη—αυτήν την λεκάνης—που ορίζεται σαν τη κορυφή-ρίζα του υπόλοιπου σκελετού και χρησιμοποιείται για τον καθορισμό του καθολικού προσανατολισμού του μοντέλου.

Το μεγαλύτερο ποσοστό των απόντων αρθρώσεων αφορούν τα δάκτυλα των χεριών. Οι χειρονομίες συνιστούν ένα εκ των σημαντικότερων μερών της γλώσσας του σώματος, ικανές να μεταβάλλουν ολοκληρωτικά το υποκείμενο νόημα. Συνεπώς η παραμετροποίησή τους καθίσταται υψίστης σημασίας για εμάς. Για τον σκοπό αυτό, επιλέξαμε το *MediaPipe Hands* [101] το οποίο παράγει για είσοδο μια εικόνα, παράγει απευθείας 21 τρισδιάστατα keypoints για κάθε χέρι, εφόσον αυτό ανιχνευτεί. Για να το επιτύχουν αυτό επιστρατεύουν μια σειρά από επί μέρους learning-based μοντέλα, σχηματίζοντας ένα *pipeline*, όχι πολύ διαφορετικό στην λογική του από αυτό που καταστρώσαμε εμείς για την τρισδιάστατη εκτίμηση στάσης του ανθρώπινου σώματος. Κατ' αρχάς, ανέπτυξαν έναν *single-shot* ανιχνευτή παλάμης—όχι χεριού. Αυτή η επιλογή έγινε δεδομένου ότι η εκτίμηση των οριοθετημένων πλαισίων άκαμπτων αντικειμέ-

| Αρθρώσεις SMPL-X |           |     |             |     |             |     |             |
|------------------|-----------|-----|-------------|-----|-------------|-----|-------------|
| idx              | Αρθρωση   | idx | Αρθρωση     | idx | Αρθρωση     | idx | Αρθρωση     |
| 0                | Pelvis    | 14  | R. Collar   | 28  | L. Middle 1 | 42  | R. Index 3  |
| 1                | L. Hip    | 15  | Head        | 29  | L. Middle 2 | 43  | R. Middle 1 |
| 2                | R. Hip    | 16  | L. Shoulder | 30  | L. Middle 3 | 44  | R. Middle 2 |
| 3                | Spine 1   | 17  | R. Shoulder | 31  | L. Pinky 1  | 45  | R. Middle 3 |
| 4                | L. Knee   | 18  | L. Elbow    | 32  | L. Pinky 2  | 46  | R. Pinky 1  |
| 5                | R. Knee   | 19  | R. Elbow    | 33  | L. Pinky 3  | 47  | R. Pinky 2  |
| 6                | Spine 2   | 20  | L. Wrist    | 34  | L. Ring 1   | 48  | R. Pinky 3  |
| 7                | L. Ankle  | 21  | R. Wrist    | 35  | L. Ring 2   | 49  | R. Ring 1   |
| 8                | R. Ankle  | 22  | Jaw         | 36  | L. Ring 3   | 50  | R. Ring 2   |
| 9                | Spine 3   | 23  | L. Eye      | 37  | L. Thump 1  | 51  | R. Ring 3   |
| 10               | L. Foot   | 24  | R. Eye      | 38  | L. Thump 2  | 52  | R. Thump 1  |
| 11               | R. Foot   | 25  | L. Index 1  | 39  | L. Thump 3  | 53  | R. Thump 2  |
| 12               | Neck      | 26  | L. Index 2  | 40  | R. Index 1  | 54  | R. Thump 3  |
| 13               | L. Collar | 27  | L. Index 3  | 41  | R. Index 2  |     |             |

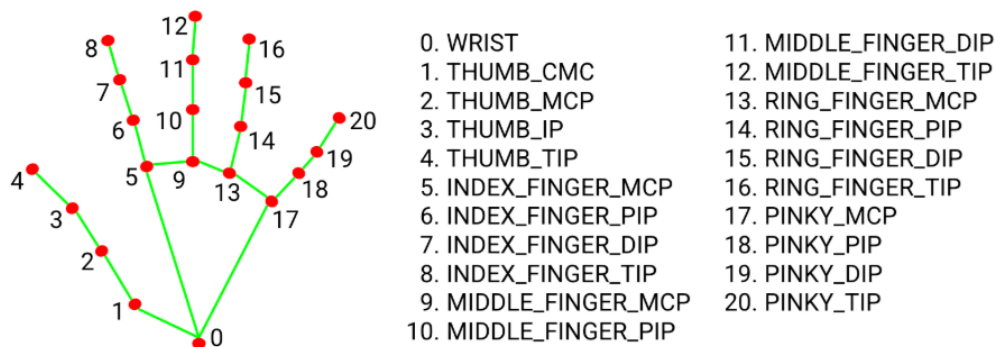
**Πίνακας 5.2:** Οι 55 Αρθρώσεις του SMPL-X: 21 για την στάση του σώματος· 30 για τα δάκτυλα των χεριών· 3 για το πρόσωπο· και μία (*pelvis*) για τον καθολικό προσανατολισμό του μοντέλου.

νων, όπως μιας παλάμης ή μιας γροθιάς, είναι αισθητά ευκολότερη από την ανίχνευση χεριών με ευκίνητα και ανεξάρτητα δάκτυλα. Επιπλέον, λόγω του σχετικά μικρού μεγέθους τους, ο αλγόριθμος *non-maximum suppression*, που χρησιμοποιείται για να αποφεύγονται επικαλυπτόμενα αποτελέσματα, εξακολουθεί να λειτουργεί καλά, ακόμη και σε περιπτώσεις έντονων επικαλύψεων, όπως, παραδείγματος χάριν, σε μια χειραψία. Τέλος, οι παλάμες μπορούν να μοντελοποιηθούν με τετράγωνα πλαίσια, δίχως να απαιτείται η εξέταση περίπλοκων σχημάτων, μειώνοντας, καθ' αυτόν τον τρόπο, τον αριθμό των πιθανών *anchors* κατά τρεις έως πέντε φορές. Στην συνέχεια, χρησιμοποιείται ένας μηχανισμός *encoder-decoder* εξαγωγής χαρακτηριστικών. Στόχος της εκπαίδευσης αποτελεί η ελαχιστοποίηση του *focal loss*, μιας συνάρτησης κόστους που βοηθά το μοντέλο να διαχειρίζεται αποτελεσματικά μεγάλο αριθμό από *anchors* που προκύπτουν λόγω της μεγάλης ποικιλίας μεγεθών στις εικόνες. Μετά την ανίχνευση της παλάμης σε ολόκληρη την εικόνα, χρησιμοποιούν ένα μοντέλο εξαγωγής *Hand Landmarks* που πραγματοποιεί ακριβή εντοπισμό 21 τρισδιάστατων συντεταγμένων των αρθρώσεων του χεριού εντός των προηγούμενων ανιχνευμένων οριοθετημένων πλαισίων. Σε αντίθεση με την δική μας μέθοδο που πρώτα εντοπίζουμε τα δισδιάστατα *keypoints* κι εν συνεχεία τα ανυψώνουμε στην τρίτη διάσταση, εν προκειμένω, οι ερευνητές προβλέπουν απευθείας τις τρισδιάστατες συντεταγμένες. Το μοντέλο μαθαίνει μια συνεπή εσωτερική αναπαράσταση της στάσης του χεριού και είναι ανθεκτικό σε μερικώς ορατά χέρια ή αυτοεπικαλύψεις.

Καθώς αναφέρθηκε στην υποενότητα 5.5.1, οι ερευνητές του *MediaPipe* έχουν επίσης δημοσιεύσει ένα μοντέλο εκτέλεσης *3D body pose estimation* [91]. Τότε κατονομάσαμε την απουσία εικονογραφημένων χαρακτήρων στο σύνολο δεδομένων εκπαίδευ-

σης ως την αιτία που δεν το επιλέξαμε. Ο λόγος, που το αντίστοιχο μοντέλο τους για την εκτίμηση στάσης χεριών είναι αποδεκτό για την χρήση που επιθυμούμε εμείς να κάνουμε, αποτελεί το γεγονός ότι, αυτή τη φορά, τις εικόνες πραγματικών χεριών συνόδευαν συνθετικές εικόνες ενός μοντέλου χεριού υψηλής ποιότητας σε διαφορετικά υπόβαθρα. Επιβεβαιώσαμε, εμπειρικά, ότι το εν λόγω μοντέλο δύναται να διαχειριστεί τους έντονα στυλιζαρισμένους χαρακτήρες με τους οποίους ασχολούμαστε.

Μετά την επιτυχή ανά frame εξαγωγή των landmarks των χεριών, τα περιστρέψουμε, κατ' αρχάς, έτσι ώστε να ακολουθούν τους συμβατικούς καρτεσιανούς άξονες. Στην συνέχεια, προκειμένου να τα ενσωματώσουμε στον ήδη υπάρχων τρισδιάστατο σκελετό του υπόλοιπου σώματος, παρατηρήσαμε ότι τόσο το *MediaPipe Hands*, όσο και το *Human3.6M* περιλαμβάνουν τις αρθρώσεις του δεξιού και αριστερού καρπού. Οπότε, δεν είχαμε παρά να μετατοπίσουμε τα keypoints των χεριών κάθε frame κατά τρόπο που οι συντεταγμένες των εκάστοτε καρπών να ταυτίζονται.



Σχήμα 5.5: Οι 21 αρθρώσεις του χεριού που παράγει το *MediaPipe Hands* [101].

Σε αυτό το σημείο, αξίζει να σχολιαστεί ότι εξετάστηκε και η ενδεχόμενη προσθήκη κάποιου μοντέλου στην διαδικασία μας που να εξειδικεύεται στην εξαγωγή τρισδιάστατων keypoints του προσώπου. Εξάλλου το *SMPL-X* [54] στις  $\theta$  παραμέτρους του, συμπεριλαμβάνει τρεις σχετικές αρθρώσεις: τα δύο μάτια και το πιγούνι. Αυτά δεν καθορίζουν τις εκφράσεις του προσώπου· αυτές δίδονται από το διάνυσμα  $\psi \in \mathbb{R}^{100}$ , που, λόγω των αναγκών του *AMUSE* [1], μηδενίζεται. Τουναντίον, καθορίζουν, απλώς, την σχετική θέση του μέσα στον σκελετό. Αυτό το σημείο θα αναφερθεί εκ νέου στην επόμενη υποενότητα, στην οποία θα επεξηγηθεί, επακριβώς, ο λόγος για τον οποίο αποφανθήκαμε να μην λάβουμε, τελικώς, υπ' όψιν, τα εν λόγω keypoints.

Στο μεταξύ, χάριν πληρότητας, αλλά και για να παράσχουμε ένα ικανοποιητικό αφηγηματικό framework για μελλοντική εργασία, θα κάνουμε μια σύντομη αναφορά των αποτελεσμάτων της έρευνάς μας. Η τρισδιάστατη εκτίμηση των landmarks του προσώπου εικονογραφημένων χαρακτήρων συνιστά ένα ιδιαίτερα απαιτητικό πρόβλημα, ειδικά στην περίπτωση των Anime. Ένα από τα πλέον ειδοποιά χαρακτηριστικά του είδους, είναι το καλλιτεχνικό στυλ που διαπνέει τα πρόσωπα των χαρακτήρων. Γιγαντιαία εκφραστικά μάτια που καλύπτουν το μεγαλύτερο ποσοστό της επιφάνειας του προσώπου, σε συνδυασμό με λιγότερο έντονα περιγράμματα μυτών, χειλιών, και άλλων δομών, καθιστούν τους χαρακτήρες Anime ιδιαίτερα αγαπητά στους θεατές, αλλά, συνάμα, αρκετά μακρινά από τα ρεαλιστικά πρόσωπα των πραγματικών ανθρώπων.

Ως εκ τούτου, θα πρέπει είτε να βρούμε κάποιο εργαλείο που να εξειδικεύεται στην ανίχνευση και εκτίμηση landmarks προσώπων anime, είτε να συμβιβαστούμε με κάποιο μοντέλο εκπαιδευμένο σε δεδομένα του πραγματικού κόσμου που αναμένεται να υπολειτουργεί. Όσον αφορά το πρώτη προσέγγιση, επιβεβαιώσαμε, για ακόμη μια φορά, την σχετική έλλειψη στην υπάρχουσα βιβλιογραφία. Παρόλο που εντοπίσουμε ορισμένα πολλά υποσχόμενα μοντέλα για δισδιάστατη εκτίμηση, μεταξύ των οποίων το *anime-face-detector* [102] που παράγει 28 2D landmarks προσώπου, αυτό μας το εύρημα δεν συνοδεύτηκε από κάποιο μοντέλο που να τα ανυψώνει στην τρίτη διάσταση, ή που να παράγει απευθείας τρισδιάστατα keypoints. Συνεπώς, ακολουθήσαμε στην πορεία την δεύτερη προσέγγιση. Σε αυτή την περίπτωση, υφίσταται μια αρκετά ικανοποιητική ποικιλία δημοσίως διαθέσιμων μοντέλων [103, 104, 105]. Εμείς, ως επί το πλείστον, εξερευνήσαμε τον *Dense-Head-Pose-Estimation* [105], όπου για την λειτουργία εξαγωγής αραιών landmarks, δύναται να παραγάγει 68 τρισδιάστατα σημεία. Ωστόσο το μοντέλο, όπως αναμενόταν, συχνά δεν παρήγαγε αποδεκτά αποτελέσματα. Ακόμα και η εκ των προτέρων βοηθητική αποκοπή του οριοθετημένου πλαισίου του προσώπου, παρόλο που βελτίωσε εν γένει την επίδοση, το “μπέρδεμα” του μοντέλου πάνω σε έντονα στυλιζαρισμένα χαρακτηριστικά, όπως μαλλιά και είδη ρουχισμού, δεν ήταν διόλου ενθαρρυντικό. Στην τελική, εμείς, δεν χρειαζόμασταν 68 landmarks, αλλά μόνον 3 συν 1—την μύτη, για την ενσωμάτωσή τους στον γενικότερο σκελετό, κατά αναλογία με την χρήση των καρπών στην περίπτωση των χεριών. Για τα μάτια, που δεν δίδεται κάποιο συγκεκριμένο keypoint, πχ. για τις κόρες, αλλά δίδονται για το περίγραμμά τους, υπολογίσαμε το γεωμετρικό τους μέσο. Το τελικό αποτέλεσμα, στις περισσότερες περιπτώσεις, ήταν ένα άκρως ασύμμετρο σύνολο keypoints που συχνά φάνταζε τυχαίο.

#### 5.5.4 Υπολογισμός 6D Αναπαραστάσεων Περιστροφής

Ως προείπαμε, τελικό βήμα για την ολοκλήρωση της διαδικασίας εκτίμησης του τρισδιάστατου πλέγματος, αποτελεί το *fitting* του διαφορίσιμου ανθρώπινου μοντέλου *SMPL-X* [54] και ο υπολογισμός των 6D αναπαραστάσεων περιστροφής.

Ο υπολογισμός των παραπάνω αναπαραστάσεων δύναται να γίνει αρκετά περίπλοκος σε ορισμένες περιπτώσεις, εντούτοις είναι ντετερμινιστικός και δεν απαιτεί απαραίτητα την χρήση κάποιου learning-based μοντέλου. Ωστόσο, αξίζει να σχολιάσουμε ότι, αρχικά, εξερευνήσαμε αυτήν την προοπτική. Το *MeshLifter* [106] αποτελεί μια προσέγγιση ασθενούς εποπτείας που επιστρατεύει μια απλοϊκή διάταξη *encoder-decoder* για να παραγάγει παραμέτρους *SMPL* [63] με είσοδο μια ακολουθία δισδιάστατων keypoints που ακολουθούν το πρότυπο *Human3.6M* [96]. Ωστόσο, ενώ σίγουρα η υιοθέτηση μιας τέτοιας μεθόδου θα ήταν σίγουρα βολική, ειδικά λαμβάνοντας υπ’ όψιν ορισμένες από τις δυσκολίες που θα αντιμετωπίσουμε στην πορεία, αποφανθήκαμε να μην προχωρήσουμε με την ιδέα. Κατ’ αρχάς, οι παραγόμενες παράμετροι αφορούν το *SMPL*, πρόγονο του *SMPL-X* με μόλις 24 αρθρώσεις αντί για 55. Εξάλλου, υπενθυμίζουμε ότι το *SMPL-X* υπήρξε το αποτέλεσμα της συγχώνευσης τριών πρότερων μοντέλων, του *SMPL* [63], του *MANO* [64] και του *FLAME* [65]. Αυτό θα συνεπαγόταν ότι η πλειοψηφία των ζητούμενων  $\theta$  παραμέτρων θα εκμηδενίζονταν, συμπεριλαμβανομένων των χεριών που θεωρούμε εξαιρετικά σημαντικά για την ορθή μετάδοση των μηνυμάτων της γλώσσας του σώματος. Επιπροσθέτως, θεωρούμε πως το γεγονός ότι το *MeshLifter*



κατά την είσοδο δεν έχει πρόσβαση σε κρίσιμα σημεία, όπως το σαγόνι και τα μάτια, τα οποία, όμως, καλείται να εκτιμήσει, δεν συνάδει με την λογική την οποία θα πρέπει να διέπει την κατασκευή ενός συνόλου δεδομένων εκπαίδευσης. Στόχος μας αποτελεί η αξιοποίηση των καλύτερων διαθέσιμων εργαλείων και τεχνικών για την όσο το δυνατόν κοντινότερη προσέγγιση της *ground truth* των πραγματικών αφετηριακών δεδομένων. Η επινόνηση τιμών δεν αποτελεί προσέγγιση αλλά δημιουργία νέων δειγμάτων.

Θα υπολογίσουμε, συνεπώς, χειροκίνητα τις αναπαραστάσεις περιστροφής. Το *SMPL-X* αντιμετωπίζει τις αρθρώσεις του πίνακα 5.2 σαν δέντρο. Χωρίζονται, επομένως, σε τρεις διακριτές κατηγορίες για καθεμία εκ των οποίων ακολουθείται ελαφρώς διαφορετικός αλγόριθμος: οι αρθρώσεις που βρίσκονται στον μέσον του δέντρου, που κατέχουν, δηλαδή, τόσο άρθρωση-πατέρα όσο και αρθρώσεις-παιδιά· τις αρθρώσεις-φύλλα· και την άρθρωση-ρίζα, που στην προκειμένη περίπτωση την συνιστά η λεκάνη και καθορίζει τον καθολικό προσανατολισμό του μοντέλου.

Ας ξεκινήσουμε την ανάλυσή μας με την ρίζα του δέντρου αρθρώσεων. Οι τιμές που θα λάβει η λεκάνη καθορίζουν τον καθολικό προσανατολισμό· ήτοι θα πρέπει να ορίζει τους άξονες αναφοράς και τις κατευθύνσεις τους. Για το σκοπό αυτό, θα χρειαστούμε τρία βοηθητικά *keypoints* που συνδέονται με την λεκάνη. Συνήθως, επιλέγονται το δεξί και αριστερό ισχίο και η σπονδυλική στήλη. Μέσω αυτών θα ορίσουμε το ορθοκανονικό μας σύστημα αξόνων  $\mathbf{s}$ ,  $\mathbf{f}$ ,  $\mathbf{u}$ . Οπότε, αν το  $\mathbf{j}$  συμβολίζει ένα διάνυσμα θέσης μιας άρθρωσης, τότε το πλάγιο μοναδιαίο διάνυσμα  $\mathbf{s}$  δίδεται από:

$$\mathbf{s} = \frac{\mathbf{j}_{R.Hip} - \mathbf{j}_{L.Hip}}{\|\mathbf{j}_{R.Hip} - \mathbf{j}_{L.Hip}\|} \quad (5.1)$$

Για τον προς τα επάνω άξονα  $\mathbf{u}$ , προκειμένου να εγγυηθούμε την ορθογωνιότητα, χρησιμοποιούμε την μέθοδο *Gram-Schmidt*, ως εξής:

$$\mathbf{u}_0 = \mathbf{j}_{Spine} - \mathbf{j}_{Pelvis} \quad (5.2)$$

$$\mathbf{u}_1 = \mathbf{f}_0 - (\mathbf{f}_0 \cdot \mathbf{s}) \mathbf{s} \quad (5.3)$$

$$\mathbf{u} = \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|} \quad (5.4)$$

Τέλος, για τον προς τα εμπρός άξονα  $\mathbf{f}$ , αρκεί να υπολογίσουμε το εξωτερικό γινόμενο των δύο προηγούμενων μοναδιαίων διανυσμάτων, δηλαδή:

$$\mathbf{f} = \frac{\mathbf{s} \times \mathbf{u}}{\|\mathbf{s} \times \mathbf{u}\|} \quad (5.5)$$

Ο ορισμός των τριών άνωθι ορθοκανονικών αξόνων μας δίδει την δυνατότητα να ορίσουμε τον *Πίνακα Βάσης*, ή διαφορετικά τον ορθοκανονικό πίνακα περιστροφής:

$$\mathbf{B} = [\mathbf{s}, \mathbf{f}, \mathbf{u}] \quad (5.6)$$

Αυτός ο πίνακας αφορά το σύστημα που προκύπτει από τις τρισδιάστατες συντεταγμένες που έχουμε εξαγάγει έως αυτό το σημείο. Επαναλαμβάνοντας την διαδικασία, αντικαθιστώντας τις αντίστοιχες συντεταγμένες των αρθρώσεων σε κατάσταση ηρε-

μίας, λαμβάνουμε τον πίνακα:

$$\mathbf{B}_{Rest} = [\mathbf{s}_{Rest}, \mathbf{f}_{Rest}, \mathbf{u}_{Rest}] \quad (5.7)$$

Τα δύο μητρώα  $\mathbf{B}$  και  $\mathbf{B}_{Rest}$  μπορούν να χρησιμοποιηθούν για να μετατρέψουν ένα διάνυσμα της τρέχουσας στάσης ή της στάσης αναφοράς, αντιστοίχως, έστω  $\mathbf{v}_{local}$ , που είναι εκφρασμένο στο τοπικό σύστημα της λεκάνης, στο καθολικό σύστημα,  $\mathbf{v}_{global}$ , και αντίστροφα:

$$\mathbf{v}_{global}^{(current)} = \mathbf{B} \mathbf{v}_{local}^{(current)}, \quad \mathbf{v}_{global}^{(rest)} = \mathbf{B}_{Rest} \mathbf{v}_{local}^{(rest)} \quad (5.8)$$

$$\mathbf{v}_{local}^{(current)} = \mathbf{B}^\top \mathbf{v}_{global}^{(current)}, \quad \mathbf{v}_{local}^{(rest)} = \mathbf{B}_{Rest}^\top \mathbf{v}_{global}^{(rest)} \quad (5.9)$$

Χρησιμοποιούμε, τώρα, τους δύο πίνακες βάσης για να ορίσουμε το μητρώο περιστροφής:

$$\mathbf{R} = \mathbf{B}\mathbf{B}_{Rest}^\top \quad (5.10)$$

Ενώ, το  $\mathbf{B}$  δίδει τον μετασχηματισμό  $local \rightarrow global$ , το  $\mathbf{R}$  δίδει τον μετασχηματισμό  $rest \rightarrow current$ . δίδει, δηλαδή, την πραγματική περιστροφή της λεκάνης. Τώρα, το μοντέλο *SMPL-X* δέχεται παραμέτρους της μορφής 3D περιστροφών γωνίας άξονα, που λαμβάνονται με τον κάτωθι τρόπο:

$$\mathbf{r}(\theta) = \frac{1}{2 \sin \theta} \begin{bmatrix} R_{32} - R_{23} \\ R_{13} - R_{31} \\ R_{21} - R_{12} \end{bmatrix}, \quad \theta = \arccos\left(\frac{\text{tr}(\mathbf{R}) - 1}{2}\right) \quad (5.11)$$

Ωστόσο, όπως έχουμε ήδη αιτιολογήσει, εμείς θα ακολουθήσουμε την μέθοδο των *Zhou et al.* [66] που πρότειναν 6D συνεχή αναπαράσταση περιστροφής. Αυτή ορίζεται, πολύ απλά, διατηρώντας τις δύο πρώτες σειρές του μητρώου περιστροφής  $\mathbf{R}$  ως εξής:

$$\mathbf{r} = [R_{11}, R_{21}, R_{31}, R_{12}, R_{22}, R_{23}]^\top \quad (5.12)$$

Από την άλλη πλευρά, ο αντίστοιχος υπολογισμός μιας άρθρωσης  $\mathbf{j}$  που βρίσκεται στην μέση του σκελετού-δέντρου, αντί για τρεις βοηθητικές γειτνιάζουσες αρθρώσεις, κάνει χρήση των συντεταγμένων του πατέρα και του παιδιού του,  $\mathbf{p}$  και  $\mathbf{c}$  αντίστοιχα. Έτσι, έστω οι κατευθύνσεις προς τον γονέα και προς το παιδί:

$$\mathbf{u} = \frac{\mathbf{j} - \mathbf{p}}{\|\mathbf{j} - \mathbf{p}\|} \quad (5.13)$$

$$\mathbf{v} = \frac{\mathbf{c} - \mathbf{j}}{\|\mathbf{c} - \mathbf{j}\|} \quad (5.14)$$

Με την βοήθεια αυτών ορίζουμε ένα κάθετο επίπεδο:

$$\mathbf{n} = \frac{\mathbf{u} \times \mathbf{v}}{\|\mathbf{u} \times \mathbf{v}\|} \quad (5.15)$$

$$\mathbf{b} = \frac{\mathbf{n} \times \mathbf{v}}{\|\mathbf{n} \times \mathbf{v}\|} \quad (5.16)$$

Κτίζουμε, συνεπώς, την τοπική ορθοκανονική βάση της άρθρωσης:

$$\mathbf{B} = [\mathbf{b}, \mathbf{n}, \mathbf{v}] \quad (5.17)$$

Κατά αντιστοιχία, δε, με την περίπτωση της λεκάνης, οφείλουμε να επαναλάβουμε την διαδικασία για την εξαγωγή του μητρώου βάσης της στάσης αναφοράς και, εν συνεχεία, του καθολικού μητρώου περιστροφής:

$$\mathbf{B}_{Rest} = [\mathbf{b}_{Rest}, \mathbf{n}_{Rest}, \mathbf{v}_{Rest}] \quad (5.18)$$

$$\mathbf{R}_j = \mathbf{B}\mathbf{B}_{Rest}^\top \quad (5.19)$$

Το μητρώο περιστροφής  $\mathbf{R}_j$  χαρακτηρίζεται ως καθολικό διότι δείχνει πώς το τοπικό σύστημα αξόνων  $\mathbf{B}$  της άρθρωσης είναι τοποθετημένο στον κόσμο. Αυτό είναι το επιθυμητό στην περίπτωση της λεκάνης, που όπως έχουμε εμπεδώσει ορίζει τον καθολικό προσανατολισμό του μοντέλου· ωστόσο, όταν διαχειριζόμαστε κάποια ενδιάμεση άρθρωση σε έναν 3D σκελετό που εφαρμόζεται κάποια ιεραρχία, όπως στην περίπτωση μας, η περιστροφή που απαιτείται δεν είναι η καθολική, αλλά η σχετική, ήτοι, πώς η εν λόγω άρθρωση περιστρέφεται σε σχέση με την θέση του γονέα. Αυτή η σχετική περιστροφή υπολογίζεται, με την βοήθεια του καθολικού μητρώου του πατέρα  $\mathbf{R}_p$ , ως εξής:

$$\mathbf{R} = \mathbf{R}_p^\top \mathbf{R}_j \quad (5.20)$$

Η 6D συνεχής αναπαράσταση περιστροφής αποκτάται με τον ίδιο τρόπο όπως στην (5.12), για το σχετικό μητρώο περιστροφής.

Τέλος, έχουμε την περίπτωση των αρθρώσεων-φύλλων. Ο υπολογισμός τους είναι αρκετά παρόμοιος με τον αντίστοιχο των ενδιάμεσων αρθρώσεων· η διαφορά τους έγκειται στο ότι, πλέον, δεν διαθέτουμε κάποιο παιδί. Αυτό έχει την επιπλοκή ότι δεν μπορούμε εξ' αρχής να ορίσουμε ένα αρχικό επίπεδο  $(\mathbf{u}, \mathbf{v})$ , όπως στις (5.13) και (5.14)—διαθέτουμε μόνον το  $\mathbf{u}$ :

$$\mathbf{u} = \frac{\mathbf{j} - \mathbf{p}}{\|\mathbf{j} - \mathbf{p}\|} \quad (5.21)$$

Υπό αυτές τις συνθήκες, για να εγγυηθούμε την παραγωγή ενός ορθοκανονικού συστήματος αξόνων, θα ακολουθήσουμε την προσέγγιση της ρίζας εφαρμόζοντας ορθοκανονικοποίηση *Gram-Schmidt* γύρω από τον άξονα  $\mathbf{u}$ . Προς αυτόν τον σκοπό, θα πρέπει να επιστρατεύσουμε το διάνυσμα  $\mathbf{b}_p$  (5.16) που φέρει πληροφορία για τον προσανατολισμό του άμεσου προγόνου της άρθρωσης. Το εν λόγω διάνυσμα προβάλλεται στο επίπεδο κάθετο προς τον  $\mathbf{u}$ , ώστε να διασφαλιστεί η ορθογωνιότητά του. Έτσι:

$$\mathbf{b} = \frac{\mathbf{b}_p - (\mathbf{b}_p \cdot \mathbf{u}) \mathbf{u}}{\|\mathbf{b}_p - (\mathbf{b}_p \cdot \mathbf{u}) \mathbf{u}\|} \quad (5.22)$$



Οπότε, ορίζουμε και το τρίτο διάνυσμα κατεύθυνσης με απλό εξωτερικό γινόμενο:

$$\mathbf{n} = \frac{\mathbf{u} \times \mathbf{v}}{\|\mathbf{u} \times \mathbf{v}\|} \quad (5.23)$$

Έχοντας, σε αυτό το σημείο, ορίσει τους τρεις τοπικούς ορθοκανονικούς άξονες, δύναται, πλέον, να ακολουθήσουμε την ίδια διαδικασία με την περίπτωση των ενδιάμεσων αρθρώσεων. Κοινώς:

$$\mathbf{B} = [\mathbf{b}, \mathbf{n}, \mathbf{v}] \quad (5.24)$$

$$\mathbf{B}_{Rest} = [\mathbf{b}_{Rest}, \mathbf{n}_{Rest}, \mathbf{v}_{Rest}] \quad (5.25)$$

$$\mathbf{R}_j = \mathbf{B}\mathbf{B}_{Rest}^T \quad (5.26)$$

$$\mathbf{R} = \mathbf{R}_p^T \mathbf{R}_j \quad (5.27)$$

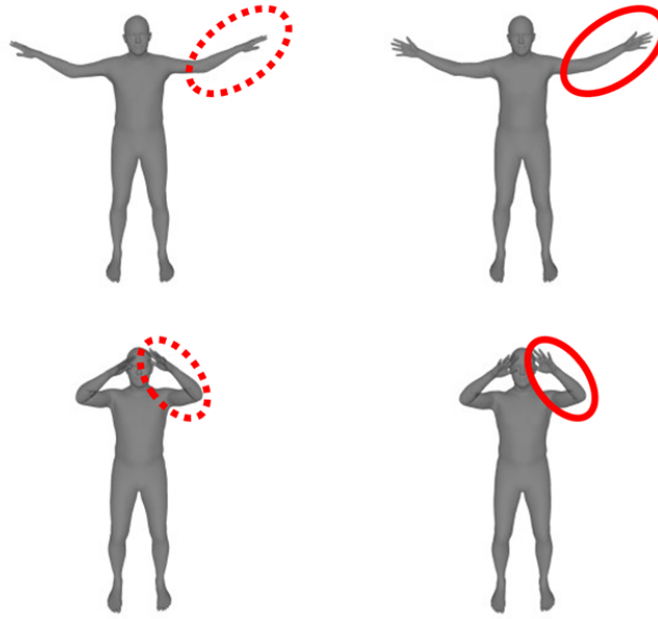
Και τέλος, η 6D συνεχής αναπαράσταση περιστροφής που δίδεται από την κάτωθι σχέση:

$$\mathbf{r} = [R_{11}, R_{21}, R_{31}, R_{12}, R_{22}, R_{23}]^T \quad (5.28)$$

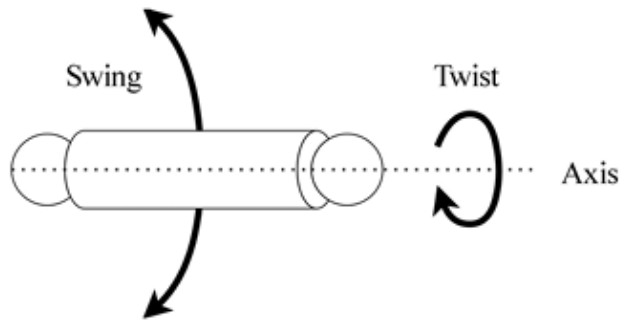
Ωστόσο, σε αυτό το σημείο οφείλουμε να σχολιάσουμε ότι οι τρισδιάστατες καρτεσιανές συντεταγμένες, τις οποίες κατέχουμε, δεν είναι αρκετές για να περιγράψουν εξ' ολοκλήρου την περιστροφική κίνηση. Όπως διακρίνεται στην εικόνα 5.6, δύναται να οριστούν διαφορετικές στάσεις με τις ίδιες σχετικές θέσεις των αρθρώσεων. Αυτό το φαινόμενο είναι γνωστό ως *Ασάφεια Στρέψης* (Twist Ambiguity), και αποτελεί ένα γνωστό πρόβλημα στο πεδίο παραγωγής παραμέτρων στάσης [107]. Δυστυχώς, η έρευνά μας δεν μας υπέδειξε κάποια μέθοδο με την οποία να μπορούμε να καθορίσουμε επακριβώς τις στρέψεις των αρθρώσεων από μονοφθαλμικές εικόνες anime. Ωστόσο, αυτό που μπορούμε να κάνουμε είναι να τις δεσμεύσουμε εντός ενός εύλογου και ρεαλιστικού εύρους.

Πιο συγκεκριμένα, παρατηρήσαμε ότι αν αρκούμασταν στους υπολογισμούς που παρατέθηκαν προ ολίγου, παρόλο που οι σχετικές θέσεις των αρθρώσεων ήταν σύμφωνες με τον εξαχθέν τρισδιάστατο σκελετό, πολλές φορές η γωνία στρέψης γύρω από την εκάστοτε άρθρωση λάμβανε ακραίες μη ρεαλιστικές τιμές—έως και 360°. Για να διορθώσουμε αυτό το πρόβλημα θα πρέπει να φέρουμε την γωνία στρέψης εντός λογικών πλαισίων δίχως να μεταβάλουμε την σχετική θέση των αρθρώσεων. Για τον υπολογισμό της θα εφαρμόσουμε τον τύπο *αποσύνθεσης ταλάντωσης-στρέψης του Dobrowolski* (Dobrowolski's swing-twist decomposition formula) [108].

Υποθέτουμε μία διάταξη σαν αυτήν την εικόνα 5.7: ήτοι, δύο αρθρώσεις—γονέα και παιδί—που συνδέονται μεταξύ τους με ένα άκρο. Η περιστροφική κίνηση του γονέα δύναται να αναλυθεί στις επιμέρους περιστροφές ταλάντωσης (*swing*) και στρέψης (*twist*). Για τον υπολογισμό της θα πρέπει να μετατρέψουμε την γνωστή από κοινού περιστροφική κίνηση σε τετραδόνιο (*quaternion*). Αυτό γίνεται, απλοϊκά δια της κατ' αρχάς μετατροπής του μητρώου περιστροφής  $\mathbf{R}$  σε αναπαράσταση άξονα-γωνία,  $\mathbf{r} =$



**Σχήμα 5.6:** Ασάφεια Στρέψης: Διαφορετικές στάσεις δύναται να οριστούν με τις ίδιες σχετικές θέσεις αρθρώσεων [107].



**Σχήμα 5.7:** Σχηματική αναπαράσταση μιας διάταξης δύο αρθρώσεων κι ενός άκρου. Η αριστερή σφαίρα συνιστά την άρθρωση-γονέα και η δεξιά την άρθρωση-παιδί. Η περιστροφή της γονικής άρθρωσης αναλύεται σε περιστροφές ταλάντωσης (*swing*) και στρέψης (*twist*) [107].

$[r_x, r_y, r_z]$ , βάσει της (5.11), και την μετέπειτα μετατροπή της τελευταίας σε τετραδόνιο,  $\mathbf{q} = [q_x, q_y, q_z, q_w]$ , βάσει της παρακάτω σχέσης:

$$\mathbf{q} = \left[ r_x \sin\left(\frac{\theta}{2}\right), r_y \sin\left(\frac{\theta}{2}\right), r_z \sin\left(\frac{\theta}{2}\right), \cos\left(\frac{\theta}{2}\right) \right] \quad (5.29)$$

Σημειώνουμε ότι η αντίστροφη διαδικασία μετατροπής από αναπαράσταση τετραδονίου σε άξονα-γωνία δίδεται από την σχέση (5.30), όπου—καθώς προκύπτει από την

(5.29)—η γωνία  $\theta = 2 \arccos(q_w)$ :

$$\mathbf{r} = \begin{cases} \frac{(q_x, q_y, q_z)}{\sin\left(\frac{\theta}{2}\right)}, & \text{αν } \sin\left(\frac{\theta}{2}\right) \neq 0 \\ \text{οποιοσδήποτε μοναδιαίος άξονας (π.χ. (1, 0, 0))}, & \text{αν } \sin\left(\frac{\theta}{2}\right) = 0 \end{cases} \quad (5.30)$$

Από τούδε και στο εξής θα συμβολίζουμε τον μετασχηματισμό από αναπαράσταση άξονα-γωνίας σε τετραδόνιο (5.29) ως  $\phi$ , και την αντίστροφη μετατροπή (5.30) ως  $\psi$ . Ας υποθέσουμε, λοιπόν,  $\mathbf{Q}_j$  την περιστροφή της άρθρωσης  $j$  σε αναπαράσταση τετραδονίου. Η περιστροφή  $\mathbf{Q}_j$  μπορεί να αναλυθεί στις επιμέρους περιστροφές ταλάντωσης  $\mathbf{S}_j$  και στρέψης  $\mathbf{T}_j$ , όπως η υποδεικνύει η σχέση:

$$\mathbf{Q}_j = \mathbf{T}_j \cdot \mathbf{S}_j \quad (5.31)$$

Έστω, τώρα,  $\mathbf{v}_j^{(rest)}$  το κατευθυντήριο διάνυσμα από την άρθρωση  $j$  στην άρθρωση-παιδί του στην στάση αναφοράς. Περιστρέφοντας το  $\mathbf{v}_j^{(rest)}$  κατά  $\mathbf{Q}_j$  ή  $\mathbf{S}_j$  λαμβάνουμε την περιστραμμένο κατευθυντήριο διάνυσμα  $\mathbf{v}_j^{(current)}$ :

$$\mathbf{v}_j^{(current)} = \mathbf{Q}_j \begin{bmatrix} \mathbf{v}_j^{(rest)} \\ 0 \end{bmatrix} \mathbf{Q}_j^{-1} = \mathbf{S}_j \begin{bmatrix} \mathbf{v}_j^{(rest)} \\ 0 \end{bmatrix} \mathbf{S}_j^{-1} \quad (5.32)$$

Όμως η  $\mathbf{S}_j$  μπορεί να υπολογιστεί από τα  $\mathbf{v}_j^{(rest)}$  και  $\mathbf{v}_j^{(current)}$  ως εξής:

$$\mathbf{S}_j = \phi \left( \arccos(\mathbf{v}_j^{(rest)} \cdot \mathbf{v}_j^{(current)}) \cdot \frac{\mathbf{v}_j^{(rest)} \times \mathbf{v}_j^{(current)}}{\|\mathbf{v}_j^{(rest)} \times \mathbf{v}_j^{(current)}\|} \right) \quad (5.33)$$

Μπορούμε, συνεπώς, να λάβουμε την περιστροφή στρέψης  $\mathbf{T}_j$ :

$$\mathbf{T}_j = \mathbf{Q}_j \cdot \mathbf{S}_j^{-1} \quad (5.34)$$

εκ της οποίας υπολογίζουμε την ζητούμενη γωνία στρέψης βάσει της (5.35):

$$\theta_j = \begin{cases} \|\psi(\mathbf{T}_j)\|, & \text{αν } \frac{\psi(\mathbf{T}_j)}{\|\psi(\mathbf{T}_j)\|} = \mathbf{v}_j^{(current)} \\ -\|\psi(\mathbf{T}_j)\|, & \text{διαφορετικά} \end{cases} \quad (5.35)$$

Τελικό βήμα συνιστά ο περιορισμός της γωνίας  $\theta_j$  εντός ενός εύρους  $(-\pi, \pi)$ .

Για κάθε άρθρωση, υπολογίζουμε την γωνία στρέψης κατά *Dobrowolski* και, επιπλέον, ορίζουμε ένα επιτρεπτό εύρος τιμών για αυτήν. Σε κάθε άρθρωση ορίζεται ένα ξεχωριστό εύρος αναλόγως με τον βαθμό ελευθερίας κίνησής της στην πραγματική ζωή. Φερ' ειπείν, θέσαμε για τους αγκώνες, που εν γένει παρουσιάζουν μεγάλη ευκινησία, επιτρεπτό εύρος  $(-\frac{\pi}{2}, \frac{\pi}{2})$ , ενώ για τις αρθρώσεις των δακτύλων, όπου η ελευθερία στρέψης τους είναι αρκετά περιορισμένη, το αντίστοιχο εύρος τέθηκε σε  $(-\frac{\pi}{6}, \frac{\pi}{6})$ . Σε περίπτωση που η υπολογιζόμενη γωνία στρέψης βρεθεί εκτός του επιτρεπτού ορίου,

θα πρέπει να την επαναπροσδιορίσουμε αυθαίρετα. Μια πρώτη σκέψη ήταν, απλώς, να την αντικαταστήσουμε με την κοντινότερη επιτρεπτή τιμή. Μια τέτοια προσέγγιση, σίγουρα, θα ήταν έγκυρη δεδομένου ότι δεν διαθέτουμε κάποια επιπρόσθετη πληροφορία για την πραγματική στρέψη της άρθρωσης. Εντούτοις, τα τελικά αποτελέσματα δεν ήταν ικανοποιητικά· λογικό, αφού για μεγάλο πλήθος των αρθρώσεων η γωνία στρέψης οριζόταν στην μέγιστη δυνατή. Αντί αυτού, προκειμένου να διατηρήσουμε, υπό κάποια έννοια, την πληροφορία που φέρει ο αρχικός υπολογισμός της περιστροφικής κίνησης, ανακλούμε, κατ' αρχάς την γωνία στρέψης γύρω από το  $\pm\pi$ , και αν εξακολουθεί να είναι εκτός του οριοθετημένου επιτρεπτού εύρους, τότε την ανάγουμε στην κοντινότερη επιτρεπτή τιμή. Δηλαδή:

$$\theta'_j = \begin{cases} \theta_j - \pi, & \text{αν } \theta_j \geq 0 \\ -\theta_j - \pi, & \text{αν } \theta_j < 0 \end{cases} \quad (5.36)$$

$$\theta_j^{(clamped)} = \min(\max(\theta'_j, -r), r) \quad (5.37)$$

όπου  $[-r, r]$  το επιτρεπτό εύρος στρέψης της άρθρωσης  $j$ .

Κατόπιν, περιορισμού της γωνίας στρέψης εντός ρεαλιστικών ορίων, υπολογίζουμε την ορθή πλέον 6D συνεχή αναπαράσταση περιστροφής, ακολουθώντας τα αντίστροφα βήματα του αλγορίθμου του *Dobrowolski*:

$$\mathbf{T}_j^{(clamped)} = \phi\left(\theta_j^{(clamped)} \cdot \mathbf{v}_j^{(rest)}\right) \quad (5.38)$$

$$\mathbf{Q}'_j = \mathbf{T}_j^{(clamped)} \cdot \mathbf{S}_j \quad (5.39)$$

Αν  $\mathbf{Q}'_j = [q'_x, q'_y, q'_z, q'_w]$  η νέα περιστροφική κίνηση σε μορφή τετραδονίου, υπολογίζουμε το νέο  $3 \times 3$  μητρώο περιστροφής ως εξής:

$$\mathbf{R}'_j = \begin{bmatrix} 1 - 2(q_y'^2 + q_z'^2) & 2(q'_x q'_y - q'_z q'_w) & 2(q'_x q'_z - q'_y q'_w) \\ 2(q'_x q'_y - q'_z q'_w) & 1 - 2(q_x'^2 + q_z'^2) & 2(q'_y q'_z - q'_x q'_w) \\ 2(q'_x q'_z - q'_y q'_w) & 2(q'_y q'_z - q'_x q'_w) & 1 - 2(q_x'^2 + q_y'^2) \end{bmatrix} \quad (5.40)$$

Από την οποία εξάγουμε την ζητούμενη 6D συνεχή αναπαράσταση περιστροφής δια της γνωστής σχέσης (5.28).

Η άνωθι μέθοδος δεν αποτελεί σε καμία περίπτωση την μοναδική που μπορεί να καταστρωθεί για τον έλεγχο της γωνίας στρέψης· εντούτοις, επιβεβαιώσαμε ότι παράγει αισθητά καλύτερα και πιο ρεαλιστικά δείγματα. Βέβαια, οφείλουμε να αναγνωρίσουμε ότι δεν αντιμετωπίζει το σύνολο των προβληματικών περιπτώσεων. Η περίπλοκη φύση της ανθρώπινης κινησιολογίας συνεπάγεται ότι πολλές αρθρώσεις εξακολουθούν να ακολουθούν μη-ρεαλιστική συμπεριφορά η οποία δεν είναι δυνατόν να αναιρεθεί δια της απλοϊκής μεθόδου που έχουμε επιστρατεύσει. Για παράδειγμα, ο βαθμός δυνατής στρέψης μιας άρθρωσης-παιδί μπορεί να εξαρτάται από την στρέψη που έχει ήδη λάβει η άρθρωση-γονέας. Σε αυτό το σημείο έγκειται και η απόφασή μας να παραλείψουμε τον υπολογισμό των παραμέτρων προσώπου. Σε συνδυασμό με τον προαναφερθέντα λόγο

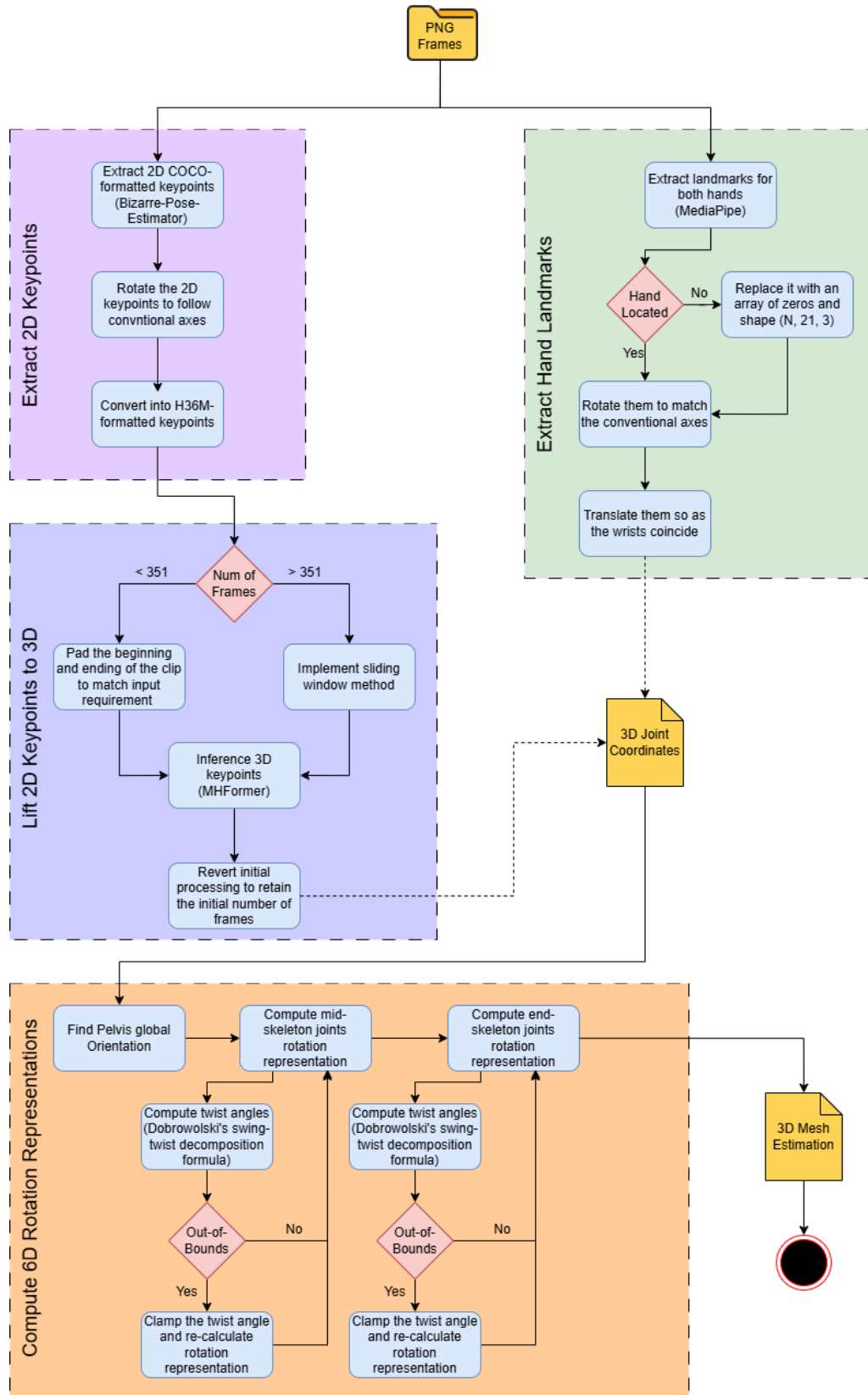
περί υπολειτουργίας των διαθέσιμων μοντέλων και της μη-συμμετρικότητας των εξαχθέντων χαρακτηριστικών, τα keypoints του προσώπου αποτελούν αρθρώσεις φύλλα στο *SMPL-X*. Αυτό σημαίνει ότι η σχετική τους θέση (η πληροφορία που θα λαμβάναμε για αυτά από κάποιο 3D pose estimator) καθορίζεται εξ' ολοκλήρου από την περιστροφή της άρθρωσης-πατέρα τους. Η δικές τους παράμετροι καθορίζουν μονάχα την γωνία στρέψης τους η οποία, ωστόσο, θα έπρεπε είναι αρκετά περιορισμένη για φυσιολογικούς λόγους αλλά λαμβάνει συχνά μη-ρεαλιστικές τιμές λόγω του *Twist Ambiguity*. Συνεπώς, αποφασίσαμε να μηδενίσουμε τις εν λόγω τιμές, γεγονός που αυτομάτως αποδίδει στις εν λόγω αρθρώσεις τις θέσεις που θα είχαν σε κατάσταση ηρεμίας—στην στάση αναφοράς. Η ίδια προσέγγιση ακολουθήθηκε και στις επιπρόσθετες αρθρώσεις που απουσιάζουν από τον σκελετό του *Human3.6M*.

Ωστόσο, με αυτό το τελευταίο βήμα διόρθωσης της γωνίας στρέψης, ολοκληρώνεται η διαδικασία υπολογισμού των  $\theta$  παραμέτρων του *SMPL-X* [54], και, ως εκ τούτου, ολόκληρης της διαδικασίας εκτίμησης του πλέγματος από μονοφθαλμικά σύντομα βίντεο κλιπ Anime. Η εν λόγω διαδικασία αναπαρίσταται σχηματικά στην εικόνα 5.8.

## 5.6 Λεπτομέρειες Εκτέλεσης και Αποτελέσματα

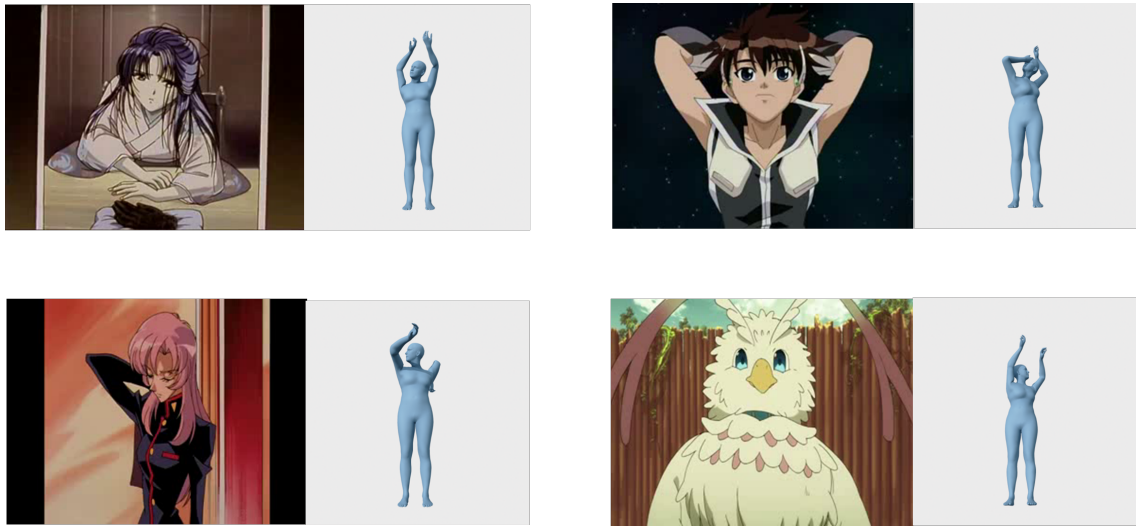
Όπως αναφέρθηκε στην ενότητα 5.3, το επιλεγμένο σύνολο δεδομένων περιλαμβάνει 437 χιλιάδες σύντομα βίντεο κλιπ Anime. Τελικά, εξ' αυτών, εκτελέσαμε την μέθοδο που περιεγράφηκε στο παρόν Κεφάλαιο σε ένα υποσύνολο 66 χιλιάδων βίντεο κλιπς. Τους λόγους συνιστούσαν μια σειρά διαφορετικών πρακτικών ζητημάτων, μεταξύ των οποίων η έλλειψη χρόνου, και οι δυσκολίες που θα αντιμετωπίζαμε κατά την διαδικασία εκπαίδευσης (Κεφάλαιο 6) σε περίπτωση μεγάλου συνόλου δεδομένων εκπαίδευσης. Από αυτά προέκυψαν 3545 διαφορετικά δείγματα που περιλαμβάνουν ένα WAV και ένα NPZ αρχείο. Εξάλλου, υπενθυμίζουμε ότι από την διαδικασία φιλτραρίσματος περνάει μόνον περί το 5% των αρχικών βίντεο κλιπς 5.4, δεδομένου ότι δεν είναι όλα κατάλληλα για την χρήση που επιθυμούμε να κάνουμε. Η συνολική διάρκεια ανέρχεται περί τις 6 ώρες βίντεο ώρες βίντεο. Ο, δε, κώδικας που αναπτύξαμε σε *Python* για την εκπόνηση της περιγραφείσας μεθόδου διαχειρίζεται περί τα 200 βίντεο κλιπς την ώρα σε υπολογιστικό περιβάλλον που βασίζεται σε εικονική μηχανή *Amazon EC2* με τέσσερις πυρήνες *AMD EPYC 7R13*, 30 GiB μνήμης RAM, και κάρτα γραφικών *NVIDIA L40S* με 46 GiB μνήμης.

Ακολουθούν ορισμένα ενδεικτικά δείγματα (frames) που παράχθηκαν. Περισσότερα εξ' αυτών παρατίθενται στο Παράρτημα Α.



Σχήμα 5.8: Η διαδικασία εκτίμησης του τρισδιάστατου πλέγματος.





Σχήμα 5.9: Ενδεικτικά δείγματα (μεμονωμένα frames) του συνόλου δεδομένων μας. Για κάθε ζεύγος, η αριστερή εικόνα είναι το αρχικό frame και η δεξιά το τρισδιάστατο πλέγμα που εκτιμήθηκε με την διαδικασία της ενότητας 5.5.

Από την εικόνα 5.9 μπορούμε να εξαγάγουμε ορισμένα πολύ σημαντικά συμπεράσματα για την αποτελεσματικότητα της διαδικασίας εκτίμησης του τρισδιάστατου πλέγματος. Οι εικόνες της πάνω σειράς αποτελούν δύο παραδείγματα που ο αλγόριθμος λειτούργησε ως άρμοξε. Ασφαλώς, το πρόβλημα της ασάφειας περιστροφής υφίσταται και είναι αισθητό. Επιπροσθέτως, κάθε ενδιαμέσο βήμα—δισδιάστατη εκτίμηση, τρισδιάστατη ανύψωση, εκτίμηση landmarks των χεριών—απομακρύνει, ελαφρώς, τις παραμέτρους μας από την πραγματική *ground-truth*. Εντούτοις, κάτι τέτοιο είναι αναμενόμενο, δεδομένου ότι εργαζόμαστε με *in-the-wild* δείγματα, οπότε και αποσκοπούμε στην παραγωγή *pseudo-ground truth* ετικετών. Λαμβάνοντας υπ’ όψιν την υψηλή περιπλοκότητα μιας διαδικασίας εκτίμησης τρισδιάστατου πλέγματος από μονοφθαλμικά βίντεο κλιπς, την επιπλέον δυσκολία που προσθέτει η ενασχόληση με έντονα στυλιζαρισμένους χαρακτήρες Anime, καθώς και η σχετική έλλειψη απαραίτητων εργαλείων στην βιβλιογραφία, θεωρούμε ότι τα εν λόγω αποτελέσματα είναι αποδεκτής ποιότητας.

Από την άλλη πλευρά, οι εικόνες της κάτω σειράς αντιπροσωπεύουν τα δύο συχνότερα σφάλματα που εντοπίσαμε κατά την επισκόπηση των παραγόμενων δειγμάτων. Ας τα αναλύσουμε περεταίρω. Η κάτω αριστερά εικόνα είναι ενδεικτική του αποτελέσματος που προκύπτει στην περίπτωση που ο *Bizarre-Pose-Estimator* [94] δεν κατορθώσει να εντοπίσει κάποιο άκρο. Αυτό συνιστά, ενδεχομένως, το πιο συχνό σφάλμα που καλούμαστε να αντιμετωπίσουμε. Εξάλλου, κατ’ ομολογίαν των συγγραφέων, το *Bizarre-Pose-Estimator* δεν κατασκευάστηκε ώστε να διαχειρίζεται τυχόν επικαλύψεις. Παρατηρούμε, συνεπώς, στην κάτω αριστερά εικόνα, το δεξί χέρι του απεικονιζόμενου χαρακτήρα να έχει εντοπιστεί και εκτιμηθεί καθώς πρέπει· το αριστερό, όμως, όχι. Αυτό οδήγησε τον *Bizarre-Pose-Estimator* να επιστρέψει την εκτίμηση ενός φασματικού άκρου που, τελικά, οδήγησε στο διπλανό αποτέλεσμα. Η, δε, κάτω δεξιά εικόνα της 5.9 αντιπροσωπεύει την περίπτωση που έντονες δομές, παρούσες στο εκάστοτε frame, που ανήκουν είτε στο ίδιο το *avatar* είτε στο υπόβαθρο, “μπερδεύουν” τον *Bizarre-Pose-*



*Estimator*, ο οποίος τις συγχέει με τα άκρα του απεικονιζόμενου χαρακτήρα. Αφέλειες μαλλιών, κολώνες κτηρίων, σπαθιά και κλαδιά αποτελούν μονάχα ορισμένες από τέτοιες δομές που έχουμε εντοπίσει να οδηγούν ενίοτε τον δισδιάστατο εκτιμητή σε εσφαλμένες παραγωγές. Με παρόμοιο τρόπο, οι κεραίες της ανθρωπόμορφης κουκουβάγιας της κάτω δεξιάς εικόνας θεωρήθηκαν ως χέρια οδηγώντας στο απεικονιζόμενο αποτέλεσμα.

Σε αυτό το σημείο, βέβαια, οφείλουμε να σχολιάσουμε ότι τέτοιου είδους σφάλματα είναι πλήρως αναμενόμενα σε οιονδήποτε αλγόριθμο αυτόματης ή ημιαυτόματης παραγωγής συνόλων δεδομένων και δεν θα έπρεπε να θεωρηθούν ως μετρικές σχολιασμού της λειτουργικότητάς τους. Επόμενο λογικό βήμα σε κάθε δημιουργία dataset από in-the-wild δεδομένα συνιστά η χειρωνακτική επισκόπηση όλων των παραχθέντων δειγμάτων και η απόρριψη των λανθασμένων. Ωστόσο, η περάτωση μίας τέτοιας διαδικασίας δεν αποτελεί δεν ήταν ρεαλιστική για τα πλαίσια της παρούσης διπλωματικής εργασίας.

## Κεφάλαιο 6

### Εκπαίδευση

---

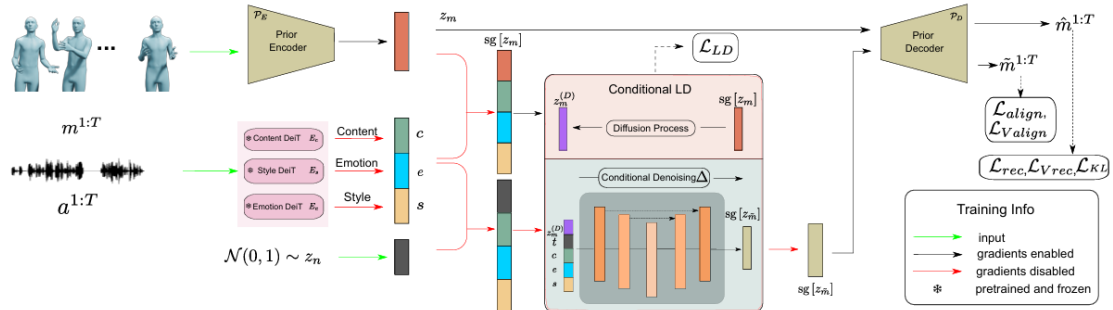
Κατόπιν ολοκλήρωσης της διαδικασίας δημιουργίας του συνόλου δεδομένων εκπαίδευσης που περιεγράφηκε στο προηγούμενο Κεφάλαιο, επόμενο φυσικό βήμα αποτελεί η χρήση του για την εκπαίδευση ή *fine-tuning* ενός μοντέλου διάχυσης για την παραγωγή τρισδιάστατων κινούμενων σχεδίων βάσει της ομιλίας και του συναισθήματος, που συνιστά και τον δεύτερο ερευνητικό μας στόχο, όπως αυτοί εκφράστηκαν στο Κεφάλαιο 4. Πρώτο βήμα, για την περάτωση αυτού μας του στόχου, αποτελεί η επιλογή της καταλληλότερης αρχιτεκτονικής. Υπενθυμίζουμε ότι η σχετική βιβλιογραφία επί του θέματος εξετάστηκε και παρατέθηκε στο Κεφάλαιο 3. Εξ' αυτών διακρίναμε δύο αρχιτεκτονικές οι οποίες θεωρούνται *State-of-the-Art* και οι οποίες, αδιαφιλονίκητα, έχουν αναδείξει εντυπωσιακά αποτελέσματα. Πρόκειται για το *DiffSHEG* [27] και το *AMUSE* [1].

Το *DiffSHEG* (*Diffusion-based Speech-driven Human Expression and Gesture Generator*) [27] χωρίζεται σε δύο κύρια μέρη: τους *Audio Encoders* και τον *UniEG Transformer Generator*. Αρχικά το ηχητικό σήμα εισάγεται στους κωδικοποιητές ήχου όπου μετατρέπεται τόσο σε χαρακτηριστικά χαμηλού επιπέδου, μέσω ενός *Mel-Spectrogram Converter*, όσο και χαρακτηριστικά υψηλού επιπέδου, μέσω του προεκπαιδευμένου *HuBERT Encoder* [109]. Στην συνέχεια, αμφότερα διανύσματα των χαρακτηριστικών του ήχου συνενώνονται μεταξύ τους και τροφοδοτούνται στον *UniEG Transformer Generator*. Ο τελευταίος διακρίνεται, επιπλέον, σε δύο επιμέρους δομικά μπλοκς: το *Motion-Speech Fusion Residual Block* και το *Style-aware Transformer Block*. Το συνενωμένο διάνυσμα ήχου περνά, κατ' αρχάς, από το πρώτο μπλοκ το οποίο, επίσης, ενώνει τα ακουστικά χαρακτηριστικά με πληροφορίες κίνησης, και κατόπιν από δύο διακριτούς *Transformer Encoders*—έναν για τις εκφράσεις του προσώπου και έναν για τις κινήσεις του σώματος. Ο μηχανισμός διάχυσης προσθέτει θόρυβο στα δεδομένα κίνησης και τον αφαιρεί βηματικά μέσω των *Style-aware Transformer Blocks* που χρησιμοποιούνται προκειμένου να διατηρήσουν το στυλ της ομιλίας και της κίνησης. Η καινοτομία του *DiffSHEG* έγκειται στην υιοθέτηση στην αρχιτεκτονικής του μιας μονοκατευθυνόμενης ροής πληροφορίας από τις εκφράσεις στις κινήσεις. Ο λόγος για αυτήν τους την απόφαση αποτελεί η αντίληψη ότι οι εκφράσεις του προσώπου και οι κινήσεις των χεριών δεν είναι ασύνδετες μεταξύ τους· αλλά επηρεάζουν άμεσα η μία την άλλη. Η εν λόγω μονοκατευθυ-

νόμηση ροή πληροφορίας υλοποιείται μέσω της εξάρτησης της διαδικασίας διάχυσης του *Transformer Encoder* των χειρονομιών από το εκάστοτε προβλεπόμενο διάνυσμα έκφρασης και χρησιμοποιείται με σκοπό να μαθευτούν κοινές κατανομές και χρονική συνοχή μεταξύ εκφράσεων προσώπου και κινήσεων του σώματος.

Ωστόσο, όπως καθίσταται σαφές από την άνωθεν περιγραφή της αρχιτεκτονικής του, το *DiffSHEG* εξαρτάται σε πολύ μεγάλο βαθμό από τις εκφράσεις του προσώπου. Το μοντέλο χρησιμοποιεί και αυτό παραμέτρους *SMPL-X* [54]· εντούτοις, πέρα από τις παραμέτρους  $\theta$  της στάσης του σώματος τις οποίες παράγουμε με την μεθοδολογία του Κεφαλαίου 5, απαιτεί και τις παραμέτρους  $\psi$ , τις εκφράσεις, δηλαδή, προσώπου. Δυστυχώς, η έρευνά μας δεν ανέδειξε κάποιο learning-based εργαλείο ικανό να παραγάγει τέτοιου είδους παραμέτρους σε στυλιζαρισμένους χαρακτήρες, πόσο μάλλον σε χαρακτήρες Anime που φημίζονται για τα έντονα χαρακτηριστικά του προσώπου τους. Ως εκ τούτου, η επιστράτευση του *DiffSHEG* για την υλοποίηση του δικού μας ερευνητικού στόχου καθίσταται μη ρεαλιστική. Τουναντίον, το *AMUSE* [1] επικεντρώνεται στο ίδιο το σώμα και την κινησιολογία του, εφαρμόζει μια πιο διαισθητική αρχιτεκτονική, και αξιοποιεί μοντέλα διάχυσης καθοδηγούμενα τόσο από την ομιλία όσο και από το συναίσθημα και το προσωπικό στυλ του ομιλητή. Αποτελεί, συνεπώς, μια φυσική επιλογή για εμάς.

## 6.1 Η Αρχιτεκτονική του AMUSE



Σχήμα 6.1: Σχηματική αναπαράσταση της αρχιτεκτονικής του AMUSE [1].

Το μοντέλο *AMUSE*, δημιουργήθηκε από τους *Chhatre et al.* [1], το 2024, οπότε κι έγινε δεκτό στο διεθνές συνέδριο *CVPR* τον ίδιο χρόνο. Η αρχιτεκτονική του, όπως αυτή διακρίνεται στην εικόνα 6.1, απαρτίζεται από δύο διακριτά επιμέρους δίκτυα, τα οποία εκπαιδεύονται ξεχωριστά. Κατ' αρχάς, υφίσταται το μπλοκ αποσύνθεσης του ήχου (*audio disentanglement module*), σκοπός του οποίου αποτελεί η κωδικοποίηση της ομιλίας-εισόδου σε τρία διαφορετικά λανθάνοντα διανύσματα που αντιπροσωπεύουν το περιεχόμενο της ομιλίας, το συναίσθημα, και το προσωπικό στυλ του ομιλητή. Κατόπιν, συναντάται το κατ' εξοχήν δομικό μπλοκ του *AMUSE*, υπεύθυνο για την παραγωγή των τρισδιάστατων κινουμένων σχεδίων. Απαρτίζεται από ένα 3D *motion prior* του ανθρώπινου σώματος και από ένα μοντέλο λανθάνουσας διάχυσης. Λαμβάνει για είσοδο τυχαίο θόρυβο συνδυασμένο με τα λανθάνοντα διανύσματα του μπλοκ αποσύνθεσης του ήχου και εξάγει μια ακολουθία ανθρώπινης κίνησης.

### 6.1.1 Μπλοκ Αποσύνθεσης Ήχου

Αναμφίβολα, τρεις από τους πιο σημαντικούς παράγοντες που καθορίζουν τις εκάστοτε κινήσεις του σώματος αποτελούν το περιεχόμενο, ήτοι οι χειρονομίες σε σχέση με τον ρυθμό ομιλίας και το ειπωμένες λέξεις, η συναισθηματική κατάσταση και το προσωπικό στυλ του ομιλητή. Οι συγγραφείς του AMUSE διαπίστωσαν ότι αυτά τα τρία χαρακτηριστικά είναι διαχωρίσιμα. Σκοπός τους αποτελεί η “παραγοντοποίηση” του ήχου-εισόδου σε τρεις διακριτές και διαχωρισμένες λανθάνουσες αναπαραστάσεις των εν λόγω χαρακτηριστικών.

Για να το επιτύχουν αυτό, κατέστρωσαν μία εξειδικευμένη αρχιτεκτονική *Encoder-Decoder*, με τρεις διαφορετικούς *Encoders*—έναν για κάθε λανθάνοντα χώρο. Αν  $a$  το ηχητικό δείγμα εισόδου περασμένο από ένα *filterbank* (ενότητα 6.2), τότε συμβολίζουμε τους τρεις λανθάνοντες χώρους  $c$ ,  $e$ ,  $s$  του περιεχομένου, συναισθήματος, και στυλ, αντιστοίχως, ως εξής:

$$c = E_c(a), \quad e = E_e(a), \quad s = E_s(a) \quad (6.1)$$

όπου  $E_c$ ,  $E_e$ , και  $E_s$  οι αντίστοιχοι *Encoders*. Κατόπιν, ο *Decoder* λαμβάνει τα τρία διανύσματα και επιχειρεί να ανακατασκευάσει το δεδομένο *filterbank*. Επισήμως:

$$D(c, e, s) = \hat{a} \quad (6.2)$$

Ο *Decoder* απαρτίζεται από ένα *fusion module* και ένα *transformer-decoder*.

Ας τα εξετάσουμε πιο αναλυτικά. Όπως ήδη αναφέραμε, το εισαχθέν *filterbank* περνάει ταυτοχρόνως από τους *encoders*  $E_c$ ,  $E_e$ , και  $E_s$  παράγοντας τρία διαχωρισμένα λανθάνοντα διανύσματα περιεχομένου, συναισθήματος και στυλ. Για την υλοποίησή τους, οι συγγραφείς ακολούθησαν την μέθοδο των Gong et al. [110, 111] όπου κατέστρωσαν μια προσέγγιση τύπου transfer learning για το πέρασμα από το οπτικό μέσον στο ακουστικό με την βοήθεια προεκπαιδευμένων *DeiT* (*Data-efficient Image Transformers*)· *Visual Transformers*, δηλαδή, που εκπαιδεύονται για το πρόβλημα της ταξινόμησης εικόνων δίχως συνέλιξη [112]. Ουσιαστικά, μέσω της τροφοδότησης του ακατέργαστου ήχου σε *filterbank*, τα χαρακτηριστικά του λαμβάνουν την μορφή “εικόνας”, αφού μετατρέπονται σε έναν δισδιάστατο πίνακα με άξονες τον χρόνο και την συχνότητα, επιτρέποντάς τους, καθ’ αυτόν τον τρόπο να χρησιμοποιηθούν ως είσοδοι σε έναν *ViT*. Καθένας εκ των τριών *encoders* απαρτίζεται από έναν τέτοιο *DeiT* ο οποίος τροφοδοτείται από διαδοχικά τμήματα του *filterbank* εισόδου που αντιστοιχούν σε ένα frame. Τελική έξοδος αποτελεί το εκάστοτε μονοδιάστατο λανθάνον διάνυσμα. Ειδικά για τους *encoders* του συναισθήματος και του στυλ, οι ερευνητές υπολογίζουν τον μέσο όρο των εκάστοτε εξαγόμενων tokens κλάσης και απόσταξης (*class token* και *distillation token*) του *DeiT*—τμήματα του τελικού διανύσματος—προκειμένου να ταξινομήσουν το ηχητικό δείγμα σε 8 διαφορετικά συναισθήματα και 30 ετικέτες στυλ. Εξάλλου, υπενθυμίζουμε ότι το AMUSE είναι εκπαιδευμένο επί του συνόλου δεδομένων BEAT [45] που φέρει τα άνωθι χαρακτηριστικά. Σε αυτό το σημείο, αξίζει να σημειωθεί ότι παρά την εκ πρώτης όψεως αντιφατική απόφαση διαλογής οπτικών μετασχηματιστών για την επεξεργασία ηχητικών δειγμάτων, η επιλογή αυτή δίδει την δυνατότητα εκμετάλλευσης της ικανότητας των *DeiT* να μαθαίνουν πλούσιες και αφηρημένες δομές, ακόμη και κατά την μεταφορά στον χώρο του ήχου, καθιστώντας, έτσι, την διαδικασία εκπαίδευσης

πιο αποδοτική και σταθερή.

Εν συνεχεία, προκειμένου να διασφαλιστεί ότι οι τρεις λανθάνοντες χώροι που παρήχθησαν από τους *Transformer Encoders* είναι πραγματικά διαχωρισμένοι, εισάγεται ένας μηχανισμός *cross reconstruction*. Ο μηχανισμός αυτός βασίζεται στην ανταλλαγή λανθανουσών αναπαραστάσεων προερχόμενων από διαφορετικά ηχητικά δείγματα και υφίσταται έτσι ώστε να επιβάλλεται σε κάθε χώρο να κωδικοποιεί μονάχα το αντίστοιχο χαρακτηριστικό που έχει αναλάβει. Η εν λόγω ανταλλαγή ενδεχομένως να αφορά είτε το περιεχόμενο (*content swap*) είτε το συναίσθημα και το στυλ (*emotion/style swap*). Η διαδικασία αυτή επαναλαμβάνεται για κάθε ηχητικό δείγμα στο σύνολο δεδομένων εκπαίδευσης, δημιουργώντας, καθ' αυτόν τον τρόπο, ένα πλούσιο πλήθος *cross-reconstruction* περιπτώσεων.

Τέλος, ακολουθεί ο *Decoder*, που όπως έχουμε αναφέρει απαρτίζεται από ένα *fusion module* και ένα *transformer-decoder*. Οι αρχιτεκτονικές και των δύο αυτών μπλοκ βασίζονται σε μετασχηματιστές. Το πρώτο είναι υπεύθυνο για την δημιουργία μιας ενιαίας λανθάνουσας αναπαράστασης εφαρμόζοντας *cross attention* στις τριπλέτες εισόδου ( $c, e, s$ ). Ο, δε, *transformer-decoder*, λαμβάνει την νέα ενιαία, συμπιεσμένη λανθάνουσα αναπαράσταση και επιχειρεί να ανακατασκευάσει το αρχικό *filterbank*. Η συνολική αναπαράσταση της αρχιτεκτονικής του Μπλοκ Αποσύνθεσης Ήχου διακρίνεται στην εικόνα 6.2.

Όσον αφορά τις λεπτομέρειες εκπαίδευσης, τώρα, οι συγγραφείς κάνουν χρήση πολλαπλών όρων απώλειας προκειμένου να πετύχουν τα καλύτερα δυνατά αποτελέσματα. Πέραν της κλασικής απώλειας ανακατασκευής του *autoencoder*,  $\mathcal{L}_{self}$ , εισάγουν, επιπρόσθετα, τρεις όρους απώλειας διασταυρούμενης ανακατασκευής, που συνδέονται άμεσα με τον προαναφερθέντα μηχανισμό *cross reconstruction*,  $\mathcal{L}_{xcon}$ ,  $\mathcal{L}_{xemo}$ ,  $\mathcal{L}_{xsty}$ , αλλά και τρεις απολεστικούς όρους στα εκτιμώμενα λανθάνοντα διανύσματα των χαρακτηριστικών που εξάγονται,  $\mathcal{L}_{con}$ ,  $\mathcal{L}_{emo}$ ,  $\mathcal{L}_{sty}$ . Αναλυτικότερα, κατά την διάρκεια της εκπαίδευσης, ένα δείγμα αποτελεί, κατ' ουσίαν, μία τετράδα διαφορετικών ηχητικών δεδομένων που παρουσιάζουν δύο διαφορετικά περιεχόμενα (σενάρια)  $c_1, c_2$ , δύο διαφορετικά στυλ  $s_1, s_2$ , και το ίδιο συναίσθημα  $e_1$ . έστω:

$$a_1 = a_{c_1, e_1, s_1}, \quad a_2 = a_{c_2, e_1, s_1}, \quad a_3 = a_{c_1, e_1, s_2}, \quad a_4 = a_{c_2, e_1, s_2} \quad (6.3)$$

Αν, λοιπόν,  $E$  ο εκάστοτε Encoder και  $D$  ο Decoder, η απώλεια αυτοανακατασκευής (*self-reconstruction*), που υπολογίζεται βάσει της *L1 νόρμας*, εγγυάται την ικανότητα ανακατασκευής του αρχικού δείγματος από τους τρεις εξαχθέντες λανθάνοντες χώρους  $c, e, s$ :

$$\mathcal{L}_{self} := \sum_{k=1}^4 \|D(E_c(a_k), E_e(a_k), E_s(a_k)) - a_k\|_1 \quad (6.4)$$

Από την άλλη πλευρά, η απώλεια περιεχομένου εξασφαλίζει ότι δύο λανθάνοντα περιεχόμενα προερχόμενα από δύο διαφορετικά ηχητικά δείγματα ίδιου περιεχομένου,

αλλά διαφορετικού στυλ, ταιριάζουν μεταξύ τους:

$$\mathcal{L}_{con} := \sum_{k=1}^2 \|\mathbf{E}_c(\mathbf{a}_k) - \mathbf{E}_c(\mathbf{a}_{k+2})\|_1 \quad (6.5)$$

Όσον αφορά, δε, τις απώλειες ταξινόμησης του συναισθήματος και του στυλ, αυτές προκύπτουν δια της προβολής τους σε έναν διανυσματικό ταξινομητή, μέσω μιας γραμμικής κεφαλής προβολής, και του μετέπειτα υπολογισμού της διασταυρωμένης απώλειας εντροπίας της ταξινόμησης:

$$\mathcal{L}_{emo} := - \sum_{1 \leq l_e \leq n_e} y_{l_e} \log p_{l_e} \quad (6.6)$$

$$\mathcal{L}_{sty} := - \sum_{1 \leq l_s \leq n_s} y_{l_s} \log p_{l_s} \quad (6.7)$$

όπου  $n$  το πλήθος των εκάστοτε κατηγοριών ( $n_e = 8$  για το συναίσθημα και  $n_s = 30$  για το προσωπικό στυλ),  $y_l$  η πραγματική one-hot ετικέτα του δείγματος για την εκάστοτε ταξινομική κατηγορία  $l$ , και  $p_l$  η προβλεπόμενη πιθανότητα το δείγμα να ανήκει στην κατηγορία  $l$ . Τέλος, εισάγονται οι cross-reconstruction απώλειες για την διασφάλιση της πλήρους αποσύνθεσης των τριών λανθανόντων αναπαραστάσεων. Αυτές, ορίζονται ως εξής:

$$\mathcal{L}_{xcon} := \sum_{k=1}^4 \mathbf{D}(\mathbf{E}_c(\mathbf{a}_{j(k)}), \mathbf{E}_e(\mathbf{a}_k), \mathbf{E}_s(\mathbf{a}_k)) - \mathbf{a}_k, \quad j(k) = [(1+k) \bmod 4] + 1 \quad (6.8)$$

$$\mathcal{L}_{xemo} := \sum_{k=1}^4 \mathbf{D}(\mathbf{E}_c(\mathbf{a}_k), \mathbf{E}_e(\mathbf{a}_{j(k)}), \mathbf{E}_s(\mathbf{a}_k)) - \mathbf{a}_k, \quad j(k) = [(6-k) \bmod 4] + 1 \quad (6.9)$$

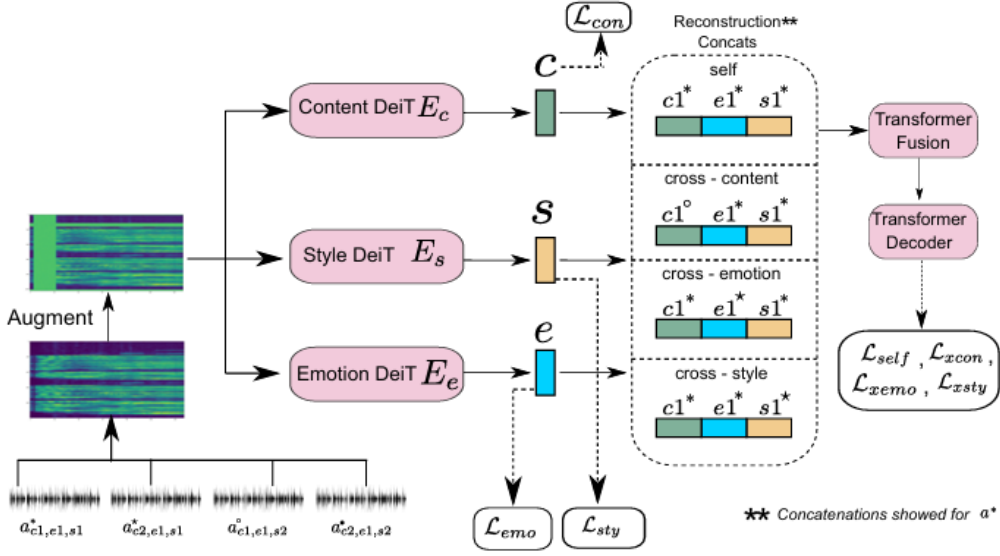
$$\mathcal{L}_{xsty} := \sum_{k=1}^4 \mathbf{D}(\mathbf{E}_c(\mathbf{a}_k), \mathbf{E}_e(\mathbf{a}_k), \mathbf{E}_s(\mathbf{a}_{j(k)})) - \mathbf{a}_k, \quad j(k) = [(6-k) \bmod 4] + 1 \quad (6.10)$$

Έχοντας, σε αυτό το σημείο, ορίσει όλα τα επιμέρους σφάλματα, η συνολική απώλεια του ήχου ορίζεται, απλοϊκά, ως εξής:

$$\mathcal{L}_{dis} := \mathcal{L}_{self} + \mathcal{L}_{con} + \mathcal{L}_{emo} + \mathcal{L}_{sty} + \mathcal{L}_{xcon} + \mathcal{L}_{xemo} + \mathcal{L}_{xsty} \quad (6.11)$$

Το Μπλοκ Αποσύνθεσης Ήχου εκπαιδεύεται σε ηχητικά δείγματα διάρκειας της τάξεως των 10 δευτερολέπτων, προερχόμενα από το σύνολο δεδομένων BEAT [45]. Κατόπιν περάτωσης της διαδικασίας εκπαίδευσης παράγει τρία πλήρως αποσυντεθημένα λανθάνοντα διανύσματα περιεχομένου, συναισθήματος και προσωπικού στυλ, τα οποία και τροφοδοτούνται στο επακόλουθο μοντέλο διάχυσης.





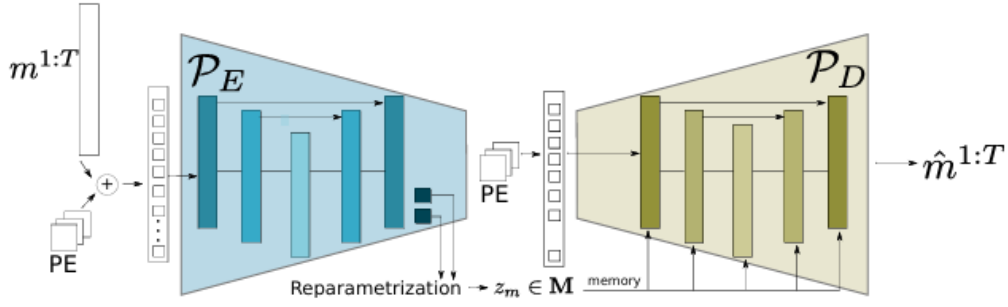
Σχήμα 6.2: Σχηματική αναπαράσταση της αρχιτεκτονικής του Μπλοκ Αποσύνθεσης Ήχου του AMUSE [1].

### 6.1.2 Μπλοκ Παραγωγής Κινησιολογίας

Κατόπιν εκπαίδευσης του Μπλοκ Αποσύνθεσης Ήχου, οι συγγραφείς διέθεταν πλέον τα μέσα για να προχωρήσουν στο κυρίως τμήμα του έργου τους που αφορά την παραγωγή 3D animations του ανθρώπινου σώματος κατευθυνόμενη από το περιεχόμενο, το συναίσθημα, και το στυλ της ομιλίας. Για τον σκοπό αυτό, συνέθεσαν το Μπλοκ Παραγωγής Κινησιολογίας, ως αποτέλεσμα δυο επιμέρους βαθέων μοντέλων: ενός *3D motion prior* για την αναπαράσταση μίας ακολουθίας κίνησης σε λανθάνοντα χώρο, και ενός μοντέλου λανθάνουσας διάχυσης που αναλαμβάνει κατ' εξοχήν την παραγωγική διαδικασία.

Για την σύνθεση του *motion prior*, ακολουθείται μια μέθοδος παρόμοια με τους [113, 114]. Πρόκειται, δηλαδή, για μία αρχιτεκτονική μετασχηματιστή VAE με έναν *encoder*, έστω  $P_E$ , και έναν *decoder*, έστω  $P_D$ . Αμφότεροι ακολουθούν μία δομή τύπου *U-Net* [14] με συνδέσεις παράκαμψης (*skip connections*) μεταξύ των μπλοκ μετασχηματιστών. Η μορφή του *motion prior* αναπαρίσταται στην εικόνα 6.3. Τα *positional embeddings* είναι παραμετροποιήσιμα και εισάγονται σε κάθε *multi-head attention layer* των μετασχηματιστών. Επί της ουσίας, ο κωδικοποιητής λαμβάνει μια ακολουθία  $T$  frames από παραμέτρους στάσης *SMPL-X* [54] σε συνεχή 6D αναπαράσταση περιστροφής [66], έστω  $\mathbf{m}^{1:T} \in \mathbb{R}^{6 \cdot 55 \times T}$ , και τα δύο πρώτα *tokens* της εξόδου του,  $\mu \in \mathbb{R}^{256}$  και  $\Sigma \in \mathbb{R}^{256 \times 256}$ , χρησιμοποιούνται για τον σχηματισμό της λανθάνουσας αναπαράστασης της κινησιολογίας,  $\mathbf{z}_m \in \mathbb{R}^{256}$ , μέσω της τεχνικής επαναπαραμετροποίησης, όπου  $\mu$  η μέση τιμή και  $\Sigma$  η διασπορά της κατανομής της λανθάνουσας μεταβλητής. Ο αποκωδικοποιητής, από την άλλη πλευρά, δεν παίρνει τα πραγματικά *positional embeddings* των frames ως *query*. Τουναντίον, χρησιμοποιεί μηδενικά *positional encodings* για *query*, με το λανθάνον της κίνησης,  $\mathbf{z}_m$ , να τροφοδοτείται ως μνήμη σε κάθε *cross-attention* επίπεδο μετασχηματιστών. Αυτό συνεπάγεται ότι για κάθε χρονικό βήμα ο decoder αρχίζει από ένα





Σχήμα 6.3: Σχηματική αναπαράσταση του Motion Prior Μπλοκ [1].

ουδέτερο query και μαθαίνει να παράγει την έξοδο βασιζόμενος αποκλειστικά στην πληροφορία που του δίδει το latent. Καθ' αυτόν τον τρόπο, αποτρέπεται η εξάρτησή του από πληροφορία θέσης της αρχικής εισόδου, ενώ ενθαρρύνεται η εκμάθηση της χρονικής δομής μέσω του latent και των παραμετροποιήσιμων *positional embeddings*. Στο τέλος της διαδικασίας, ο αποκωδικοποιητής παράγει την ανακατασκευασμένη ακολουθία κινήσεων,  $\hat{\mathbf{m}}^{1:T}$ .

Κατόπιν κωδικοποίησης—και επακόλουθης αποκωδικοποίησης—της εισαγόμενης ακολουθίας κίνησης από τον *3D motion prior*, εφαρμόζεται μια διαδικασία διάχυσης στο gradient-detached λανθάνον διάνυσμα κίνησης,  $\text{sg}[\mathbf{z}_m]$ , που παρήγαγε ο encoder. Εν αντιθέσει με τις αρχικές έρευνες επί των *diffusion models* που αφορούσαν, ως επί το πλείστον, την επεξεργασία και παραγωγή εικόνων σε επίπεδο pixel, στην προκειμένη περίπτωση η διάχυση πραγματοποιείται σε έναν λανθάνοντα χώρο, γεγονός που προσδίδει και τον χαρακτηρισμό “μοντέλο λανθάνουσας διάχυσης” (latent diffusion model, LDM) [115]. Η ανάπτυξή τους προέκυψε ως αποτέλεσμα της παρατήρησης ότι οι παραδοσιακές μέθοδοι είναι εξαιρετικά ακριβές σε υπολογιστικούς πόρους, και ότι η μεταφορά στον λανθάνοντα χώρο δύναται να ελαττώσει το εν λόγω μειονέκτημα δίχως να θυσιάζει την ποιότητα των τελικών δειγμάτων. Παρά αυτόν τον διαχωρισμό, ωστόσο, οι βασικές αρχές που αναλύθηκαν στο Κεφάλαιο 2 παραμένουν. Μάλιστα, η τα εμπρός διαδικασία διάχυσης (forward diffusion process) μοντελοποιείται από ένα *Denoising Diffusion Probabilistic Model* (DDPM). Όμοια, λοιπόν, με τους [8, 116], επιστρατεύεται ένας προγραμματιστής θορύβου (*noise scheduler*) σταθερής διακύμανσης και γραμμικής κλιμάκωσης. Το λανθάνον διάνυσμα  $\mathbf{z}_m$  διαταράσσεται με θόρυβο, σταδιακά, κατά την διάρκεια  $D$  χρονικών βημάτων, προκειμένου να ληφθεί το διάνυσμα καθαρού θορύβου  $\mathbf{z}_m^{(D)}$ , σύμφωνα με την σχέση:

$$q(\mathbf{z}_m^{(t)} | \mathbf{z}_m^{(0)}) = \mathcal{N}(\mathbf{z}_m^{(t)}; \sqrt{\bar{\alpha}_t} \mathbf{z}_m^{(0)}, (1 - \bar{\alpha}_t) \mathbf{I}), \quad \forall 0 < t \leq D \quad (6.12)$$

όπου  $\alpha_t := 1 - \beta_t$ ,  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$ , και  $\beta_t$  η υπερπαραμέτρος που δηλώνει την διακύμανση της διαδικασίας διάχυσης. Παρατηρούμε ότι η (6.12) δεν αποτελεί παρά τον πυρήνα μετάβασης της (2.5) ανηγμένο στην περίπτωση των λανθανόντων αναπαραστάσεων.

Ακολουθεί η αντίστροφη διαδικασία αποθορυβοποίησης στην οποία έγκειται και η εξάρτηση από τα εξαχθέντα χαρακτηριστικά της ομιλίας εισόδου. Ορίζεται, συνεπώς, ένα LDM, έστω  $\Delta$ , που ακολουθεί παρόμοια δομή, τύπου *U-Net* [14], με τον κωδικο-

ποιητή  $P_E$  του motion prior, που εξετάστηκε νωρίτερα. Ως είσοδο στο  $\Delta$ , δίδεται ένα συνενωμένο διάνυσμα που περιλαμβάνει το θορυβώδες latent  $z_m^{(t)}$  που προκύπτει από την forward διαδικασία διάχυσης· τα τρία διανύσματα περιχομένου  $c$ , συναισθήματος  $e$ , και στυλ  $s$ , έξοδοι του μπλοκ αποσύνθεσης ήχου· καθώς και το  $SE(t)$  που συνιστά μια ημιτονοειδή κωδικοποίηση του βήματος διάχυσης  $t$ , και που ορίζεται από τους Vaswani *et al.* [117] ως εξής:

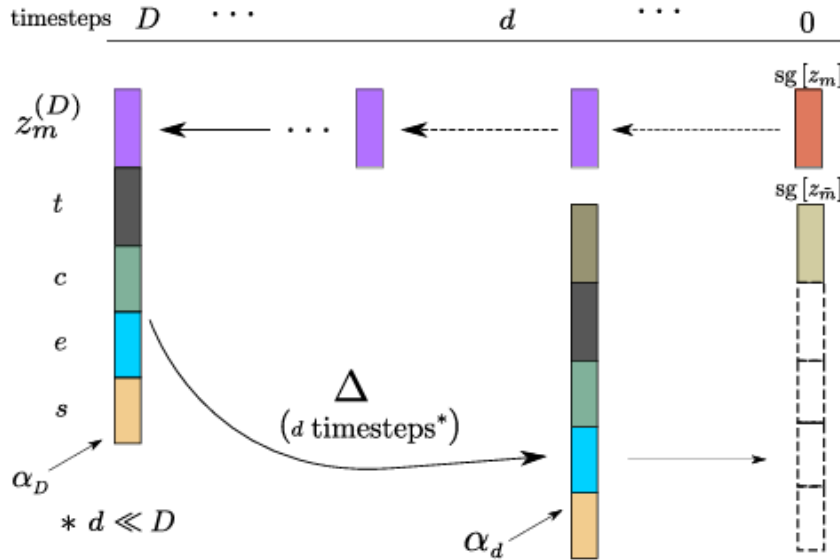
$$SE(t) := \begin{cases} \sin\left(\frac{t}{10000^{2i/d_m}}\right), & \text{για } i \text{ άρτιο} \\ \cos\left(\frac{t}{10000^{2i/d_m}}\right), & \text{για } i \text{ περιττό} \end{cases}, \quad \text{για } i = 0, 1, 2, \dots, \frac{d_m}{2} - 1 \quad (6.13)$$

όπου  $d_m = 256$  η διάσταση του διανύσματος  $SE(t)$ , και  $i$  ο δείκτης της διάστασης. Με άλλα λόγια, η είσοδος στο LDM είναι το:

$$x^{(t)} = [z_m^{(t)} \oplus SE(t) \oplus c \oplus e \oplus s] \in \mathbb{R}^{5 \times 256} \quad (6.14)$$

Το καθήκον του  $\Delta$  είναι η σταδιακή αποθορύβωση μέσω της (6.15), έως ότου ληφθεί η αποθορυβοποιημένη λανθάνουσα αναπαράσταση της κίνησης  $z_{\tilde{m}}^{1:T}$ , η οποία κι εν τέλει τροφοδοτείται στον αποκωδικοποιητή  $P_D$ , για την εξαγωγή της ανακατασκευασμένης κινησιολογικής ακολουθίας  $\tilde{m}^{1:T}$ .

$$z_m^{(t-1)} = \Delta\left([z_m^{(t)} \oplus SE(t) \oplus c \oplus e \oplus s]\right) \quad (6.15)$$



**Σχήμα 6.4:** Σχηματική αναπαράσταση της αρχής λειτουργίας του *conditional latent diffusion model* του AMUSE [1]. Κατά την προς τα εμπρός διαδικασία διάχυσης (δεξιά προς τα αριστερά), λαμβάνεται η θορυβώδης λανθάνουσα αναπαράσταση της κίνησης, ενώ κατά την αντίστροφη διαδικασία αποθορυβοποίησης (αριστερά προς τα δεξιά) εξάγεται το εξαρτώμενο αποθορυβοποιημένο λανθάνον διάνυσμα κίνησης.

Η εκπαίδευση των motion prior και conditional latent diffusion model γίνεται από κοινού προκειμένου να εξασφαλιστεί η ορθή ευθυγράμμιση της ομιλίας-εισόδου με την παραγόμενη κινησιολογία. Κατά την διάρκεια της, το Μπλοκ Αποσύνθεσης του Ήχου θεωρείται προεκπαιδευμένο και τα βάρη του παραμένουν “παγωμένα”. Όπως και στην περίπτωση του τελευταίου, έτσι και εδώ, εισάγεται μια πληθώρα διαφορετικών όρων απώλειας που προκύπτουν από μία διαδικασία προσπέλασης του μοντέλου παραγωγής χειρονομιών τριών βημάτων. Αυτή η διαδικασία δόθηκε σχηματικά στην εικόνα 6.1.

Κατά το πρώτο βήμα, ως συνηθίζεται στην περίπτωση των VAE, ανακατασκευάζεται η ακολουθία της κίνησης  $\hat{\mathbf{m}}^{1:T}$ , κατόπιν forward περάσματος από το motion prior. Το πρώτο αυτό βήμα εισάγει τρεις όρους απώλειας: κατ’ ονομασίαν, τις απώλειες ανακατασκευής των παραμέτρων στάσης,  $\mathcal{L}_{rec}$ , και των συντεταγμένων των κορυφών  $\mathcal{L}_{Vrec}$ , καθώς και την απώλεια απόκλισης Kullback-Leibler (KL),  $\mathcal{L}_{KL}$ , του motion prior. Για τους δύο πρώτους όρους, οι συγγραφείς του AMUSE, επιστράτευσαν μία ομαλή παραλλαγή της  $L1$  νόρμας, που εισήγαγαν οι Girshick και Ross [118]. Αν την συμβολίσουμε ως  $L_1^s$ , αυτή ορίζεται ως:

$$L_1^s(x, y) := \begin{cases} \frac{1}{2}(x - y)^2, & \text{αν } |x - y| < 1 \\ |x - y| - \frac{1}{2}, & \text{διαφορετικά} \end{cases} \quad (6.16)$$

και χρησιμοποιείται διότι είναι λιγότερη ευαίσθητη σε ακραίες περιπτώσεις από την κλασική απώλεια  $L1$  και συμπεριφέρεται σαν απώλεια  $L2$  για τιμές κοντά στο μηδέν. Βάσει, λοιπόν, της (6.16), οι δύο απώλειες ανακατασκευής υπολογίζονται ως:

$$\mathcal{L}_{rec} := L_1^s(\mathbf{m}^{1:T}, \hat{\mathbf{m}}^{1:T}), \quad \mathcal{L}_{Vrec} := L_1^s(\mathbf{V}^{1:T}, \hat{\mathbf{V}}^{1:T}), \quad (6.17)$$

όπου οι ακμές  $\mathbf{V}$  και  $\hat{\mathbf{V}}$  προκύπτουν κατόπιν *fitting* των παραμέτρων στάσης  $\mathbf{m}$  και  $\hat{\mathbf{m}}$ , αντίστοιχα, στο διαφορίσιμο επίπεδο SMPL-X [54] με μηδενικές παραμέτρους σχήματος  $\beta$ . Η, δε, απώλεια απόκλισης KL δίδεται από την κάτωθι σχέση:

$$\mathcal{L}_{KL} := \frac{1}{2} \left[ \sum_{i=1}^z (\mu_i^2 + \sigma_i^2) - \sum_{i=1}^z (\log(\sigma_i^2) + 1) \right] \quad (6.18)$$

όπου  $\mu$  και  $\sigma$  ο μέσος όρος και η διασπορά της λανθάνουσας κατανομής  $\mathbf{z}_m$ .

Στο δεύτερο βήμα, απενεργοποιείται ο υπολογισμός της κλίσης στον  $\mathbf{P}_E$ , οπότε και παράγεται το ενδιάμεσο latent της κίνησης,  $\text{sg}[\mathbf{z}_m]$ . Αυτό τροφοδοτείται ως είσοδος στο conditional LDM, όπου κατόπιν περάτωσης της forward διαδικασίας διάχυσης και επακόλουθης αποθορύβωσης, δύναται να υπολογιστεί η απώλεια λανθάνουσας διάχυσης,  $\mathcal{L}_{LD}$ . Πρόκειται για την κλασική αντικειμενική συνάρτηση που σχολιάστηκε στην ενότητα 2.1, δηλαδή:

$$\mathcal{L}_{LD} := \left\| \delta^{(t)} - \Delta \left( \left[ \mathbf{z}_m^{(t)} \oplus \text{SE}(t) \oplus \mathbf{c} \oplus \mathbf{e} \oplus \mathbf{s} \right] \right) \right\|_2^2 \quad (6.19)$$

Στην (6.19), ο όρος  $\delta^{(t)}$  συμβολίζει το διάνυσμα θορύβου, κατά το χρονικό βήμα  $t$ , που δειγματολήφθηκε αρχικά από τον γκαουσιανό θόρυβο  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ .

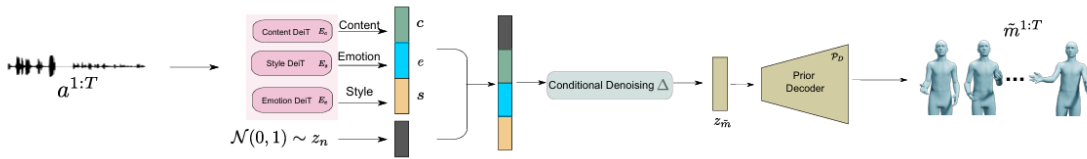
Στο τρίτο και τελευταίο βήμα, αφού έχουμε εξαγάγει το πλήρως αποθορυβοποιημένο latent  $\mathbf{z}_{\tilde{\mathbf{m}}}$  μέσω της αναδρομικής εκτέλεσης της (6.15), απενεργοποιούμε εκ νέου τον υπολογισμό της κλίσης και το τροφοδοτούμε στον αποκωδικοποιητή  $\mathbf{P}_D$ . Αποτέλεσμα είναι η λήψη της εξαρτημένης, αποθορυβοποιημένης και ανακατασκευασμένης ακολουθίας κινήσεων  $\tilde{\mathbf{m}}^{1:T} = \mathbf{P}_D(\text{sg}[\mathbf{z}_{\tilde{\mathbf{m}}}]$ ). Σε αυτό το σημείο εισάγονται και οι δύο τελευταίοι όροι απώλειας, που αφορούν την εξασφάλιση της ευθυγράμμισης των κινήσεων που δημιουργούνται από την διάχυση και του εισερχόμενου ηχητικού δείγματος. Σε πλήρη αντιστοιχία με τις απώλειες ανακατασκευής του πρώτου βήματος, γίνεται χρήση της ομαλής L1 νόρμας για τον ορισμό των απωλειών ευθυγράμμισης των παραχθέντων παραμέτρων στάσης και συντεταγμένων κορυφών. Μαθηματικώς:

$$\mathcal{L}_{align} := L_1^s(\mathbf{m}^{1:T}, \tilde{\mathbf{m}}^{1:T}), \quad \mathcal{L}_{Valign} := L_1^s(\mathbf{V}^{1:T}, \tilde{\mathbf{V}}^{1:T}) \quad (6.20)$$

Όπως και στην περίπτωση της εκπαίδευσης του Μπλοκ Αποσύνθεσης Ήχου, έτσι και εν προκειμένω, η συνολική απώλεια του Μπλοκ Παραγωγής Κινησιολογίας ορίζεται, απλοϊκά, ως το άθροισμα των επιμέρους απωλειών, που μόλις περιεγράφηκαν, ήτοι:

$$\mathcal{L}_{ges} := \mathcal{L}_{rec} + \mathcal{L}_{Vrec} + \mathcal{L}_{KL} + \mathcal{L}_{align} + \mathcal{L}_{Valign} + \mathcal{L}_{LD} \quad (6.21)$$

Ιδιαίτερο ενδιαφέρον παρουσιάζει, σε αυτό το σημείο, να αναφέρουμε ότι, για την παραγωγική διαδικασία, οι συγγραφείς του *AMUSE* επιστρατεύουν ένα *Denoising Diffusion Implicit Model*, ή αλλιώς *DDIM* (ενότητα, 2.4). Υπενθυμίζουμε ότι πρόκειται για μοντέλα διάχυσης τα οποία κάνουν χρήση μη-μαρκοβιανών αλυσίδων προκειμένου να βελτιώσουν την απόδοση της παραγωγικής διαδικασίας των DDPMs. Εδώ, δειγματοληπτείται, κατ' αρχάς, ένα διάνυσμα από την κατανομή  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ , αποθορυβοποιείται σταδιακά, και, τέλος, περνάει από τον αποκωδικοποιητή  $\mathbf{P}_D$  για την τελική εξαγωγή του διανύσματος κίνησης. Η άνωθι διαδικασία περιγράφεται σχηματικά στην εικόνα 6.5.



Σχήμα 6.5: Σχηματική αναπαράσταση της διαδικασίας παραγωγής νέων δειγμάτων του *AMUSE* [1].

## 6.2 Προεπεξεργασία Δεδομένων

Έως αυτό το σημείο, έχει πραγματοποιηθεί μία ιδιαίτερα διεξοδική και αναλυτική περιγραφή της αρχιτεκτονικής του *AMUSE* [1]. Επρόκειτο για ένα απαραίτητο βήμα για την εις βάθος κατανόηση των χαρακτηριστικών, βασικών δομών, αλλά και των αρχών λειτουργίας που διέπουν το εν λόγω μοντέλο. Μια τέτοια ενδελεχής θεώρηση αναμένεται να συντελέσει ουσιαδώς στην ορθότερη και πληρέστερη κατάστρωση των πειραμάτων μας, για την διασφάλιση ενός υψηλότερου βαθμού συνοχής και εγκυρότητας.

Προτού, όμως, προχωρήσουμε σε αυτά, καλούμαστε να υλοποιήσουμε μια σύντομη προεπεξεργασία των δεδομένων που εξαγάγαμε δια της μεθόδου του Κεφαλαίου 5. Αυτή η διαδικασία είναι αναγκαία προκειμένου να διασφαλιστεί η συμβατότητα της μορφής των δεδομένων του συνόλου μας με αυτήν που αναμένεται από το μοντέλο. Κρίνεται, λοιπόν, σκόπιμο να γίνει μια σύντομη περιγραφή της.

Όσον αφορά την προεπεξεργασία των δεδομένων της κίνησης, δεν έχουμε πολλά να κάνουμε. Τα βίντεο που χρησιμοποιούμε έχουν ήδη δειγματοληφθεί στα 30 καρέ το δευτερόλεπτο (FPS), που συνάδει με την αντίστοιχη τιμή του συνόλου BEAT [45]. Επιπροσθέτως, όπως αναφέραμε στο Κεφάλαιο 5, μετατρέπουμε επί τόπου τις παραμέτρους στάσης σε 6D συνεχείς αναπαραστάσεις περιστροφής κατά Zhou et al. [66]. Η μόνη επιπλέον επεξεργασία που οφείλουμε να υλοποιήσουμε αφορά το μήκος της δεδομένης αλληλουχίας κινήσεων. Οι συγγραφείς του AMUSE κατασκεύασαν το μοντέλο λανθάνουσας διάχυσης με σκοπό να δέχεται βίντεο των 10 δευτερολέπτων· αυτό μεταφράζεται σε ακολουθίες συνολικά 300 frames. Υπενθυμίζουμε ότι κατά την ανάπτυξη της διαδικασίας φιλτραρίσματος των βίντεο του Anim-400K [85] (ενότητα 5.4), εξετάστηκε το ενδεχόμενο απόρριψης βίντεο μικρότερων των 300 frames· εντούτοις, το εξαιρετικά μικρό πλήθος εναπομεινάντων δειγμάτων, μας οδήγησε στην απόφαση να μειώσουμε το ελάχιστο επιτρεπτό μήκος στα 100 frames. Τώρα, αντιμετωπίζουμε αυτήν την ασυμβατότητα, επαναλαμβάνοντας στην αρχή της ακολουθίας μας τις παραμέτρους του πρώτου καρέ, έως ότου συμπληρωθούν τα απαιτούμενα 300 frames. Από την άλλη πλευρά, βίντεο άνω των 10 δευτερολέπτων διαιρούνται σε διακριτά τμήματα 300 frames με τα εναπομείναντα καρέ να απορρίπτονται.

Η προεπεξεργασία των ηχητικών δειγμάτων παρουσιάζει, αν μη τι άλλο, μεγαλύτερο ενδιαφέρον. Το Μπλοκ Παραγωγής Κινησιολογίας απαιτεί τις λανθάνουσες αναπαραστάσεις του περιεχομένου, συναισθήματος και προσωπικού στυλ, ως εισόδους στο conditional LDM. Ως εκ τούτου, οφείλουμε πρώτα να περάσουμε τον ακατέργαστο ήχο από το Μπλοκ Αποσύνθεσης, το οποίο, όμως, δέχεται εικόνα σε μορφή filterbank. Συνεπώς, κάθε ηχητικό τμήμα διάρκειας 10 δευτερολέπτων μετατρέπεται σε ένα filterbank με 128 ζώνες συχνότητας *mel*, χρησιμοποιώντας παράθυρο *Hamming* 25ms και μετατόπιση frame 10ms. Ακολουθώντας την μεθοδολογία των Park et al. [119], σε κάθε δείγμα, εφαρμόζεται μάσκα μέγιστου μήκους 24 στο πεδίο της συχνότητας, και 96 στο πεδίο του χρόνου. Εν συνεχεία, κανονικοποιείται το filterbank, κατά το παράδειγμα των Gong et al. [110, 111], και εμπλουτίζεται μέσω της έγχυσης θορύβου και κυκλικής μετατόπισης. Τέλος, πριν την εισαγωγή στο Μπλοκ Αποσύνθεσης, κάθε filterbank χωρίζεται σε μια ακολουθία από 1209 σταθερά patches μεγέθους  $16 \times 16$ , με επικάλυψη 6 μονάδων στους τομείς της συχνότητας και του χρόνου.

Έχοντας, πλέον, ολοκληρώσει και την προκαταρκτική διαδικασία προετοιμασίας των δεδομένων μας για εκπαίδευση του Μπλοκ Παραγωγής Κινησιολογίας, βρισκόμαστε σε θέση να προχωρήσουμε στην κατάστρωση των πειραμάτων μας. Προτού, όμως, προχωρήσουμε στην περιγραφή τους, θα ήταν φρόνιμο να σχολιάσουμε, εν συντομία, ορισμένες λεπτομέρειες που έγκεινται πρωτίστως στην υλοποίηση των ίδιων των δομικών μοντέλων του AMUSE. Συγκεκριμένα, το Motion Prior αποτελεί έναν *transformer-based VAE encoder-decoder* με 9 επίπεδα και 4 κεφαλές. Πρόκειται για έναν μετασχηματιστή τύπου U-Net με υπολειπόμενες συνδέσεις. Την ίδια δομή ακολουθεί και ο Denoiser. Η κρυφή διάσταση όλων των επιπέδων μετασχηματιστών ορίζεται στα 1024. Οι υπό-

λοιπες λεπτομέρειες εφαρμογής παρουσιάζονται στον πίνακα 6.3.

## 6.3 Εκπαίδευση από την Αρχή

### 6.3.1 Λεπτομέρειες Εκπαίδευσης

Το πρώτο μας πρακτικό πείραμα συνιστά την από την αρχή (*from scratch*) εκπαίδευση του Μπλοκ Παραγωγής Κινησιολογίας του *AMUSE*—δίχως, δηλαδή, της χρήσης κάποιων προεκπαιδευμένων βαρών—με το δικό μας *speech-to-motion dataset* από βίντεο κλιπς *Anime*, που συνθέσαμε στο Κεφάλαιο 5.

Οι ακριβείς υπερπαραμέτροι της εκπαίδευσης διακρίνονται συνοπτικά στον πίνακα 6.3. Όσον αφορά τις πιο ουσιώδεις εξ' αυτών, ακολουθήσαμε σε μεγάλο βαθμό το παράδειγμα των *Chhatre et al.* [1]. Εξάλλου, στόχος των πειραμάτων μας δεν αποτελεί η εξερεύνηση της επιρροής της κάθε μίας υπερπαραμέτρου στο τελικό αποτέλεσμα, αλλά η ενσωμάτωση και αναπαραγωγή επαγγελματικών κινήσεων έντονα στυλιζαρισμένων χαρακτήρων. Ως εκ τούτου, θέσαμε τα χρονικά βήματα της προς τα εμπρός διαδικασίας διάχυσης στα 1000, με συντελεστή διακύμανσης θορύβου  $\beta_t \in [0.00085, 0.012]$ . Η αντίστροφη διαδικασία μοντελοποιείται από ένα DDIM για ταχύτερη και φθηνότερη υπολογιστικά παραγωγή δειγμάτων· ορίζουμε τα χρονικά βήματα της μη-μαρκοβιανής αλυσίδας στα 50. Ο, δε, ρυθμός εκμάθησης αφήνεται στο 0.0001. Ως βελτιστοποιητής επιλέγεται ο *AdamW* που αποτελεί μια παραλλαγή των *Loshchilov et al.* [120] του κλασικού αλγόριθμου *Adam*. Η κύρια υπερπαραμέτρος που μεταβάλλεται αποτελεί το μέγεθος του batch, το οποίο τίθεται στα 16, για λόγους που αφορούν ως επί το πλείστον την έλλειψη υπολογιστικών πόρων και το σχετικά μικρό μέγεθος dataset το οποίο καταφέραμε να εξαγάγουμε. Το μοντέλο εκπαιδεύτηκε για συνολικά 2900 εποχές, ωστόσο η εποχή για την οποία επιτεύχθηκε η μικρότερη απώλεια  $\mathcal{L}_{ges}$  (6.21) ήταν η 2810. Αυτά τα βάρη θα χρησιμοποιήσουμε και κατά την διαδικασία της παραγωγής νέων δειγμάτων.

| Εποχή | $\mathcal{L}_{rec}$ | $\mathcal{L}_{Vrec}$ | $\mathcal{L}_{KL}$ | $\mathcal{L}_{align}$ | $\mathcal{L}_{Valign}$ | $\mathcal{L}_{LD}$ | $\mathcal{L}_{ges}$ |
|-------|---------------------|----------------------|--------------------|-----------------------|------------------------|--------------------|---------------------|
| 2810  | 0.0020              | 0.0017               | 0.5190             | 0.0121                | 0.0107                 | 0.3181             | 0.3446              |

**Πίνακας 6.1:** Τιμές των επιμέρους ορών απώλειας του βέλτιστου μοντέλου εκπαιδευμένου από την αρχή.

### 6.3.2 Αποτελέσματα

Παρακάτω παρατίθενται ορισμένα ενδεικτικά δείγματα τα οποία παρήχθησαν με την διαδικασία της εικόνας 6.5, χρησιμοποιώντας για το Μπλοκ Παραγωγής Κινησιολογίας τα βάρη που προέκυψαν από την *from scratch* εκπαίδευση για την εποχή 2810. Περισσότερα δείγματα διατίθενται στο Παράρτημα Β.

### 6.3.3 Συμπεράσματα

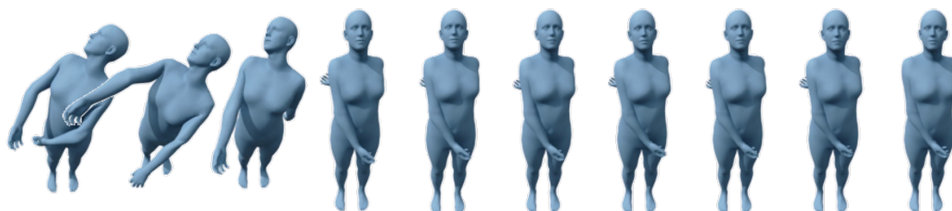
Όπως παρατηρούμε τα αποτελέσματα δεν είναι πολύ θετικά. Οι κινήσεις είναι εξαιρετικά σπασμωδικές και αφύσικες· μοιάζουν οριακά τυχαίες. Συνάμα, τα σώματα δι-



If you come and join us, you'll be able to live the life you want and deserve.



There're still inaccessible areas of the earth, where everything is possible.



When you do you daily prayers, the Earth-Mother is the goddess you pray to, right?



Hmm... just what am I going to do with you?



If this guy wants a fight, he's gonna get one!



Σχήμα 6.6: Δείγματα που παρήχθησαν από το *from scratch* εκπαιδευμένο μοντέλο. Κάθε σειρά αντιπροσωπεύει ένα διαφορετικό δείγμα και απεικονίζει μια ακολουθία ενδεικτικών frames συνοδευόμενη από μια λεζάντα του περιεχομένου της ομιλίας.



πλώνονται συχνά γύρω από τον εαυτό τους κατά τρόπο μη ανθρώπινο. Αυτά τα αποτελέσματα, εντούτοις, είναι δικαιολογίσιμα και εν μέρει αναμενόμενα. Κατ' αρχάς, το σύνολο δεδομένων μας, όπως έχουμε ήδη εξηγήσει, είναι αρκετά περιορισμένο ως προς το πλήθος των δειγμάτων που περιλαμβάνει. Η *from scratch* εκπαίδευση ενός τόσο περίπλοκου μοντέλου σαν του *AMUSE* σε τόσο μικρό σύνολο δεδομένων είναι σχεδόν εγγυημένο ότι δεν θα επιφέρει ποιοτικά αποτελέσματα. Επιπλέον, σε αυτόν τον περιορισμό έρχεται να προστεθεί και η σχετικά ελαττωμένη ποιότητα του συνόλου μας, συγκριτικά με τα καθιερωμένα της βιβλιογραφίας. Στο Κεφάλαιο 5, αναλύθηκαν διεξοδικά περιορισμοί που συχνά οδηγούν σε σφάλματα, όπως, παραδείγματος χάριν οι επικαλύψεις άκρων, και η ενδεχόμενη ύπαρξη έντονων λοιπών δομών είτε του ίδιου του απεικονιζόμενου χαρακτήρα είτε του υποβάθρου (π.χ. έντονες αφέλειες μαλλιών, δοκάρια κτηρίων κλπ.). Επεξηγήθηκε τότε, ότι επόμενο λογικό βήμα για την δημιουργία ενός συνόλου δεδομένων υψηλής ποιότητας, συνιστά η χειρωνακτική επισκόπηση των αυτόματα παραχθέντων *annotations* και η επακόλουθη απόρριψη των εσφαλμένων· διαδικασία μη-ρεαλιστική για τα πλαίσια της παρούσης εργασίας. Ως εκ τούτου, ένα υπολογίσιμο ποσοστό των δεδομένων εκπαίδευσης που χρησιμοποιήθηκε στο συγκεκριμένο πείραμα είναι θορυβώδες και λανθασμένο.

Αυτά τα δύο προβλήματα είναι αρκετά για να δικαιολογήσουν πλήρως τα αποτελέσματα του παρόντος πειράματος. Ωστόσο, η κατάσταση δεν είναι τελείως αποθαρρυντική, όσο φαίνονται εκ πρώτης όψεως. Σε ορισμένα από τα παραχθέντα δείγμα, διακρίνεται πίσω από το τρέμουλο, η ασθενής εμφάνιση κάποιων κινήσεων που θα θωρούντουσαν σημασιολογικά έγκυρες, συναρτήσει του συνοδευόμενου ήχου. Πρόκειται για μία νύξη ότι ο τελικός μας στόχος της ενσωμάτωσης κινησιολογικών μοτίβων που συναντώνται σε βίντεο *Anime* είναι πράγματι εφικτός, παρά τους περιορισμούς. Αποφασίσαμε, λοιπόν, να εκπαιδεύσουμε εκ νέου το *AMUSE*, αλλά αυτή τη φορά όχι από το μηδέν. Τουναντίον, θα εφαρμόσουμε μία τακτική *fine-tuning* στα ήδη προεκπαιδευμένα βάρη που παρέχονται από τους *Chhatre et al* [1]. Η ιδέα είναι ότι το μοντέλο θα έχει αποκτήσει εκ των προτέρων μια καλή βάση στην παραγωγή τρισδιάστατων *animations*—βάση που λόγω των άνωθι περιορισμών αδυνατεί να δώσει το δικό μας σύνολο δεδομένων—οπότε, πλέον, θα επιχειρήσει να μοντελοποιήσει μονάχα τις κινησιολογικές ιδιαιτερότητες των *Anime*.

## 6.4 Εκπαίδευση Fine-Tuning

### 6.4.1 Λεπτομέρειες Εκπαίδευσης

Η πειραματική διάταξη, τώρα, εξακολουθεί να προσομοιάζει σε μεγάλο βαθμό αυτή των *Chhatre et al* [1]. Με άλλα λόγια διατηρούμε το πλήθος των βημάτων εμπρόσθιας διάχυσης στα 1000 και τα βήματα παραγωγής του DDIM στα 50. Η εκάστοτε διακύμανση του θορύβου  $\beta_t$  κυμαίνεται στι ίδιες τιμές, και ο ρυθμός εκμάθησης παραμένει στο 0.0001, ενώ ως βελτιστοποιητής επιλέγεται, εκ νέου, ο *AdamW*. Το μέγεθος του batch ορίζεται στα 12 και εκπαιδεύουμε για 1000, αυτή τη φορά, εποχές. Πάλι, οι εν λόγω τιμές των υπερπαραμέτρων παρουσιάζονται συνοπτικά στον πίνακα 6.3.

Αυτή τη φορά, η εποχή για την οποία το μοντέλο σημείωσε το μικρότερο συνολικό σφάλμα είναι η υπ' αριθμός 820. Όπως και προηγουμένως, αυτά τα βάρη θα χρησι-

μποποιηθούν για την παραγωγική διαδικασία. Οι ακριβείς τιμές των επιμέρους όρων απωλειών παρατίθενται στον πίνακα 6.2.

| Εποχή | $\mathcal{L}_{rec}$ | $\mathcal{L}_{Vrec}$ | $\mathcal{L}_{KL}$ | $\mathcal{L}_{align}$ | $\mathcal{L}_{Valign}$ | $\mathcal{L}_{LD}$ | $\mathcal{L}_{ges}$ |
|-------|---------------------|----------------------|--------------------|-----------------------|------------------------|--------------------|---------------------|
| 820   | 0.0239              | 0.2863               | 0.8485             | 0.0302                | 0.2974                 | 0.2729             | 0.9107              |

Πίνακας 6.2: Τιμές των επιμέρους όρων απώλειας του βέλτιστου fine-tuned μοντέλου.

#### 6.4.2 Αποτελέσματα

Στην εικόνα 6.7, αναπαρίστανται ορισμένα ενδεικτικά δείγματα που υπήρξαν το αποτέλεσμα της παραγωγικής διαδικασίας κάνοντας χρήση των βαρών του fine-tuned Μπλοκ Παραγωγής Κινησιολογίας. Όμοια με προηγουμένως, προκειμένου να αποτυπώσουμε τα βίντεο στο έντυπο χαρτί, παραθέτουμε μια ακολουθία ενδεικτικών frames της κίνησης σε συνδυασμό με το περιεχόμενο της ομιλίας του ηχητικού δείγματος που την συνοδεύει. Περισσότερα δείγματα μπορούν να βρεθούν στο Παράρτημα B.

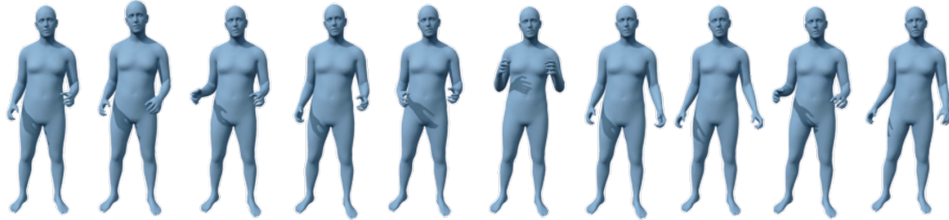
#### 6.4.3 Συμπεράσματα

Σε πλήρη αντίθεση με την περίπτωση της *from scratch* εκπαίδευσης, η ποιότητα των αποτελεσμάτων του προκειμένου πειράματος είναι αδιαφιλονίκητα εντυπωσιακή. Οι χαρακτήρες των βίντεο δεν παρουσιάζουν πλέον το τρέμουλο και τις σπασμωδικές κινήσεις που συναντούσαμε κατά κόρον προηγουμένως. Μήτε προκύπτουν κινήσεις οι οποίες θα χαρακτηρίζοντας εξόφθαλμα ως μη ανθρώπινες. Τουναντίον, όχι απλώς εμφανίζονται κινησιολογικά μοτίβα που διαπνέονται από υψηλό ρεαλισμό, αλλά επιπρόσθετα ευθυγραμμίζονται σε πολύ μεγάλο βαθμό με τις ρυθμικές και σημασιολογικές πληροφορίες που φέρει η συνοδεύουσα ομιλία. Οι χειρονομίες φαντάζουν μετρημένες και γήινες, δίχως υπερβολικά έντονες και απότομες μεταβολές. Επιβεβαιώνεται, με άλλα λόγια, η υπόθεσή μας ότι παρά τα προβλήματα που χαρακτηρίζουν το σύνολο δεδομένων μας, το τελευταίο είναι αρκετό για την εκμάθηση θεμελιωδών μοτίβων Anime, δοθείσας μια αρχικής βάσης επί της παραγωγής τρισδιάστατων κινουμένων σχεδίων.

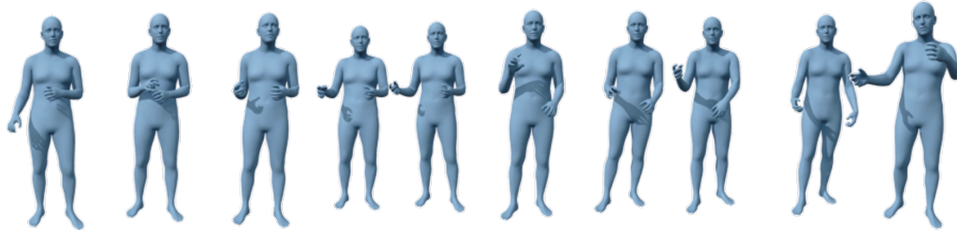
| Παράμετρος                   | Προεκπαιδευμένο  | Από την Αρχή     | Fine-Tuned       |
|------------------------------|------------------|------------------|------------------|
| Βήματα Διάχυσης              | 1000             | 1000             | 1000             |
| Βήματα Παραγωγής             | 50               | 50               | 50               |
| Διακύμανση Θορύβου $\beta_t$ | [0.00085, 0.012] | [0.00085, 0.012] | [0.00085, 0.012] |
| Βελτιστοποιητής              | AdamW            | AdamW            | AdamW            |
| Ρυθμός Εκμάθησης             | 0.0001           | 0.0001           | 0.0001           |
| Μέγεθος Batch                | 64               | 16               | 12               |
| Εποχές                       | 6000             | 2900             | 1000             |

Πίνακας 6.3: Υπερπαραμέτροι για το προεκπαιδευμένο μοντέλο των Chhatre et al. [1], το από την αρχή εκπαιδευμένο μοντέλο, και το fine-tuned μοντέλο.

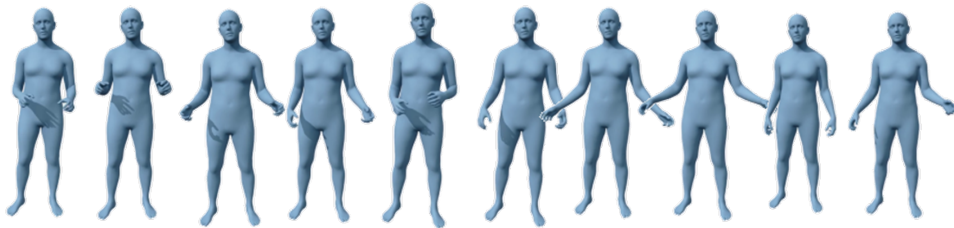
Now, even if we lose ground, we will never give up!



Look at all the evidence he left behind... like he wants to make sure we know he's here.



Ahh... I like the one with the A.I. What's the date today anyway?



I see... That must be... our dinner!



As the Lord wills it, so shall be!



Σχήμα 6.7: Δείγματα που παρήχθησαν από το *fine-tuned* μοντέλο.

## 6.5 Ποιοτική Σύγκριση

Στην συγκεκριμένη ενότητα θα πραγματοποιηθεί μία σύντομη ποιοτική σύγκριση των αποτελεσμάτων τόσο μεταξύ των δύο πειραμάτων που διεξάγαμε στα πλαίσια της παρούσης διπλωματικής εργασίας, ήτοι την *from-scratch* και την *fine-tuning* εκπαίδευση του Μπλοκ Παραγωγής Κινησιολογίας του AMUSE, όσο και με το προεκπαιδευμένο μοντέλο των *Chhatre et al.* [1] που εκπαιδεύτηκε για 6000 εποχές στο σύνολο δεδομένων BEAT [45].

| Εποχή | $\mathcal{L}_{rec}$ | $\mathcal{L}_{Vrec}$ | $\mathcal{L}_{KL}$ | $\mathcal{L}_{align}$ | $\mathcal{L}_{Valign}$ | $\mathcal{L}_{LD}$ | $\mathcal{L}_{ges}$ |
|-------|---------------------|----------------------|--------------------|-----------------------|------------------------|--------------------|---------------------|
| 6000  | 0.0023              | 0.0027               | 0.9061             | 0.0219                | 0.0515                 | 0.3052             | 0.3837              |

**Πίνακας 6.4:** Τιμές των επιμέρους ορών απώλειας του μοντέλου των *Chhatre et al.* [1] που εκπαιδεύτηκε για 6000 εποχές στο στο σύνολο δεδομένων BEAT [45].

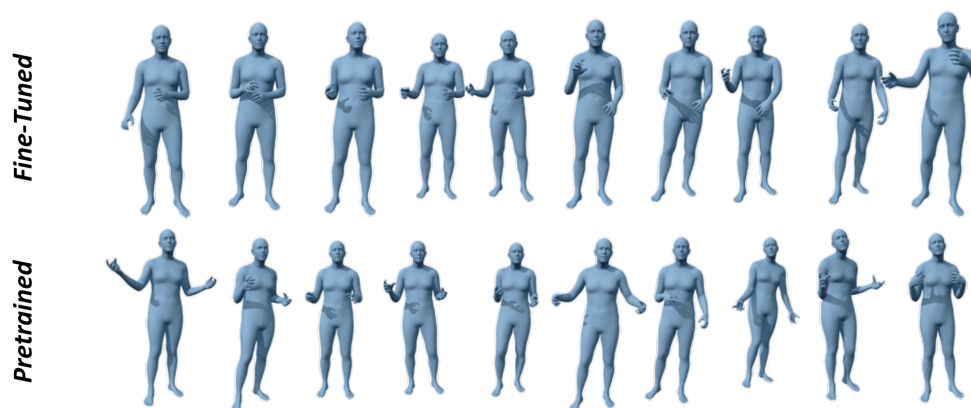
Ως διεπιστώθη στην ενότητα 6.3, τα αποτελέσματα που αναδεικνύει η *from scratch* εκπαίδευση δεν είναι ιδιαιτέρως εντυπωσιακά. Τρέμολα, αφύσικες περιστροφές των άκρων και φαινομενικά τυχαίες χειρονομίες συνθέτουν μια εικόνα η οποία δεν είναι ικανοποιητική. Είναι, μολαταύτα, αναμενόμενη, έως ένα βαθμό, λόγω των γνωστών περιορισμών του συνόλου δεδομένων εκπαίδευσης.

Η σύγκριση των δύο άλλων μοντέλων παρουσιάζει, αν μη τι άλλο, μεγαλύτερο ενδιαφέρον. Για την πραγματοποίησή της, εφαρμόσαμε και για τα προεκπαιδευμένα βάρη την παραγωγική διαδικασία με είσοδο τα ίδια ηχητικά δεδομένα εξαχθέντα από τα σύντομα βίντεο κλιπ του *Anim-400K* [85]. Το αποτέλεσμα δεν ήτο το αναμενόμενο. Αν και, αναμφισβήτητα εντυπωσιακές, οι κινήσεις του προεκπαιδευμένου μοντέλου φάνταζαν, στα σημεία, περισσότερο χαοτικές και υπερβολικές απ' ό,τι των αντίστοιχων του *fine-tuned* μοντέλου. Αυτό εκ πρώτης όψεως συνιστά αντίθεση. Το τελευταίο εκπαιδεύτηκε με δείγματα τα οποία προέκυψαν κατόπιν επεξεργασίας σύντομων βίντεο κλιπ καρτούν που απεικονίζουν έντονα στυλιζαρισμένους χαρακτήρες. Η αποτύπωσή τους είναι εικαστική—δεν διέπεται από τους φυσικούς νόμους. Ως εκ τούτου, θα ήταν λογικό να πέσει κανείς στην ψευδαίσθηση ότι και οι κινήσεις τους είναι εξίσου έντονες και μη ρεαλιστικές. Ωστόσο, κατόπιν μιας ενδελεχέστερης διερεύνησης, θα ερχόταν γρήγορα στην αντίληψη ότι κάτι τέτοιο δεν είναι διόλου απαραίτητο. Ασφαλώς, σκηνές με απότομες εναλλαγές και χειρονομίες υφίστανται· εντούτοις, η ραγδαία καταναλωτική ανάπτυξη του συγκεκριμένου υποείδους κινουμένων σχεδίων έχει οδηγήσει στην ανάγκη, από πλευράς των διαφόρων στούντιο παραγωγής, για μαζική, ταχεία και φθηνή παραγωγή τέτοιων έργων. Ως αποτέλεσμα, σε μεγάλο πλήθος επιμέρους σκηνών που απεικονίζουν κάποιον χαρακτήρα να μονολογεί—όπως στην περίπτωση μας—ο εν λόγω χαρακτήρας εμφανίζεται ακίνητος, με μεγαλύτερη έμφαση να δίδεται στις εκφράσεις του προσώπου. Αυτή ακριβώς η πραγματικότητα είναι υπεύθυνη για την μετριότητα των συχνά έντονων χειρονομιών του προεκπαιδευμένου μοντέλου κατόπιν περάτωσης της διαδικασίας *fine-tuning*.

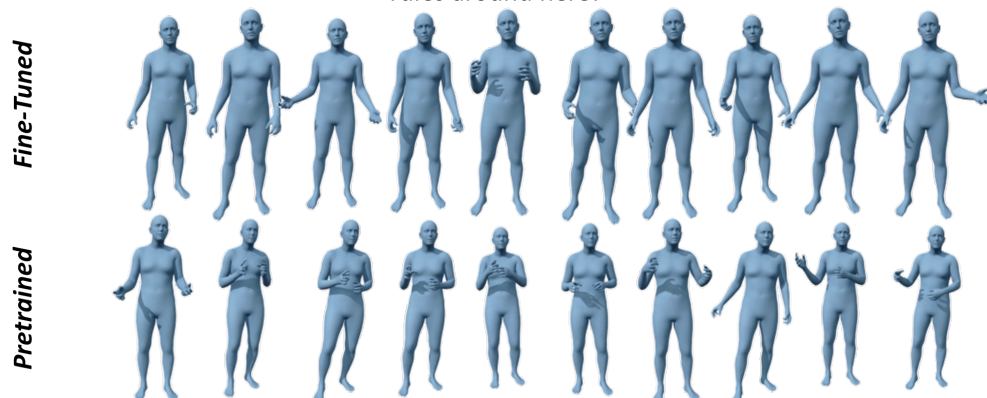
Εμείς, θεωρούμε, ότι οι πιο συνεσταλμένες κινήσεις της *fine-tuned* έκδοσης του AMUSE υπερτερούν των κατά τόπους υπερβολικών χειρονομιών του προεκπαιδευμένου μοντέλου των *Chhatre et al.* [1]. Αναγνωρίζουμε, ωστόσο, ότι αυτό αποτελεί ξεκάθαρα

ένα υποκειμενικό θέμα και ζήτημα προσωπικής προτίμησης. Στην εικόνα 6.8, διακρίνονται ορισμένα δείγματα των προεκπαιδευμένου και fine-tuned μοντέλων, παραγμένα υπό την ίδια είσοδο για άμεση σύγκριση.

Look at all the evidence he left behind... like he wants to make sure we know he's here.



Don't be sticking your nose into things you shouldn't. I'm telling you, these are the rules around here!



Σχήμα 6.8: Άμεση σύγκριση των αποτελεσμάτων του fine-tuned (πάνω) και προεκπαιδευμένου (κάτω) μοντέλου υπό την ίδια είσοδο.

## 6.6 Ποσοτική Σύγκριση

Για την περεταίρω και πιο αντικειμενική αξιολόγηση των μοντέλων, επιστρατεύσαμε, επιπρόσθετα, μια σειρά από ποσοτικές μετρικές. Δυστυχώς, δεν υφίσταται κάποια κοινά αποδεκτή μετρική για την αξιολόγηση παραγόμενων χειρονομιών. Είναι συχνό, μάλιστα, το φαινόμενο μετρικές να εμφανίζονται, σε διαφορετικές έρευνες, υπό το ίδιο όνομα, να αφορούν τον ίδιο χώρο, αλλά να υπολογίζονται με διαφορετικό τύπο.

Στα πλαίσια της παρούσης διπλωματικής εργασίας θα εφαρμόσουμε ορισμένες από τις ποσοτικές μετρικές που χρησιμοποιούν και οι συγγραφείς του *AMUSE*. Πιο συγκεκρι-



κριμένα, θα υπολογίσουμε την *Fréchet Gesture Distance* (FGD), την ποικιλία χειρονομιών *Diversity* (Div), και τον συγχρονισμό ρυθμού *Beat Align* (BA).

Το *Fréchet Gesture Distance* (FGD) είναι μία μετρική που υπολογίζει την απόσταση των χαρακτηριστικών μεταξύ των παραχθέντων και των *ground-truth* κινήσεων. Αποτελεί μια παραλλαγή της γνωστής μετρικής *Fréchet Implicit Distance* (FID) [121] για την αξιολόγηση εικόνων, ανηγμένη στην περίπτωση των ακολουθιών κίνησης. Εισήχθη από τους *Yoon et al.* [34], και ορίζεται ως η απόσταση Fréchet μεταξύ των λανθανόντων κατανομών χαρακτηριστικών των δύο ακολουθιών, που στην προκειμένη περίπτωση μπορούν να θεωρηθούν ως γκαουσιανές κατανομές. Συνεπώς, αν  $\mu_r$  και  $\Sigma_r$  είναι οι μέσος όρος και η συνδιακύμανση της λανθάνουσας κατανομής χαρακτηριστικών  $Z_r$  των *ground-truth* κινήσεων  $X$ , και, παρομοίως, τα  $\mu_s$  και  $\Sigma_s$  της λανθάνουσας κατανομής  $Z_s$  των συνθετικών κινήσεων  $\hat{X}$ , τότε η FGD προκύπτει από τον υπολογισμό της απόστασης Fréchet δύο πολυδιάστατων γκαουσιανών κατανομών, ήτοι:

$$\text{FGD}(X, \hat{X}) = \|\mu_r - \mu_s\|^2 + \text{tr}(\Sigma_r + \Sigma_s - 2\sqrt{\Sigma_r \Sigma_s}) \quad (6.22)$$

Η *Ποικιλία* (Diversity, Div), καθώς υποδεικνύει το όνομα, πρόκειται για μια μετρική που αφορά τον βαθμό ποικιλίας και διαφορετικότητας στην ακολουθία κινήσεων. Μετράει, ουσιαστικά, κατά πόσο το μοντέλο παράγει ένα εύρος διαφορετικών κινήσεων ή εάν καταρρέει σε επαναλαμβανόμενα, πανομοιότυπα μοτίβα. Για τον καθορισμό της, χρησιμοποιήσαμε την μέθοδο των *Chen et al.* [113] που υπολόγισαν την μέση διακύμανση των χαρακτηριστικών της λανθάνουσας αναπαράστασης της παραχθείσας κινησιολογίας. Μαθηματικώς, εάν  $Z$  η εν λόγω λανθάνουσα κατανομή χαρακτηριστικών, διάστασης  $D$ , της ακολουθίας κινήσεων  $X$ , τότε:

$$\text{Div}(X) = \frac{1}{D} \sum_{d=1}^D \text{var}[Z_d] \quad (6.23)$$

Ο *Συγχρονισμός Ρυθμού* (Beat Align, BA), τέλος, συνιστά μια μετρική που αξιολογεί την ευθυγράμμιση της παραγόμενης κινησιολογίας με την ομιλία εισόδου. Εκφράζει, στην πραγματικότητα, το ποσοστό των ηχητικών beats που συμπίπτουν με beats της κίνησης. Ακολουθούμε την μεθοδολογία των [45, 122]: εξάγουμε τα beats του ηχητικού δείγματος, ως ακολουθία χρονικών στιγμών  $B^a = \{t_i^a\}$ , κάνοντας χρήση των συναρτήσεων της βιβλιοθήκης *librosa* [123], και εντοπίζουμε τα κινητικά beats,  $B^m = \{t_j^m\}$ , ως τοπικά ελάχιστα της κινητικής ταχύτητας της παραχθείσας κίνησης. Τελικά, το BA ορίζεται ως η απόσταση Chamfer, μεταξύ των δύο ακολουθιών beats. Δηλαδή:

$$\text{BA}(X) = \frac{1}{D} \sum_{i=1}^D \exp\left(-\frac{\min_{t_j^m \in B^m} \|t_j^m - t_i^a\|^2}{2\sigma^2}\right) \quad (6.24)$$

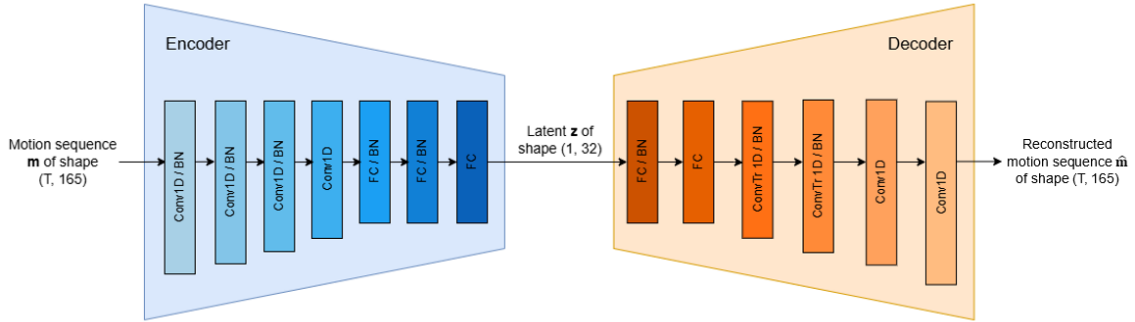
όπου  $D$  το πλήθος των ανιχνευμένων beats κίνησης, και  $\sigma$  μια παράμετρος κανονικοποίησης που ρυθμίζει πόσο επιεικής είναι η μέτρηση ως προς τη χρονική διαφορά. Θέσαμε αυτήν την παράμετρο στα 0.1 δευτερόλεπτα.

Παρατηρούμε ότι η μετρική BA κάνει χρήση της ακατέργαστης ακολουθία κίνησης· ωστόσο, οι FGD και Div απαιτούν την πρότερη έκφρασή της σε λανθάνοντα χώρο.

Learning-based εργαλεία για την εξαγωγή χαρακτηριστικών της ανθρώπινης κίνησης υφίστανται δημοσίως διαθέσιμα· εντούτοις, δεν κατορθώσαμε να εντοπίσουμε κάποιο που να προορίζεται για παραμέτρους στάσης SMPL-X [54]. Αποφασίσαμε, συνεπώς, να εκπαιδεύσουμε τον δικό μας *εξαγωγέα χαρακτηριστικών* (feature extractor) βασισμένο στο *autoencoding* [124], που δύναται να εκπαιδευτεί χωρίς επιτήρηση (unsupervised). Η αρχιτεκτονική του βασίζεται στο αντίστοιχο μοντέλο που χρησιμοποίησαν οι Yoon *et al.* [34], και αναπαρίσταται στην εικόνα 6.9. Ο εξαγωγέας απαρτίζεται από έναν συνελκτικό κωδικοποιητή, που χαρτογραφεί την ακολουθία εισόδου  $\mathbf{x}$  σε latent διάνυσμα  $\mathbf{z} = f_{\theta}(\mathbf{x})$ , και αποκωδικοποιητή, που προσπαθεί να ανακατασκευάσει  $\hat{\mathbf{x}} = g_{\phi}(\mathbf{z})$ . Εκπαιδεύουμε με το σύνολο δεδομένων μας (Κεφάλαιο 5), σε υπακολουθίες των 100 frames, βάσει της κλασικής απώλειας ανακατασκευής της εξίσωσης (6.25), για 100 εποχές.

$$\mathcal{L}_{FE} := \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \quad (6.25)$$

Στην εκατοστή εποχή επετεύχθη απώλεια περί το 0.0135. Τέλος, κάνουμε χρήση του εκπαιδευμένου encoder για την έκφραση των ακολουθιών κίνησης σε λανθάνοντα χώρο και τον επακόλουθο υπολογισμό των ζητούμενων μετρικών FGD και Div.



Σχήμα 6.9: Σχηματική αναπαράσταση της αρχιτεκτονικής του *Autoencoder Εξαγωγής Χαρακτηριστικών*, βασισμένη στους Yoon *et al.* [34]. Το BN συμβολίζει *Batch Normalization*, το FC ένα *fully connected layer*, και το ConvTr ένα επίπεδο *transpose convolution*.

Διαθέτοντας, πλέον, όλα τα απαραίτητα μέσα, υπολογίζουμε τις προαναφερθείσες ποσοτικές μετρικές—FGD, Div, και BA—για την αξιολόγηση του προεκπαιδευμένου, του from-scratch, και του fine-tuned μοντέλου. Για τον σκοπό αυτό, δειγματολήφθηκαν ομοιόμορφα 500 ηχητικά δείγματα από το σύνολο δεδομένων μας και, τελικά, υπολογίστηκαν οι μέσοι όροι. Τα αποτελέσματα παρατίθενται στον πίνακα 6.5.

Τα αποτελέσματα φαντάζουν αρκετά αντιφατικά στην αρχή. Τα καλύτερα αποτελέσματα σε όλες τις μετρικές φαίνεται να πετυχαίνει το μοντέλο που εκπαιδεύτηκε από την αρχή· μοντέλο που ποιοτικά συμπεράναμε ότι παράγει τα χειρότερα δείγματα. Τα, δε, fine-tuned και pretrained μοντέλα, εναλλάσσονται στην δεύτερη θέση. Ωστόσο, υπάρχουν λογικές εξηγήσεις πίσω από τα εν λόγω αποτελέσματα.

Κατ' αρχάς, όσον αφορά την απόσταση Fréchet, το From-Scratch μοντέλο πετυχαίνει την μικρότερη τιμή, όντας το μοναδικό που έχει εκπαιδευτεί εξ' ολοκλήρου στο δικό μας σύνολο. Είναι, συνεπώς, φυσικό να κατέχει την πιο κοντινή απόσταση στα αρχικά δεδομένα. Το Fine-Tuned, που επίσης έχει εκπαιδευτεί εν μέρει με το σύνολο δεδομέ-



| Μέθοδος         | FGD       | Div     | BA     |
|-----------------|-----------|---------|--------|
| Ground Truth    | -         | 50.751  | 0.6982 |
| Προεκπαιδευμένο | 3318.8842 | 3.0737  | 0.4147 |
| Fine-Tuned      | 2455.3832 | 2.0371  | 0.4062 |
| From-Scratch    | 423.1720  | 25.0954 | 0.4969 |

**Πίνακας 6.5:** Αποτελέσματα της ποσοτικής σύγκρισης του προεκπαιδευμένου, from-scratch, και fine-tuned μοντέλου. Για τις μετρικές Div και BA, όσο μεγαλύτερη είναι η τιμή τους, τόσο καλύτερα τα αποτελέσματα, ενώ για το FGD ισχύει το αντίστροφο. Με πράσινο χρώμα έχουμε σημειώσει το εκάστοτε μοντέλο με την καλύτερη επίδοση, ενώ με μπλε το δεύτερο καλύτερο.

νων μας, πετυχαίνει την δεύτερη καλύτερη επίδοση στην συγκεκριμένη μετρική, ενώ το προεκπαιδευμένο μοντέλο, όπως ήταν αναμενόμενο εμφανίζει τα χειρότερα αποτελέσματα. Για τις άλλες δύο μετρικές, το From-Scratch μοντέλο πετυχαίνει εκ νέου την καλύτερη επίδοση. Υποθέτουμε ότι αυτό οφείλεται στην σπασμωδική και τυχαία φύση των κινήσεων που αναδεικνύει. Υπενθυμίζουμε ότι το Div δεν είναι παρά ο μέσος όρος των διακυμάνσεων του εκάστοτε latent, ενώ το BA μετράει κατά πόσο έντονες μεταβολές στην ομιλία συνάδουν με αντίστοιχες μεταβολές στην κίνηση. Το From-Scratch μοντέλο παράγει, σχεδόν στο σύνολό τους, τέτοιες έντονες και απότομες χειρονομίες, δικαιολογώντας, καθ' αυτόν τον τρόπο, τα υψηλά score. Από την άλλη πλευρά, στην δεύτερη θέση, στις συγκεκριμένες περιπτώσεις, έρχεται πλέον το προεκπαιδευμένο μοντέλο. Η διαφοράς, ωστόσο, το Fine-Tuned μοντέλο είναι αρκετά μικρές και θεωρούμε ότι έγκεινται, κατά κύριο λόγο, στην παρατήρηση που κάναμε κατά την ποιοτική σύγκριση των δύο μοντέλων· ότι, δηλαδή, οι κινήσεις του pretrained μοντέλου είναι ελαφρώς πιο ζωηρές και, ενδεχομένως, πιο υπερβολικές, καθιστώντας τις πιο “ευαίσθητες” ως προς την ανίχνευσή τους από τις μετρικές που χρησιμοποιήσαμε.

Τα συμπεράσματά μας από αυτήν την ποσοτική σύγκριση δεν είναι τόσο το ποιο μοντέλο παρουσιάζει τα καλύτερα αποτελέσματα. Εξάλλου, δεν υφίσταται ποσοτική μετρική που να θεωρείται πιο έγκυρη από την ποιοτική αξιολόγηση. Αντ' αυτού, θα πρέπει να εκλάβουμε αυτά τα αποτελέσματα ως απόδειξη επίτευξης του βασικού στόχου μας που έγκειται στην ενσωμάτωση ορισμένων βασικών κινησιολογικών μοτίβων των Anime στην κατευθυνόμενη από ομιλία και συναίσθημα παραγωγή τρισδιάστατων κινουμένων σχεδίων του ανθρώπινου σώματος.



## Κεφάλαιο 7

### Ανακεφαλαίωση και Μελλοντική Έρευνα

---

Εν κατακλείδι, στα πλαίσια της παρούσης διπλωματικής εργασίας, μελετήθηκε το πρόβλημα της αξιοποίησης μοντέλων διάχυσης για την παραγωγή τρισδιάστατων κινουμένων σχεδίων του ανθρώπινου σώματος κατευθυνόμενη από την ομιλία και το συναίσθημα. Πρόκειται για ένα πρόβλημα που καθίσταται ολοένα και πιο κρίσιμο με την πάροδο του χρόνου και την διαρκή ανάπτυξη της τεχνολογίας. Υφιστάμενα σχετικά μοντέλα έχουν παραγάγει, αδιαμφισβήτητα, εντυπωσιακά δείγματα· εντούτοις, συχνά προκαλούν ένα έντονο συναίσθημα δυσφορίας και ενόχλησης στον θεατή που βρίσκεται ενώπιόν τους. Το συναίσθημα αυτό αναγνωρίστηκε ως το “φαινόμενο της απόκοσμης κοιλάδας” που περιγράφει το συναίσθημα που δημιουργείται από το αντίκρισμα αντικειμένων πολύ ρεαλιστικών και ανθρωπόμορφων, που όμως δεν αποδίδουν πλήρως τις φυσικές κινήσεις και εκφράσεις. Υποθέσαμε ότι αιτία για το συγκεκριμένο πρόβλημα έγκειται στην χρήση γνωστών καθιερωμένων *speech-to-motion* συνόλων δεδομένων εκπαίδευσης που, όμως, προέρχονται από βίντεο πραγματικών ανθρώπων. Τα εν λόγω βαθιά μοντέλα, στην απόπειρά τους να φθάσουν όσο πιο κοντά γίνεται στο *ground-truth*, καταλήγουν συχνά στην απόκοσμη κοιλάδα. Θεωρήσαμε, συνεπώς, πως η εκπαίδευση βάσει δεδομένων από όχι απαραίτητα ρεαλιστικές, έντονα στυλιζαρισμένες φιγούρες, δύναται να αναιρέσει τον περιορισμό.

Στο πλαίσιο αυτό, δημιουργήσαμε το πρώτο, από ό,τι γνωρίζουμε, *speech-to-motion* σύνολο δεδομένων εκπαίδευσης προερχόμενο από σύντομα βίντεο κλιπς Anime. Στην συνέχεια χρησιμοποιήσαμε το νέο αυτό, πρωτοποριακό, dataset για να εκτελέσουμε δύο πειράματα: την εκπαίδευση *from-scratch* και το *fine-tuning* του μοντέλου AMUSE [1]. Η επιτυχία του πρώτου πειράματος αποδείχθηκε σίγουρα αμφισβητήσιμη. Χρονικοί και υπολογιστικοί περιορισμοί, σε συνδυασμό με τα συχνά θορυβώδη δεδομένα του dataset, κατέστησε την από την αρχή εκπαίδευση αποτυχή. Τουναντίον, τα αποτελέσματα της *fine-tuned* περίπτωση υπήρξαν θεαματικά. Τα παραγόμενα δείγματα είναι, αναμφίβολα, πολύ πιο εκφραστικά, ρεαλιστικά και προσιτά στον θεατή. Η ποιοτική και ποσοτική σύγκριση με τα προεκπαιδευμένα βάρη των *Chhatre et al.* [1], επιβεβαίωσαν το αντίκτυπο που είχε το σύνολο μας στην παραγωγική διαδικασία. Καθ’ αυτόν τον τρόπο, επιτύχαμε τον απώτερο στόχο μας για τα πλαίσια της συγκεκριμένης εργασίας, την εξερεύνηση, δηλαδή, της δυνατότητας ενσωμάτωσης βασικών κινησιολογι-

κών μοτίβων, που χρησιμοποιούν επαγγελματίες animators Anime, στην παραγωγική διαδικασία, έτσι ώστε να εξομαλυνθούν οι εμφανίσεις του φαινομένου της “απόκοσμης κοιλάδας”.

Σε αυτό το σημείο, θα μπορούσαμε να σχολιάσουμε ποια είναι τα επόμενα, μελλοντικά βήματα για την επιπλέον ανάπτυξη του συγκεκριμένου πλαισίου. Πρώτο βήμα, συνιστά, ασφαλώς, η περεταίρω επέκταση του συνόλου δεδομένων εκπαίδευσης. Λόγω χρονικών περιορισμών, η συνολική του διάρκεια δεν ξεπέρασε της 6 ώρες, δημιουργώντας μεγάλα προβλήματα στην επακόλουθη εκπαιδευτική διαδικασία. Το αρχικό σύνολο, *Anim-400K* [85], περιλαμβάνει ακόμη μεγάλο πλήθος ακατέργαστων δεδομένων. Ως εκ τούτου, η επεξεργασία τους κατά την μέθοδο που περιεγράφηκε στο Κεφάλαιο 5 αποτελεί φυσική επιλογή.

Σε δεύτερη φάση, ακολουθεί η βελτίωση της ποιότητας των ίδιων των δεδομένων του συνόλου. Όπως έχει εξηγηθεί επανειλημμένως, τυχόν επικαλύψεις των άκρων του απεικονιζόμενου χαρακτήρα ή η παρουσία έντονων δομών στα εκάστοτε frames οδηγεί, συχνά, σε εσφαλμένα και θορυβώδη δείγματα. Αυτό είναι αναμενόμενο σε περιπτώσεις αυτόματων ή ημιαυτόματων μεθόδων παραγωγής ετικετών *pseudo-ground truth*, όπως εν προκειμένω. Για την επίλυση του συγκεκριμένου προβλήματος, καλούμαστε να επισκοπήσουμε χειρωνακτικά το σύνολο των παραχθέντων δειγμάτων και να αποφανθούμε, καθ’ αυτόν τον τρόπο, για την εγκυρότητα του καθενός εξ’ αυτών.

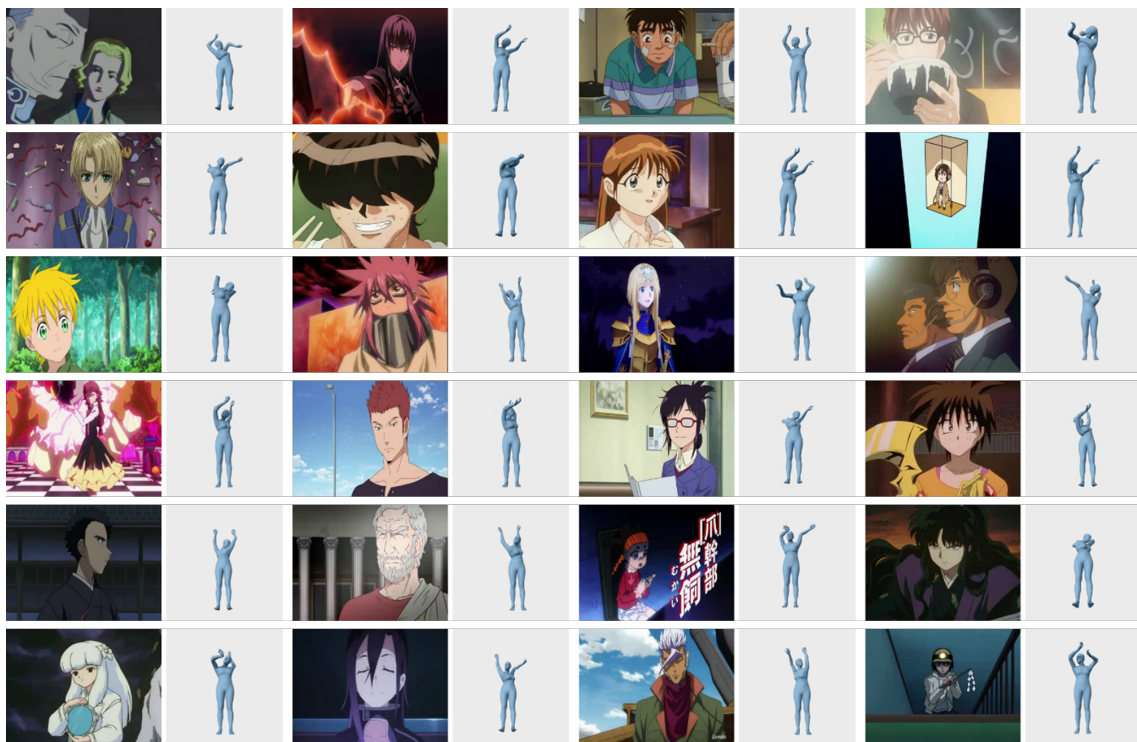
Τέλος, αφού έχουμε εξασφαλίσει ένα ικανοποιητικά εκτενές και ποιοτικό *dataset*, θα μπορούσαμε να το χρησιμοποιήσουμε για την εκπαίδευση ενός βαθύς *speech-to-motion* μοντέλου διάχυσης· όχι ενός έτοιμου, δημοσίως διαθέσιμου μοντέλου σαν του *AMUSE* [1], αλλά ενός μοντέλου του οποίου η αρχιτεκτονική θα έχει παραχθεί εξ’ ολοκλήρου από εμάς.

Η εργασία αυτή ανέδειξε τη δυναμική των μοντέλων διάχυσης στην παραγωγή ρεαλιστικής και εκφραστικής ανθρώπινης κίνησης, καθώς και τη σημασία της χρήσης στυλιζαρισμένων δεδομένων για τον μετριασμό του φαινομένου της “απόκοσμης κοιλάδας”. Παρά τους υφιστάμενους περιορισμούς, τα αποτελέσματα επιβεβαιώνουν τη βιωσιμότητα της προσέγγισης και θέτουν ισχυρά θεμέλια για περαιτέρω έρευνα και βελτιστοποίηση του πλαισίου.

# Παράρτημα Α

## Επιπλέον Δείγματα Συνόλου Δεδομένων

Στο παρόν παράρτημα παρατίθενται ορισμένα επιπρόσθετα δείγματα του συνόλου δεδομένων εκπαίδευσης που παράχθηκαν μέσω της διαδικασίας που περιεγράφηκε αναλυτικώς στο Κεφάλαιο 5. Οι εικόνες είναι οργανωμένες σε ζεύγη: η αριστερή εικόνα αποτελεί κάποιο frame του αρχικού βίντεο κλιπ και η δεξιά η αντίστοιχη εκτίμηση του τρισδιάστατου πλέγματος.

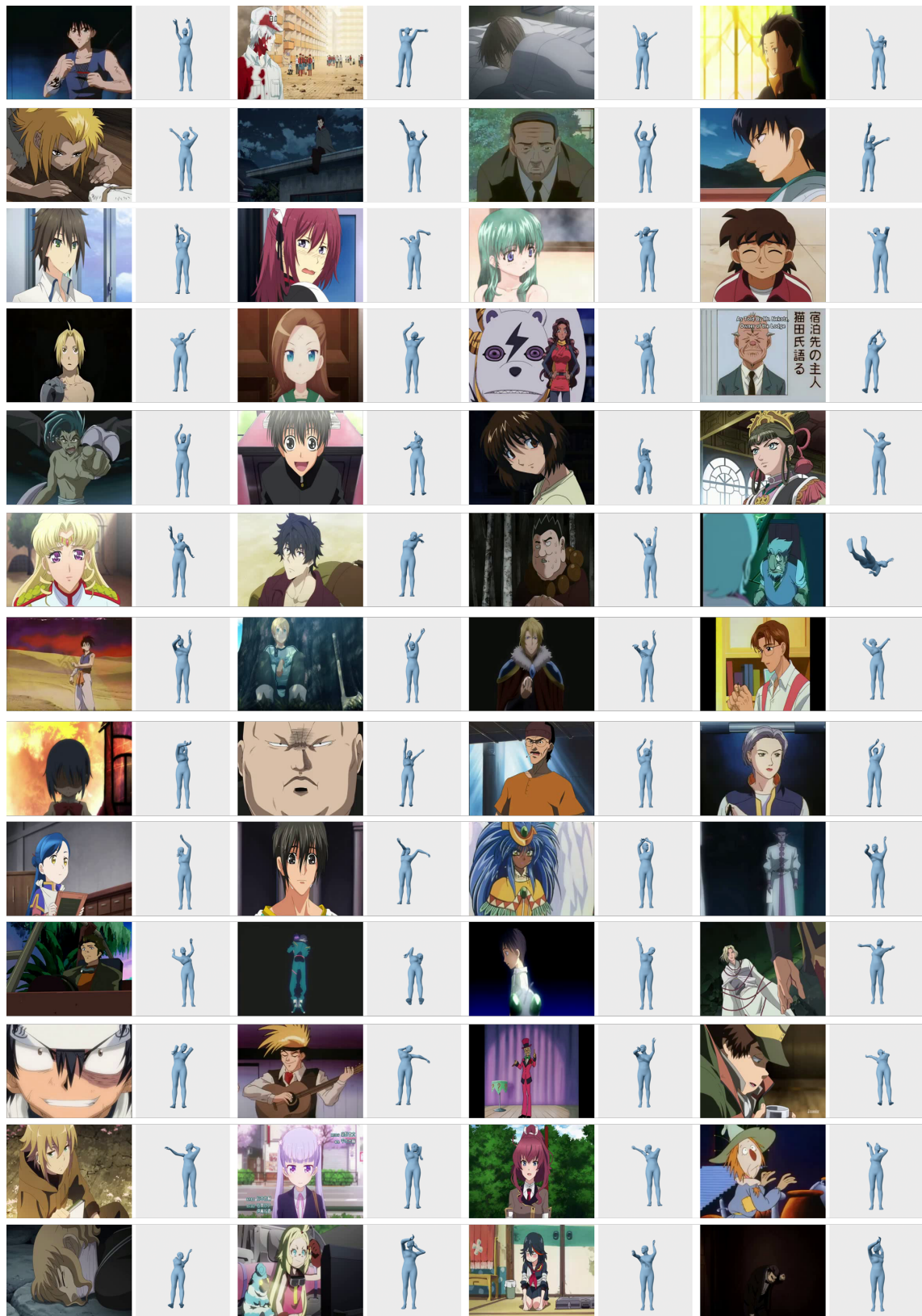


Σχήμα Α.1: Επιπλέον δείγματα του συνόλου δεδομένων εκπαίδευσης.





Σχήμα Α.2: Επιπλέον δείγματα του συνόλου δεδομένων εκπαίδευσης (συνέχεια).



Σχήμα Α.3: Επιπλέον δείγματα του συνόλου δεδομένων εκπαίδευσης (συνέχεια).





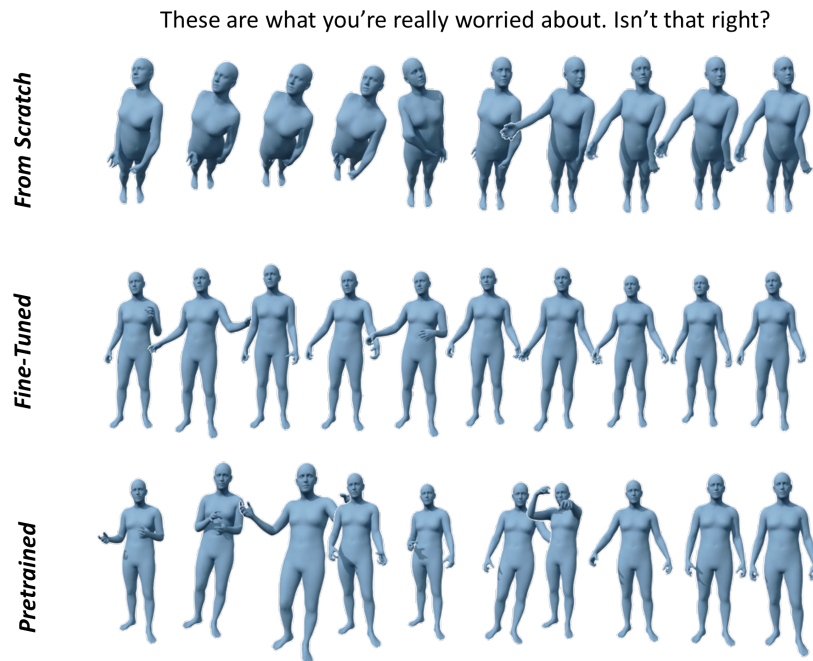
Σχήμα Α.4: Επιπλέον δείγματα του συνόλου δεδομένων εκπαίδευσης (συνέχεια).

## Παράρτημα B

### Επιπλέον Παραδείγματα Σύνθεσης

---

Στο παρόν παράρτημα παρατίθενται ορισμένα επιπρόσθετα παραδείγματα σύνθεσης των διάφορων μοντέλων που αναπτύξαμε στο Κεφάλαιο 6. Οι εικόνες είναι οργανωμένες σε τριάδες: Στην πρώτη σειρά βρίσκονται διατεταγμένα τα τρισδιάστατα πλέγματα ορισμένων καίριων frames παραχθέντα από το εκπαιδευμένο από την αρχή μοντέλο, στην δεύτερη σειρά συναντώνται τα ίδια frames αλλά για το *fine-tuned* μοντέλο, και στην τρίτη σειρά τα αντίστοιχα για τα προεκπαιδευμένα βάρη του *AMUSE* [1]. Οι τρεις σειρές συνοδεύονται από μια λεζάντα που αποτελεί το περιεχόμενο της εκάστοτε ομιλίας εισόδου. Η διάταξη αυτή γίνεται με σκοπό να παρακινηθεί ο αναγνώστης να επιβεβαιώσει ο ίδιος την ποιοτική σύγκριση που παραθέσαμε στο Κεφάλαιο 6.



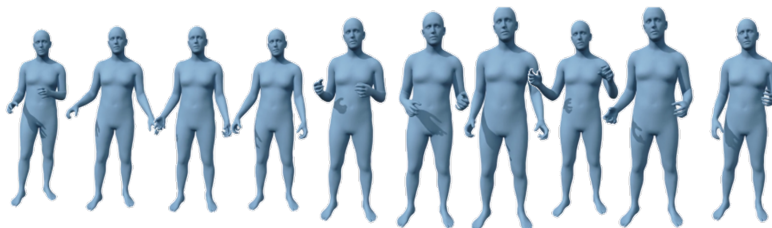
Σχήμα B.1: Επιπλέον παραδείγματα σύνθεσης.

Look at you! Greeting us in full makeup, huh? It's embarrassing when you are that eager to see us

*From Scratch*



*Fine-Tuned*



*Pretrained*

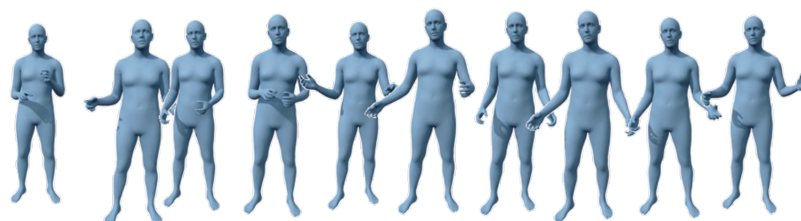


What an absolute joke this entire thing is! Might I suggest you use your authority as lord to simply abolish this worthless system!

*From Scratch*



*Fine-Tuned*



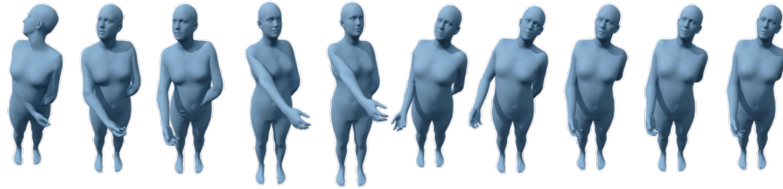
*Pretrained*



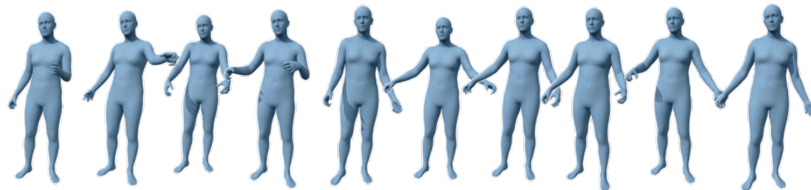
Σχήμα Β.2: Επιπλέον παραδείγματα σύνθεσης (συνέχεια).

My body doesn't grow all the way you think it would. I even have to change the places I hang out every 30 years or so, otherwise people get suspicious.

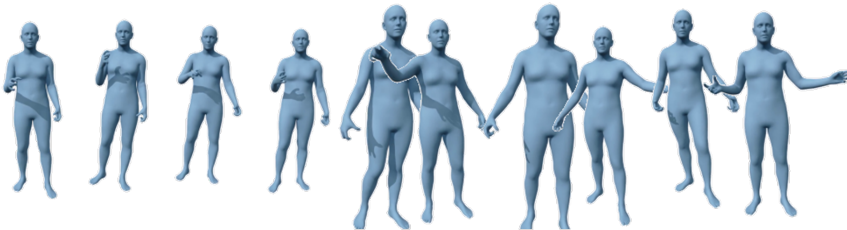
*From Scratch*



*Fine-Tuned*

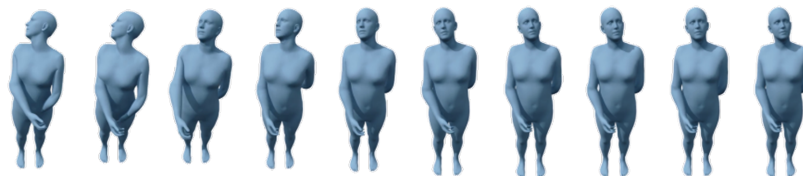


*Pretrained*

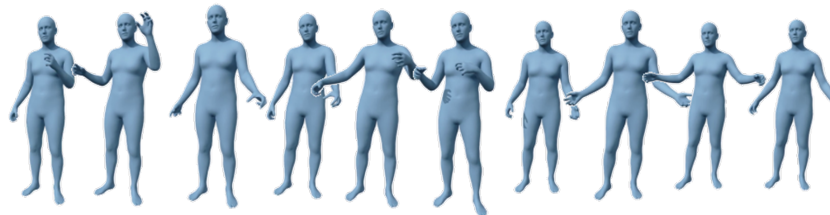


I can't die, because... I have to protect you.

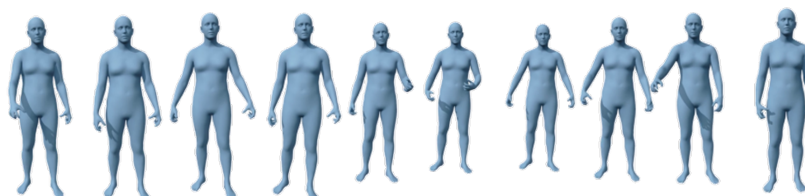
*From Scratch*



*Fine-Tuned*



*Pretrained*



Σχήμα Β.3: Επιπλέον παραδείγματα σύνθεσης (συνέχεια).



Unfortunately, it would seem that I'm rather incapable of achieving my greatest desire...



Before I take your life, I will grate you a thousand pieces!



Σχήμα Β.4: Επιπλέον παραδείγματα σύνθεσης (συνέχεια).

## Βιβλιογραφία

---

- [1] Kiran Chhatre κ.ά. «AMUSE: Emotional Speech-driven 3D Body Animation via Disentangled Latent Diffusion». Στο: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Ιούν. 2024, σσ. 1942–1953. URL: <https://amuse.is.tue.mpg.de>.
- [2] Ian Goodfellow κ.ά. «Generative adversarial networks». Στο: *Communications of the ACM* 63.11 (2020), σσ. 139–144.
- [3] Antonia Creswell κ.ά. «Generative adversarial networks: An overview». Στο: *IEEE signal processing magazine* 35.1 (2018), σσ. 53–65.
- [4] Yoshua Bengio κ.ά. «A neural probabilistic language model». Στο: *Journal of machine learning research* 3.Feb (2003), σσ. 1137–1155.
- [5] Ivan Kobyzev, Simon JD Prince και Marcus A Brubaker. «Normalizing flows: An introduction and review of current methods». Στο: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), σσ. 3964–3979.
- [6] Diederik P Kingma και Max Welling. «Auto-encoding variational bayes». Στο: *arXiv preprint arXiv:1312.6114* (2013).
- [7] Yang Song και Stefano Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. 2020. arXiv: 1907.05600 [cs.LG]. URL: <https://arxiv.org/abs/1907.05600>.
- [8] Jonathan Ho, Ajay Jain και Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG]. URL: <https://arxiv.org/abs/2006.11239>.
- [9] Sam Bond-Taylor κ.ά. «Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models». Στο: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.11 (2022), σσ. 7327–7347. DOI: 10.1109/TPAMI.2021.3116668.
- [10] Florinel-Alin Croitoru κ.ά. «Diffusion models in vision: A survey». Στο: *IEEE transactions on pattern analysis and machine intelligence* 45.9 (2023), σσ. 10850–10869.
- [11] Ling Yang κ.ά. *Diffusion Models: A Comprehensive Survey of Methods and Applications*. 2024. arXiv: 2209.00796 [cs.LG]. URL: <https://arxiv.org/abs/2209.00796>.

- [12] Mingxing Peng κ.ά. *Diffusion Models for Intelligent Transportation Systems: A Survey*. 2025. arXiv: 2409.15816 [eess.SY]. URL: <https://arxiv.org/abs/2409.15816>.
- [13] Jascha Sohl-Dickstein κ.ά. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. 2015. arXiv: 1503.03585 [cs.LG]. URL: <https://arxiv.org/abs/1503.03585>.
- [14] Olaf Ronneberger, Philipp Fischer και Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: 1505.04597 [cs.CV]. URL: <https://arxiv.org/abs/1505.04597>.
- [15] Hiroshi Sasaki, Chris G. Willcocks και Toby P. Breckon. *UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models*. 2021. arXiv: 2104.05358 [cs.CV]. URL: <https://arxiv.org/abs/2104.05358>.
- [16] Yang Song και Stefano Ermon. «Improved techniques for training score-based generative models». Στο: *Advances in neural information processing systems* 33 (2020), σσ. 12438–12448.
- [17] Aapo Hyvärinen και Peter Dayan. «Estimation of non-normalized statistical models by score matching.» Στο: *Journal of Machine Learning Research* 6.4 (2005).
- [18] Yang Song κ.ά. *Sliced Score Matching: A Scalable Approach to Density and Score Estimation*. 2019. arXiv: 1905.07088 [cs.LG]. URL: <https://arxiv.org/abs/1905.07088>.
- [19] Pascal Vincent. «A connection between score matching and denoising autoencoders». Στο: *Neural computation* 23.7 (2011), σσ. 1661–1674.
- [20] Max Welling και Yee W Teh. «Bayesian learning via stochastic gradient Langevin dynamics». Στο: *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011, σσ. 681–688.
- [21] Scott Kirkpatrick, C Daniel Gelatt Jr και Mario P Vecchi. «Optimization by simulated annealing». Στο: *science* 220.4598 (1983), σσ. 671–680.
- [22] Radford M Neal. «Annealed importance sampling». Στο: *Statistics and computing* 11.2 (2001), σσ. 125–139.
- [23] Yang Song κ.ά. *Score-Based Generative Modeling through Stochastic Differential Equations*. 2021. arXiv: 2011.13456 [cs.LG]. URL: <https://arxiv.org/abs/2011.13456>.
- [24] Brian DO Anderson. «Reverse-time diffusion equation models». Στο: *Stochastic Processes and their Applications* 12.3 (1982), σσ. 313–326.
- [25] Jiaming Song, Chenlin Meng και Stefano Ermon. *Denoising Diffusion Implicit Models*. 2022. arXiv: 2010.02502 [cs.LG]. URL: <https://arxiv.org/abs/2010.02502>.
- [26] Christopher M Bishop και Nasser M Nasrabadi. *Pattern recognition and machine learning*. Τόμ. 4. 4. Springer, 2006.



- [27] Junming Chen κ.ά. «DiffSHEG: A Diffusion-Based Approach for Real-Time Speech-driven Holistic 3D Expression and Gesture Generation». Στο: *CVPR*. 2024.
- [28] Simbarashe Nyatsanga κ.ά. «A comprehensive review of data-driven co-speech gesture generation». Στο: *Computer Graphics Forum*. Τόμ. 42. 2. Wiley Online Library. 2023, σσ. 569–596.
- [29] Isabella Poggi κ.ά. «Greta. a believable embodied conversational agent». Στο: *Multimodal intelligent information presentation*. Springer, 2005, σσ. 3–25.
- [30] Stefan Kopp κ.ά. «Towards a common framework for multimodal generation: The behavior markup language». Στο: *International workshop on intelligent virtual agents*. Springer. 2006, σσ. 205–217.
- [31] Justine Cassell, Hannes Högni Vilhjálmsson και Timothy Bickmore. «Beat: the behavior expression animation toolkit». Στο: *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 2001, σσ. 477–486.
- [32] Marcus Thiebaux κ.ά. «Smartbody: Behavior realization for embodied conversational agents». Στο: *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1*. 2008, σσ. 151–158.
- [33] Stacy Marsella κ.ά. «Virtual character performance from speech». Στο: *Proceedings of the 12th ACM SIGGRAPH/Eurographics symposium on computer animation*. 2013, σσ. 25–35.
- [34] Youngwoo Yoon κ.ά. «Speech gesture generation from the trimodal context of text, audio, and speaker identity». Στο: *ACM Transactions on Graphics (TOG)* 39.6 (2020), σσ. 1–16.
- [35] Yuanzhi Liang κ.ά. «Seeg: Semantic energized co-speech gesture generation». Στο: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, σσ. 10473–10482.
- [36] Chaitanya Ahuja κ.ά. «No gestures left behind: Learning relationships between spoken language and freeform gestures». Στο: *Findings of the association for computational linguistics: EMNLP 2020*. 2020, σσ. 1884–1895.
- [37] Taras Kucherenko κ.ά. «Gesticulator: A framework for semantically-aware speech-driven gesture generation». Στο: *Proceedings of the 2020 International Conference on Multimodal Interaction*. ICMI '20. Virtual Event, Netherlands: Association for Computing Machinery, 2020, σσ. 242–250. ISBN: 9781450375818. DOI: 10.1145/3382507.3418815. URL: <https://doi.org/10.1145/3382507.3418815>.
- [38] Jacob Devlin κ.ά. «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding». Στο: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, σσ. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: <https://www.aclweb.org/anthology/N19-1423/>.

- [39] Vladislav Korzun, Ilya Dimov και Andrey Zharkov. *The FineMotion entry to the GENE Challenge 2020*. Οκτ. 2020. DOI: 10.5281/zenodo.4088609. URL: <https://doi.org/10.5281/zenodo.4088609>.
- [40] Taras Kucherenko κ.ά. «A Large, Crowdsourced Evaluation of Gesture Generation Systems on Common Data: The GENE Challenge 2020». Στο: *26th International Conference on Intelligent User Interfaces*. IUI '21. College Station, TX, USA: Association for Computing Machinery, 2021, σσ. 11–21. ISBN: 9781450380171. DOI: 10.1145/3397481.3450692. URL: <https://doi.org/10.1145/3397481.3450692>.
- [41] Shivam Mehta κ.ά. «Diff-TTSG: Denoising probabilistic integrated speech and gesture synthesis». Στο: *arXiv preprint arXiv:2306.09417* (2023).
- [42] Chaitanya Ahuja, Dong Won Lee και Louis-Philippe Morency. «Low-resource adaptation for personalized co-speech gesture generation». Στο: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, σσ. 20566–20576.
- [43] Masahiro Mori, Karl F. MacDorman και Norri Kageki. «The Uncanny Valley [From the Field]». Στο: *IEEE Robotics & Automation Magazine* 19.2 (2012), σσ. 98–100. DOI: 10.1109/MRA.2012.2192811.
- [44] Roger K Moore. «A Bayesian explanation of the ‘Uncanny Valley’ effect and related psychological phenomena». Στο: *Scientific reports* 2.1 (2012), σ. 864.
- [45] Haiyang Liu κ.ά. «Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis». Στο: *European conference on computer vision*. Springer. 2022, σσ. 612–630.
- [46] Kenta Takeuchi κ.ά. «Creating a gesture-speech dataset for speech-based automatic gesture generation». Στο: *International Conference on Human-Computer Interaction*. Springer. 2017, σσ. 198–202.
- [47] Ylva Ferstl και Rachel McDonnell. «Investigating the use of recurrent motion modelling for speech gesture generation». Στο: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. 2018, σσ. 93–98.
- [48] Cheng-hsin Wu κ.ά. *Multiface: A Dataset for Neural Face Rendering*. 2023. arXiv: 2207.11243 [cs.CV]. URL: <https://arxiv.org/abs/2207.11243>.
- [49] Gabriele Fanelli κ.ά. «A 3-d audio-visual corpus of affective communication». Στο: *IEEE Transactions on Multimedia* 12.6 (2010), σσ. 591–598.
- [50] Daniel Cudeiro κ.ά. «Capture, learning, and synthesis of 3D speaking styles». Στο: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, σσ. 10101–10111.
- [51] Shiry Ginosar κ.ά. «Learning individual styles of conversational gesture». Στο: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, σσ. 3497–3506.

- [52] Ikhsanul Habibie κ.ά. «Learning speech-driven 3d conversational gestures from video». Στο: *Proceedings of the 21st ACM international conference on intelligent virtual agents*. 2021, σσ. 101–108.
- [53] Hongwei Yi κ.ά. «Generating holistic 3d human motion from speech». Στο: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, σσ. 469–480.
- [54] Georgios Pavlakos κ.ά. «Expressive body capture: 3d hands, face, and body from a single image». Στο: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, σσ. 10975–10985.
- [55] Ollie Johnston και Frank Thomas. *The illusion of life: Disney animation*. Disney Editions New York, 1981.
- [56] Thomas P Thesen. «Reviewing and updating the 12 principles of animation». Στο: *Animation* 15.3 (2020), σσ. 276–296.
- [57] European Parliament and Council. *Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights*. Official Journal of the European Union, L 372, pp. 12–18. Accessed: 2025-10-09. 2006. URL: <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32006L0116>.
- [58] United States Congress. *Copyright Term Extension Act of 1998 (Sonny Bono Act)*. Public Law 105–298, 112 Stat. 2827. Accessed: 2025-10-09. 1998. URL: <https://www.govinfo.gov/content/pkg/PLAW-105publ298/pdf/PLAW-105publ298.pdf>.
- [59] U.S. Copyright Office. *Copyright Basics (Circular 1)*. U.S. Library of Congress. Accessed: 2025-10-09. 2024. URL: <https://www.copyright.gov/circls/circ01.pdf>.
- [60] Robin E Brenner. *Understanding manga and anime*. Bloomsbury Publishing USA, 2007.
- [61] Rayna Denison. *Anime: A critical introduction*. Bloomsbury Publishing, 2015.
- [62] Youngwoo Yoon κ.ά. «Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots». Στο: *2019 International Conference on Robotics and Automation (ICRA)*. IEEE. 2019, σσ. 4303–4309.
- [63] Matthew Loper κ.ά. «SMPL: A Skinned Multi-Person Linear Model». Στο: *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34.6 (Οκτ. 2015), 248:1–248:16.
- [64] Javier Romero, Dimitrios Tzionas και Michael J. Black. «Embodied Hands: Modeling and Capturing Hands and Bodies Together». Στο: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*. 245:1–245:17 36.6 (Νοέ. 2017).
- [65] Tianye Li κ.ά. «Learning a model of facial shape and expression from 4D scans». Στο: *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36.6 (2017), 194:1–194:17. URL: <https://doi.org/10.1145/3130800.3130813>.

- [66] Yi Zhou κ.ά. «On the continuity of rotation representations in neural networks». Στο: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, σσ. 5745–5753.
- [67] OpenAI ChatGPT. *YesChat: AI-based Character Identifier and Reference Generator*. Anime/Video Game Character Identifier. 2025. URL: [https://www.yeschat.ai/gpts-20TolWWMNi-Anime-video-game-Character-Identifier?utm\\_source=chatgpt.com](https://www.yeschat.ai/gpts-20TolWWMNi-Anime-video-game-Character-Identifier?utm_source=chatgpt.com).
- [68] Martinus Grady Naftali, Jason Sebastian Sulistyawan και Kelvin Julian. *AniWho: A Quick and Accurate Way to Classify Anime Character Faces in Images*. 2023. arXiv: 2208.11012 [cs.CV]. URL: <https://arxiv.org/abs/2208.11012>.
- [69] Christian Szegedy κ.ά. «Rethinking the Inception Architecture for Computer Vision». Στο: *CoRR abs/1512.00567* (2015). arXiv: 1512.00567. URL: <http://arxiv.org/abs/1512.00567>.
- [70] Christian Szegedy κ.ά. *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. 2016. arXiv: 1602.07261 [cs.CV].
- [71] Mark Sandler κ.ά. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2019. arXiv: 1801.04381 [cs.CV]. URL: <https://arxiv.org/abs/1801.04381>.
- [72] Mingxing Tan και Quoc V. Le. «EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks». Στο: *ArXiv abs/1905.11946* (2019).
- [73] SmilingWolf. *WD ConvNext Tagger v3*. <https://huggingface.co/SmilingWolf/wd-convnext-tagger-v3>. Model repository on Hugging Face. 2024.
- [74] Gwern Branwen, Aaron Gokaslan Zhang κ.ά. *Danbooru2021: A Large-Scale Crowdsourced and Tagged Anime Illustration Dataset*. <https://gwern.net/danbooru2021>. Accessed: 2025-10-13. 2021.
- [75] Robert Plutchik. «The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice». Στο: *American scientist* 89.4 (2001), σσ. 344–350.
- [76] James A Russell. «Core affect and the psychological construction of emotion.» Στο: *Psychological review* 110.1 (2003), σ. 145.
- [77] Paul Ekman, Tim Dalgleish και M Power. «Basic emotions». Στο: *San Francisco, USA* (1999).
- [78] Maria Chiara Caschera, Patrizia Grifoni και Fernando Ferri. «Emotion classification from speech and text in videos using a multimodal approach». Στο: *Multimodal Technologies and Interaction* 6.4 (2022), σ. 28.
- [79] Wafa Mellouk και Wahida Handouzi. «Facial emotion recognition using deep learning: review and insights». Στο: *Procedia Computer Science* 175 (2020), σσ. 689–694.
- [80] Sharmeen M Saleem Abdullah Abdullah κ.ά. «Multimodal emotion recognition using deep learning». Στο: *Journal of Applied Science and Technology Trends* 2.01 (2021), σσ. 73–79.

- [81] Albert Mehrabian και Susan R. Ferris. «Inference of Attitudes from Nonverbal Communication in Two Channels». Στο: *Journal of Consulting Psychology* 31.3 (1967), σσ. 248–252. DOI: 10.1037/h0024648.
- [82] justfoolingaround. *AnimDL: A highly efficient, fast, powerful and light-weight anime downloader and streamer*. Έκδ. 1.7.27. 2023. URL: <https://github.com/justfoolingaround/animdl>.
- [83] Gary Bradski και Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. ” O’Reilly Media, Inc.”, 2008.
- [84] Caption Emporium. *anime-caption-danbooru-2021-sfw-5m-hq*. <https://huggingface.co/datasets/CaptionEmporium/anime-caption-danbooru-2021-sfw-5m-hq>. 2024.
- [85] Kevin Cai, Chonghua Liu και David M. Chan. *ANIM-400K: A Large-Scale Dataset for Automated End-To-End Dubbing of Video*. 2024. arXiv: 2401.05314 [eess.AS]. URL: <https://arxiv.org/abs/2401.05314>.
- [86] FFmpeg Developers. *FFmpeg*. Available: <https://ffmpeg.org/>. 2023.
- [87] Alec Radford κ.ά. *Robust Speech Recognition via Large-Scale Weak Supervision*. 2022. arXiv: 2212.04356 [eess.AS]. URL: <https://arxiv.org/abs/2212.04356>.
- [88] Guanshuo Xu. *imgutils: A Collection of Image Processing Utilities*. <https://github.com/deepghs/imgutils>. GitHub repository. 2023.
- [89] Wu Liu κ.ά. «Recent advances of monocular 2D and 3D human pose estimation: A deep learning perspective». Στο: *ACM Computing Surveys* 55.4 (2022), σσ. 1–41.
- [90] Zhe Cao κ.ά. *OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields*. 2019. arXiv: 1812.08008 [cs.CV]. URL: <https://arxiv.org/abs/1812.08008>.
- [91] Google LLC. *MediaPipe*. Available: [https://ai.google.dev/edge/mediapipe/solutions/vision/pose\\_landmarker](https://ai.google.dev/edge/mediapipe/solutions/vision/pose_landmarker). 2019.
- [92] Hao-Shu Fang κ.ά. «AlphaPose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time». Στο: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
- [93] Yuxin Wu κ.ά. *Detectron2*. <https://github.com/facebookresearch/detectron2>. 2019.
- [94] Shuhong Chen και Matthias Zwicker. *Transfer Learning for Pose Estimation of Illustrated Characters*. 2021. arXiv: 2108.01819 [cs.CV]. URL: <https://arxiv.org/abs/2108.01819>.
- [95] Tsung-Yi Lin κ.ά. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: 1405.0312 [cs.CV].
- [96] Catalin Ionescu κ.ά. «Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments». Στο: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.7 (Ιούλ. 2014), σσ. 1325–1339.



- [97] Julieta Martinez κ.ά. *A simple yet effective baseline for 3d human pose estimation*. 2017. arXiv: 1705.03098 [cs.CV]. URL: <https://arxiv.org/abs/1705.03098>.
- [98] Tianlang Chen κ.ά. *Anatomy-aware 3D Human Pose Estimation with Bone-based Pose Decomposition*. 2021. arXiv: 2002.10322 [cs.CV]. URL: <https://arxiv.org/abs/2002.10322>.
- [99] Yujun Cai κ.ά. «Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks». Στο: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, σσ. 2272–2281.
- [100] Wenhao Li κ.ά. «MHFormer: Multi-Hypothesis Transformer for 3D Human Pose Estimation». Στο: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, σσ. 13147–13156.
- [101] Fan Zhang κ.ά. *MediaPipe Hands: On-device Real-time Hand Tracking*. 2020. arXiv: 2006.10214 [cs.CV]. URL: <https://arxiv.org/abs/2006.10214>.
- [102] hysts. *Anime Face Detector*. <https://github.com/hysts/anime-face-detector>. 2021.
- [103] Adrian Bulat και Georgios Tzimiropoulos. «How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks)». Στο: *International Conference on Computer Vision*. 2017.
- [104] Erroll Wood κ.ά. «3d face reconstruction with dense landmarks». Στο: *European Conference on Computer Vision*. Springer. 2022, σσ. 160–177.
- [105] Jianzhu Guo κ.ά. «Towards fast, accurate and stable 3d dense face alignment». Στο: *European Conference on Computer Vision*. Springer. 2020, σσ. 152–168.
- [106] Sunwon Jeong και Ju Yong Chang. «MeshLifter: Weakly Supervised Approach for 3D Human Mesh Reconstruction from a Single 2D Pose Based on Loop Structure». Στο: *Sensors* 20.15 (2020), σ. 4257.
- [107] Chihiro Nakatsuka, Jianfeng Xu και Kazuyuki Tasaka. «Learning Joint Twist Rotation for 3D Human Pose Estimation from a Single Image.» Στο: *VISIGRAPP (5: VISAPP)*. 2021, σσ. 379–386.
- [108] Przemysław Dobrowolski. *Swing-twist decomposition in Clifford algebra*. 2015. arXiv: 1506.05481 [cs.R0]. URL: <https://arxiv.org/abs/1506.05481>.
- [109] Benjamin van Niekerc κ.ά. «A Comparison of Discrete and Soft Speech Units for Improved Voice Conversion». Στο: *ICASSP*. 2022.
- [110] Yuan Gong, Yu-An Chung και James Glass. «Ast: Audio spectrogram transformer». Στο: *arXiv preprint arXiv:2104.01778* (2021).
- [111] Yuan Gong κ.ά. «Ssast: Self-supervised audio spectrogram transformer». Στο: *Proceedings of the AAAI Conference on Artificial Intelligence*. Τόμ. 36. 10. 2022, σσ. 10699–10709.
- [112] Hugo Touvron κ.ά. «Training data-efficient image transformers & distillation through attention». Στο: *International conference on machine learning*. PMLR. 2021, σσ. 10347–10357.



- [113] Xin Chen κ.ά. «Executing your commands via motion diffusion in latent space». Στο: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, σσ. 18000–18010.
- [114] Mathis Petrovich, Michael J Black και Gül Varol. «Action-conditioned 3d human motion synthesis with transformer vae». Στο: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, σσ. 10985–10995.
- [115] Robin Rombach κ.ά. «High-resolution image synthesis with latent diffusion models». Στο: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, σσ. 10684–10695.
- [116] Alexander Quinn Nichol και Prafulla Dhariwal. «Improved denoising diffusion probabilistic models». Στο: *International conference on machine learning*. PMLR. 2021, σσ. 8162–8171.
- [117] Ashish Vaswani κ.ά. «Attention is all you need». Στο: *Advances in neural information processing systems* 30 (2017).
- [118] Ross Girshick. «Fast r-cnn». Στο: *Proceedings of the IEEE international conference on computer vision*. 2015, σσ. 1440–1448.
- [119] Daniel S Park κ.ά. «SpecAugment: A simple data augmentation method for automatic speech recognition». Στο: *arXiv preprint arXiv:1904.08779* (2019).
- [120] Ilya Loshchilov και Frank Hutter. «Decoupled weight decay regularization». Στο: *arXiv preprint arXiv:1711.05101* (2017).
- [121] Martin Heusel κ.ά. «Gans trained by a two time-scale update rule converge to a local nash equilibrium». Στο: *Advances in neural information processing systems* 30 (2017).
- [122] Ruilong Li κ.ά. «Ai choreographer: Music conditioned 3d dance generation with aist++». Στο: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, σσ. 13401–13412.
- [123] Brian McFee κ.ά. «librosa: Audio and music signal analysis in python.» Στο: *SciPy 2015* (2015), σσ. 18–24.
- [124] David E Rumelhart, Geoffrey E Hinton και Ronald J Williams. *Learning internal representations by error propagation*. Αδημοσίευτη ερευνητική εργασία. 1985.