National Technical University of Athens

School of Electrical and Computer Engineering

DIVISION OF SIGNALS, CONTROL AND ROBOTICS

Photorealistic Sign Language Production from Text using Transformer Networks and Neural Rendering

DIPLOMA THESIS

of

Chrysa Pratikaki

Supervisor: Petros Maragos
Professor NTUA

COMPUTER VISION, SPEECH COMMUNICATION AND SIGNAL PROCESSING GROUP

Athens, March 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Signals, Control and Robotics
Computer Vision, Speech Communication and Signal Processing Group

Photorealistic Sign Language Production from Text using Transformer Networks and Neural Rendering

DIPLOMA THESIS

of

Chrysa Pratikaki

Supervisor: Petros Maragos
Professor NTUA

Co-Supervisor: Anastasios Roussos
Principal Researcher FORTH ICS

Co-Supervisor: Panagiotis Filntisis
Researcher Athena Research Center

Approved by the examination committee on 7th March, 2025.

.....	.....	.....
Petros Maragos	Athanasios Rontogiannis	Ioannis Kordonis
Professor NTUA	As. Professor NTUA	As. Professor NTUA

Athens, March 2025

.....

Chrysa Pratikaki
Graduate of Electrical and
Computer Engineering NTUA

Περίληψη

Οι νοηματικές γλώσσες αποτελούν την κύρια μορφή επικοινωνίας για εκατοντάδες εκατομμύρια ανθρώπους που ανήκουν στις κοινότητες των κωφών ανά τον κόσμο, ενώ υπολογίζεται ότι υπάρχουν πάνω από 200 διαφορετικές νοηματικές γλώσσες και διάλεκτοι. Με σκοπό την γεφύρωση της επικοινωνίας μεταξύ γνωστών της νοηματικής γλώσσας και μη, ιδιαίτερα στην σύγχρονη ψηφιακή πραγματικότητα, θα ήταν καθοριστική η ύπαρξη ενός καθολικού συστήματος, ικανού να μεταφράζει από και προς την ομιλούμενη και την νοηματική γλώσσα. Ένα τέτοιο σύστημα βρίσκει άμεση εφαρμογή σε βασικότατους τομείς της καθημερινότητας κωφών ατόμων, όπως είναι η εκπαίδευση, τα μέσα ενημέρωσης, η ψυχαγωγία και ο πολιτισμός.

Για το σκοπό αυτό, η συνεχής μετάφραση και παραγωγή της νοηματικής γλώσσας από κείμενο είναι τα δύο απαραίτητα στοιχεία για τη δημιουργία ενός τέτοιου συστήματος. Τις τελευταίες δεκαετίες, η μετάφραση της νοηματικής γλώσσας, η μετατροπή δηλαδή των μηνυμάτων από βίντεο νοηματικής γλώσσας σε κείμενο, έχει συγκεντρώσει σημαντικό ερευνητικό ενδιαφέρον στο χώρο της Όρασης Υπολογιστών, με πληθώρα δημοσιεύσεων να εξερευνούν αυτή την πρόκληση χρησιμοποιώντας τεχνολογίες Μηχανικής Μάθησης. Ωστόσο, συγκεκριμένα η αυτόματη παραγωγή ρεαλιστικών βίντεο νοηματικής γλώσσας από κείμενο θεωρείται ένα από τα πιο απαιτητικά ανοιχτά προβλήματα όσον αφορά τις τεχνολογίες νοηματικών γλωσσών. Οι πιο πρόσφατες τεχνολογίες για την παραγωγή νοηματικής γλώσσας αντιμετωπίζουν το πρόβλημα αξιοποιώντας βαθιά νευρωνικά δίκτυα όπως οι Μετασχηματιστές και ποικιλία παραγωγικών μοντέλων, και παρόλο που αυτές οι μέθοδοι δείχνουν ενθαρρυντικά αποτελέσματα, υπάρχει βέβαιη δυνατότητα για περαιτέρω εξέλιξη.

Ο κεντρικός στόχος αυτής της διπλωματικής είναι να αναπτύξει ένα μοντέλο Μηχανικής Μάθησης για την αυτόματη παραγωγή ρεαλιστικών βίντεο νοηματικής γλώσσας από κείμενο. Χωρίζουμε την λύση μας σε δύο βασικά στάδια: Πρώτα, χρησιμοποιώντας δίκτυα Μετασχηματιστών (Transformers) μεταφράζουμε το κείμενο σε σκελετικές ακολουθίες ανθρώπινης πόζας (MediaPipe). Στη συνέχεια συνθέτουμε τον παραγόμενο ρεαλιστικό σκελετό μέσω αρχιτεκτονικής βασισμένη σε Generative Adversarial Networks, το οποίο οδηγεί στο τελικό συνθετικό βίντεο νοηματιστή. Αξιολογούμε την αποτελεσματικότητα του προτεινόμενου συστήματος σε τρία διαφορετικά σύνολα δεδομένων μέσω εκτεταμένων συγκριτικών πειραμάτων και αξιολόγησης χρηστών.

Τμήμα της εργασίας έγινε δεκτό στο 18ο συνέδριο PErvasive Technologies Related to Assistive Environments (PETRA 2025), με τίτλο "A Transformer-Based Framework for Greek Sign Language Production using Extended Skeletal Motion Representations" [49] και συγγραφείς τους Χρύσα Πρατικάκη, Παναγιώτη Φιλντίση, Αθανάσιο Κατσαμάνη, Αναστάσιο Ρούσσο και Πέτρο Μαραγκό.

Λέξεις κλειδιά — Παραγωγή Νοηματικής Γλώσσας, Βαθιά Μάθηση, Transformers, Generative Adversarial Networks, Neural Rendering, Εκτίμηση Πόζας

Abstract

Sign Languages are the primary form of communication for Deaf communities across the world. It is estimated that more than 70 million people make part of the deaf and hard-of-hearing (DHH) community, while there are more than 200 Sign Languages across the world. To break the communication barriers between the DHH and the hearing communities, it is imperative to build systems capable of translating the spoken language into sign language and reciprocally. To this end, continuous sign language translation and production are the two necessary components for making such machine-learning based system. Over the past three decades, Sign Language Translation has gained significant interest, resulting in a plethora of publications exploring various technologies to address the challenge. However, Sign Language Production is considered to be one of the most challenging open problems regarding Sign Language technologies. The most recent suggested technologies for SLP tackle the synthesis of photorealistic sign language videos with neural machine translation and a variety of generative models, and although these methods show encouraging results, there remains potential for further adaptations and innovation.

Building on insights from previous research, the central objective of this thesis is to develop a robust deep learning model for Sign Language Production (SLP). We tackle this task by utilizing a transformer-based architecture that enables the translation from text input to human pose keypoints. Furthermore, we explore the photerealistic aspect of the problem, aiming to create a complete SLP pipeline that transforms text directly into realistic human SL videos. For the photorealistic module, we harness Generative Adversarial Networks (GANs) to perform neural rendering on the pose sequences generated by the transformer model. Finally, we evaluate the effectiveness of the proposed pipeline on three different datasets through an extensive series of comparative analyses, ablation studies, and user studies.

Part of our work was accepted at the 18th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA 2025), titled "A Transformer-Based Framework for Greek Sign Language Production using Extended Skeletal Motion Representations" [49] with the authors being Chrysa Pratikaki, Panagiotis Filntisis, Athanasios Katsamanis, Anastasios Roussos and Petros Maragos.

Keywords — Sign Language Production, Sign Language Translation, Deep Learning, Transformers, LLMs, Generative Adversarial Networks, Neural Rendering, Pose Estimation

Ευχαριστίες

Η εργασία αυτή σημαίνει αισίως την ολοκλήρωση των σπουδών μου στην Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του ΕΜΠ. Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή Πέτρο Μαραγκό για την ευκαιρία να εκπονήσω την διπλωματική μου εργασία στο εργαστήριο Όρασης Υπολογιστών και Επεξεργασίας Σήματος, γύρω από ένα θέμα με έντονες ακαδημαϊκές απαιτήσεις αλλά και κοινωνικό αντίκτυπο.

Στη συνέχεια θα ήθελα να ευχαριστήσω ολόψυχα τους Δρ. Αναστάσιο Ρούσσο, κύριο ερευνητή στο ΙΤΕ και Δρ. Παναγιώτη Φιλντίση, ερευνητή στο Ερευνητικό Κέντρο Αθηνά, για την εξαιρετική συνεργασία και συνεπίβλεψή τους. Με την καθοδήγηση τους κατάφερα να φέρω εις πέρας κάθε πρόκληση και να καλλιεργήσω το ενδιαφέρον μου για την έρευνα.

Φυσικά, θα ήθελα να ευχαριστήσω όλες τις φίλες και τους φίλους μου για τις όμορφες εμπειρίες που κάνουν τα φοιτητικά μου χρόνια πολύτιμα και αξέχαστα.

Τέλος δεν θα μπορούσα να μην ευχαριστήσω τους γονείς μου, Ελένη και Σάββα, για την συνεχή υποστήριξη και πίστη τους σε εμένα από μικρό παιδί.

Χρύσα Πρατικάκη
Μάρτιος 2025

Table of contents

List of Acronyms

ΕΝΓ	Ελληνική Νοηματική Γλώσσα
ΝΓ	Νοηματική Γλώσσα
AD	Autoregressive Decoding
BT	Back Translation
CNN	Convolutional Neural Network
DHH	Deaf, Hard-of Hearing
GAN	Generative Adversarial Network
MSE	Mean Squared Error
MP	Media-Pipe
NLP	Natural Language Processing
PT	Progressive Transformer
SL	Sign Language
SLR	Sign Language Recognition
SLP	Sign Language Production
SLT	Sign Language Translation
TF	Teacher Forcing

List of figures

List of tables

Εκτεταμένη Περίληψη στα ελληνικά

Εισαγωγή

Οι νοηματικές γλώσσες αποτελούν την κύρια μορφή επικοινωνίας δεκάδων εκατομμυρίων ανθρώπων ανά τον κόσμο. Εμφανίζονται από την αρχαιότητα ενώ συναντάμε εκατοντάδες διαλέκτους με μεγάλη ποικιλομορφία σε γραμματική, εκφράσεις και κινήσεις.

Για τα κωφά άτομα η ορθή εκμάθηση νοηματικής γλώσσας είναι κρίσιμη, διότι η χρήση της είναι απαραίτητη προϋπόθεση για την πλήρη συμμετοχή τους στην εκπαίδευση, στον εργασιακό βίο αλλά και σε κοινωνικά θέματα. Ωστόσο, μεγάλη πλειοψηφία των κωφών/ βαρήκοων ανθρώπων ανήκουν είτε στην τρίτη ηλικία είτε σε ευπαθείς κοινωνικές ομάδες, με αποτέλεσμα η εκμάθηση νοηματικής γλώσσας να μην αποτελεί τυπικό μέρος της εκπαίδευσής τους σε νεαρή ηλικία, οδηγώντας σε αποσύνδεση που μπορεί να εμποδίσει την κοινωνική αλληλεπίδραση και την πρόσβαση σε διάφορες υπηρεσίες. Αυτό το κενό, αλλά και η ανάγκη πολλαπλασιασμού των ανθρώπων που γνωρίζουν την νοηματική γλώσσα, υπογραμμίζει τη σημασία της ανάπτυξης ηλεκτρονικών τεχνολογιών που μπορούν να διευκολύνουν την επικοινωνία σε πραγματικό χρόνο μεταξύ των χρηστών της νοηματικής γλώσσας σε βασικότατους τομείς όπως η εκπαίδευση, τα μέσα ενημέρωσης και η ψυχαγωγία.

Οι νοηματικές γλώσσες είναι πολύπλοκες γλώσσες που βασίζονται κυρίως σε οπτικοκινητικά στοιχεία και όχι σε ήχους. Έχουν δομή που συγκρίνεται με τις προφορικές γλώσσες, περιλαμβάνοντας φωνολογία, μορφολογία, σύνταξη, σημασιολογία και πραγματολογία. Αντί να χρησιμοποιούν φωνητικά στοιχεία που παράγονται με τη φωνητική άρθρωση, οι νοηματικές γλώσσες χρησιμοποιούν ταυτόχρονες χειρονομίες, εκφράσεις προσώπου και κινήσεις του σώματος για να μεταδώσουν νόημα. Αυτή η πολυτροπική φύση κάνει την υπολογιστική μοντελοποίηση της νοηματικής γλώσσας πιο δύσκολη, ειδικά όταν χρησιμοποιούνται τεχνικές βαθιάς μάθησης για την αυτόματη αναγνώριση ή παραγωγή επικοινωνίας με χειρονομίες. Η βασική διαφορά μεταξύ των προφορικών και των νοηματικών γλωσσών είναι ότι οι πρώτες χρησιμοποιούν γραμμικά φωνήματα, ενώ οι δεύτερες χρησιμοποιούν παράλληλους διαύλους πληροφοριών, κάτι που απαιτεί διαφορετικές γλωσσικές και υπολογιστικές προσεγγίσεις. Παρά την έλλειψη φωνημάτων, οι νοηματικές γλώσσες έχουν ένα αφωνολογικό σύστημα βασισμένο σε χειροκίνητες (manual) παραμέτρους, οι οποίες είναι οι εξής:

- Μορφή χειρονομίας
- Θέση των χεριών στο χώρο
- Κίνηση χεριών (είδος και κατεύθυνση)
- Προσανατολισμός χεριών στο χώρο

Εκτός από τις χειρονομίες, οι νοηματικές γλώσσες βασίζονται και σε μη χειροκίνητους δείκτες, που

συμβάλλουν στη σημασία της πρότασης. Μεταξύ αυτών συμπεριλαμβάνονται:

- Εκφράσεις του προσώπου
- Εκφράσεις του στόματος
- Αλλαγές στην πόζα του σώματος

Οι νοηματικές γλώσσες δραστηριοποιούνται στον τρισδιάστατο χώρο υπογραφής για την κωδικοποίηση τους μέσω μιας συγκεκριμένης χωρικής γραμματικής. Σε αντίθεση με τις προφορικές γλώσσες, που βασίζονται στη διαδοχική σειρά λέξεων και στις προθέσεις, οι νοηματικές γλώσσες επιτρέπουν στους χρήστες να ανατοποθετούν λέξεις και να επαναφέρουν αργότερα νοήματα με κινήσεις κατεύθυνσης. Έτσι διευκολύνεται η αναπαράσταση σύνθετων εννοιών, αλλά παράλληλα δημιουργεί προκλήσεις για την υπολογιστική μοντελοποίηση της νοηματικής γλώσσας, ειδικά όταν χρησιμοποιούνται τεχνικές βαθιάς μάθησης.

Η γλωσσική πολυπλοκότητα των νοηματικών γλωσσών όπως περιγράφηκε, δημιουργεί προκλήσεις για τα μοντέλα βαθιάς μάθησης. Σε αντίθεση με τα συστήματα που βασίζονται σε κείμενο ή ομιλία, η παραγωγή νοηματικής γλώσσας απαιτεί πολυτροπική προσέγγιση που ενσωματώνει αποτελεσματικά την ταυτόχρονη παρακολούθηση κινήσεων των χεριών, του προσώπου και του σώματος, προκειμένου να παραχθούν νοηματικά συνεπείς προτάσεις ΝΓ.

Πρόσφατες εξελίξεις στην βαθιά μάθηση και στην εκτίμηση της ανθρώπινης στάσης έχουν συνεισφέρει σημαντικά στον τομέα της αναγνώρισης και παραγωγής νοηματικής γλώσσας. Ωστόσο, παραμένουν αρκετές προκλήσεις, όπως η γενίκευση σε διαφορετικές νοηματικές γλώσσες και η προσαρμογή σε διαλεκτικές παραλλαγές. Σε αυτή τη διπλωματική εργασία, θα εξετάσουμε λεπτομερώς αυτές τις προκλήσεις, παρουσιάζοντας μεθοδολογίες για την παραγωγή νοηματικής γλώσσας βασισμένες σε βαθιά μάθηση και αξιολογώντας την αποτελεσματικότητά τους σε πραγματικές εφαρμογές.

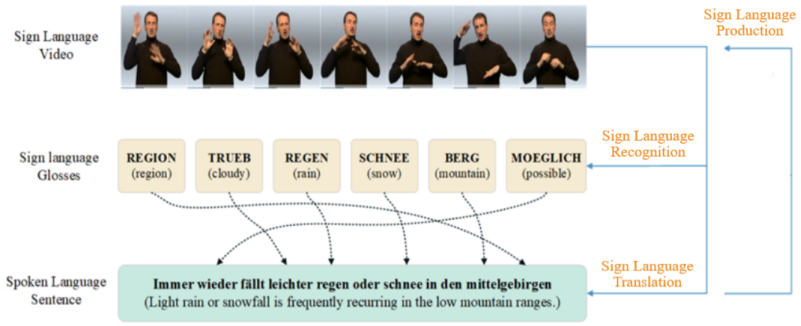


Figure 0.0.1: Μεταφράσεις που υποστηρίζει ένα σύστημα μετάφρασης Νοηματικής Γλώσσας

Ο τομέας της Επεξεργασίας Νοηματικής Γλώσσας (Sign Language Processing) αγγίζει την διασταύρωση τομέων της γλωσσολογίας, της όρασης υπολογιστών (computer vision) και της επεξεργασίας φυσικής γλώσσας (NLP) καθώς συνδυάζει την παρακολούθηση του σώματος με την κειμενική αναπαράσταση. Τα πιο βασικά στοιχεία ενός διαδραστικού ηλεκτρονικού συστήματος ΝΓ είναι η Αναγ-

νώριση Νοηματικής Γλώσσας (Sign Language Recognition - SLR), η Μετάφραση Νοηματικής Γλώσσας (Sign Language Translation - SLT) και η Παραγωγή Νοηματικής Γλώσσας από κείμενο (Sign Language Production - SLP).

Η Αναγνώριση Νοηματικής Γλώσσας (SLR) εστιάζει στην ερμηνεία της νοηματικής γλώσσας (βίντεο → κείμενο) και στην αντιστοίχισή της σε σειριακές κωδικές λέξεις (glosses) που αναπαριστούν το νόημα κάθε διακριτής κίνησης. Η Μετάφραση Νοηματικής Γλώσσας (SLT) είναι προέκταση της απλής αναγνώρισης, και επιτρέπει την μετατροπή των δεδομένων βίντεο σε απλό κείμενο (σε αντίθεση με τα glosses) . Αυτή η δυνατότητα είναι ζωτικής σημασίας για την σύγχρονη επικοινωνία μεταξύ ακουστικών και κωφών ατόμων, παραδειγματικά μέσω πλατφορμών τηλεδιάσκεψης, είτε μέσω δημόσιων υπηρεσιών.

Μία από τις πιο δύσκολες και λιγότερο εξερευνημένες περιοχές στην Επεξεργασία Νοηματικής Γλώσσας είναι η Παραγωγή Νοηματικής Γλώσσας από κείμενο (Sign Language Production - SLP). Η SLP αναφέρεται στη δημιουργία κινούμενων βίντεο ΝΓ ακριβείας με δεδομένα εισόδου κείμενο. Σε αντίθεση με την αναγνώριση και την μετάφραση, που εστιάζουν στην κατανόηση της νοηματικής γλώσσας, η παραγωγή από κείμενο απαιτεί απ' το δίκτυο την ικανότητα να παράγει συνθετικό περιεχόμενο νοηματικής γλώσσας που είναι ταυτόχρονα γλωσσικά ακριβές και οπτικά πειστικό.

Τα πρωτοεμφανιζόμενα συστήματα SLP βασίζονται κυρίως σε απλή αντιστοίχιση λέξεων σε ακολουθίες κινήσεων. Πρόσφατες εξελίξεις στη βαθιά μάθηση, ιδιαίτερα στα νευρωνικά δίκτυα και στις παραγωγικές (generative) αρχιτεκτονικές, έχουν ξεκλειδώσει πολλά υποσχόμενες δυνατότητες για τη δημιουργία πιο φωτορεαλιστικού περιεχομένου νοηματικής γλώσσας. Παρά αυτές τις εξελίξεις, οι τρέχουσες λύσεις βρίσκονται ακόμα στα αρχικά στάδια, με σημαντικό χώρο για βελτίωση στην ευελιξία και την ακρίβεια των παραγόμενων βίντεο νοηματικής γλώσσας.

Το βασικό κίνητρο αυτής της διπλωματικής εργασίας είναι η αξιοποίηση σύγχρονων τεχνικών μηχανικής μάθησης με σκοπό την ανάπτυξη ενός βελτιωμένου συστήματος παραγωγικής βίντεο ΝΓ από κείμενο. Στόχος μας είναι να αντιμετωπίσουμε τις υπαρκτές προκλήσεις που ενδέχεται να αντιμετωπίσει ένα τέτοιο σύστημα, μεταξύ των οποίων βρίσκονται η επίτευξη ικανοποιητικού ρεαλισμού και φυσικότητας, η σημασιολογική ακρίβεια αλλά και η έλλειψη διαθέσιμων δεδομένων ΕΝΓ.

Για το σκοπό αυτό, αυτή η διπλωματική εργασία διερευνά μια προσέγγιση που συνδυάζει αρχιτεκτονικές βασισμένες σε δίκτυα γλωσσικών μετασχηματιστών (transformers) για τη γλωσσική επεξεργασία και σε τεχνικές νευρωνικής απόδοσης (neural rendering) για ρεαλιστική οπτική σύνθεση. Πιο συγκεκριμένα, αναλύουμε την παραγωγή βίντεο ΝΓ από κείμενο σε δύο βήματα: Πρώτα, χρησιμοποιώντας αρχιτεκτονική βασισμένη σε μετασχηματιστές (transformers) μετατρέπουμε το κείμενο (είσοδος) σε σκελετικές αναπαραστάσεις ανθρώπινης πόζας (keypoints, έξοδος). Στη συνέχεια, χρησιμοποιούμε Παραγωγικά Νευρωνικά Δίκτυα (Generative Adversarial Networks - GANs) για να πραγματοποιήσουμε απόδοση των σκελετικών ακολουθιών σε ένα φωτο-ρεαλιστικό συνθετικό νοηματιστή. Τέλος, αξιολογούμε την αποτελεσματικότητα της προτεινόμενης μεθόδου σε τρία διαφορετικά σύνολα δεδομένων, μέσω μιας σειράς πειραμάτων και μελετών χρηστών.

Θεωρητικό Υπόβαθρο

Αυτή η διπλωματική χρησιμοποιεί αρχιτεκτονικές βασισμένες στη Βαθιά Μάθηση - Transformers και Generative Adversarial Networks - και προκειμένου να κατανοήσουμε την λειτουργία της προτεινόμενης μεθοδολογίας παρέχουμε μια σύντομη επεξήγηση της λειτουργίας ορισμένων εξ αυτών.

Αρχικά, τα Feed Forward Networks (FFNs) είναι από τις απλούστερες αρχιτεκτονικές νευρωνικών δικτύων, όπου η πληροφορία ρέει μόνο προς τα εμπρός, από τις εισόδους προς τις εξόδους. Ωστόσο, για δεδομένα με διαδοχική εξάρτηση, όπως στη Φυσική Γλώσσα (NLP), απαιτούνται πιο πολύπλοκες αρχιτεκτονικές, όπως τα Επαναλαμβανόμενα Νευρωνικά Δίκτυα (RNNs). Τα RNNs έχουν μια κυκλική δομή που επιτρέπει στην πληροφορία να διατηρείται από το ένα βήμα της ακολουθίας στο επόμενο, κάνοντάς τα ιδανικά για εργασίες μοντελοποίησης γλώσσας. Μια ειδική κατηγορία RNNs είναι τα Δίκτυα LSTMs, τα οποία επιλύουν το πρόβλημα των vanishing gradients. Χρησιμοποιούν πύλες (gates) για να ελέγχουν ποια πληροφορία θα διατηρηθεί ή θα απορριφθεί κατά την επεξεργασία της ακολουθίας.

Ένα βήμα μετά τα RNNs, βρίσκονται οι Μετασχηματιστές (Transformers) που προτάθηκαν το 2017 και βασίζονται σε μηχανισμό προσοχής (self-attention). Ο μηχανισμός προσοχής συνήθως προσδίδει την ιδιότητα καλύτερης εκμάθησης των εξαρτήσεων μεταξύ των στοιχείων μιας ακολουθίας αλλά και παραλληλισμό στην εκπαίδευση. Η αρχιτεκτονική ενός Transformer αποτελείται από κωδικοποιητές (encoders) και αποκωδικοποιητές (decoders), οι οποίοι χρησιμοποιούν πολλαπλές κεφαλές προσοχής (multi-head attention) για την εξαγωγή πληροφοριών από διαφορετικά υποσύνολα αναπαραστάσεων. Στην εικόνα 0.0.2 βλέπουμε αναλυτικά την βασική αρχιτεκτονική ενός Transformer. Ο πιο συνηθισμένος τύπος προσοχής είναι η Scaled Dot-Product Attention, η οποία περιγράφεται από την εξίσωση:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{0.0.1}$$

όπου:

- Q: Είναι ο πίνακας των ερωτημάτων (queries), που αντιπροσωπεύει την τρέχουσα εργασία (π.χ. μια λέξη που θέλουμε να μεταφράσουμε).
- K: Είναι ο πίνακας των κλειδιών (keys), που αντιπροσωπεύει τα στοιχεία της ακολουθίας εισόδου.
- V: Είναι ο πίνακας των τιμών (values), που περιέχει τις πληροφορίες που θέλουμε να εξάγουμε από την ακολουθία εισόδου.
- d_k: Είναι η διάσταση των κλειδιών (keys), σταθερά κανονικοποίησης

Μια άλλη σημαντική αναφορά σε αυτό το σύντομο θεωρητικό υπόβαθρο είναι τα Συνελικτικά Νευρωνικά Δίκτυα (CNNs). Μετά την περιγραφή των πιο βασικών αρχιτεκτονικών που χρησιμοποιούνται στην Επεξεργασία Φυσικής Γλώσσας (NLP), αναφέρουμε επίσης τα CNNs, καθώς είναι εξαιρετικά σχετικά με την έρευνά μας, τόσο σε επίπεδο neural rendering και GAN όσο και σε επίπεδο pose estima-

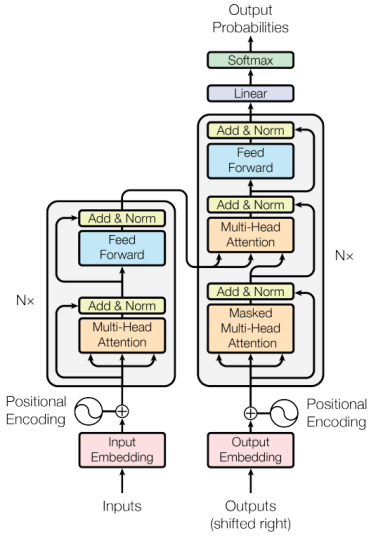


Figure 0.0.2: Βασική αρχιτεκτονική Μετασχηματιστή από την δημοσίευση [67].

tion. Τα CNNs είναι μια κατηγορία μοντέλων βαθιάς μάθησης που χρησιμοποιούνται κυρίως για την επεξεργασία εικόνας, αν και μπορούν να εφαρμοστούν και σε άλλες εργασίες. Το βασικό συστατικό ενός CNN είναι ίσως το στρώμα συνελίξεων (convolutional layer), το οποίο χρησιμοποιεί φίλτρα, καθένα από τα οποία έχει ένα ξεχωριστό σύνολο learnable βαρών. Αυτοί τα φίλτρα συνελίσσονται με την εικόνα εισόδου, δημιουργώντας τους λεγόμενους χάρτες χαρακτηριστικών (feature maps). Καθώς το δίκτυο εμβαθύνει, αυτά τα στρώματα καταγράφουν σταδιακά πιο πολύπλοκα χαρακτηριστικά, ξεκινώντας από την ανίχνευση απλών γραμμών και υφών στα αρχικά στρώματα έως πιο περίπλοκες αναπαραστάσεις στα βαθύτερα στρώματα. Για τη μείωση της διάστασης των χαρτών χαρακτηριστικών και του αριθμού των παραμέτρων, χρησιμοποιούνται συχνά στρώματα pooling layers, όπως το max pooling ή το average pooling. Τα τελικά στρώματα ενός CNN είναι συνήθως πλήρως συνδεδεμένα στρώματα (fully connected layers), τα οποία μετατρέπουν την έξοδο από τα προηγούμενα στρώματα σε έναν μονοδιάστατο διάνυσμα για την ταξινόμηση.

Τα CNNs είναι ισχυρά μοντέλα, ευρέως χρησιμοποιούμενα για εργασίες όπως η ταξινόμηση εικόνων, η ανίχνευση αντικειμένων και άλλες εργασίες οπτικής αναγνώρισης. Στο πλαίσιο της Επεξεργασίας Νοηματικής Γλώσσας, τα CNNs έχουν αποδειχθεί αποτελεσματικά στην ανίχνευση βασικών στοιχείων της νοηματικής, όπως τα σχήματα των χεριών, τα στοιχεία του προσώπου και η πόζα του σώματος. Στη μετάφραση νοηματικής γλώσσας με βάση βίντεο, τα CNN έχουν συνδυαστεί με Transformers στην βιβλιογραφία εκτεταμένα. Επιπλέον, τα CNNs παίζουν κρίσιμο ρόλο στην εκτίμηση χειρονομιών και στάσεων, αξιοποιώντας προ-εκπαιδευμένα μοντέλα όπως το OpenPose ή το MediaPipe για την ανίχνευση χαρακτηριστικών σημείων της ανθρώπινης πόζας. Τέλος, πέρα από την αναγνώριση και τη μετάφραση, τα CNNs έχουν χρησιμοποιηθεί και στην παραγωγή νοηματικής γλώσσα. Τα Παραγωγικά Νευρωνικά Δίκτυα (Generative Adversarial Networks - GANs) που βασίζονται σε αρχιτεκτονικές CNNs χρησιμοποιούνται συχνά στην έρευνα για τη δημιουργία ρεαλιστικών συνθετικών βίντεο ΝΓ.

Βιβλιογραφική Ανασκόπιση

Στη συνέχεια ακολουθεί μια ανάλυση της υπάρχουσας βιβλιογραφίες γύρω από τεχνολογίες ΝΓ και ειδικότερα γύρω από αυτές που χρησιμοποιούν τεχνικές Μηχανικής Μάθησης.

Αναγνώριση και Μετάφραση Νοηματικής Γλώσσας

Η Αναγνώριση Νοηματικής Γλώσσας ορίζεται ως η διαδικασία απόδοσης των νοημάτων ενός βίντεο ΝΓ σε χαρακτηριστικές λέξεις-νοήματα που το ερμηνεύουν σειριακά. Αυτές οι λέξεις αναφέρονται ως Sign Language Glosses στη βιβλιογραφία και απλοποιούν την απόδοση νοημάτων σε βίντεο νοηματικής γλώσσας. Από την άλλη πλευρά, η Μετάφραση Νοηματικής Γλώσσας αφορά την μετάφραση των βίντεο ΝΓ σε προτάσεις προφορικής γλώσσας, προς την κατανόηση και από ανθρώπους που δεν έχουν οικειότητα με τις νοηματικές γλώσσες.

Τις τελευταίες δεκαετίες η αναγνώριση νοηματικής γλώσσας έχει προσεγγιστεί από μια πληθώρα αρχιτεκτονικών μηχανικής μάθησης, συμπεριλαμβανομένων των Επαναλαμβανόμενων Νευρωνικών Δικτύων ([3], [8]), των LSTMs ([14]), GRUs ([35]) και των Transformers Transformers ([10], [9]), συνήθως σε συνδυασμό με τη χρήση Συνελικτικών Νευρωνικών Δικτύων (CNNs) για την εξαγωγή χωρικών χαρακτηριστικών από τα βίντεο εισόδου. Σε αυτήν την ενότητα θα παρουσιάσουμε πιο αναλυτικά κυρίως τις δημοσιεύσεις των Camgöz et al. [10] σχετικά με τη Μετάφραση Νοηματικής Γλώσσας (Sign Language Translation - SLT) με Μετασχηματιστές λόγω της καινοτομίας της, του υψηλού αριθμού παραπομπών και της σχετικότητάς της με αυτήν τη διπλωματική.

Αρχικά στο [14] προτάθηκε μια αρχιτεκτονική βαθιάς μάθησης για το πρόβλημα της αναγνώρισης νοηματικής γλώσσας, βασισμένη σε μικρά εξειδικευμένα υποδίκτυα, τα οποία ονομάζονται SubUNets. Η αρχιτεκτονική με SubUNets βασίζεται στην αποσύνθεση του πολύπλοκου προβλήματος της αναγνώρισης νοηματικής γλώσσας μέσω τριών βασικών στοιχείων: Συνελικτικά Νευρωνικά Δίκτυα (CNNs) για την εξαγωγή χωρικών χαρακτηριστικών από τα video frames εισόδου, BLSTMs για την διαχείριση των χωρικών εξαρτήσεων των δεδομένων με την πάροδο του χρόνου και Connectionist Temporal Classification (CTC Loss για την διαδικασία εκμάθησης του μοντέλου σε σύνθετο seq2seq πρόβλημα (με μεταβλητά μήκη ακολουθιών εισόδου και εξόδου).

Στο [8] τυποποιήθηκε η Μετάφραση Νοηματικής Γλώσσας (SLT) ως πρόβλημα μάθησης ακολουθίας-σε-ακολουθία (seq2seq). Αυτή η προσέγγιση χρησιμοποιεί CNNs για την εξαγωγή χωρικών χαρακτηριστικών από βίντεο νοηματικής γλώσσας, τα οποία στη συνέχεια τροφοδοτούνται σε ένα πλαίσιο κωδικοποιητή-αποκωδικοποιητή με μηχανισμό προσοχής για τη δημιουργία μεταφράσεων προφορικής γλώσσας. Τα πειράματα έγιναν σε τρεις διαφορετικές διαδικασίες: Gloss-to-Text (G2T), end-to-end Sign-to-Text (S2T) και Sign2Gloss2Text (S2G2T), η οποία χρησιμοποιεί τα gloss annotations ως ενδιάμεσο στρώμα.

Τέλος, στο [10] χρησιμοποιήθηκαν Transformers τόσο για την αναγνώριση όσο και για τη μετάφραση νοηματικής γλώσσας. Οι κωδικοποιητές επεξεργάζονται ακολουθίες βίντεο νοηματικής γλώσσας για να παράγουν embeddings που καταγράφουν τόσο χωρικά όσο και χρονικά χαρακτηριστικά, ενώ

οι αποκωδικοποιητές παράγουν προτάσεις προφορικής γλώσσας. Η απώλεια CTC χρησιμοποιείται για να διευκολύνει τη μάθηση χωρίς ρητά δεδομένα ευθυγράμμισης, συνδέοντας την αναγνώριση των glosses με τη δημιουργία κειμένου. Τα πειραματικά αποτελέσματα των προαναφερθέντων εργασιών αποδεικνύουν ότι η χρήση πληροφοριών των glosses ως ενδιάμεσο βήμα για τη μετάφραση προφορικής γλώσσας βελτιώνει την απόδοση του μοντέλου, ωστόσο η εξάρτηση από αυτά τα σχόλια (gloss annotations) μπορεί να είναι περιοριστική σε μεγαλύτερα σύνολα δεδομένων, καθώς απαιτούν ανθρώπινη επίβλεψη για την δημιουργία τους.

Ανωνυμοποίηση Βίντεο Νοηματικής Γλώσσας

Η ανωνυμοποίηση βίντεο νοηματικής γλώσσας ορίζεται ως η ηλεκτρονική κάλυψη της ταυτότητας των νοηματιστών, όπου αυτή κρίνεται απαραίτητη. Η ανωνυμοποίηση βίντεο ΝΓ βρίσκει εφαρμογή σε διάφορα πλαίσια, όπως η προστασία της ιδιωτικής ζωής των Κωφών και Βαρήκοων ατόμων σε διαδικτυακές πλατφόρμες, ακαδημαϊκές έρευνες και νομικές περιπτώσεις όπου συζητούνται ευαίσθητες πληροφορίες. Οι νοηματικές γλώσσες είναι γλώσσες οπτικές και βασίζονται σε μεγάλο βαθμό στις εκφράσεις του προσώπου, τις κινήσεις του σώματος και τα σχήματα των χεριών για τη μετάδοση του νοήματος, κάνοντας τις τυπικές τεχνικές ανωνυμοποίησης αναποτελεσματικές. Προηγούμενες έρευνες [32] διακρίνουν τους διαφορετικούς τρόπους με τους οποίους ένα βίντεο μπορεί να ανωνυμοποιηθεί σε δύο κατηγορίες: αυτές που αποκρύπτουν ολόκληρο ή μέρος του βίντεο και αυτές που παράγουν ένα συνθετικό βίντεο. Η απόκρυψη μπορεί να επιτευχθεί μέσω της θόλωσης τμημάτων της εικόνας ή με την εφαρμογή ενός φίλτρου pixelation στα χέρια και το στόμα του χρήστη κατά τη διάρκεια της σχετικής νοηματικής έκφρασης. Οι προσεγγίσεις αναπαραγωγής περιλαμβάνουν την ολοκληρωτική επανασύνθεση του βίντεο είτε με ηθοποιούς είτε με υπολογιστικά avatars.

Στην βιβλιογραφία εντοπίζουμε αρκετές ενδιαφέρουσες μεθόδους που προσεγγίζουν την ανωνυμοποίηση ΝΓ μέσω αναπαραγωγής χρησιμοποιώντας τεχνικές deep learning. Η μέθοδος Cartoonized Anonymization των Tze el al. [65] προτείνει τη χρήση μοντέλων εκτίμησης στάσης για την αυτόματη δημιουργία χαρακτήρων τύπου avatar που εκτελούν νοήματα. Η διαδικασία περιλαμβάνει την εξαγωγή σκελετικών ακολουθιών από το αρχικό βίντεο και την επανατοποθέτησή τους στον χαρακτήρα άβαταρ μέσω ενός αλγορίθμου μεταφοράς σκελετού που στοχεύει στη διατήρηση της ορθής σκελετικής δομής. Ξεφεύγοντας από τα avatars, ο Neural Sign Reenactor [64] εισάγει μια τεχνική αναπαραγωγής βίντεο βασισμένη σε GANs, ειδικά σχεδιασμένη για βίντεο νοηματικής γλώσσας. Μεταφέρει τις εκφράσεις του προσώπου, τις στάσεις του κεφαλιού και τις κινήσεις του σώματος από ένα βίντεο-πηγή σε ένα βίντεο-στόχο, διασφαλίζοντας τη διατήρηση των λεπτομερειών των χειρονομιών και βελτιώνοντας την φωτορεαλιστική αναπαράσταση σε εφαρμογές ανωνυμοποίησης νοηματικής γλώσσας. Σε μια άλλη δημοσίευση, προτείνεται το ANONYSIGN [55], μια αρχιτεκτονική που συνδυάζει Variational Autoencoders και (VAEs) και GANs για την αφαίρεση της εμφάνισης του αρχικού χρήστη και αντικατάστασης με συνθετικό νοηματιστή.

Παραγωγή Νοηματικής Γλώσσας

Η διαδικασία της Παραγωγής Νοηματικής Γλώσσας από κείμενο (Sign Language Production - SLP) που μας απασχολεί σε αυτή την διπλωματική μπορεί να οριστεί ως εξής: Δεδομένης μιας πρότασης κειμένου προφορικής γλώσσας (είσοδος), το μοντέλο μηχανικής μάθησης παράγει το αντίστοιχο βίντεο νοηματικής γλώσσας (έξοδος). Σε αυτή την ενότητα, εξετάζουμε συνοπτικά προηγούμενες έρευνες γύρω από την παραγωγή ΝΓ από κείμενο.

Οι πρώτες τεχνολογίες παραγωγής νοηματικής γλώσσας βασίζονταν κυρίως σε συστήματα αναζήτησης φράσεων και αντιστοίχισης προτάσεων, καθώς και σε υπολογιστικά παραγόμενα avatars για την παραγωγή των βίντεο νοηματικής γλώσσας. Παραδείγματα από τέτοια avatars είναι η Tessa και ο Simon, της βρετανικής νοηματικής γλώσσας. Παρόλου που τέτοιες τεχνικές έχουν την δυνατότητα να πετυχαίνουν υψηλό ρεαλισμό, συνήθως τις αποφεύγουμε επειδή η ακρίβειά τους εξαρτάται σε μεγάλο βαθμό από χρονοβόρα annotations και αποδόσεις νοημάτων, ενώ τα περιορισμένα σύνολα προ-καταγεγραμμένων δεδομένων αποτελούν την βασικότερη πρόκληση. Παράλληλα, πρέπει να αναφερθούμε και στα διάφορα συστήματα γραφής των νοηματικών γλωσσών όπως ASCII Stokoe, HamNoSys και SiGML που έχουν επίσης χρησιμοποιηθεί στην παραγωγή συστημάτων SLP, όπως στο [34]. Στη συνέχεια, στο άρθρο [54], ο Saunders εισάγει την πρώτη αρχιτεκτονική βασισμένη σε Transformers για την end-to-end παραγωγή νοηματικής γλώσσας από δεδομένο κείμενο. Η συνολική αρχιτεκτονική των Μετασχηματιστών ΝΓ φαίνεται στην εικόνα 0.0.3. Έπειτα από την πρωτοποριακή δουλεία του Saunders έχουν ακολουθήσει αρκετές δουλειές που επιχειρούν την παραγωγή ΝΓ από κείμενο μέσω των Transformers, όπως οι [56], [60], [43].

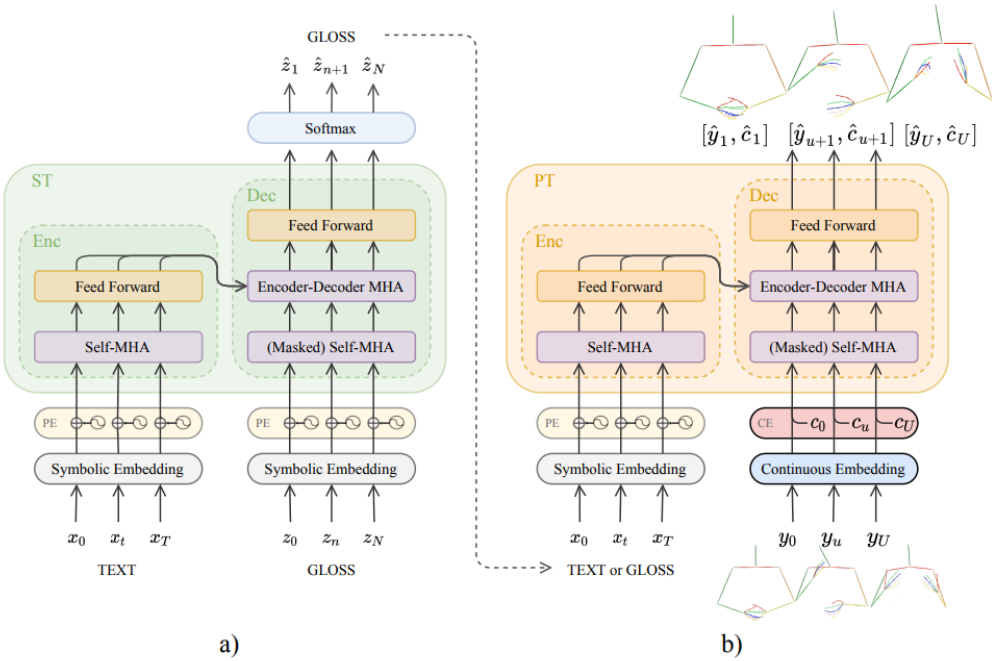


Figure 0.0.3: Αρχιτεκτονική Progressive Transformers: Πρώτη αρχιτεκτονική για SLP με μετασχηματιστές, χρησιμοποιεί ως ενδιάμεσο βήμα τα glosses.

Μετά τη δημιουργία της ακολουθίας καρέ, προηγούμενες έρευνες χρησιμοποιούν παραγωγικές αρχιτεκτονικές όπως GANs, VAEs ή Diffusion models για τη σύνθεση ρεαλιστικών βίντεο. Για παράδειγμα, στο [60], το δίκτυο pose-to-vid συνδυάζει έναν κωδικοποιητή εικόνας με GAN για τη δημιουργία ρεαλιστικών βίντεο από ανθρώπινες στάσεις. Στα δίκτυα SignDiff [18] και Neural Sign Actors [2] χρησιμοποιούνται Diffusion Models για τη δημιουργία ρεαλιστικών 3D avatar που εκτελούν νοήματα. Για την αξιολόγηση της ποιότητας της παραγωγής νοηματικής γλώσσας, χρησιμοποιούνται μετρικές όπως η BLEU και ROUGE από τον τομέα της επεξεργασίας φυσικής γλώσσας (NLP), οι οποίες μετρούν την ομοιότητα μεταξύ της παραγόμενης πρότασης και της αναφοράς. Επίσης, χρησιμοποιείται και η τεχνική Dynamic Time Wrapping (DTW) για την εύρεση της βέλτιστης ευθυγράμμισης μεταξύ της παραγόμενης ακολουθίας και της ακολουθίας αναφοράς.

Προτεινόμενη Μεθοδολογία

Στο κεφάλαιο αυτό θα περιγράψουμε την προτεινόμενη μεθοδολογία για την Παραγωγή βίντεο Νοηματικής Γλώσσας από κείμενο. Στόχος μας είναι η δημιουργία ενός μοντέλου μηχανικής μάθησης που θα δέχεται ως είσοδο μια πρόταση κειμένου και θα παράγει το αντίστοιχο βίντεο νοηματικής γλώσσας. Απ' όσο γνωρίζουμε, αυτή είναι η πρώτη δουλεία για την παραγωγή Ελληνικής Νοηματικής Γλώσσας, βασισμένη σε Βαθιά Μάθηση. Για να υλοποιήσουμε το σύστημά μας, δημιουργούμε δύο διακριτά βήματα παραγωγής ΕΝΓ. Αρχικά, χρησιμοποιούμε δίκτυα Transformer για τη δημιουργία δισδιάστατων σκελετικών ακολουθιών από το κείμενο εισόδου. Στη συνέχεια, χρησιμοποιούμε έναν Neural Renderer βασισμένο σε GANs για να μετατρέψουμε τη δημιουργημένη σκελετική ακολουθία σε ένα συνθετικό φωτορεαλιστικό βίντεο νοηματικής γλώσσας, το οποίο χρησιμοποιεί νοηματιστές από το αρχικό σύνολο δεδομένων. Στην Εικόνα 0.0.4 απεικονίζονται τα κύρια στοιχεία του προτεινόμενου δικτύου.

Αρχικά, το στάδιο προεπεξεργασίας δεδομένων περιλαμβάνει τη δημιουργία ζευγών κειμένου και ακολουθιών πόζας που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου Transformer. Χρησιμοποιούμε διαθέσιμα σύνολα δεδομένων νοηματικής γλώσσας, όπως το How2Sign (Αμερικανικά Αγγλικά), το Elementary23 (Ελληνικά) και το PHOENIX14T (Γερμανικά). Η εξαγωγή χαρακτηριστικών γίνεται χρησιμοποιώντας το MediaPipe Holistic σε κάθε σύνολο δεδομένων, και τα αρχικά 578 σημεία αναφοράς (landmarks) μειώνονται σε 191, όπως εξηγείται σε επόμενη ενότητα.

Τα δημιουργημένα ζεύγη στη συνέχεια χρησιμοποιούνται για την εκπαίδευση ενός δικτύου βαθιάς μάθησης βασισμένου στην αρχική αρχιτεκτονική Transformer. Έχουμε ενσωματώσει μια νέα απώλεια μετάφρασης από βίντεο σε κείμενο (pose-to-text SL translation loss) από τη χρήση ενός προεκπαιδευμένου μοντέλου SLT βασισμένου στο state-of-the-art μοντέλο [10]. Η χρήση των τεχνητών γλωσσικών σχολιασμών (gloss annotations) μέσω μεγάλων γλωσσικών μοντέλων (LLMs) φαίνεται γενικά να ενισχύει την απόδοση του μοντέλου και να μειώνει την λεξική πολυπλοκότητα των δεδομένων.

Έπειτα, η αρχιτεκτονική για φωτορεαλιστική σύνθεση λαμβάνει ως είσοδο τις ευθυγραμμισμένες εικόνες με χρωματικό κώδικα NMFC και δημιουργεί ένα βίντεο νοηματιστή που εκτελεί την φράση,

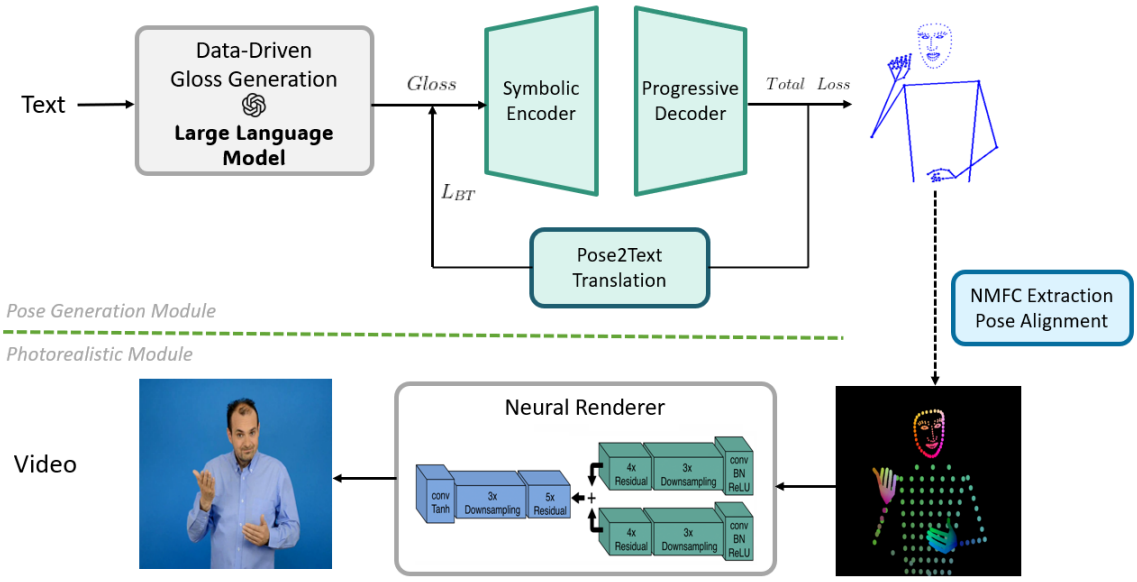


Figure 0.0.4: Προτεινόμενη διαδικασία παραγωγής βίντεο ΝΓ από κείμενο: (Πάνω) **Παραγωγή σκελετικής πόζας**: Χρησιμοποιούμε transformers για να παράξουμε μια πόζα νοηματικής γλώσσας σε ενδιάμεση αναπαράσταση με σημεία ενδιαφέροντος MediaPipe. Το δίκτυο εκπαιδεύεται από ένα άθροισμα MSE και pose-to-text απωλειών. (Κάτω) **Φωτορεαλιστική Σύνθεση**: Ακολουθώντας την ακολουθία πόζας πραγματοποιούμε neural rendering και συνθέτουμε το φωτο-ρεαλιστικό βίντεο νοηματιστή.

ανάλογα με τους χειριστές του αρχικού συνόλου δεδομένων. Ο neural renderer χρησιμοποιεί συνελικτικά νευρωνικά δίκτυα (CNNs) με επίπεδα υποδειγματοληψίας (downsampling) και residual στρώματα για τη σύνθεση ενός βίντεο υψηλής ποιότητας από τα δεδομένα πόζας.

Από το κείμενο στο βίντεο

Πρώτο βήμα της μεθόδου μας είναι οι μετατροπή των αρχικών text embeddings της εισόδου σε σκελετικές ακολουθίες νοηματικής γλώσσας. Για να υλοποιήσουμε αυτό το κομμάτι της αρχιτεκτονικής θα χτίσουμε πάνω στο open source Progressive Transformers δίκτυο [54]. Τροποποιούμε την αρχιτεκτονική ώστε να έχει την εξής λειτουργία: Αρχικά, το κείμενο κωδικοποιείται μέσω του Encoder του μετασχηματιστή. Στη συνέχεια, το κωδικοποιημένο input περνάει μέσα από τον Progressive Decoder, ο οποίος χρησιμοποιείται για την παραγωγή της συνεχούς ακολουθίας καρέ. Ο progressive decoder είναι ένα auto-regressive μοντέλο που παράγει ένα καρέ σκελετικής πόζα σε κάθε χρονικό βήμα, μαζί με μια τιμή μετρητή που δηλώνει θέση στο συνολικό βίντεο. Το Σχήμα 0.0.5 δείχνει την αρχιτεκτονική που χρησιμοποιούμε για την μετατροπή των εισόδων κειμένου σε σύνολα σκελετικών ακολουθιών. Η έξοδος του προοδευτικού αποκωδικοποιητή μπορεί να περιγραφεί από την εξίσωση:

$$[\hat{y}_{u+1}, \hat{c}_{u+1}] = D_P(\hat{j}_u | \hat{j}_{1:u-1}, r_{1:T}) \tag{0.0.2}$$

όπου $[\hat{y}_{u+1}, \hat{c}_{u+1}]$ το παραγόμενο καρέ και τιμή μετρητή καρέ στο χρονικό βήμα u+1, και $\hat{j}_u$ οι παράμετροι των προηγούμενων frames. Στη συνέχεια, το συνολικό μοντέλο εκπαιδεύεται βάσει Μέσου Τετραγωνικού Σφάλματος (MSE error) μεταξύ διαδοχικών καρέ σύμφωνα με την εξίσωση:

$$L_{MSE} = \frac{1}{U}\sum_{i=1}^{u}(y^*_{1:U} - \hat{y}_{1:U})^2 \tag{0.0.3}$$

όπου $y^*_{1:U}$ η ακολουθία ΝΓ αναφοράς και $\hat{y}_{1:U})^2$ η παραγόμενη ακολουθία.

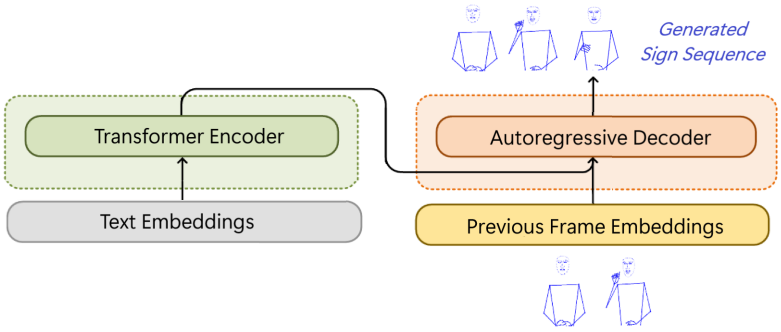


Figure 0.0.5: Transformer-Based Sign Language Production (Text-to-Video) Module

Teacher Forcing και Auto-regressive Decoding

Σε προηγούμενες μεθόδους, τα μοντέλα Transformers εκπαιδεύονταν χρησιμοποιώντας την τεχνική teacher forcing. Κατά αυτή την τεχνική παρέχουμε στο μοντέλο τα ground-truth embeddings των προηγούμενων καρέ προκειμένου να παράξει την πρόβλεψη του τρέχοντος καρέ. Χρησιμοποιώντας τα embeddings αναφοράς ως είσοδο, γίνεται εφικτή η παραλληλοποίηση της εκπαίδευσης καθώς πολλές αναπαραστάσεις είναι ήδη διαθέσιμες ταυτόχρονα από τις εξαγωγές που είχαμε κάνει στην διαμόρφωση του συνόλου δεδομένων. Παρόλο που η τεχνική teacher forcing έχει δείξει ικανοποιητικές επιδόσεις σε σύνολα δεδομένων με περιορισμένο λεξιλόγιο, όπως το PHOENIX14T, στην περίπτωσή μας δυσκολεύεται με το ευρύτερο και πιο ποικίλο σύνολο δεδομένων της Ελληνικής Νοηματικής Γλώσσας Elementary23. Γενικά, ενώ το teacher forcing παρέχει καλύτερη σταθερότητα στην εκπαίδευση και εξασφαλίζει ευθυγράμμιση μεταξύ εισόδων και εξόδων, ειδικά στα αρχικά στάδια της εκπαίδευσης, παρουσιάζει συχνά παλινδρόμηση στη μέση τιμή σφάλματος και το δίκτυο δεν είναι σε θέση να ανακάμψει από τα δικά του σφάλματα πρόβλεψης.

Αντίθετα, στην auto-regressive decoding εκπαίδευση, οι ακολουθίες των frames δημιουργούνται διαδοχικά κατά τη διάρκεια της εκπαίδευσης. Σε αυτή την προσέγγιση, το μοντέλο προβλέπει κάθε frame βασιζόμενο στα embeddings που έχει προηγουμένως δημιουργήσει, και όχι στις αναφορές. Πριν από την εφαρμογή της συνάρτησης απώλειας MSE (Mean Squared Error), ολόκληρη η ακολουθία νοημάτων δημιουργείται από τα embeddings κειμένου, μιμούμενη αποτελεσματικά τη διαδικασία του inference. Αυτό επιτρέπει στο μοντέλο να μάθει να διορθώνει τα δικά του σφάλματα αντί να βασίζεται σε πραγματικές εισόδους. Ωστόσο, αυτή η διαδικασία εκπαίδευσης είναι σημαντικά πιο χρονοβόρα σε σχέση με το teacher forcing λόγω της σειριακής δημιουργίας frames και της απουσίας παραλληλοποίησης.

Για να ισορροπήσουμε μεταξύ αποδοτικότητας και αποτελεσματικότητας, χρησιμοποιήσαμε μια υβριδική προσέγγιση, εκπαιδεύοντας το μοντέλο χρησιμοποιώντας τις τεχνικές teacher forcing και

auto-regressive decoding για ένα υποσύνολο εποχών την κάθε μια. Συγκεκριμένα, ξεκινήσαμε την εκπαίδευση με teacher forcing για να αξιοποιήσουμε τη σταθερότητα και την ταχύτητα κατά τα κρίσιμα αρχικά στάδια της εκπαίδευσης. Αυτό εξασφαλίζει ότι το μοντέλο μαθαίνει αποτελεσματικά τις βασικές εξαρτήσεις μεταξύ δεδομένων. Στη συνέχεια, μεταβαίνουμε σε auto-regressive decoding, επιτρέποντας στο μοντέλο να μάθει να διορθώνει τα δικά του σφάλματα. Αυτή η στρατηγική συνδυάζει τα πλεονεκτήματα και των δύο μεθόδων, οδηγώντας σε βελτιωμένη απόδοση σε σύγκριση με τη χρήση κάθε μεθόδου ξεχωριστά. Σε αυτή τη διπλωματική, στόχος μας είναι να πραγματοποιήσουμε πειράματα με διαφορετικές ισορροπίες και των δύο μεθόδων, προκειμένου να επιτύχουμε τα καλύτερα δυνατά αποτελέσματα στην παραγωγή νοηματικής γλώσσας.

Τεχνητή Παραγωγή Glosses με Data-driven μέθοδο

Στη συνέχεια στο πλαίσιο της έρευνάς μας, εξερευνήσαμε την δυνατότητα να χρησιμοποιήσουμε off-the-shelf μεγάλα γλωσσικά μοντέλα για να δημιουργήσουμε gloss annotations για το κειμενικό κομμάτι των συνόλων δεδομένων μας. Αυτή η τεχνική μειώνει αισθητά την λεξική πολυπλοκότητα του κειμένου, αγνοώντας μικρές λέξεις όπως άρθρα η αντωνυμίες που δεν εκφράζονταν στο βίντεο ΝΓ, διατηρώντας το συνολικό νόημα. Στον πίνακα 1 φαίνονται κάποια τέτοια παραδείγματα εξαγωγής glosses, τα οποία συγκεκριμένα έχουν προκύψει από το gpt-4o API. Σαν ενδιάμεσο βήμα, θα δείξουμε στα πειράματά μας ότι μπορεί να επιφέρει θετικά αποτελέσματα στο δίκτυο παραγωγής ΝΓ.

Table 1: Παραδείγματα Εξαγωγής Gloss από κείμενο με γλωσσικά μοντέλα

Prompt	Transform this Greek sentence into Greek Sign Language gloss: "ο άξονας συμμετρίας χωρίζει ένα σχήμα σε δύο ίσα μέρη"
Gloss	ΑΞΟΝΑΣ ΣΥΜΜΕΤΡΙΑ ΧΩΡΙΖΕΙ ΣΧΗΜΑ ΔΥΟ ΙΣΑ ΜΕΡΗ
Prompt	Transform this Greek sentence into Greek Sign Language gloss: "συμπληρώνω τον πίνακα υπολογίζοντας πρώτα τις τιμές στο περίπου ελέγχω στη συνέχεια τους υπολογισμούς μου"
Gloss	ΣΥΜΠΛΗΡΩΝΩ ΠΙΝΑΚΑΣ ΥΠΟΛΟΓΙΖΩ ΠΡΩΤΑ ΤΙΜΕΣ ΠΕΡΙΠΟΥ ΕΛΕΓΧΩ ΜΕΤΑ ΥΠΟΛΟΓΙΣΜΟΙ ΜΟΥ
Prompt	Transform this Greek sentence into Greek Sign Language gloss: "παρατηρώ και συνεχίζω τα μοτίβα "
Gloss	ΠΑΡΑΤΗΡΩ ΣΥΝΕΧΙΖΩ ΜΟΤΙΒΑ

Από το βίντεο στο κείμενο

Για την υλοποίηση αυτού του τμήματος χτίζουμε πάνω στο open source Sign Language Transformers και πραγματοποιούμε τις εξής τροποποιήσεις: Για να απλοποιήσουμε τη συνολική διαδικασία εκπαίδευσης που εκτελεί τόσο την αναγνώριση όσο και τη μετάφραση της νοηματικής γλώσσας, διατηρούμε μόνο την απώλεια που σχετίζεται με τη μετάφραση, με στόχο να επιτύχουμε τα επιθυμητά αποτελέσματα μέσω ενός άμεσου μοντέλου sign2text (από βίντεο σε κείμενο). Ο κύριος στόχος του τμήματος αυτού είναι να βελτιώσει την ακρίβεια και να αποτρέψει το μοντέλο από την παλινδρόμηση στη μέση τιμή της πόζας (mean pose), κάτι που συμβαίνει συχνά όταν η εκπαίδευση γίνεται μόνο με τη συνάρτηση απώλειας MSE (Mean Squared Error). Επιπλέον, στόχος είναι να αποδειχθεί η ικανότητα του μοντέλου να ενισχύει την ποιότητα της ευθύτερης μετάφρασης. Το Σχήμα 0.0.6 δείχνει την αρχιτεκτονική που χρησιμοποιούμε για την μετατροπή των σκελετικών ακολουθιών σε

κείμενο. Η απώλεια μετάφρασης, η οποία είναι απαραίτητη τόσο για την εκπαίδευση όσο και για την αξιολόγηση, διαμορφώνεται ως εξής:

$$L_T = 1 - \prod_{u=1}^{U} \sum_{d=1}^{D} p(\hat{w}_u^d) p(w_u^d | h_u) \tag{0.0.4}$$

όπου $p(\hat{w}_u^d)$ είναι η πιθανότητα της λέξης w^d στο βήμα αποκωδικοποίησης u και D το μέγεθος του λεξιλογίου, ενώ υπολογίζουμε το $\prod_{u=1}^{U} p(w_u^d | h_u)$ με διαδοχική εφαρμογή της απώλειας CTC σε επίπεδο frame για κάθε λέξη.

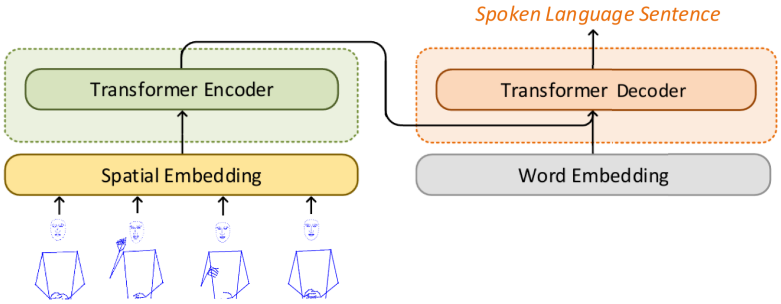


Figure 0.0.6: Transformer-Based Sign Language Translation

Συνολικά, έχουμε δημιουργήσει ένα δυαδικό σύστημα μετάφρασης που μπορεί να παράγει επιτυχώς σκελετικές ακολουθίες νοηματικής γλώσσας από κείμενο και στη συνέχεια να μεταφράζει αυτές τις παραγόμενες ακολουθίες πίσω σε μορφή κειμένου, το οποίο χρησιμοποιείται τόσο για εκπαιδευτικό όσο και για αξιολογικό σκοπό. Το Σχήμα 0.0.7 δείχνει τη λογική ροή της περιγραφόμενης διαδικασίας μετάφρασης, χρησιμοποιώντας αποκλειστικά Transformers.

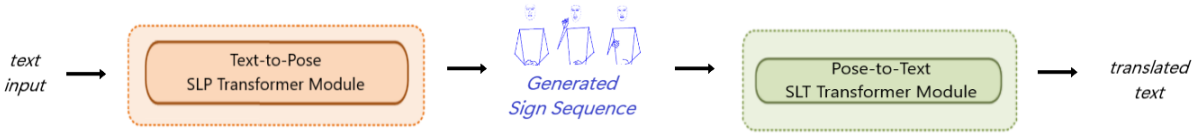


Figure 0.0.7: Αρχιτεκτονική μετάφρασης δύο κατευθύνσεων

Φωτορεαλιστική σύνθεση

Στη συνέχεια, σαν επόμενο βήμα θα θέλαμε να μετατρέψουμε τις σκελετικές ακολουθίες σε ένα ρεαλιστικό βίντεο νοηματιστή για να είναι πιο αποδεκτό από τον χρήστη. Για τον σκοπό αυτό ακολουθούμε την Head2head [37] αρχιτεκτονική για neural rendering, η οποία είχε αποδειχθεί αποτελεσματική και στην δουλειά για ανωνυμοποίηση νοηματικής γλώσσας [64]. Το σύστημα, συνοπτικά, αποτελείται από τα εξής τμήματα:

- Generator (G): Συνθέτει το t-οστο frame χρησιμοποιώντας το τωρινό input, καθώς και τα inputs και outputs των δυο προηγούμενων χρονικών βημάτων.

- Image Discriminator (D_I): Διακρίνει μεταξύ αληθινών και συνθετικών καρέ.
- Sign Features Discriminators: Ξεχωριστοί Discriminators για τα μάτια, στόμα και χέρια, οι είσοδος πραγματοποιείται με περικοπή γύρω από τα σημεία ενδιαφέροντος.
- Dynamic Discriminator (D_D): Χρησιμοποιείται για να διακρίνει τις μη φυσικές πτικές μεταβάσεις ανά τα καρέ.

Η συνολική συνάρτηση απώλειας για να εκπαιδευτεί το δίκτυο ώστε να ενθαρρύνει την διάκριση μεταξύ αληθινής- ψεύτικης εξόδου και να παράγει ικανοποιητικό αποτέλεσμα, είναι η εξής:

$$\mathcal{L}_G = \mathcal{L}_G^{\text{adv}} + \lambda_{\text{vgg}}\mathcal{L}_G^{\text{vgg}} + \lambda_{\text{feat}}\mathcal{L}_G^{\text{feat}} + \lambda_{\text{face}}\mathcal{L}_G^{\text{face}} \tag{0.0.5}$$

Πειράματα

Στο κεφάλαιο αυτό θα μπούμε σε μεγαλύτερη λεπτομέρεια για τα διάφορα τεχνικά χαρακτηριστικά (εκπαίδευση δικτύων, σύνολα δεδομένων) της μεθόδου ενώ επίσης θα παρουσιάσουμε αναλυτικά τα αποτελέσματά μας.

Σύνολα Δεδομένων

Παρόλο που υπάρχουν κάποια large-scale σύνολα δεδομένων νοηματικής γλώσσας διαθέσιμα για μετάφραση (Open-ASL [57], Youtube-ASL [66]), μόνο λιγοστά από αυτά έχουν χρησιμοποιηθεί και για τη συνεχή παραγωγή νοηματικής γλώσσας από κείμενο. Μετά από προσεκτική εξέταση των επιλογών μας για δημοσίως διαθέσιμα σύνολα ΝΓ, καταλήξαμε στις ακόλουθες επιλογές:

- Ελληνική Νοηματική Γλώσσα: Σχετικά με την ΕΝΓ, το Elementary23 [68] περιλαμβάνει ζεύγη μεταφράσεων βασισμένα στο επίσημο πρόγραμμα σπουδών του Ελληνικού Δημοτικού Σχολείου και έχει χρησιμοποιηθεί για τη μετάφραση νοηματικής γλώσσας με μετασχηματιστές (transformer-based SLT). Το περιεχόμενό του διαχωρίζεται με βάση το αντικείμενο του προγράμματος σπουδών (π.χ. Μαθηματικά, Ελληνική Γλώσσα), γεγονός που μας επιτρέπει να πραγματοποιήσουμε εκπαίδευση παραγωγής νοηματικής γλώσσας ειδικά για κάθε αντικείμενο. Δεν έχει χρησιμοποιηθεί ακόμη στην εργασία Παραγωγής Νοηματικής Γλώσσας και είναι ένα από τα λίγα large-scale σύνολα στην Ελληνική Γλώσσα.
- Αμερικανική Νοηματική Γλώσσα: Το How2Sign [16] είναι ένα multimodal σύνολο δεδομένων συνεχούς Αμερικανικής Νοηματικής Γλώσσας (ASL). Αποτελείται από ένα παράλληλο corpus 80 ωρών βίντεο νοηματικής γλώσσας που περιλαμβάνουν ομιλία, αγγλικές μεταγραφές και βάθος. Το περιεχόμενό του αποτελείται κυρίως από εκπαιδευτικά βίντεο και tutorials.
- Γερμανική Νοηματική Γλώσσα: Το PHOENIX14T [36] είναι ένα σύνολο δεδομένων της Γερμανικής Νοηματικής Γλώσσας (GSL) που περιλαμβάνει εκπομπές πρόγνωσης καιρού από το 2009 έως το 2011, συνοδευμένες από gloss annotations. Είναι το σύνολο δεδομένων που χρησιμοποιήθηκε στις προηγούμενες εργασίες των Saunders και Stoll ([54], [56], [60], [61]), οι οποίες

ήταν οι πρώτες που αντιμετώπισαν την μετάφραση και παραγωγή ΝΓ χρησιμοποιώντας νευρωνικά δίκτυα για neural machine translation.

Dataset	Language	Year	Video	Text	Gloss
PHOENIX14T [36]	German SL	2014	✓	✓	✓
How2Sign [16]	ASL	2021	✓	✓	✗
Elementary23 [68]	Greek SL	23k	✓	✓	✗

Table 2: Σύνολα Δεδομένων για νοηματικές γλώσσες

Εξαγωγή Χαρακτηριστικών

Χρησιμοποιούμε το MediaPipe Holistic για την εξαγωγή σκελετικών σημείων αναφοράς σε κάθε βίντεο ΝΓ από τα σύνολα δεδομένων που προαναφέρθηκαν. Το MediaPipe Holistic χρησιμοποιεί μια διαδικασία επεξεργασίας διαφορετικών περιοχών ενδιαφέροντος (ROIs) μέσα σε μια εικόνα για να υπολογίσει συνολικά έως και 543 σημεία αναφοράς. Αυτά περιλαμβάνουν 33 σημεία αναφοράς για τη στάση του σώματος, 468 σημεία αναφοράς για το πρόσωπο και 42 σημεία αναφοράς για τα χέρια (21 ανά χέρι). Το MediaPipe Holistic λειτουργεί μόνο με την CPU, απαιτώντας περίπου 3 δευτερόλεπτα για την επεξεργασία κάθε καρέ.

Για να επιταχύνουμε τη διαδικασία εκπαίδευσης, υποδειγματοληπτούμε τόσο τα σημεία αναφοράς της στάσης του σώματος όσο και του προσώπου. Για τα σημεία αναφοράς της πόζας, επιλέγουμε τα 8 χαρακτηριστικά σημεία που περιλαμβάνουν τα μέρη του σώματος που είναι απαραίτητα για ένα βίντεο SL, όπως ο κορμός, οι αγκώνες και οι καρποί. Για τα σημεία αναφοράς του προσώπου, επιλέγουμε 141 αντί για 468 σημεία αναφοράς, τα οποία περιέχουν όλες τις απαραίτητες πληροφορίες για το πρόσωπο, όπως το στόμα, τα μάτια, η μύτη και η περίμετρος του προσώπου. Για κάθε χέρι, διατηρούμε και τα 21 σημεία αναφοράς. Αυτό μας φέρνει σε ένα σύνολο 191 σημείων αναφοράς, αντί για τα αρχικά 543 σημεία αναφοράς του MP, που αποτελεί σημαντική μείωση. Τέλος, η συνολική ακολουθία σημείων αναφοράς που εξάγεται για κάθε καρέ ορίζεται ως εξής:

$$\mathbf{P}_f = [\mathbf{a}_{left\ hand}||\mathbf{a}_{right\ hand}||\mathbf{a}_{face}||\mathbf{a}_{pose}||c_f] \tag{0.0.6}$$

όπου το P_f είναι η ακολουθία σημείων αναφοράς για το f-οστό καρέ, το c_f είναι η τιμή του μετρητή που κυμαίνεται από 0 έως 1 και υποδεικνύει τη σχετική θέση του καρέ, και το || είναι το σύμβολο της συνένωσης.

Διαχείρηση Πόζας

Μετά το αρχικό βήμα της εκτίμησης στάσης και τη διαμόρφωση του συνόλου δεδομένων για τα μοντέλα Text-to-Pose, πρέπει επίσης να προετοιμάσουμε τα δεδομένα για τη διαδικασία Neural Rendering. Δεδομένου ότι κάθε σύνολο δεδομένων περιέχει δύο ή περισσότερους διαφορετικούς χειριστές, όταν πραγματοποιούμε μεταφορά κίνησης από τον έναν στον άλλον, πρέπει να λάβουμε υπόψη πιθανές διαφορές στη σωματική τους ανατομία ή διαφορές στη θέση της κάμερας σε σχέση με το σώμα στα αρχικά βίντεο.

Για το λόγο αυτό, προσαρμόζουμε μεθόδους μεταφοράς πόζας για τα σημεία αναφοράς του προσώπου και του σώματος ξεχωριστά, ακολουθώντας τις μεθόδους Ανάλυσης Procrustes που χρησιμοποιούνται στο Head2head για το πρόσωπο και στο Neural Sign Reenactor για τα χέρια και το σώμα. Για την μεταφορά του προσώπου, εστιάζουμε σε σταθερές περιοχές του προσώπου που επηρεάζονται λιγότερο από παραμορφώσεις λόγω εκφράσεων. Χρησιμοποιώντας ένα υποσύνολο των σημείων, ευθυγραμμίζουμε το πρόσωπο της πηγής και του στόχου σε ένα μέσο πρότυπο προσώπου μέσω ανάλυσης Procrustes. Για την μεταφορά των σημείων του σώματος, συμπεριλαμβανομένου του κορμού και των χεριών, υιοθετούμε επίσης μια παρόμοια προσέγγιση βασισμένη στην Ανάλυση Procrustes.

Αφού πραγματοποιήσουμε την αναδιάταξη όπου αυτό είναι απαραίτητο, δημιουργούμε τα αντίστοιχα καρέ με χρωματικό κώδικα, τα οποία χρησιμοποιούνται για να ρυθμίσουν τον neural rendering. Αυτά είναι εικόνες RGB, όπου κάθε σημείο αναφοράς απεικονίζεται ως ένας δίσκος σταθερής ακτίνας με ένα μοναδικό χρώμα που έχει εκχωρηθεί μέσω ενός προκαθορισμένου σχήματος χρωματικού κωδικοποίησης. Κάθε άρθρωση διατηρεί ένα σταθερό χρώμα σε όλους τους χειριστές, διασφαλίζοντας μια συνεπή και σημασιολογική αναπαράσταση στο χώρο RGB, η οποία βοηθά τον renderer να μάθει την αντιστοίχιση με τις εικόνες εξόδου.

Αξιολόγηση Αποτελεσμάτων

Όπως αναφέρθηκε, στόχος μας ήταν να αξιολογήσουμε την προτεινόμενη διαδικασία όσο το δυνατόν πιο εκτενώς, πραγματοποιώντας πειράματα και στα τρία σύνολα δεδομένων που χρησιμοποιήσαμε. Συγκρίνουμε τη μέθοδό μας με τουλάχιστον τρία προηγούμενα benchmarks που υπάρχουν για την παραγωγή ΝΓ από κείμενο, εφόσον αυτά είναι διαθέσιμα. Τα benchmarks που παρουσιάζονται επιλέχθηκαν με βάση τη σχετικότητα και την ομοιότητά τους με τη μέθοδό μας. Είναι σημαντικό να διευκρινιστεί σε ποια εργασία (Μετάφραση ή Παραγωγή) αναφέρεται κάθε benchmark, καθώς διαφέρουν σε δυσκολία υλοποίησης.

Ξεκινάμε την διαδικασία αξιολόγησής μας εκπαιδεύοντας σε ολόκληρα τα σύνολα δεδομένων, πρώτα στην Μετάφραση και στη συνέχεια στην Παραγωγή ΝΓ. Είναι σημαντικό να σημειωθεί ότι τα μοντέλα που προκύπτουν σε αυτή την ενότητα προέρχονται από μοντέλα που εκπαιδεύτηκαν αποκλειστικά σε ακολουθίες σκελετικών δεδομένων βασισμένων σε ground-truth. Αυτό συνήθως μπορεί να αυξάνει την απόδοση των μοντέλων και ταυτόχρονα ακολουθεί με μεγαλύτερη ακρίβεια τα benchmarks που παρουσιάζονται στην βιβλιογραφία. Ο Πίνακας 3 δείχνει τα αποτελέσματά μας για το SLT στα τρία σύνολα δεδομένων: How2Sign (ASL), PHOENIX14T (Γερμανική Νοηματική) και Elementary23 (Ελληνική Νοηματική) ενώ αντίστοιχα ο Πίνακας 4 δείχνει τα αποτελέσματα SLP στα ίδια σύνολα.

Η επόμενη μελέτη απομόνωσης (ablation study), που παρουσιάζεται στο Elementary23 Greek Language Subset 5 και στο Math Subset 6, δείχνει ότι η συμπερίληψη της απώλειας pose-to-text και των σχολιασμών Gloss επηρεάζει επιδραστικά την απόδοση. Αν και τα σκορ BLEU-4 βελτιώνονται ανεξάρτητα (4.42 dev, 4.55 test), ο συνδυασμός με τα Gloss δίνει ανάμεικτα αποτελέσματα, μειώνοντας ελαφρά το BLEU-4 στο dev set (4.06) αλλά βελτιώνοντας το στο test set (4.32). Αυτή η αλληλεπίδραση υποδηλώνει ότι ενώ τα glosses απλοποιούν τη γλωσσική πολυπλοκότητα, ενώ η υπερβολική εξάρτηση

Method	Dev		Test	
	BLEU-4↑	**ROUGE↑**	**BLEU-4↑**	**ROUGE↑**
How2Sign (SLT) [52]	-	-	14.9	36.0
ours	9.01	22.34	8.53	25.22
Progressive Transformers [54] (SLT)	20.23	55.41	19.10	54.55
Sign Language Transformers [10] (SLT)	22.38	-	21.32	-
ours	21.53	-	21.22	-
Elementary23 SLT (Voskou et al. [68])	6.67	-	5.69	-
Elementary23 SLT (ours)	8.34	32.36	**8.2**	32.16

Table 3: SLT Evaluation Banchmarks

Method	Dev		Test	
	BLEU-4↑	**ROUGE↑**	**BLEU-4↑**	**ROUGE↑**
Neural Sign Actors [2]	13.12	47.55	13.12	47.55
SignDiff [18]	16.92	49.74	15.92	46.57
MS2SL [42]	4.26	16.38	4.26	16.38
ours	4.5	12.36	**4.48**	12.16
Progressive Transformers [54]	11.82	33.18	10.51	32.46
There and Back again [61]	17.10	40.42	16.91	40.22
ours	12.34	33.98	12.51	34.21
Elementary23 SLT [68]	6.67	-	5.69	-
Elementary23 Math (ours)	7.58	15.11	**7.69**	15.26
Elementary23 Greek Lang (ours)	5.63	14.56	**5.52**	14.23

Table 4: SLP Evaluation Banchmarks

από αυτά μπορεί να περιορίσει την προσαρμοστικότητα. Στο Elementary23 Math Subset και πάλι, τα αποτελέσματα δείχνουν ότι το Video-to-Text loss βοηθάει στην αύξηση των επιδόσεων BLUE4 ενώ η ελαφρά μείωση με την χρήση των glosses υποδηλώνει ότι μπορεί να υπάρχει κάποια παραποίηση των κειμένων μέσω της επεξεργασίας από γλωσσικά μοντέλα.

$L_{Video2Text}$	*Gloss*	Dev **BLEU-4↑**	Test **BLEU-4↑**
✗	✗	4.17	4.15
✗	✓	3.56	3.44
✓	✗	4.42	**4.55**
✓	✓	4.06	**4.32**

Table 5: Ablation Study on the Elementary23 Greek Language SL Dataset

$L_{Video2Text}$	*Gloss*	Dev **BLEU-4↑**	Test **BLEU-4↑**
✗	✗	3.17	3.15
✗	✓	4.36	4.44
✓	✗	5.42	**5.55**
✓	✓	5.12	**5.06**

Table 6: Ablation Study on the Elementary23 Math SL Dataset

Στη συνέχεια, εστιάζουμε στη διεξαγωγή πειραμάτων στη σύγκριση των μεθόδων εκπαίδευσης των transformers στο σύνολο Elementary23. Επιλέγουμε συγκεκριμένα ολόκληρο το υποσύνολο Math και το υποσύνολο Ελληνικής Γλώσσας. Η μελέτη, που παρουσιάζεται στον πίνακα 7, συγκρίνει την εκπαίδευση με Teacher Forcing (TF), Auto-regressive Decoding (AD) και τον συνδυασμό τους (TF+AD), υπογραμμίζοντας τα οφέλη της χρήσης μιας υβριδικής προσέγγισης. Ενώ με το Auto-regressive Decoding επιτυγχάνονται σημαντικά υψηλότερα σκορ BLEU-4 και ROUGE (5.4 και 14.5 στο dev set, αντίστοιχα) σε σύγκριση με το Teacher Forcing (1.69 και 8.52), το υβριδικό μοντέλο TF+AD παρέχει ισορροπία μεταξύ υπολογιστικής αποδοτικότητας και ακρίβειας. Το υβριδικό μοντέλο επιτυγχάνει

την υψηλότερη συνολική απόδοση, τόσο στο υποσύνολο Ελληνικής Γλώσσας όσο και στο υποσύνολο Μαθηματικών, επικυρώνοντας την αποδοτικότητα της λύσης μας.

Subset	Method	Epochs	Time/ Epoch (s)	**Dev BLEU-4↑**	**Test BLEU-4↑**
Greek	Teacher Forcing, (PT [54])	2500	5	0.49	0.35
	Autoregressive Dec	2500	30	4.3	4.13
	TF + AD	1250 + 1250	5, 30	**4.67**	**4.46**
Math	Teacher Forcing, (PT [54])	2500	5	1.69	1.46
	Autoregressive Dec	2500	30	5.4	5.3
	TF + AD	1250 + 1250	5, 30	**5.69**	**5.59**

Table 7: Σύγκριση μεταξύ Teacher Forcing, Auto-regressive decoding και υβριδικής προσέγγισης

Ακόμη ένας τρόπος να αξιολογήσουμε την οπτική ποιότητα των αποτελεσμάτων είναι η μετρική DTW (Dynamic Time Wrapping) για εύρεση της μαθηματικά βέλτιστης ευθυγράμμισης μεταξύ της παραγόμενης σκελετικής ακολουθίας και της ακολουθίας αναφοράς. Στην διάγραμμα 0.0.8 συγκρίνουμε την τιμή της DTW στα τελευταία βήματα της εκπαίδευσης για τις μεθόδους TF, AD και TF+AD. Βλέπουμε πως αρχικά, με χρήση μόνο Teacher Forcing η τιμή κυμαίνεται στο φάσμα 15-25, ενώ με τον υβριδικό αλγόριθμο φτάνει σε ικανοποιητικές τιμές ευθυγράμμισης, δηλαδή κάτω από 10.

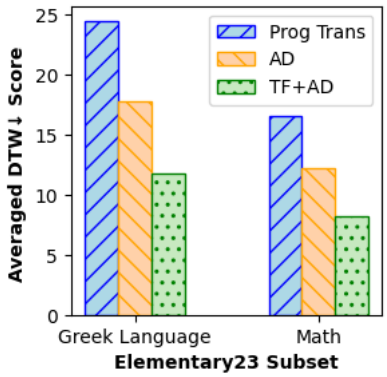


Figure 0.0.8: Σύγκριση DTW για τις μεθόδους TF, AD και Hybrid TF,AD στα τελευταία στάδια της εκπαίδευσης

Αξιολόγηση Χρηστών

Οι γλωσσικές μετρικές BLUE και ROUGE καθώς και η μέθοδος αντιστοίχισης DTW επιβεβαιώνουν την ορθή λειτουργία του συστήματος μόνο υπολογιστικά. Για την ουσιαστική αξιολόγηση ενός συστήματος επικοινωνίας ΝΓ χρειάζεται και αξιολόγηση από γνώστες και ειδικούς ΕΝΓ. Για το λόγο αυτό, δημιουργήσαμε ένα ερωτηματολόγιο χρηστών (web-based) το οποίο απαντήθηκε από 8 ειδικούς της ελληνικής νοηματικής γλώσσας, μέσω της συνεργασίας μας με το Ερευνητικό Κέντρο Αθηνά.

Στο πρώτο μέρος του ερωτηματολογίου, επιλέξαμε προσεκτικά 14 προτάσεις (και τα αντίστοιχα βίντεο νοηματικής γλώσσας) από το σύνολο δεδομένων Math Elementary23, συμπεριλαμβανομένων προτάσεων τόσο από το test όσο και από το dev set. Για κάθε βίντεο, ζητήσαμε από τους συμμετέχοντες να επιλέξουν την πιο ταιριαστή πρόταση από τρεις επιλογές, με επιπλέον δυνατότητα

απάντησης "κανένα από τα παραπάνω" εάν καμία δεν ανταποκρινόταν. Ακολούθησαν δύο ερωτήσεις σχετικές με τη σαφήνεια και την ακρίβεια της νοηματικής έκφρασης: "Πόσο εύκολο ήταν να κατανοήσετε το νόημα του βίντεο;" και "Πόσο καλά πιστεύετε ότι υπογραμμίστηκε το νόημα;" Στη συνέχεια, οι συμμετέχοντες αξιολόγησαν σε κλίμακα 1–5 την οπτική ποιότητα των βίντεο (πρόσωπο, χέρια, συνολική ποιότητα). Αυτό το πλαίσιο αξιολόγησης συνδύασε γλωσσικά και οπτικά κριτήρια για ολοκληρωμένη ανάλυση.

Ο Πίνακας 8 δείχνει ότι η προτεινόμενη μέθοδος ξεπέρασε την μέθοδο PT κατά 35 μονάδες, επιδεικνύοντας μεγαλύτερη ακρίβεια στην παραγωγή νοηματικής γλώσσας. Ωστόσο, το υπολειπόμενο 26% υπογραμμίζει πιθανές ασάφειες στις χειρονομίες ή περιορισμούς στα δεδομένα εκπαίδευσης. Το Σχήμα 0.0.9 παρουσιάζει αναλυτικά τις απαντήσεις σχετικές με την οπτική ποιότητα αλλά και την κατανόηση. Οι χρήστες βαθμολόγησαν και τις δύο μεθόδους σχετικά χαμηλά (1–3), πιθανώς λόγω ασαφών χειρονομιών ή χαμηλής ανάλυσης, στα οποία και θα εστιάσουμε σε μελλοντικές επεκτάσεις.

Table 8: Συγκριτικά Αποτελέσματα ερωτήσεων κατανόησης νοηματισμού

Method	Answered Choices	Accuracy
Proposed Pipeline	40/56	**71.42%**
PT Baseline SLP Pipeline	20/56	**35.71%**

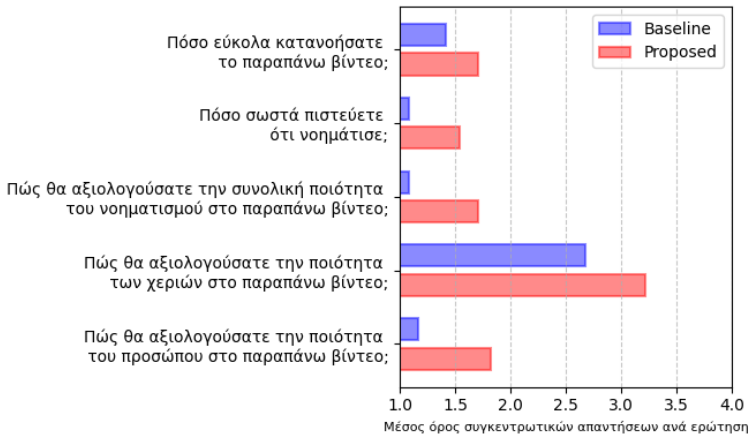


Figure 0.0.9: User study results: Picture Quality and Signer- related questions

Στο δεύτερο μέρος, επιλέξαμε 6 προτάσεις από το ίδιο σύνολο δεδομένων και ζητήσαμε από τους συμμετέχοντες να συγκρίνουν βίντεο από την προτεινόμενη μέθοδο και τους Progressive Transformers ως baseline. Η προτεινόμενη μέθοδος προτιμήθηκε από 91.66% των συμμετεχόντων, επιδεικνύοντας ανώτερο ρεαλισμό και φυσικότητα (Πίνακας 9).

Table 9: Συγκριτικά Αποτελέσματα ερωτήσεων ρεαλισμού

Method	Answered Choices	Accuracy
Proposed Pipeline	44/48	**91.66%**
PT Baseline SLP Pipeline	4/48	**8.33%**

Οπτικά Αποτελέσματα

Τέλος, αυτή η ενότητα περιλαμβάνει απεικονίσεις που καταγράφουν κάθε βήμα της διαδικασίας παραγωγής ΝΓ με επεξηγηματικό τρόπο. Το Σχήμα 0.0.10 δείχνει ένα παράδειγμα της συνθετικής παραγωγής βίντεο, μαζί με τις απεικονίσεις της ενδίαμεσης σκελετικής ακολουθίας καθώς και των καρέ με χρωματική κωδικοποίηση. Τέλος, στον πίνακα 10 παρέχουμε παραδείγματα της ενότητας video-to-text που εκτελεί την Μετάφραση Νοηματικής Γλώσσας. Συγκρίνουμε μεταφρασμένες ακολουθίες από τα ground-truth landmarks και τα landmarks της μεθόδου μας.

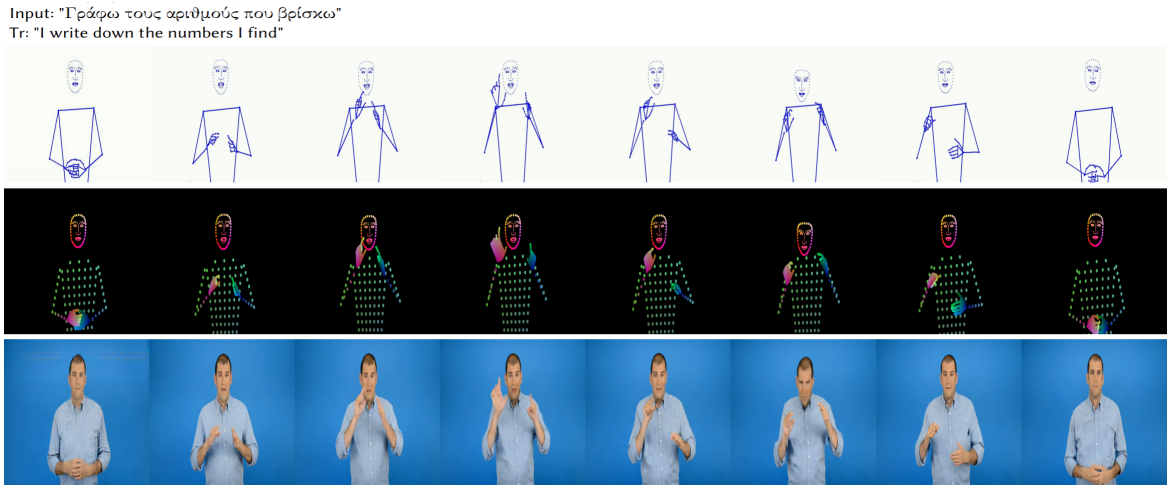


Figure 0.0.10: Πλήρες οπτικό αποτέλεσμα από την προτεινόμενη μέθοδο παραγωγής ΝΓ. (Πάνω) Η παραγόμενη πόζα από τον transformer, (μέση) Η επεξεργασμένη εικόνα για είσοδο στον renderer και (κάτω) το συνθετικό βίντεο νοηματιστή.

1	**Ref:** οι αριθμοί από το 6 μέχρι το 10 **Prod w/ GT:** οι αριθμοί από το 6 μέχρι το 10 **Prod w/ SLP:** οι αριθμοί από το 6 μέχρι το 10
2	**Ref:** φτιάχνω γεωμετρικά σχήματα **Prod w/ GT:** γεωμετρικά σχήματα **Prod w/ SLP:** φτιάχνω γεωμετρικά σχήματα
3	**Ref:** αν σε κάθε φύλλο του άλμπουμ βάλει 8 αυτοκόλλητα , πόσα φύλλα θα χρησιμοποιήσει; **Prod w/ GT:** αν σε κάθε φύλλο του άλμπουμ βάλει 10 αυτοκόλλητα, πόσα φύλλα θα χρησιμοποιήσει; **Prod w/ SLP:** αν σε κάθε φύλλο του άλμπουμ βάλει 10 αυτοκόλλητα, πόσα φύλλα θα χρησιμοποιήσει;
4	**Ref:** κάνω τις διαιρέσεις και γράφω το αποτέλεσμα **Prod w/ GT:** κάνω τις παρακάτω πράξεις **Prod w/ SLP:** κάνω τις προσθέσεις και γράφω το αποτέλεσμα
5	**Ref:** συνδέω τα σχήματα με το όνομα τους **Prod w/ GT:** διηγούμαι ένα πρόβλημα **Prod w/ SLP:** συνδέω με μια γραμμή τα κομμάτια

Table 10: Παραδείγματα από την μετάφραση ΝΓ. με Ref αναφερόμαστε στο κείμενο-αναφορά από το σύνολο δεδομένων, με Prod w/ GT στο κείμενο που έχει μεταφραστεί από τους ground-truth σκελετούς και με Prod w/ SLP στο κείμενο που έχει μεταφραστεί από τους σκελετούς που έχουν παραχθεί με την μέθοδό μας. Από πάνω προς τα κάτω υπογραμμίζουμε, με πράσινο τις επιτυχημένες μεταφράσεις, με πορτοκαλί τις μεταφράσεις που διατηρούν σε ικανοποιητικό βαθμο το νόημα και με κόκκινο τις λάθος μεταφράσεις.

Συμπέρασμα

Περίληψη Εργασίας

Σε αυτή τη διπλωματική εξερευνήσαμε τον τομέα στη διασταύρωση Όρασης Υπολογιστών και Νοηματικών Γλωσσών, δίνοντας έμφαση στην παραγωγή Νοηματικής Γλώσσας από κείμενο. Παρουσιάσαμε μια εκτεταμένη βιβλιογραφική ανάλυση γύρω από τις υπάρχουν τεχνολογίες Μηχανικής Μάθησης για νοηματικές γλώσσες. Βασίσαμε τα πειράματά μας σε μια σύγχρονη αρχιτεκτονική Μετασχηματιστή που αρχικά μετατρέπει δοσμένες προτάσεις κειμένου στις αντίστοιχες σκελετικές αλληλουχίες νοημάτων σε βίντεο. Σύμφωνα με την υπάρχουσα βιβλιογραφία, το σύστημά μας είναι η πρώτη εφαρμογή παραγωγής Ελληνικής Νοηματικής Γλώσσας, βασισμένη σε τεχνολογία Βαθιάς Μάθησης. Η αρχιτεκτονική μας εισάγει ποικίλα καινοτόμα στοιχεία στη διαδικασία παραγωγής ΝΓ, όπως η προσθήκη αντίστροφης απώλειας βίντεο-προς-κείμενο, η τεχνητή παραγωγή γλωσσικών νοημάτων με μεγάλα γλωσσικά μοντέλα και ο υβριδικός αλγόριθμος εκπαίδευσης εναλλαγής teacher-forcing και auto-regressive decoding. Έπειτα, εξερευνούμε την φωτορεαλιστική απόδοση του προβλήματος παραγωγής βίντεο, μετατρέποντας τις σκελετικές ακολουθίες του προηγούμενου βήματος σε συνθετικά βίντεο νοηματιστή. Για την πραγματοποίηση της νευρωνικής απόδοσης χρησιμοποιούμε παραγωγικά δίκτυα GANs τα οποία εκπαιδεύονται ώστε να μιμούνται τα οπτικά χαρακτηριστικά κάποιου εκ των αρχικών νοηματιστών από το σύνολο δεδομένων. Τέλος, αξιολογήσαμε την προτεινόμενη μεθοδολογία μέσω εκτενούς σειράς πειραμάτων και αξιολόγησης χρηστών, πετυχαίνοντας ανταγωνιστικά αποτελέσματα.

Υπολογιστικοί Περιορισμοί

Η δουλειά μας, παρά την ανταγωνιστική επίδοση σε πολλά πειράματα, συναντά και ορισμένους περιορισμούς. Αρχικά επειδή η διαδικασία εκπαίδευσης είναι χρονικά και υπολογιστικά ακριβή, τα μοντέλα μας συνήθως υπολείπονται σε άγνωστες ή μεγάλες σε μήκος προτάσεις κειμένου. Επίσης, η ξεχωριστή εκπαίδευση των μοντέλων transformer και του neural renderer βάζει ακόμη μία χρονική απαίτηση στο πρόβλημά μας. Τέλος, η υπάρχουσα προσέγγιση δεν εκμεταλλεύεται την τρίτη συνιστώσα που παρέχει η MediaPipe (άξονας z), η οποία συνήθως περιέχει σημαντική πληροφορία για το βάθος της εικόνας και την απόσταση από την κάμερα.

Μελλοντικές Επεκτάσεις

Έχοντας αναφέρει τους περιορισμούς, στρέφουμε την προσοχή μας σε πιθανούς τρόπους μελλοντικής εξέλιξης της δουλειάς μας:

- **Αρχιτεκτονικές VQ-GAN, VQ-VAE:** Πρόσφατες δημοσιεύσεις υποστηρίζουν ότι η ενσωμάτωση δικτύων variational autoencoders στις παραγωγικές αρχιτεκτονικές μπορεί να ενισχύσει σημαντικά την απόδοση των συστημάτων ΝΓ. Με την αντιστοίχιση των κωδικοποιημένων ακολουθιών νοηματικής γλώσσας με τον πλησιέστερο γείτονα από έναν εκπαιδεύσιμο πίνακα κωδικών (codebook), θα μπορούσαμε να εξελίξουμε αισθητά το υπάρχον δίκτυο. Η

ενσωμάτωση τέτοιων προσεγγίσεων στην αρχιτεκτονική Encoder-Decoder σε μελλοντικές υλοποιήσεις θα μπορούσε να ενισχύσει τον ρεαλισμό των παραγόμενων νοημάτων, μειώνοντας ταυτόχρονα τις υπολογιστικές απαιτήσεις του auto-regressive decoding κατά την εκπαίδευση.

- **3D Ανακατασκευή:** Σε αυτή την εργασία χρησιμοποιήθηκαν 2D αναπαραστάσεις για την κωδικοποίηση βίντεο ΝΓ, καθώς τόσο οι Transformers όσο και ο Neural Renderer λειτουργούν αποτελεσματικά σε αυτό το πλαίσιο. Ωστόσο, υπάρχει ακόμη η δυνατότητα εξαγωγής 3D αναπαραστάσεων με την MediaPipe. Οι 3D αναπαραστάσεις περιέχουν χρήσιμες πληροφορίες σχετικά με το βάθος της εικόνας και τη θέση της κάμερας, ειδικά κατά τη διαδικασία neural rendering. Από την άλλη πλευρά, πρόσφατες δημοσιεύσεις εξετάζουν τη δημιουργία μοντέλων νοηματικής γλώσσας που χρησιμοποιούν ρεαλιστικά 3D σχήματα, συχνά βασισμένα σε πιο σύνθετα πλαίσια όπως το SMPL-X.

- **Σύνολα Δεδομένων:** Ένας αξιοσημείωτος περιορισμός των υφιστάμενων συνόλων δεδομένων νοηματικής γλώσσας είναι οι περιορισμένες λεξιλογικές τους δυνατότητες. Για παράδειγμα, το Elementary23 εστιάζει σε σχολικά βιβλία δημοτικού, το How2Sign σε εκπαιδευτικά βίντεο και tutorials, και το PHOENIX14T σε πρόγνωση καιρού. Ως αποτέλεσμα, τα περισσότερα μοντέλα SLP βασίζονται στην αναπαραγωγή προτάσεων από το ήδη περιορισμένο λεξιλόγιο εκπαίδευσης της νοηματικής γλώσσας, και επομένως μπορεί να μην καλύπτουν τις ανάγκες ενός πλήρους εκπαιδευτικού συστήματος νοηματικής γλώσσας. Η πιθανή ανάπτυξη μιας μεγαλύτερης και ενοποιημένης βάσης δεδομένων νοηματικής γλώσσας θα ωφελούσε σημαντικά την κοινότητα των κωφών, ειδικά στην ψηφιακή εποχή.

- **Καθολική εκπαίδευση:** Σε αυτή τη διπλωματική εργασία, η παραγωγή ΝΓ αντιμετωπίζεται σε δύο διακριτά βήματα: Πρώτον, η δημιουργία σκελετικών ακολουθιών από κείμενο και στη συνέχεια η απόδοση των σκελετών σε φωτορεαλιστική μορφή. Επιπλέον, μέσα στη μονάδα Text-to-Pose, εκπαιδεύονται ξεχωριστοί Μετασχηματιστές για την εμπρόσθια και την αντίστροφη μετάφραση. Η μετάβαση σε μια πιο ολιστική προσέγγιση που δεν απαιτεί την εκπαίδευση διαφορετικών μοντέλων θα ήταν υπολογιστικά και σημασιολογικά ωφέλιμη.

Εφαρμογές και Κοινωνικό Αντίκτυπο

Η ψηφιακή παραγωγή νοηματικής γλώσσας είναι ένας τρόπος να γεφυρωθεί η επικοινωνία μεταξύ κωφών και ομιλούντων ατόμων. Τα ψηφιακά συστήματα νοηματικής γλώσσας μπορούν να έχουν πολύτιμες εφαρμογές στην καθημερινή ζωή των ανθρώπων που ενδέχεται να τα χρειάζονται. Για παράδειγμα, ρεαλιστικά δημιουργημένοι νοηματιστές θα μπορούσαν να λειτουργήσουν ως ενεργά εκπαιδευτικά εργαλεία, διαδραστικοί οδηγοί σε μουσεία ή παρουσιαστές σε εκπομπές ειδήσεων.

Είναι εξαιρετικά σημαντικό να σημειωθεί ότι αυτές οι τεχνολογίες δεν έχουν σκοπό να αντικαταστήσουν τους διερμηνείς της νοηματικής γλώσσας· αντίθετα, στοχεύουν να ενισχύσουν την προσβασιμότητα υποστηρίζοντας την εκπαιδευτική διαδικασία της νοηματικής γλώσσας και καθιστώντας τέτοιους πόρους πιο προσβάσιμους. Τέλος, τεχνητά συστήματα σαν και αυτό, δηλαδή που αναπαράγουν συνθετική εικόνα ανθρώπου, θα πρέπει πάντα να αναπτύσσονται με γνώμονα την αναντικατάσ-

τατη αξία της ανθρώπινης επικοινωνίας, ενώ ταυτόχρονα να δίνουν προτεραιότητα στις πραγματικές ανάγκες και τις προτιμήσεις των κοινοτήτων των κωφών και βαρήκοων ατόμων.

Chapter 1

Introduction

Contents

1.1 Sign Languages

Sign Languages are the primary form of communication for Deaf communities across the world. It is estimated that more than 70 million people make part of the deaf and hard-of-hearing (DHH) community, while there are more than 200 Sign Languages across the world. American Sign Language (ASL), British Sign Language (BSL), and Chinese Sign Language (CSL) are just a few examples, each with unique grammar, syntax, and cultural nuances.

The widespread use of sign languages highlights their critical role in enabling full participation in education, work, and society for Deaf individuals. However, for many DHH individuals, learning sign language is not a standard part of their education, leading to a disconnect that can impede social interaction and access to services. This gap underscores the importance of developing technologies that can facilitate real-time communication between sign language users and those who do not understand it, as well as making information more accessible to DHH individuals in various public domains, such as education, news media, and entertainment.

Sign languages are fully-fledged natural languages that rely on a visual-manual modality rather than an auditory-vocal one. They exhibit structural complexity comparable to spoken languages, encompassing phonology, morphology, syntax, semantics, and pragmatics (Stokoe, 1960 [59]; Sandler & Lillo-Martin, 2006 [53]). Unlike spoken languages, where linguistic elements are produced sequentially through vocal articulation, sign languages employ simultaneous articulations involving manual gestures, facial expressions, and body movements to convey meaning. This multimodal nature presents unique challenges in the computational modeling of sign language production, particularly when leveraging deep learning techniques for automatic generation and recognition of sign-based communication.

A fundamental distinction between spoken and signed languages is their reliance on visual-spatial articulation. Spoken languages organize phonemes linearly, whereas sign languages utilize multiple parallel channels of information, requiring distinct linguistic and computational approaches [44]. Despite the absence of phonemes in the traditional sense, sign languages exhibit a phonological system composed of key manual parameters:

- Handshape
- Location (placement of the hands in signing space)
- Movement (direction and type of motion)
- Palm orientation

Along with the different manual parameters that occur in Sign Languages across the world, SL encodes referents, relationships, and events within a three-dimensional signing space utilizing spatial grammar. Unlike spoken languages, which rely on sequential word order and prepositions, sign languages use spatial grammar, enabling signers to place referents in signing space and later refer to them through pointing or directional movements. This modality allows for efficient representation of complex concepts while introducing computational challenges in deep learning-based sign language synthesis.

Beyond manual articulations, sign languages also heavily on **non-manual markers**, which play syntactic and semantic roles. These may include:

- Facial expressions: Essential for conveying grammatical markers, affective meaning, and emphasis. In ASL, raised eyebrows indicate yes/no questions, while head tilts and facial movements mark conditionals and topicalization.
- Mouth gestures: Modify lexical meaning, functioning similarly to adverbial markers in spoken languages.
- Body posture and shifts: Used to indicate speaker role shifts or emphasize discourse structure.

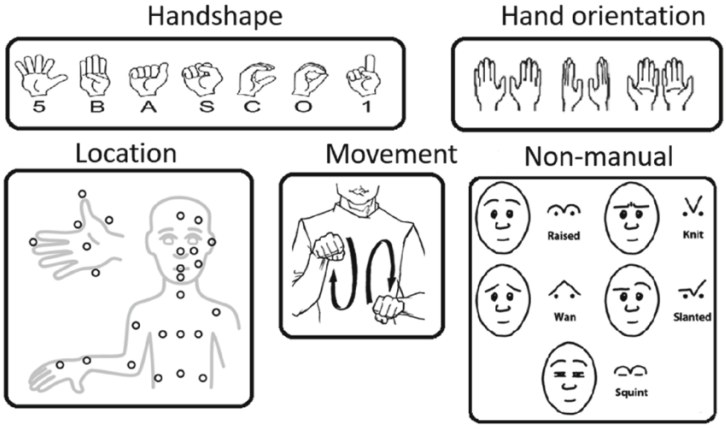


Figure 1.1.1: Overview of the different characteristics of SL

The integration of both manual and non-manual components makes sign languages highly expressive and complex, requiring computational models that can accurately capture spatial and multimodal articulations for effective sign language recognition and synthesis.

1.2 Sign Language Processing

The linguistic complexity of sign languages, as outlined in the previous section, poses unique challenges for deep learning models designed for sign language production. Unlike text or speech-based systems, sign language generation demands a multimodal approach that effectively integrates hand motion tracking, spatial representations, and non-manual features to produce grammatically coherent and natural signing.

Recent advancements in deep learning and human pose estimation have contributed significantly to the fields of sign language recognition and generation. However, several challenges persist, including generalization across different sign languages, adaptation to dialectal variations, and the accurate modeling of complex morphological and syntactic structures. The following sections of this thesis will explore these challenges in detail, outlining methodologies for deep learning-based sign language generation and evaluating their effectiveness in real-world applications.

The field of Sign Language Processing (SLP) emerges at the intersection of linguistics, computer vision,

and machine learning, to address communication barriers in the digital age. Sign Language Processing encompasses a variety of tasks aimed at bridging the gap between DHH and hearing communities by enabling the automatic recognition, translation, and generation of sign languages. The most critical components of an effective interactive SLP system are Sign Language Recognition (SLR), Sign Language Translation (SLT), and Sign Language Production (SLP).

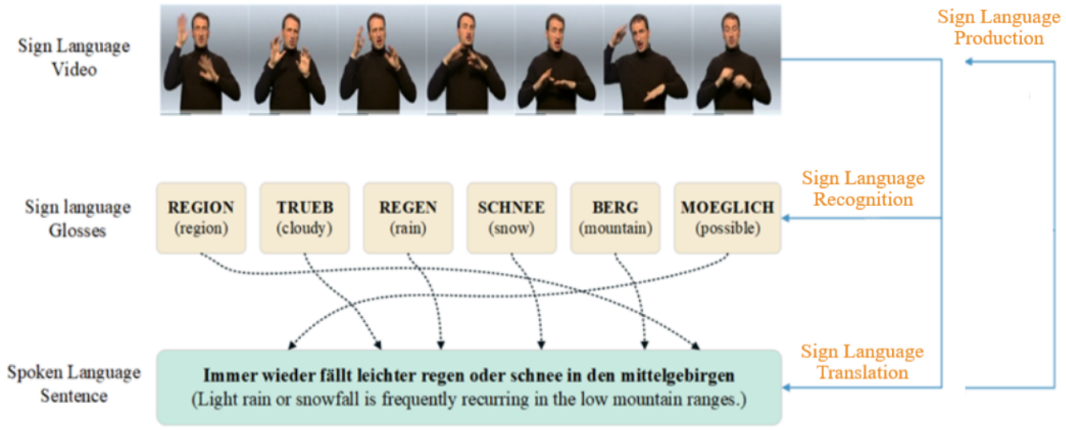


Figure 1.2.1: Overview of Sign Language Processing Technologies

Sign Language Recognition (SLR) focuses on interpreting sign language input (e.g., video or motion capture data) and mapping it to the corresponding meaning in a spoken or written language. Sign Language Translation (SLT) builds on this recognition by enabling two-way communication—translating spoken language into sign language and vice versa. This bi-directional capability is vital for real-time communication between hearing and DHH individuals, whether through video conferencing platforms, public services, or media content.

One of the most challenging and underexplored areas in Sign Language Processing is Sign Language Production (SLP). SLP refers to the generation of accurate, fluent, and natural-looking sign language animations or videos based on textual or spoken language input. Unlike SLR and SLT, which focus on understanding sign languages, SLP requires the ability to synthetically produce sign language content that is both linguistically accurate and visually convincing. The complexity of SLP stems from the intricate movements and expressions involved in sign languages, which must be replicated realistically to convey the correct meaning and cultural nuances.

Existing SLP systems have primarily relied on basic animation techniques or rule-based models, which often fail to capture the subtleties of human motion and natural language. Recent advancements in deep learning, particularly neural machine translation models and generative networks, have opened new possibilities for generating more photorealistic sign language content. Despite these advances, current solutions are still in their early stages, with significant room for improvement in the fluidity and accuracy of the produced sign language videos.

1.3 Research Motivation and Contributions

The motivation behind this thesis is to advance the field of Sign Language Production by leveraging cutting-edge deep learning techniques. Current methods for SLP often struggle with:

- Realism and Naturalness: Existing sign language avatars and animations frequently lack the fluidity and expressiveness of human signers.
- Linguistic Accuracy: Producing semantically accurate and grammatically correct sign sequences remains a major challenge.
- Data Scarcity: The limited availability of large-scale, high-quality sign language datasets hinders the development of robust models.
- Generalization Across Sign Languages: Most SLP models are trained on specific sign languages, making cross-linguistic adaptation difficult.

To address these challenges, this thesis explores an approach that integrates transformer-based architectures for linguistic processing and neural rendering techniques for realistic visual synthesis. In greater detail, we tackle sign language production by utilizing a transformer-based architecture that enables the translation from text input to extended skeletal pose representations. Furthermore, we explore the photerealistic aspect of the problem, aiming to create a complete SLP pipeline that transforms text directly into realistic human SL videos. For the photorealistic module, we harness Generative Adversarial Networks (GANs) to perform neural rendering on the pose sequences generated by the transformer model. Finally, we evaluate the effectiveness of the proposed pipeline on three different datasets through an extensive series of comparative analyses, ablation studies, and user studies.

1.4 Thesis Outline

This thesis is organized in chapters as follows:

Chapter 2: Review of fundamental deep learning architectures that have been applied to Sign Language related tasks and the pose estimation frameworks, OpenPose and MediaPipe.

Chapter 3: Review of the Neural Machine Translation related methods on Sign Language processing including Sign Language Recognition and Translation (SLT), Sign Language Video Anonymization Techniques and a review of all existing methods on Sign Language Production (SLP).

Chapter 4: Thorough description of proposed methodology for sign language production using a dual architecture with Transformers and GANs.

Chapter 5: Presentation of performed experiments, ablations and user studies.

Chapter 6: Conclusion of thesis results, discussion of limitations and future work.

Chapter 2

Deep Learning Background

Contents

2.1 Background on Deep Learning Architectures

We start the main text by presenting a brief introduction on the background technical knowledge needed to develop a proper understanding for the fields of Deep Learning, Pose Estimation and Generative Neural Architectures.

2.1.1 Neural Machine Translation Networks

2.1.1.1 Recurrent Neural Networks (RNNs)

The traditional Feed Forward Network, sometimes referred to as the "vanilla" neural network, is one of the simplest neural architecture, where the information only flows in the forward direction, from the input nodes, through the hidden layers, and to the output nodes. Feed-forward neural networks are designed to handle independent data points. Thought they are proven useful in tasks like Image Classification, more complex input-dependent architectures are required to handle sequential data for NLP tasks.

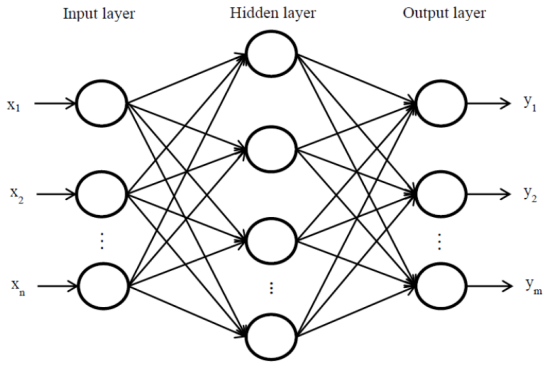


Figure 2.1.1: One Hidden layer Feed Forward Network Architecture. Figure from [1]

Unlike feedforward neural networks, **Recurrent Neural Networks (RNNs)** have a looped architecture that allows information to persist from one step of the sequence to the next. This makes them particularly well-suited for language modeling related tasks.

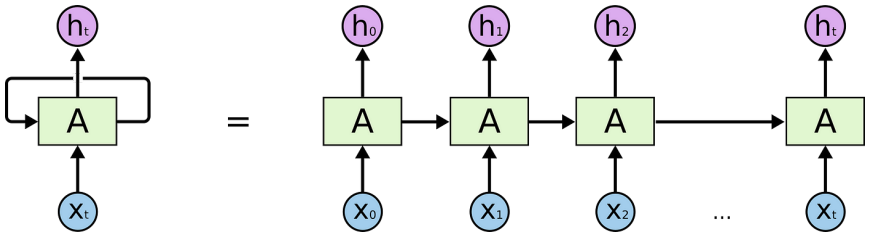


Figure 2.1.2: Unrolled Recurrent Neural Network Architecture

RNNs can be categorized depending on the input and output dimensions, as shown in figure 2.1.3. In the last case, many-to-many RNNs receive a sequence of inputs and generate a sequence of outputs while maintaining sequential dependencies and can be thus used in Machine Translation.

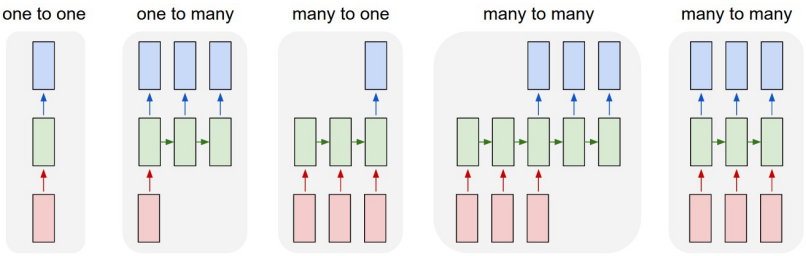


Figure 2.1.3: Different types of RNN Architectures

Long Short-Term Memory networks (LSTMs) are a special kind of RNN, capable of learning long-term dependencies. They were introduced by Hochreiter and Schmidhuber (1997) [31] to specifically address the vanishing gradient problem found in traditional RNNs. The LSTM contains an internal state called the cell state, which runs through the entire chain of the network. This state uses gating as shown in figure 2.1.4 to control which (relevant) information carries throughout the processing of the sequence.

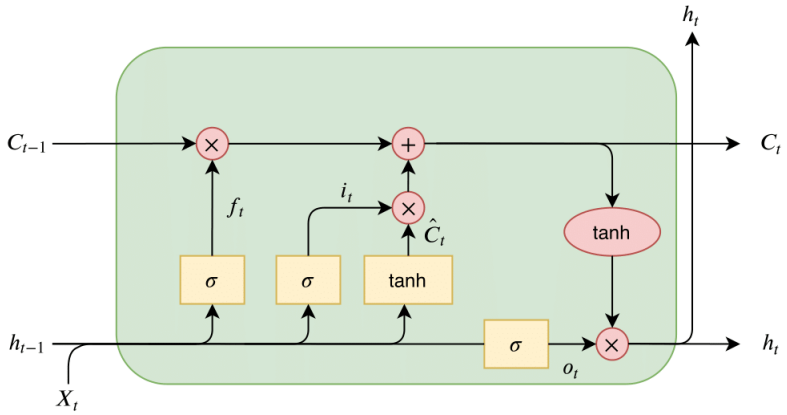


Figure 2.1.4: LSTM Cell Architecture: The Input Gate decides which values from the input should be used to update the memory state. The Forget Gate determines what portions of the memory to retain or discard from block to block, while the Output Gate focuses on output generated based on input and the memory of the block.

2.1.1.2 The Transformer Network

Transformers were originally proposed in the paper "Attention Is All You Need" [67] in 2017, and since then have been extensively used in many NLP applications, including machine translation, text summarization and question answering. Unlike RNNs, Transformers do not rely on recurrence but instead operate on self-attention, drawing global dependencies between the input and output. Transformers scale naturally to very large models (e.g., GPT, BERT) and datasets, making them suitable for tasks requiring billions of parameters. Also, by leveraging parallel computation, Transformers are faster to train than RNNs and can handle larger datasets effectively.

The full Transformer architecture follows an encoder-decoder framework, where both components are constructed using repeated layers of self-attention and feed-forward neural networks. The complete transformer architecture is shown in figure 2.1.5.

Each encoder layer includes a multi-head self-attention mechanism to capture token-to-token dependencies, a position-wise feed-forward network to transform the representation, and layer normalization with residual connections to stabilize and enhance training. The decoder, on the other hand, produces the output sequence by attending to both the encoder's output and its own previous outputs. Each decoder layer features a masked multi-head self-attention mechanism for autoregressive decoding, a cross-attention mechanism to integrate information from the encoder, and a feed-forward network with residual connections for additional transformations. Encoders and decoders can have multiple layers and multiple attention heads for parallel computing.

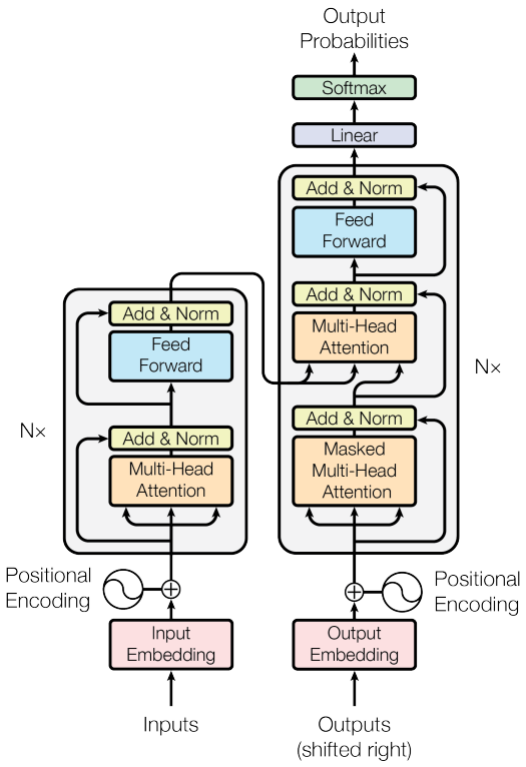


Figure 2.1.5: The Transformer Architecture. Figure from [67]

Self-Attention: The Scaled Dot-Product Attention Mechanism aims to encode the in-sentence dependencies. The input consists of queries and keys of dimension d_k, and values of dimension d_v. We compute the dot products of the query with all keys, divide each by $\sqrt{d_k}$, and apply a softmax function to obtain the weights on the values. In practice, this is done by packing multiple inputs into the matrices Q, K and V for parallel computation. The output attention matrix is computed as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1.1}$$

Multi-Head: The Multi-head attention mechanism allows the model to jointly attend to information from different representation subspaces at different positions. The scaled dot-product attention function is applied concurrently to each of the projected results, producing h output values, known as

'heads'. The heads are finally concatenated and transformed using an output weight matrix, following the equation mentioned below:

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(head_1, \ldots, head_h)W^O \\ \text{where } head_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.1.2}$$

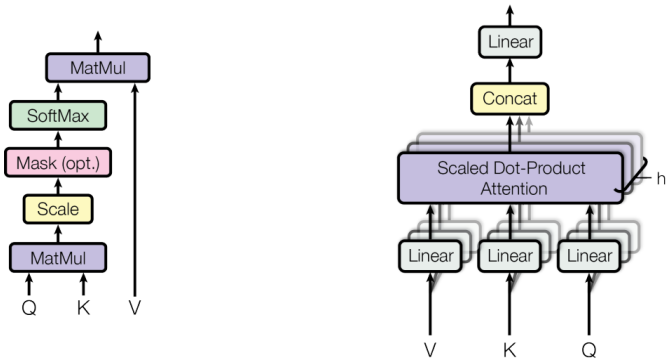


Figure 2.1.6: Self-Attention (left) and multi-head Attention (right). Figure from [67]

2.1.2 Convolutional Neural Networks

Another notable mention in this brief Machine Learning background are Convolutional Neural Networks (CNNs). After describing the most essential architectures used in NLP, we briefly mention CNNs as well, as they are highly relevant to our research. Convolutional Neural Networks are a class of deep learning models primarily used for image processing, though they can be applied to other tasks as well. The core component of a CNN is the convolutional layer, which employs multiple filters, each having a distinct set of learnable weights. These filters are convolved with the input image, creating feature maps by computing the dot product between the filter parameters and corresponding regions of the input. The size of the filters is typically smaller than the input image, and they are moved across the image to extract local features. As the network deepens, these layers progressively capture more complex features, starting from simple edge and texture detection in the early layers to more intricate representations in the deeper layers. To reduce the dimensionality of the feature maps and the number of parameters, pooling layers such as max pooling or average pooling are often used. These layers compress the feature maps by taking the maximum or average values from small patches, respectively. The final layers of a CNN are usually fully connected layers, which convert the output from the previous layers into a 1D vector for classification. The last fully connected layer classifies the image based on the features extracted by the preceding convolutional and pooling layers. CNNs are powerful models, widely used for tasks like image classification, object detection, and other visual recognition tasks.

In the context of Sign Language Processing, since it contains inherently visual tasks, CNNs have been proven effective in detecting key elements of sign communication, such as hand shapes, facial expressions, and body movements. In video-based sign language translation, CNNs are often combined with Recurrent Neural Networks (RNNs) or Transformers to capture temporal dependencies across frames.

Additionally, CNNs play a crucial role in hand gesture and pose estimation, leveraging pre-trained models such as OpenPose to detect hand and body keypoints for multiple SL pipelines. Beyond recognition and translation, CNNs have been employed in sign language production and video anonymization. Generative Adversarial Networks (GANs) utilizing CNN-based architectures are often used in research, described in the following sections, to generate realistic sign language avatars for automated translation, without compromising linguistic integrity.

2.1.3 Generative Adversarial Networks

2.1.3.1 Definition

Since their introduction by Goodfellow et al. in 2014 [24], Generative Adversarial Networks (GANs) have revolutionized the way we approach generative modeling, offering powerful capabilities in generating realistic images, videos, and audio.

GANs consist of two neural networks - the generator and the discriminator - that contest with each other in a game-theoretic scenario. The generator's role is to create new data samples, such as images, that resemble the data from a real dataset. It takes random noise (usually from a simple distribution like Gaussian or uniform) as input and transforms it into synthetic data. The goal of the generator is to generate data so realistic that the discriminator cannot distinguish it from real data. The discriminator on the other hand distinguishes between real data (training set) and fake data (by the generator). It outputs a probability indicating whether a given input is real or fake. The discriminator is a binary classifier trained to correctly identify real samples from generated ones. In the original paper [24], it is usually assumed that the generator moves first, and the discriminator moves second, thus giving the following equation for the value function $V(D, G)$, in the form of a minimax game:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.1.3}$$

where:

- G is the generator,
- D is the discriminator,
- x is a data point from the dataset,
- z is a point from the generator's input noise distribution,
- p_{data} is the data distribution,
- p_z is the noise distribution.

2.1.3.2 Applications

Image-to-Image translation covers a large variety of Computer Vision and Graphics related tasks, including style transfer, resolution improvement, image restoration etc. The need for output customiza-

tion and fine-tuning is solved by conditional GANs by extending the original GAN so that both the generator G and the discriminator D are conditioned on the same auxiliary information **y**. Depending on the type of the additional information, cGANs may be used in a variety of applications, including text-to-image synthesis, video generation and image-to-image translation. In cGANs, the objective function of the minimax game becomes:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(\mathbf{x}|\mathbf{y})] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(\mathbf{z}|\mathbf{y})))] \tag{2.1.4}$$

Pix2pix is a popular cGAN architecture developed by Isola et al. [33] for general purpose image-to-image translation. Similarly to the vanilla GAN, the pix2pix architecture involves a generator G and a discriminator D competing against each other. In addition to the noise vector z, the generator is fed with an input image y from the source domain and learns to translate it into the corresponding ground truth image x from the target domain. In the pix2pix model, the authors add an L1 reconstruction loss to the standard training objective, to stabilize training, which is defined as follows:

$$L_{L_1}(\theta^{(G)}) = E_{x,y,z}||x - G(z|y)||_1 \tag{2.1.5}$$

Pix2pix experiments also prove the usage of noise z ineffective because the generator simply ignores it. Its architecture follows the general shape of a U-Net, which is an encoder-decoder model with skip connections between symmetrical parts of the encoder and decoder stack. The promising results presented in pix2pix are shown in figure 2.1.7.

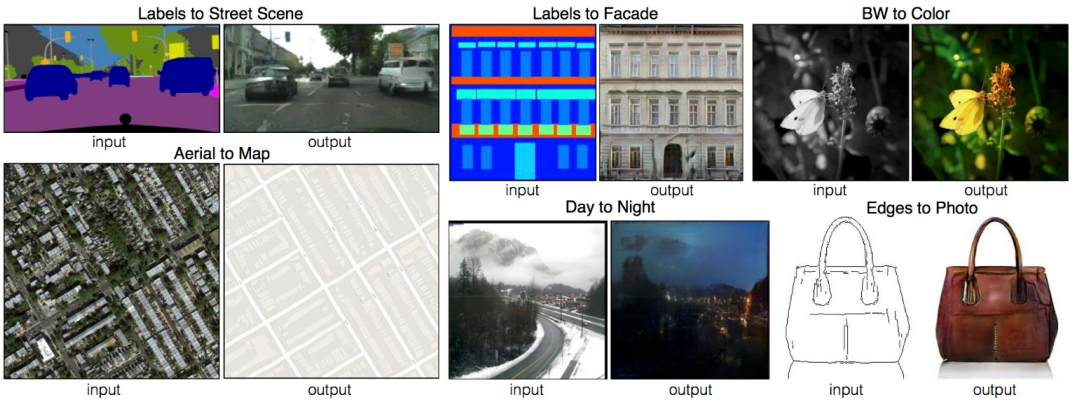


Figure 2.1.7: Example results of pix2pix on various tasks. An input image from a source domain is fed into the generator, which then converts it into the target.

Video-to-video Synthesis: Wang et al. propose vid2vid [69], which is a general purpose cGAN for generating photorealistic videos from input source videos, such as sequences of semantic segmentation masks. The authors introduce a spatio-temporal learning objective that ensures the photorealism and temporal coherence of the generated videos. The model is capable of producing videos with diverse appearances from the same input by conditioning on varied semantic representations, including segmentation masks, sketches, and poses. The generator G operates sequentially, taking a sequence of source frames (such as semantic segmentation masks) and then generating a sequence of video frames

that mimic real video frames. This sequential generation is guided by a Markov assumption, meaning the generation of a current frame only depends on a limited history of previous frames. The core of the training objective is to match the conditional distribution of the synthesized videos given the input videos to that of real videos. This is expressed mathematically as:

$$p(\tilde{x}_{T1}|s_{T1}) = p(x_{T1}|s_{T1}) \tag{2.1.6}$$

where s_{T1} represents the input sequence and x_{T1} and $\tilde{x}_{T1}$ are the real and synthesized video sequences, respectively.

The minimax optimization problem is defined as follows:

$$\max_D \min_G \mathbb{E}_{x_{T1}, s_{T1}}[\log D(x_{T1}, s_{T1})] + \mathbb{E}_{s_{T1}}[\log(1 - D(G(s_{T1}), s_{T1}))] \tag{2.1.7}$$

Figure 2.1.8: Generation of a photorealistic video from an input segmentation map video from Cityscapes. Top left: input. Top right: pix2pixHD [70] and Bottom right: vid2vid. Figure from [69]

Another interesting approach to the video-to-video synthesis task is the do as I do Motion Transfer method presented in the Everybody Dance Now publication [13]. Similar to the previously mentioned cGAN-based works, this model utilizes c-GANs for realistic Pose to Video Translation. The video generation pipeline is seperated into three stages, pose detection with OpenPose, global pose normalization and finally the video translation by mapping from normalized pose sequences to the target subject. Both the training and transfer objectives are shown in figure 2.1.9.

GANs in Neural Rendering: Neural rendering refers to the use of neural networks to generate or manipulate images, videos, or 3D scenes in a way that mimics traditional rendering methods used in computer graphics. Unlike traditional rendering techniques that rely on physics-based simulations of light and materials, neural rendering uses deep learning models to learn the mappings between 3D representations and 2D images. Conditional GANs are often used in the neural rendering frameworks. A great example of this is the Head2head network [37], which uses GANs to generate realistic frame videos from their senmantic representations.

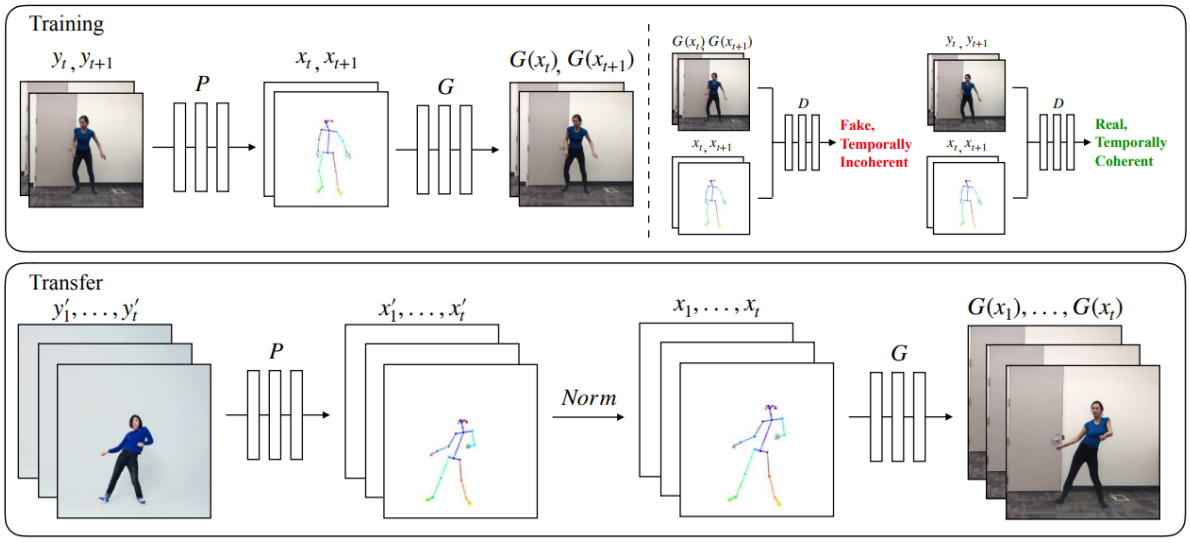


Figure 2.1.9: Top: Training - The discriminator D attempts to between the real and fake correspondences, $(x_t, x_{t+1}), (y_t, y_{t+1})$ and $(x_t, x_{t+1}), (G(x_t), G(x_{t+1}))$, respectively. Bottom: Transfer - The pose detector P is again used to obtain pose joints for the source person and are then normalized. The pre-trained mapping G is applied onto the normalized joints to generate the final output. Figure from [13]

2.2 Background on Pose Estimation

2.2.1 Pose Estimation with OpenPose

OpenPose ([71], [12]) is one the most widely used frameworks for full body 2D skeleton joint prediction. An RGB monocular image is used as an input and the model jointly predicts hand, body and face 2D anatomical keypoints. The typical OpenPose output consists of 21 keypoint locations, corresponding to each hand, 25 keypoints for the body and 70 keypoints for the face. In the context of sign language processing, OpenPose plays a crucial role by enabling detailed tracking of hand, body, and face movements. Sign languages, such as American Sign Language (ASL) and others, rely heavily on these physical gestures, and OpenPose's ability to track fine-grained details of hand poses, facial expressions, and body movements makes it a valuable tool for recognizing and analyzing sign language.

The model is a two-branch multi-stage Convolutional Neural Network. The input image is fed to a CNN, that extracts image features, namely the first 10 layers of VGG-19. At each stage the top branch passes the image features through a series of convolutional layers in order to obtain a set of confidence map S of body part locations. The bottom branch, also consisting of a stack of convolutions predicts the part affinity fields L, which are a set of flow fields that encode pairwise relationships between body parts. In each subs equent stage the predictions from both branches along with the original image features, are concatenated and used to produce more refined predictions. Two loss functions are applied at the end of each stage, one at each branch respectively. A standard L2 loss between the estimated predictions and ground truth maps and fields is used.

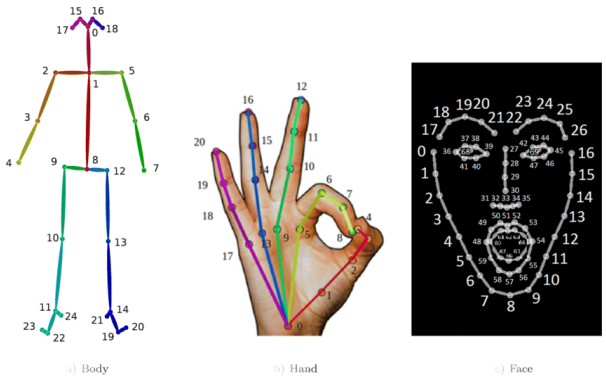


Figure 2.2.1: OpenPose Landmarks

2.2.2 Pose Estimation with the MediaPipe Framework

MediaPipe (MP) is a comprehensive, open-source framework devised by Google for constructing multimodal and cross-platform machine learning pipelines. This framework supports a broad range of applications, from video and audio processing to real-time data analysis across different platforms. It allows developers to build complex pipelines using a graph of modular components known as Calculators, which handle tasks ranging from data pre-processing and inference to post-processing and annotation. It also provides developers with a suite of pre-trained models and the flexibility to create custom calculators tailored to specific needs.

A particular implementation of MediaPipe is the MediaPipe Holistic model, which integrates several of the framework's capabilities to perform comprehensive human pose estimation. This model seamlessly combines pose, face, and hand landmarks to offer a holistic approach to motion tracking. MediaPipe Holistic employs a graph-based pipeline that processes different regions of interest (ROIs) within an image to estimate a total of up to 543 landmarks. These include 33 body pose landmarks, up to 468 facial landmarks, and 42 hand landmarks (21 per hand). [27]

The MediaPipe Holistic model operates in several phases. Initially, human pose is detected using the BlazePose detector, a sophisticated model that provides an initial set of body landmarks. These landmarks help to define the crop bounds for the face and hands, which are crucial for detailed landmark detection in subsequent steps. In scenarios where pose detection is less accurate, the model can apply additional hand and face re-crop models to ensure the precision of landmark estimation.

For each detected region (face and hands), specific models—such as the MediaPipe Face Mesh for facial landmarks and MediaPipe Hands for hand landmarks are applied. These models are designed to perform optimally on the respective cropped images, ensuring high fidelity in landmark detection. The final step in the MediaPipe Holistic pipeline merges the landmarks from all models, presenting a unified set of body, face, and hand landmarks. This comprehensive landmark detection facilitates advanced applications such as augmented reality, gesture recognition, and interactive applications where real-time human interaction with machines is required.

MediaPipe Hands is a cutting-edge solution offered by Google's MediaPipe framework, designed specifi-

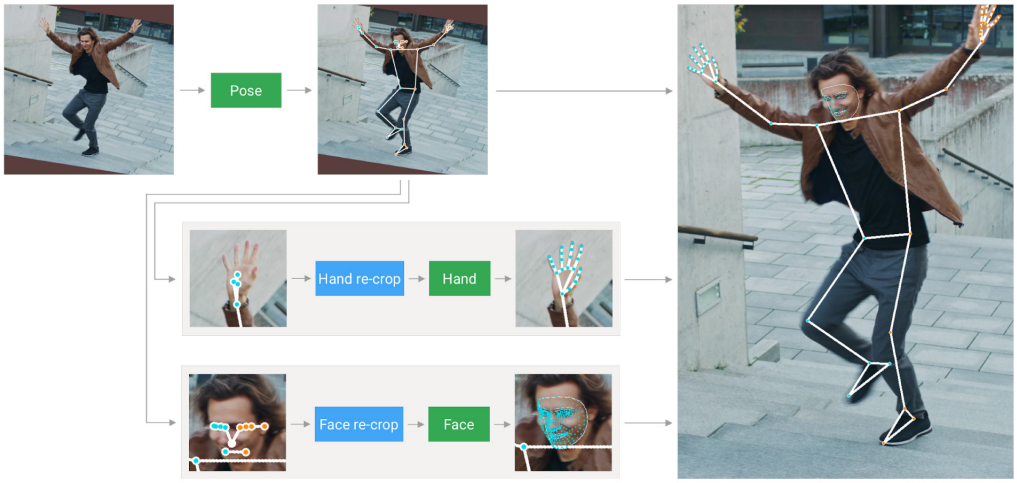
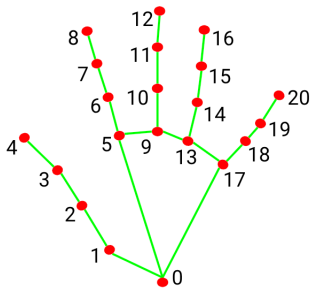


Figure 2.2.2: MediaPipe Holistic Pipeline. Figure from [26]

cally for real-time hand tracking and landmark detection. This solution is notable for its ability to infer 21 3D hand landmarks from a single RGB camera frame with high accuracy and low latency, which makes it applicable to a broad range of applications, including **sign language interpretation**.

The architecture of MediaPipe Hands consists of two main components: the palm detector and the hand landmark model. The palm detector, known as **BlazePalm**, is a crucial first step in the hand tracking process. It utilizes a single-shot detector (SSD) approach to analyze the full input image and identify the orientation and location of the hand via a bounding box. This detection is generally performed only once at the beginning or when there is no hand present in the frame, leveraging previous frame data to track the hand in subsequent frames. Following the detection of the palm, the hand landmark model comes into play. This model focuses on the region of interest (ROI) specified by the BlazePalm detector, and processes it to output 21 three-dimensional coordinates corresponding to key points across the hand, such as finger joints and the wrist. The robustness of the hand landmark model is ensured through its training on a diverse dataset comprising both real-world images and synthetic data, which helps it handle various challenging scenarios like partial visibility and complex hand gestures.

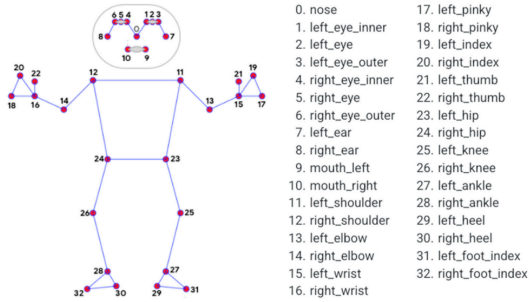


0. WRIST
1. THUMB_CMC
2. THUMB_MCP
3. THUMB_IP
4. THUMB_TIP
5. INDEX_FINGER_MCP
6. INDEX_FINGER_PIP
7. INDEX_FINGER_DIP
8. INDEX_FINGER_TIP
9. MIDDLE_FINGER_MCP
10. MIDDLE_FINGER_PIP
11. MIDDLE_FINGER_DIP
12. MIDDLE_FINGER_TIP
13. RING_FINGER_MCP
14. RING_FINGER_PIP
15. RING_FINGER_DIP
16. RING_FINGER_TIP
17. PINKY_MCP
18. PINKY_PIP
19. PINKY_DIP
20. PINKY_TIP

Figure 2.2.3: MediaPipe Hand Landmarks. Figure from [25]

Following the same logic as MP Hands, MediaPipe Pose uses a Pose Detector and a Pose Landmark Model for performing estimation of human poses.

The Pose Detector, known as BlazePose, is pivotal for the initial localization of the human subject within the video frame. BlazePose is inspired by MediaPipe's earlier model, BlazeFace, which was designed for face detection. It identifies a region of interest (ROI) around the person by predicting person-specific parameters such as the midpoint of the hips and the overall body orientation. Once the ROI is established by BlazePose, the Pose Landmark Model calculates the precise 3D coordinates of 33 landmarks on the human body. This model employs a sophisticated network that combines heatmaps and offset regression to predict landmark positions with high accuracy. The network's design ensures that it captures both global pose features and finer local details necessary for accurate pose reconstruction.



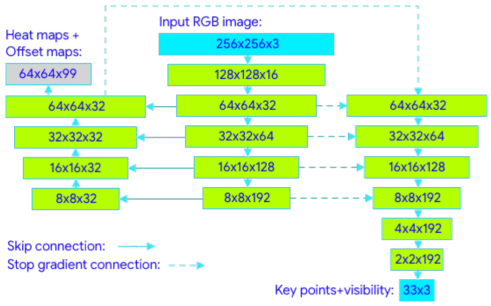


Figure 2.2.4: MediaPipe Pose Landmarks [27] (left), BlazePose Architecture [5] (right)

Chapter 3

Literature Review

Contents

3.1 Sign Language Recognition (SLR) and Translation (SLT)

Sign languages serve as the primary form of communication for DHH individuals and enable them to engage with their communities. Unlike spoken languages, which rely on auditory signals, sign languages are visual-manual languages that use hand gestures, facial expressions, and body movements to convey meaning. Despite their richness and complexity, sign languages remain underrepresented in technological advancements. The development of Sign Language Recognition (SLR), Sign Language Translation (SLT) and Production (SLP) systems, leverages deep learning architectures to bridge the communication gap. In this chapter we provide a systematic review of previous approaches to SLR, SLT and SLP and especially addressing deep learning methods.

Sign Language Recognition (SLR) involves the task of interpreting sign language gestures from video inputs and converting them into a structured representation, such as glosses (annotated labels for signs) or spoken language text. Continuous Sign Language Recognition (CSLR) extends this task to continuous signing streams, where the model must segment and recognize a sequence of signs from a video.

SLR has been the first approach to address the challenge of automating sign language understanding. Generally, this field focuses on developing methods to extract meaningful features from sign language videos and classify them into discrete signs. For instance, S. Theodorakis, V. Pitsikalis and P. Maragos [62] introduced a dynamic-static unsupervised sequentiality approach for SLR, which leveraged statistical subunits and lexicons to improve recognition accuracy. This work laid the foundation for subsequent research by demonstrating the importance of combining dynamic and static features in SLR. Similarly, A. Roussos, S. Theodorakis, V. Pitsikalis and P. Maragos [51] proposed dynamic affine-invariant shape-appearance handshape features for SLR, which addressed the challenge of recognizing handshapes in sign language videos and introduced a novel feature extraction method, robust to variations in hand appearance and orientation. These contributions highlighted the potential of machine learning techniques for automating sign language recognition.

On the other hand, as an extended task on SLR, Sign Language Translation (SLT) focuses on translating sign language videos into spoken or written language, instead of SL glosses. Recent advancements in deep learning have revolutionized the field of SLR and SLT. Instead of relying on traditional machine learning algorithms, recent work uses several Deep Learning based Approaches to tackle the task, including RNNs ([3], [8]), LSTMs ([14]), GRUs ([35]), and Transformers ([10], [9]), after using CNNs for spatial feature extraction from the input video frames. CNNs have become the standard for extracting spatial features, including hand shapes, facial expressions, and body postures, from sign language videos. On the other hand, temporal modeling using RNNs or Transformers has further improved the performance of SLR and SLT systems by enabling the models to learn the temporal text dependencies in sign language sequences. In the following section we mostly present Camgöz's work on SLT due to its novelty, high number of citations and relevance to this work.

Camgöz et al. [14] proposed a novel deep learning architecture for video to sequence learning problems based on small specialized sub-networks, called SubUNets. The SubUNets framework is built around

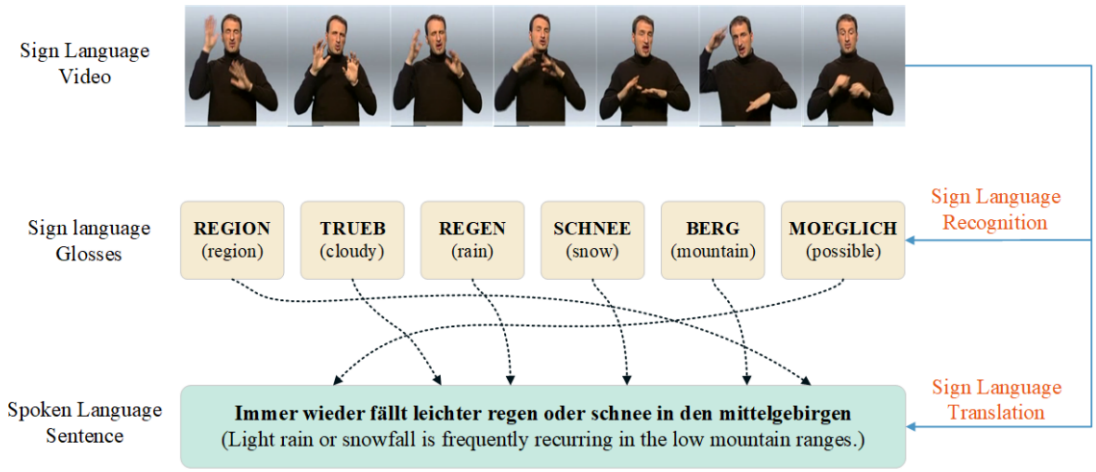


Figure 3.1.1: Sign Language Recognition vs Sign Language Translation. SLR usually refers to the conversion of signs to gloss while SLT usually refers to the translation to spoken language.

decomposing the complex problem of sign language recognition into manageable sub-problems. Each SubUNets three main components:

- Convolutional Neural Networks (CNNs), which are responsible for spatial feature extraction from input images.
- Bidirectional Long Short-Term Memory (BLSTM) Layers, which manage the temporal aspects by processing the spatial features over time, using both past and future context to make predictions at each time step.
- Connectionist Temporal Classification (CTC) Loss Layer: This layer is crucial for sequence-to-sequence learning, allowing the model to handle varying lengths of input and output sequences by introducing a 'blank' label for alignment.

Camgöz et al. [8] formalized SLT as a sequence-to-sequence (seq2seq) learning problem. This approach employs CNNs for spatial feature extraction from sign language videos, which are then fed into an attention-based encoder-decoder framework to generate spoken language translations. These experiments where made on three different pipelines, Gloss2Text (G2T), end-to-end Sign2Text (S2T) and Sign2Gloss2Text (S2G2T) which uses gloss annotations as an intermediate layer.

In another work, Camgöz et al. [10] use transformer models for both the recognition (SLRT) and translation (SLTT) pipelines. The encoders process sign video sequences to produce embeddings that capture both spatial and temporal features, while the decoders generate spoken language sentences. CTC Loss is used to facilitate learning without explicit alignment data, tying the recognition of sign glosses to the generation of text. Experimental results of the previously mentioned works prove that using gloss information as an intermediate step to spoken language translation improves the performance of the model, however relying on gloss annotations can be limiting on larger datasets since they require professional annotation.

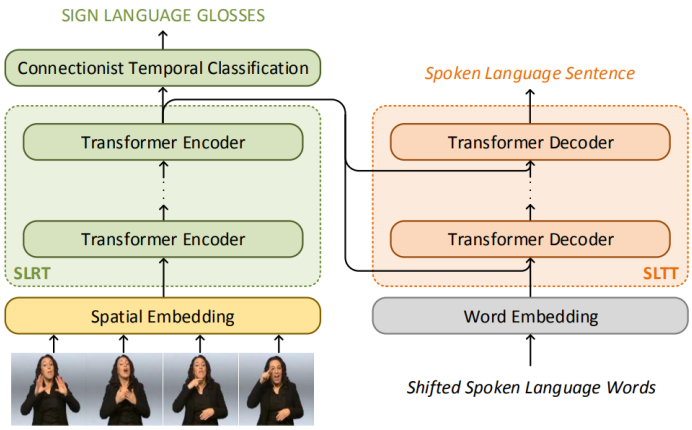


Figure 3.1.2: Transformer Based Architecture for both SLR and SLT tasks. [10]

3.2 Sign Language Video Anonymization

3.2.1 Methods

Sign language video anonymization is essential in various contexts, such as protecting the privacy of DHH individuals in online platforms, academic research, and legal settings where sensitive information is discussed. Unlike vocal-auditory languages, sign languages are visual-manual and heavily rely on facial expressions, body movements, and hand shapes to convey meaning, making standard anonymization techniques ineffective. Previous research [32] divides the different ways in which video can be anonymised into two categories, those which conceal all or part of a video, and those which reproduce a video. Concealment can be achieved through **blackening** sections of the image or by applying a **pixellation** filter to the signer's hands and mouth for the duration of the relevant sign. Reproduction approaches involve the anonymization of entire corpora by either human actors or computer-generated avatars.

3.2.2 Deep Learning Approaches for Anonymization

In [38] the authors experimented with three prototypes (with-torso, without-torso and tiger face) for anonymizing the face of ASL signers. Particularly,the with-torso prototype is based on image-to-video transformation works (ex. [63]) and swaps the face of a signer with a target face from an input image. However, [38] shows some limitations as the extent of the anonymization is not complete, since only the face is replaced, while the rest of the pose remains the same as in the original video.

In another work, the authors propose ANONYSIGN [55], an automatic method that achieves visual anonymisation of sign language which is built upon a combination of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) (cVAE-GAN [4]). In order to remove the original signers appearance, a skeleton pose sequence is first estimated from the source sign language video.

Cartoonized Anonymization [65] by C. O. Tze, P. P. Filntisis, A. Roussos, and P Maragos, proposes the use of pose estimation models to automatically generate cartoon-like characters to sign. The process

involves extracting skeleton sequences from the original video and the reference cartoon figure. These sequences are then processed through a recursive kinematic tree-based algorithm that adapts the cartoon's bone structures to match the human signer's poses.

The Neural Sign Reenactor [64] by C. O. Tze, P. P. Filntisis, A.-L. Dimou, A. Roussos, and P Maragos, introduces a GAN-based neural rendering pipeline designed for sign language videos. It transfers the facial expressions, head poses, and body movements from a source video to a target video, ensuring the retention of manual sign nuances and improving photorealistic representations in sign language anonymization applications.

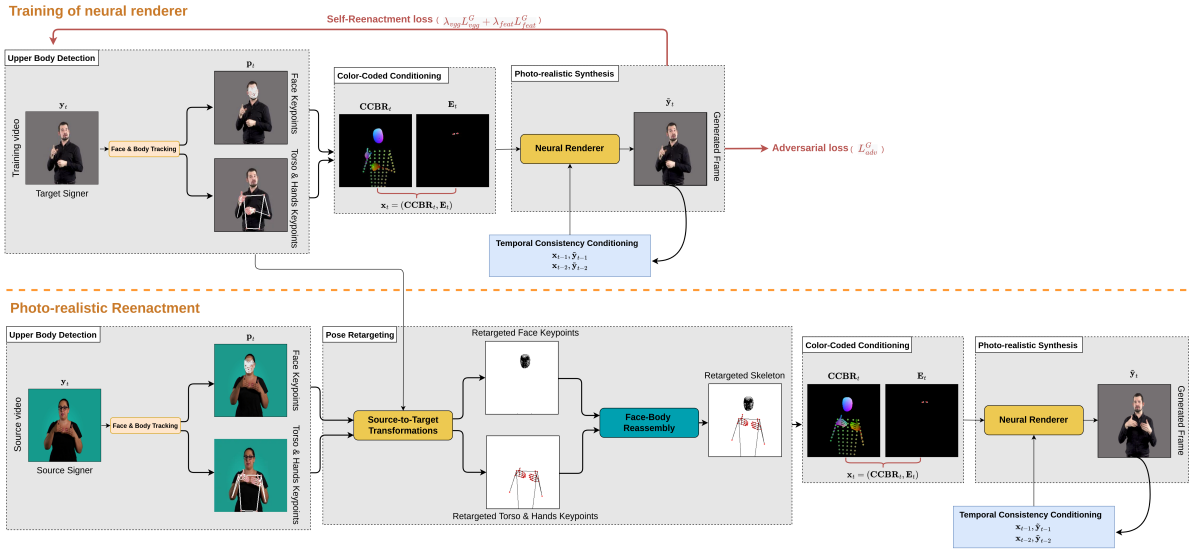


Figure 3.2.1: Neural Sign Reenactor. Figure from [65]

In [74], Xia et al. designed a sign video anonymization pipeline by using an asymmetric encoder-decoder structured image generator for high resolution image generation and designing a custom loss function for better generation of hand gestures and facial expressions. More recent work by the authors [75] called DiffSLVA, explores a novel approach to anonymizing sign language videos by utilizing diffusion models enhanced with ControlNet for spatial guidance. The method aims to preserve linguistic content while anonymizing the signer's identity effectively, addressing challenges in maintaining facial expressions and other linguistic features critical for sign language communication.

3.3 Sign Language Production (SLP)

The task of Sign Language Production (SLP) reviewed in this work can be described as follows: Given a spoken language text sentence (input) the model used generates a corresponding video of M frames (output). In this section, we briefly examine previous work regarding the Sign Language Production task.

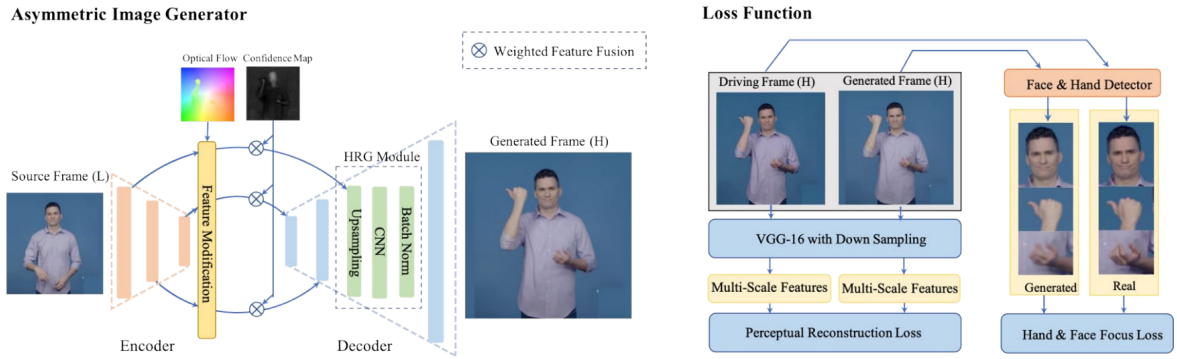


Figure 3.2.2: Encoder-Decoder architecture for video anonymization. Figure from [74]

3.3.1 SLP Prior to Deep Learning

3.3.1.1 Symbolic Notation Systems for Sign Languages

Pre-NMT work mainly uses phrase lookup and direct sentence matching for SLP as well as computer generated avatar sign videos for a realistic animated output. These avatars are programmed to generated sign language videos accurately, making them useful tools for education, accessibility, and entertainment for the deaf and hard-of-hearing communities. The effectiveness of these avatars largely depends on the accuracy and naturalness of the sign language they produce while they usually rely on manual labor-intensive work [73].

The ASCII Stokoe system represents the symbols of the Stokoe phonemic notation as ASCII characters, which makes the notation of signs compatible with computer processing. Grieve Smith took advantage of this to develop an early sign authoring system as a Web-based avatar display [28]. SignWriting is another written notation system for Sign Languages, and is the first to adequately represent facial expressions and shifts in posture, and to accommodate representation of series of signs longer than compound words and short phrases. Researchers have developed an XML-compliant format called SignWriting Markup Language (SWML) to aid in processing SignWriting texts and creating dictionaries and sign avatars [6].

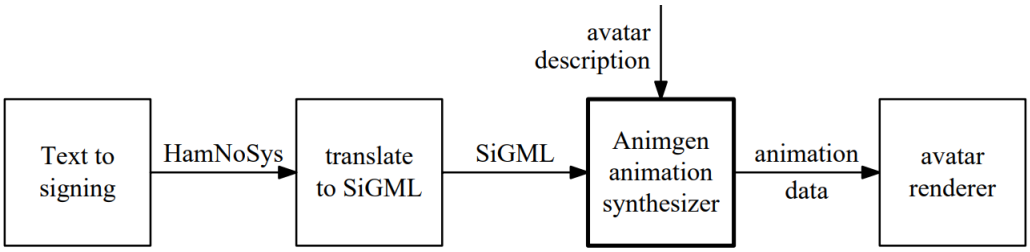


Figure 3.3.1: ViSiCAST [34]: Block diagram of the generation of synthetic signing animation

HamNoSys is a transcription system for all sign languages, with a direct correspondence between symbols and gesture aspects, such as hand location, shape and movement. Although HamNoSys is machine-readable, researchers developed SiGML (Signing Gesture Markup Language), as an XML-compliant for-

mat that is more amenable to computer processing [34]. The purpose of SiGML is to support signed language generation through avatars, and so required additional information beyond the data required for linguistic analysis. To the original specification of SiGML, Glauert and Elliott [24] added a framework that specifies timings internal to a sign. The framework incorporates Johnson and Liddell's SLPA model which decomposes signs into a series of consecutive segments, each of which lists any changes occurring to articulators during the segment [23]. An advantage of this approach was that changes to articulators were no longer constrained to sign boundaries but were constrained to segment boundaries. Several avatar-based applications have utilized SiGML or HamNoSys including the eSign Editor [30], the ViSiCAST animation component [34], and JASigning [17].

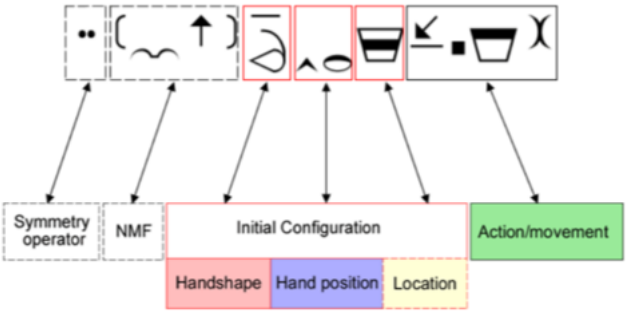


Figure 3.3.2: HamNoSys linear organization. Figure from [29]

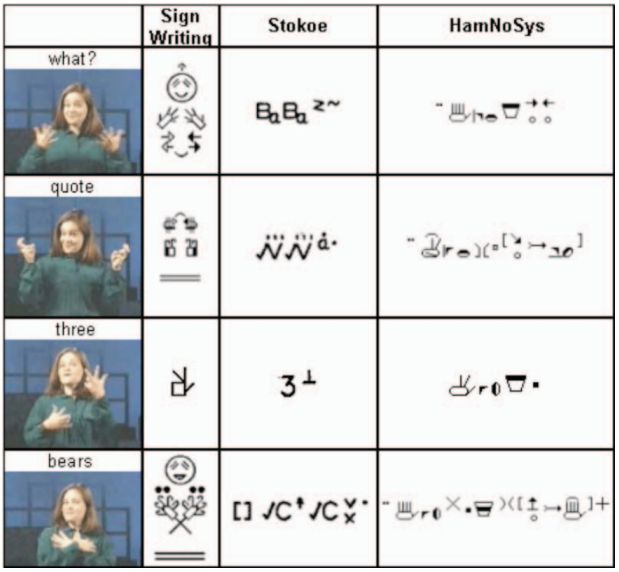


Figure 3.3.3: ASL signs in Stokoe Notation, HamNoSys and SignWriting. Figure from [29]

3.3.1.2 Data-driven Signing Avatars

Data-Driven Techniques for Sign Language Avatar Synthesis

Data-driven synthesis approaches, which utilize motion capture (MoCap) to animate avatars, are less commonly employed than either hand-crafted or procedural synthesis techniques, despite their ability to achieve a high degree of realism that procedural methods often cannot match. In MoCap, the posi-

tions of markers placed on a human performer are recorded to deduce the positions and orientations of the joints during post-processing. The crucial step of annotation involves segmenting the continuous motion flow into smaller sections and labeling these segments. This is essential for the editing process, aiming to identify linguistic features and establish precise temporal boundaries between linguistic units. The complexities associated with annotation are discussed in [72] and [22], highlighting that manual annotation is labor-intensive and its automation has been extensively explored [77], [41].

It is technically impossible to capture all the signs in all the possible contexts due the size of the SL vocabulary, the multiple inflection mechanisms of signs and sentences in SL, the need for various Deaf participants, and time and memory constraints. Therefore, synthesizing signs using a limited set of pre-recorded signing sequences is the major challenge of data-driven techniques. Essentially, the quality of the resulting avatar animations depends on the granularity of the annotation and on the size and content of the initial corpus.

Some of the data-driven avatars for sign language generation include:

- Tessa [58], a BSL avatar, and Simon [48], a Sign Supported English avatar, which both take advantage of the play-back technique
- The Sign3D project combines play-back and sign synthesis editing techniques [39]
- Sign360 (by MoCapLab), a French SL avatar driven by pre-recorded gestures [7]

(a) Tessa **(b)** Simon **(c)** Sign3D **(d)** Sign360

Figure 3.3.4: Mentioned data-driven sign avatars

3.3.2 Text-to-Pose using Neural Machine Translation

In the paper [54] the authors introduce the first Transformer-Based architecture for end-to-end SLP from given text. The architecture consists of two sequential transformers. First, the text is encoded through the Symbolic Transformer, which follows the architecture of the classic transformer model [67]. Then, the encoded input is passed through the Progressive Transformer, which is used for producing the continuous frame sequence. The Progressive Transformer uses a Counter Embedding ranging from 0 to 1 as shown in figure 3.3.6 which indicates the start and the finish of the generated frame sequence respectively. The progressive decoder is an auto-regressive model produces a sign pose frame at each time-step, along with the frame's counter value. The progressive decoder's output can be described by the equation:

$$[\hat{y}_{u+1}, \hat{c}_{u+1}] = D_P(\hat{j}_u | \hat{j}_{1:u-1}, r_{1:T}) \tag{3.3.1}$$

Finally, the overall training objective of the Progressive Transformer is concluded by calulating the Mean Squared Error (or another selected type of loss) between the groundtruth $y^*_{1:U}$ landmark sequence and the predicted landmark sequence $\hat{y}_{1:U}$:

$$L_{MSE} = \frac{1}{U}\sum_{i=1}^{u}(y^*_{1:U} - \hat{y}_{1:U})^2 \tag{3.3.2}$$

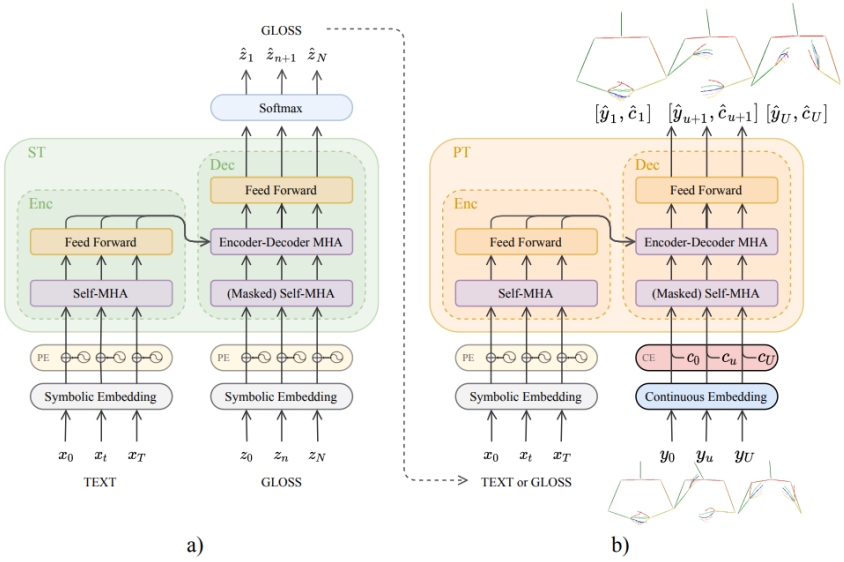


Figure 3.3.5: Total Transformer Architecture proposed in [54] for Text2Gloss2Pose and Text2Pose. Text2Gloss2Pose uses both the Symbolic (a) and the Progressive (b) Transformer, while Text2Pose essentially only uses the Progressive (b) Transformer.

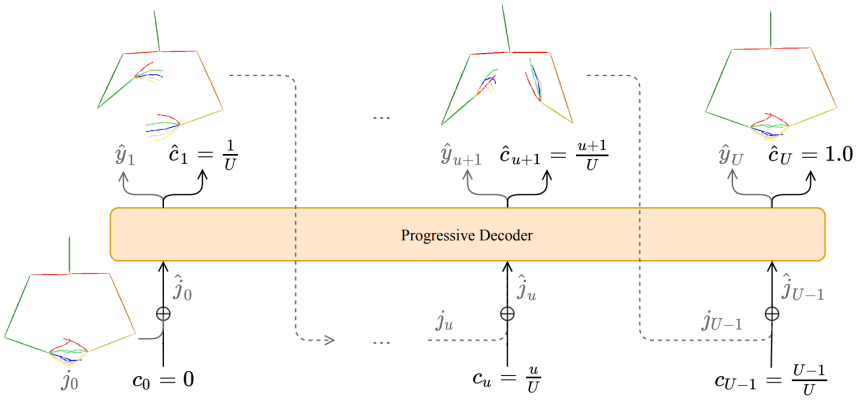


Figure 3.3.6: Progressive Decoder proposed in [54]

Prior work to the Progressive Transformer [60] employed Motion Graphs (MG) to generate realistic 2D skeletal poses corresponding to the gloss sequence following the text2pose NMT model. Motion graphs are structured as finite directed graphs with motion primitives, allowing for dynamic and smooth transitions between different sign gestures based on the sequence of glosses generated. Saunders' method can directly translate a source sequence of glosses/text into a target sequence of sign poses and thus

introducing the Transformer innovation into the SLP task.

In more recent work Saunders et al. [56] (2022) improved the previously mentioned model by introducing FS-NET, a frame selection network based on the original Encoder Arcitecture. In this work the original encoder - decoder architecture is used to train a text2gloss model. The gloss sentences are translated to pose using a pre-made set of dictionary signs and then linear interpolation between different signs. Finally, FS-NET (built as a transformer encoder [67]) is used to improve the temporal alignment of interpolated dictionary signs. This is a monotonic sequence-to-sequence task, due to the matching order of signing and the different sequence lengths. FS-NET predicts a discrete sparse monotonic temporal alignment path $\hat{\mathcal{A}}$ which contains binary decisions representing either frame selection or skipping:

$$\hat{\mathcal{A}} = FSNET(\mathcal{R}, h_{1:\mathcal{W}}) \tag{3.3.3}$$

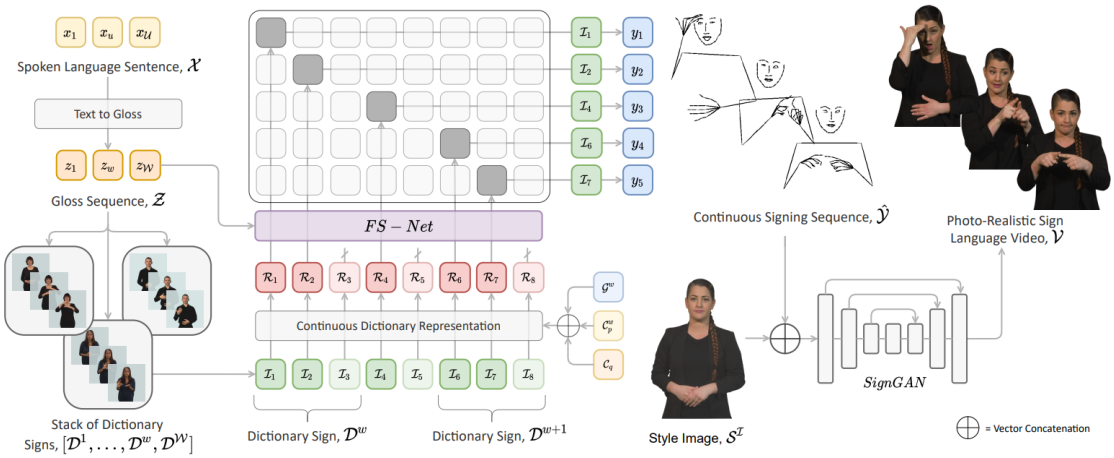


Figure 3.3.7: SLP Architecture proposed in [56]

In addition, in [61] the authors introduce a back-translation approach to generating 3D sign language animations from text. This research focuses on generating realistic 3D mesh sequences of sign language from textual inputs. The technique of back-translation is adapted to improve the generation process by refining the accuracy of the sign language animations. By translating the generated signs back into text and comparing it with the original input, the model can self-correct and enhance its output.

CasDual-Transformer: Since the publication of the Progressive Transformer for SLP [54] these works have been based on the proposed architecture and reproduce the results on the PHOENIX14T dataset with some modifications. For instance, the dual - decoder Transformer module [43] aims to resolve the regression to the mean hand position, where the original SLP architecture suffers. The source representation obtained from the encoder is fed into a hand pose decoder that generates only the manual sign pose sequence. The obtained manual representations along with the full-channel sign embedding and text representations are passed to a global decoder which produces the full sign pose sequence. The model uses a spatio - temporal loss to align the feature maps of both manual and full sign representa-

tions, which is defined as follows:

$$L_{Spatio} = L_{Spatio}^{Hands} + L_{Spatio}^{Sign} = \frac{1}{U}\sum_{i=1}^{u}(s_{1:U}^{h} - \hat{s}_{1:U}^{h})^2 + \frac{1}{U}\sum_{i=1}^{u}(s_{1:U} - \hat{s}_{1:U})^2 \tag{3.3.4}$$

3.3.3 Pose-to-Video using GANs

Following the landmark sequence generation, previous works ([56], [60]) utilize GANs to synthesize a realistic output. The pose2vid [60] network combines a convolutional image encoder and a Generative Adversarial Network (GAN) to generate photo-realistic video sequences from human poses. The generator (G) functions as an encoder-decoder, conditioned on human pose and appearance. It converts input data into realistic video frames by encoding this data into a latent space, which can be either a fixed-size one-dimensional vector or a variable-size block of residual layers. The discriminator (D) assesses whether the outputs of the generator match the distribution of real training data, helping train the generator through a minimax game. The overall training objective employs both adversarial loss (to train G against D) and L1 loss (to minimize the difference between generated and real images). The full architecture is shown in figure 3.3.8.

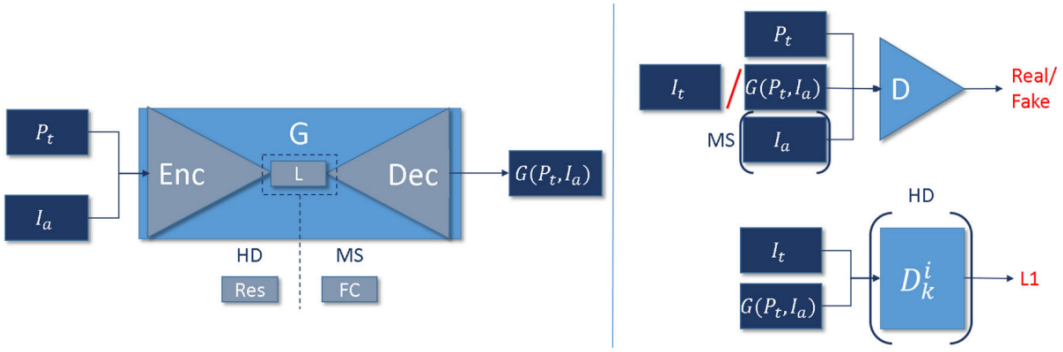


Figure 3.3.8: GAN Architecture used in Text2Sign [60]

In [56] Saunders proposes SignGAN for sign language video co-articulation. SignGAN follows a generator - discriminator architecture. The discriminator aims to evaluate the quality of the generated sign image. SignGAN enables style-controllable video-to-video signer generation given a signer target style image S, taking inspiration from [13]. The training objective combines multiple loss functions including a GAN Adversarial Loss L_{GAN}, a Feature-Matching Loss L_{FM}, a Perceptual Reconstruction Loss L_{VGG} which compares features extracted by a pre-trained VGG-Net from both generated and real images and a Hand Keypoint Loss. The full architecture is shown in figure 3.3.9.

Table 3.1 compares two GAN-based models, for the Pose-to-video SLP task, namely Text2Sign and SignGAN, using the visual-based metrics for evaluation. SSIM is a perceptual metric that measures the similarity between the generated video frames and the ground truth frames, considering three key factors of frame quality, luminance, contrast and structure. Hand SSIM is a specialized version of SSIM that focuses specifically on the hand regions in the generated videos. Hand Pose Error measures the

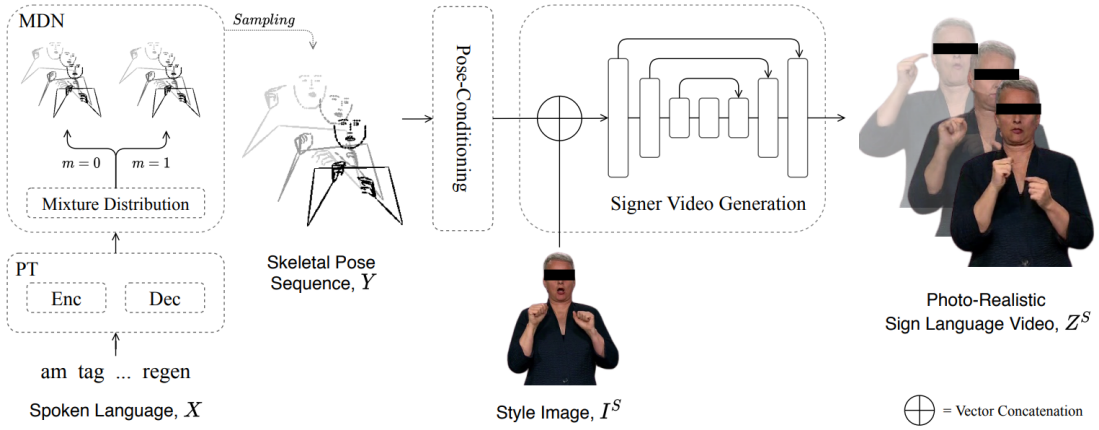


Figure 3.3.9: SignGAN [54]

deviation of the predicted hand keypoints from the ground truth keypoints, using Mean Squared Error (MSE). Lower values indicate more accurate hand pose generation. Lastly, FID measures the similarity between the feature distributions using a pre-trained Inception network.

	SSIM ↑	Hand SSIM ↑	Hand Pose ↓	FID ↓
Text2Sign [60]	0.727	0.533	23.17	64.01
SignGAN [56]	0.759	0.605	22.05	27.75

Table 3.1: Comparison results of photo-realistic sign language video generation

3.3.4 Evaluation Metrics on the SLP pipeline

The standardized evaluation of SLP systems is a critical aspect of assessing their performance and effectiveness. Since SLP involves both natural language processing and computer vision tasks, a combination of metrics from these domains is typically used. These metrics measure the quality of the generated sign language sequences, the accuracy of the translation, and the naturalness of the produced videos. Below, we provide a detailed discussion of the most commonly used evaluation metrics in SLP.

Lingual evaluation metrics are primarily borrowed from NLP and are used to assess the quality of the generated sign language sequences in terms of their linguistic accuracy and fluency. Some of the **lingual** evaluation metrics used to measure the output quality include the BLUE [46] and ROUGE [40] metrics.

In the BLEU@N metric [46], the matched N-grams between the machine-generated and the ground-truth answer are utilized to compute the precision score. BLEU@N metric is calculated for N = 1 to 4, where shorter N-grams are used to fulfill the adequacy and longer N-gram matching accounts for fluency.

$$BLUE = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3.3.5}$$

where p_n is the precision for n-grams of length n, w_n are optional weights and N is the maximum

n-gram length considered, in our case from BLUE1 to BLUE4.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation), is a set of metrics and a software package specifically designed for evaluating automatic summarization, but that can be also used for machine translation [40].

$$ROUGEL_{precision} = \frac{Length\ of\ LCS}{Length\ of\ generated\ sequence\ (\#words)} \tag{3.3.6}$$

In both metrics higher scores indicate higher similarity between the produced sentence and the reference. BLEU focuses on precision (how much the n-grams in the model output appear in the ground-truth sequence) while ROUGE focuses on recall (how much the n-grams in the ground-truth sequence appear in the model output) and usually a precision-recall trade-off is observed.

Table 3.2 shows the performance of SLP models on the Text2Gloss task, where the goal is to generate gloss sequences from spoken language text. The results indicate that SignGAN [56] achieves the highest BLEU-4 and ROUGE scores, outperforming earlier models. Table 3.3 shows the performance of SLP models on the Text2Pose task, where the goal is to generate skeletal poses from spoken language text. The results are retrieved by translating the generated poses back to text.

		Dev		Test	
Model	**Year**	**BLEU-4↑**	**ROUGE↑**	**BLEU-4↑**	**ROUGE↑**
Text2Sign [60]	2020	16.34	48.42	15.26	48.10
End-to-end [54]	2020	20.23	55.41	19.10	54.55
SignGAN [56]	2022	21.93	57.25	20.08	56.63

Table 3.2: SLP Model Evaluation - Performance Text2Gloss

		Dev		Test	
Model	**Year**	**BLEU-4↑**	**ROUGE↑**	**BLEU-4↑**	**ROUGE↑**
SignGAN [56]	2022	11.93	34.01	10.43	32.02
There-and-Back [61]	2023	17.10	40.42	16.91	40.12

Table 3.3: SLP Model Evaluation - Performance Text2Pose using Back Translation

3.3.5 Other Works on Sign Language Production

In addition to the Transformer-based approaches to SLP discussed earlier, recent publications have introduced innovative methods that leverage Machine Learning disciplines, such as vector quantization, 3D body reconstruction, and diffusion models. These approaches aim to address the limitations of earlier systems, such as the lack of realism in generated videos, the complexity of pose estimation, and the challenges of achieving high lingual accuracy. Below, we provide a brief summary of some of these cutting-edge methods and their contributions, seeing how they could benefit our work when approach photo-realistic SL synthesis.

3.3.5.1 Vector Quantization Architectures

VQ-GAN: In more recent work, Xie et al. [76] choose to completely discard the landmark sequence generation (pose estimation) and directly generate realistic **word-level** sign videos with VQ-GAN. This

work uses a two stage generation process. In the first stage, the 3D VQ-GAN with a motion transformer enables video understanding while the second stage uses sentence-to-sentence attention to perform autoregression on the flattened latent codes. In the VQ-GAN the encoded features are passed through the vector quantizer and are mapped s to the nearest discrete representations in the codebook. This quantization process introduces a bottleneck in the latent space, which helps in learning a more structured and efficient latent representation. The latent transformer performs auto-regressive training on the quantized codes to produce more accurate sequences.

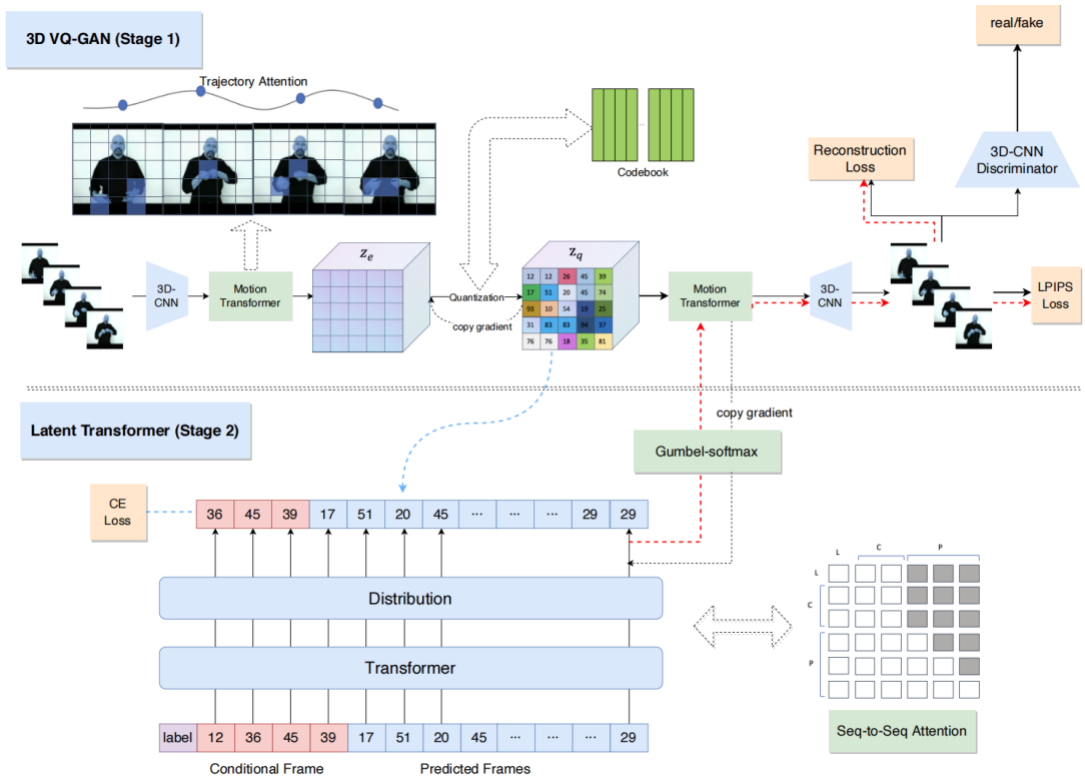


Figure 3.3.10: Figure from [76]

3.3.5.2 3D Body Reconstruction

Besides the pose-to-video frameworks that we mentioned above, recent advances in SLP enable the generation of realistic 3D signing avatars by leveraging deep learning techniques. Unlike traditional motion capture methods, modern approaches use 3D body reconstruction and diffusion models to synthesize natural signing from video or text. This section explores key methods in 3D sign avatar generation, specifically SGNify, There and Back again and Neural sign actors, which employ linguistic constraints and diffusion-based 3D generation respectively.

SGNify: The authors introduce a method called SGNify, which leverages linguistic priors to enhance the reconstruction of 3D signing avatars from monocular SL videos. SGNify emphasizes capturing detailed hand poses, facial expressions, and body movements automatically from SL videos captured in natural settings. The effectiveness of SGNify was validated through a commercial motion-capture system, showing superior performance in generating 3D avatars compared to other state-of-the-art 3D

body-pose and shape-estimation methods. A key finding from their perceptual study indicates that avatars reconstructed using SGNify are more comprehensible and appear more natural than those created using previous technologies, aligning closely with the source videos in terms of quality. SGNify is built upon SMPL-X [47] and SPECTRE [20] for 3D body parameter and facial expression extraction.

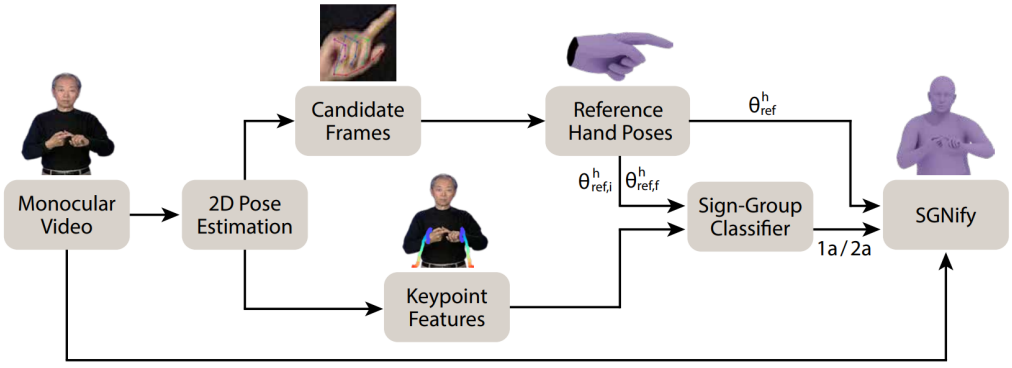


Figure 3.3.11: Figure from [21]

There and back again: The Pose2Mesh network in this paper [61] converts the predicted 2D pose sequences into 3D signing avatars using an extended version of SMPLify-X. The Pose2Mesh model uses temporal consistency between frames to ensure smooth motion and reduce artifacts and a video compression method (I and P frames) during training.

Neural Sign Actors: In [2] Baltatzis et al. propose a Diffusion based architecture that produces 3D SMPL-X avatars from text, in the architecture shown in fig 3.3.12. The model encodes text sentences with CLIP, pose embeddings with MLPs and performs autoregressive LSTM decoding to produce the 3D avatar.

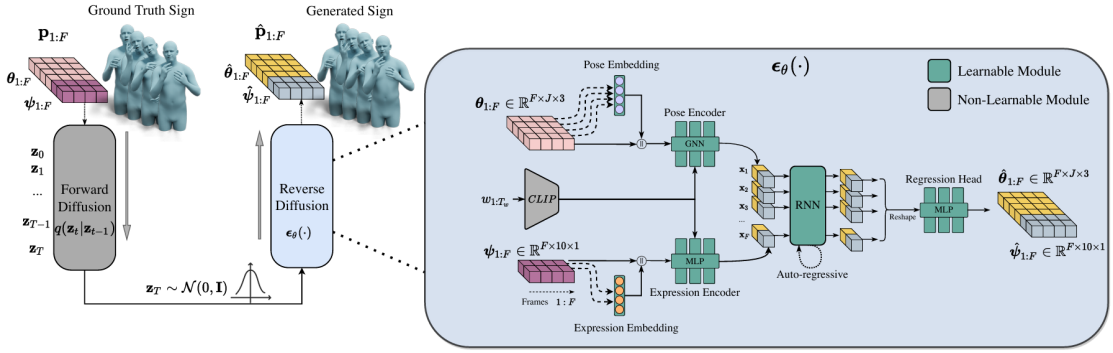


Figure 3.3.12: Neural Sign Actors Diffusion Training Process [2]

3.3.5.3 Diffusion Models

Diffusion models have emerged as a powerful tool for high-quality video generation, and their application to SLP has shown significant promise. These models work by iteratively refining noisy inputs to produce realistic outputs, making them well-suited for the sign language production tasks.

Sign-Diff: Fang et al. [18] use a Diffusion Model Architecture following the pose estimation step for photorealistic SL video synthesis. The first stage of Sign-Diff involves generating skeletal poses from the input text. This module is based on the Progressive Transformer architecture proposed by Saunders et al. [54]. The core of Sign-Diff is its use of Control-Net[78], a framework for controlled stable diffusion. Control-Net allows the model to generate high-quality videos while incorporating conditional inputs, such as skeletal poses, to guide the generation process. Control-Net operates by iteratively refining a noisy input to produce a realistic output. At each step of the diffusion process, the model incorporates the conditional inputs (e.g., skeletal poses) to ensure that the generated video adheres to the constraints imposed by the input text. Finally, Sign-Diff's method includes the use of FR-Net, a frame refinement network that selects the diffusion process inputs to achieve temporal consistency in the generated poses.

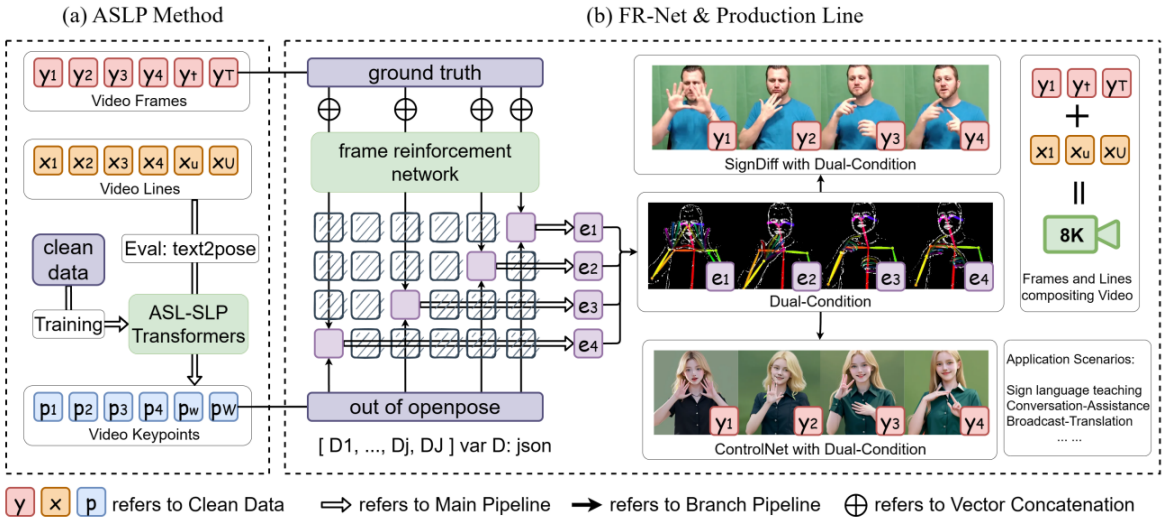


Figure 3.3.13: Figure from [18]

Neural Sign Actors: Mentioned above in the 3D reconstruction chapter, the Neural Sign Actors training process involves a novel diffusion-based architecture for sign language production. The diffusion model is trained using a combination of adversarial loss and feature-matching loss. The adversarial loss ensures that the generated avatars are realistic, while the feature-matching loss ensures that they are consistent with the input text.

MS2SL: Another recent interesting work [42], proposes the use of diffusion models for multimodal continuous sign language production, including text and speech. It uses a sequence diffusion model with embeddings from text and speech to create sign language sequences. The framework also includes an embedding-consistency learning strategy that leverages cross-modal consistency to enhance model training, even with missing audio data. The approach is validated on the How2Sign and PHOENIX14T datasets, showing competitive performance in sign language production.

Chapter 4

Proposed Method: SL Production using Transformers and Neural Rendering

Contents

4.1 Method Overview

This chapter thoroughly describes the proposed pipeline for end-to-end sign language production. The goal of our method is to introduce a novel system that allows for realistic sign language video generation from a single line of text input. To our knowledge, this is the first work on Greek Sign Language Production. To achieve our goal, we create two separate modules towards realistic SLP. First, we utilize Transformer Networks to generate (2D) Sign poses from text input. Then, we use a Neural Rendering framework to disguise the generated skeletal motion figure in a synthetic photo-realistic SL video, that uses signers from the original publicly availiable SL dataset. Figure 4.1.1 illustrates the main components of the proposed SLP network.

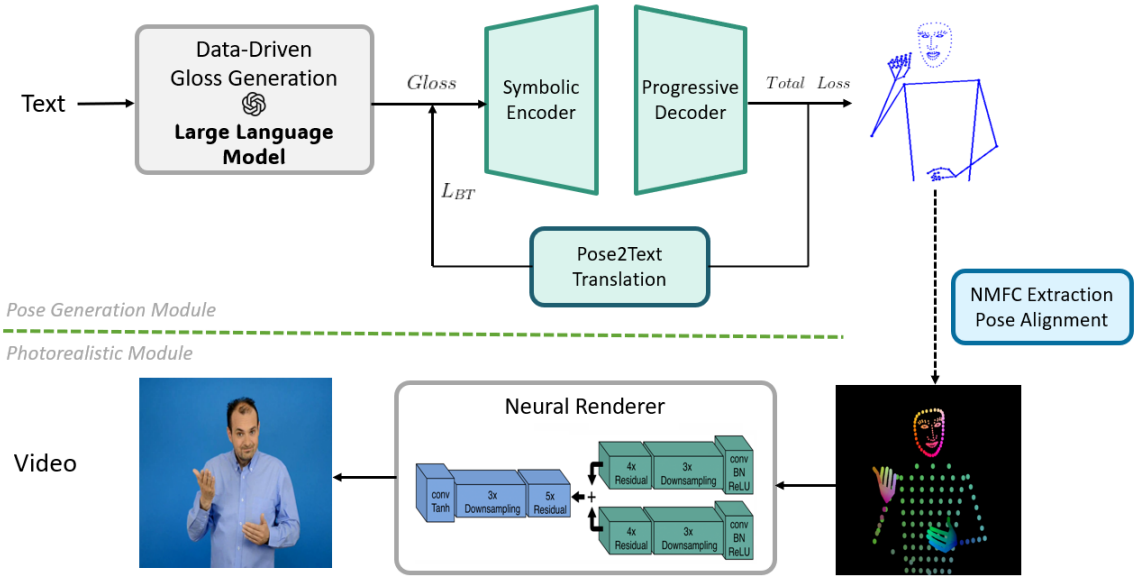


Figure 4.1.1: Overview of the proposed inference. (Top) **SLP Module**: Transformer architecture used to produce extended skeletal motion sequences. The total loss objective consists of a MSE pose loss and a pose-to-text SL translation loss. (Bottom) **Photorealistic Module**: Following the skeletal pose generation, the neural rendering framework, allows for high-quality video synthesis with respect to the original dataset sign actor, taking as input just the transformer generated sequences.

Initially, the data preprocessing stage involves creating text and pose sequence pairs which will be used to train the transformer model. We utilize existing and publicly available sign datasets, such as How2Sign (American English), Elementary23 (Greek) and PHOENIX14T (German). Feature extraction is done using MediaPipe Holistic on each dataset, and the original 578 landmarks are down-sampled to 191, as explained in section 5.

The generated pairs are then used to train a deep learning network based on the original Transformer architecture [67], that also uses the Progressive Transformer implementation [54] as the baseline model. We have incorporated a novel pose-to-text SL translation loss from inferencing on a pre-trained SLT model based on the state of the art model [10]. The optional use of data-driven generated gloss annotations (through of-the-self LLMs) seems to generally boost the performance of the model and reduce the lexical diversity of our dataset.

The photorealistic module takes the aligned NMFC color coded images and generates a video of a person performing the sign language, respective to the sign actors in the original dataset. The neural renderer uses convolutional neural networks (CNNs) with residual and downsampling layers to synthesize a high-quality video from the pose data.

4.2 Sign Language Production Module

4.2.1 Text to Video Production Module

The most critical component of the proposed SL Production pipeline arguably is the Text to Video Module. To implement this component we built upon publicly available Progressive Transformer network:

In the paper [54] Saunders et al. introduce the first Transformer-Based architecture for end-to-end SLP from given text. The architecture consists of two sequential transformers. First, the text is encoded through the Symbolic Transformer, which follows the architecture of the classic transformer model [67]. Then, the encoded input is passed through the Progressive Transformer, which is used for producing the continuous frame sequence. The Progressive Transformer uses a Counter Embedding ranging from 0 to 1 as shown in figure 3.3.6 which indicates the start and the finish of the generated frame sequence respectively. The progressive decoder is an auto-regressive model produces a sign pose frame at each time-step, along with the frame's counter value. The progressive decoder's output can be described by the equation:

$$[\hat{y}_{u+1}, \hat{c}_{u+1}] = D_P(\hat{j}_u | \hat{j}_{1:u-1}, r_{1:T}) \tag{4.2.1}$$

Finally, the overall training objective of the Progressive Transformer is concluded by calulating the Mean Squared Error (or another selected type of loss) between the groundtruth $y^*_{1:U}$ landmark sequence and the predicted landmark sequence $\hat{y}_{1:U}$:

$$L_{MSE} = \frac{1}{U} \sum_{i=1}^{u} (y^*_{1:U} - \hat{y}_{1:U})^2 \tag{4.2.2}$$

4.2.2 Teacher Forcing vs Auto-regressive Decoding

Teacher forcing is widely used technique, used during the training phase of sequence-to-sequence models like RNNs and Transformers, where the model is trained to predict the next output token given the ground truth token as input, rather than relying on its own previous predictions. During training, at each time step, the true target output from the training data is fed into the model as the input for predicting the next token. Since the correct target sequence is provided during training we are lead to faster convergence as the model is directly learning the mapping from input to the correct output sequence. We also have to be mindful of the possible exposure bias during inference. If the model is being trained solely on ground truth landmarks sequences, and is dependent on making inference

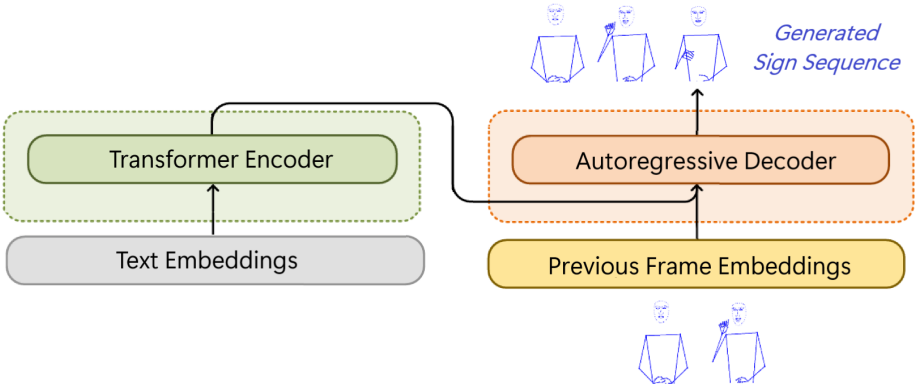


Figure 4.2.1: Transformer-Based Sign Language Production (Text-to-Video) Module

predicts in a true auto-regressive matter, we may observe poor performance in the inference on the test set. For this reason, we will explore the concept of harnessing the benefits of auto-regressive decoding during training below.

Autoregressive Decoding is a strategy used during inference where the model generates the output landmark sequence frame by frame, using its own previous predictions as inputs for future time steps. This is a widely used method with models like RNNs, Transformers during the inference for sequence generation purposes. During inference, the model generates the output sequence step by step. At each time step t, the model predicts the next token based on the previously generated tokens. Since autoregressive decoding reflects the actual inference procedure of generating outputs sequentially, it reduces the risk of exposure bias that teacher forcing introduces. The model learns to cope with its own predictions, even if they are imperfect, making it more robust during inference. The main disadvantage of using during training is the inherently slower inference since the model generates one token at a time.

In previous methods [54], transformer models were trained using teacher forcing. This approach involves providing the model with the ground truth spatial embeddings from the previous frame during sequence generation. By using the correct embeddings as input, this method enables parallel training with known outputs. While teacher forcing has demonstrated satisfactory results on limited vocabulary datasets such as PHOENIX14T [54, 10], it struggles with the broader and more diverse Greek Sign Language dataset. In general, while teacher forcing provides better training stability and ensures alignment between inputs and outputs—particularly in the earlier stages of training—it suffers from error compounding during inference, as the network is unable to recover from its own prediction errors.

On the contrary, autoregressive training generates frame sequences sequentially during training as well. In this approach, the model predicts each frame by conditioning on the spatial embeddings it has previously generated. Before applying the MSE loss, the entire sign sequence is generated from the text embeddings, effectively mimicking the inference process. This allows the model to learn to correct its own errors rather than relying on ground-truth inputs. However, this training process is considerably more time-consuming than teacher forcing due to the sequential nature of frame generation.

To balance efficiency and effectiveness, we employed a hybrid approach, training the model using teacher forcing and autoregressive generation for half of the epochs each. Specifically, we began training with teacher forcing to leverage its stability and strong input-output alignment during the critical early stages of training. This ensures the model effectively learns the foundational relationships in the data. We then switched to autoregressive training, allowing the model to learn to correct its own errors and better handle the challenges of inference. This strategy combines the strengths of both methods, resulting in improved performance compared to using either method independently. In this thesis, we aim to perform experiments with different balances of both teacher forcing and auto-regressive training in order to achieve the best possible results on SLP.

4.2.3 Data-driven Gloss Generation

Based on insights from previous research, we also explored the use of off-the-shelf large language models (LLMs) to automatically generate gloss annotations for the SL dataset. This approach effectively reduces the lexical diversity of the dataset by condensing commonly used words, such as articles and connective phrases, while preserving the overall meaning of the sentences. Table 4.1 presents examples of these translations along with the corresponding prompts used to generate them. Given the established benefits in previous literature, where gloss annotations as an intermediate step have been shown to enhance model performance, we anticipate observing similar improvements in our experiments.

Table 4.1: LLM Gloss Generation Examples

Prompt	Transform this Greek sentence into Greek Sign Language gloss: "ο άξονας συμμετρίας χωρίζει ένα σχήμα σε δύο ίσα μέρη"
Gloss	ΑΞΟΝΑΣ ΣΥΜΜΕΤΡΙΑ ΧΩΡΙΖΕΙ ΣΧΗΜΑ ΔΥΟ ΙΣΑ ΜΕΡΗ
Prompt	Transform this Greek sentence into Greek Sign Language gloss: "συμπληρώνω τον πίνακα υπολογίζοντας πρώτα τις τιμές στο περίπου ελέγχω στη συνέχεια τους υπολογισμούς μου"
Gloss	ΣΥΜΠΛΗΡΩΝΩ ΠΙΝΑΚΑΣ ΥΠΟΛΟΓΙΖΩ ΠΡΩΤΑ ΤΙΜΕΣ ΠΕΡΙΠΟΥ ΕΛΕΓΧΩ ΜΕΤΑ ΥΠΟΛΟΓΙΣΜΟΙ ΜΟΥ
Prompt	Transform this Greek sentence into Greek Sign Language gloss: "παρατηρώ και συνεχίζω τα μοτίβα "
Gloss	ΠΑΡΑΤΗΡΩ ΣΥΝΕΧΙΖΩ ΜΟΤΙΒΑ

4.2.4 Video-to-Text Translation Module

An integral component of the proposed training pipeline is the implementation of the Video-to-Text Translation Loss. This approach entails the pre-training of a distinct Translation model that maps MediaPipe features to text, which is subsequently utilized during the training of the text2pose model with the objective of enhancing accuracy. The inclusion of this step was a natural progression in the construction of the training pipeline, given that an SLT pose2text model is already required for evaluation purposes, such as computing BLEU scores. Therefore, it is only logical to assume that incorporating a Video-to-Text translation loss to the training process is going to only affect positevely the results of the SL production module.

In order to implement the back translation model we built on the state-of-the-art, publicly availiable network Sign Language Transformers by Camgoz et al. [10]. Simplifying the overall training process

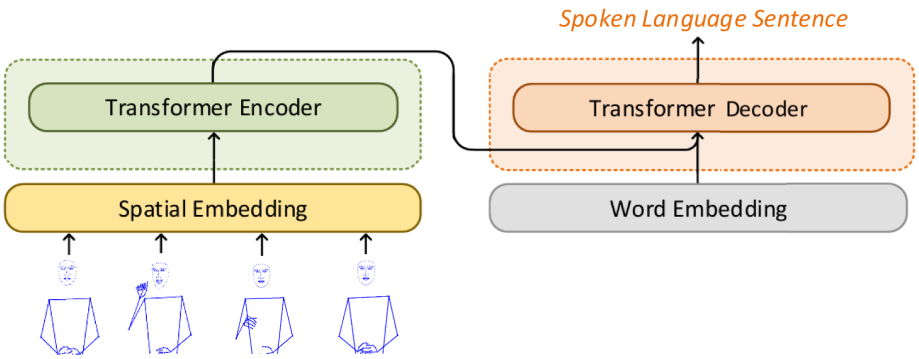


Figure 4.2.2: Transformer-Based Sign Language Translation

that performs both SL recognition and translation, we keep solely the translation loss objective, aiming to achieve the desired results through a direct sign2text model.

The core objective of the Video-to-Text module is to enhance accuracy and prevent the model from regressing to mean pose, which often happens when only training with MSE loss, and also prove its ability to reinforce the quality of the forward translation. The translation loss, essential for both training and evaluation, is formulated following [10] as follows:

$$L_T = 1 - \prod_{u=1}^{U} \sum_{d=1}^{D} p(\hat{w}_u^d) p(w_u^d | h_u) \tag{4.2.3}$$

where $p(\hat{w}_u^d)$ is the probability of word w^d at decoding step u, while D is the vocabulary size. $\prod_{u=1}^{U} p(w_u^d | h_u)$ is calculated by sequentially applying CTC Loss on a frame level for each word.

Overall, we have created a two-way translation system that can successfully produced sign sequences from text and then translate those generated sign sequences back to a text format, used both for training and evaluation purposes. Figure 4.2.3 shows the logical flow of the described translation process, using only transformer modules:

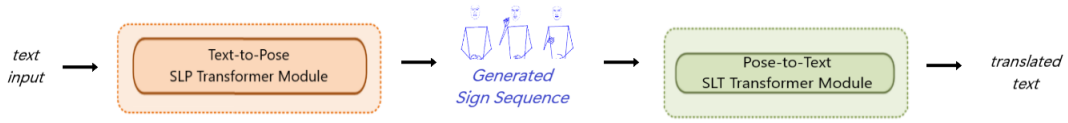


Figure 4.2.3: Transformer-Based Sign Language Translation

4.3 Photo-realistic Module

4.3.1 Head2head GAN Architecture

We built a person-specific neural rendered for SL realistic video generation based on the publicly available Head2head [37],[15] network, which was also proven highly effective in Tze's anonymization

network [64]. The network consists the following components, which are briefly described below:

- **Generator G:** Synthesizes the t-th frame, at time step t, given the current conditional input x_t, as well as the 2 previous inputs x_{t-1}, x_{t-2} along with their 2 previous generated outputs y_{t-1}, y_{t-2}. Thus the generated frame y_t at time step t is formulated as follows:

$$\tilde{y}_t = G(x_{t-2:t}, y_{t-2:t-1}) \tag{4.3.1}$$

- **Image Discriminator D_I:** Inspired form pix2pixHD [70] the Image Discriminator is used to separate the generated pair $(x_t, \tilde{y}_t)$ from the real pair (x_t, y_t) at each time step t, classifying it as either real or fake.

- **Sign Features Discriminators:** Following the architecture of the Image Discriminator, we also define separate discriminators, for the hands D_h, mouth D_m and eyes D_e. The corresponding regions are fed to the network using cropping around the predefined body and face landmarks, hoping to enhance high-detailed hand and face generation.

- **Dynamic Discriminator D_D:** The dynamics discriminator is trained to detect videos with unrealistic temporal dynamics. This network receives a set of K consecutive real frames $y_{t:t+K-1}$ or fake frames $\tilde{y}_{t:t+K-1}$ as its input, which are randomly drawn from the video. Given the optical flow $w_{1:T-1}$ and the ground truth video $y_{1:T}$, the Dynamic Discriminator ensures that the flow $w_{t:t+K-2}$ corresponds to the given video clip, by classifying the pair $(w_{t:t+K-2}, y_{t:t+K-1})$ as real and the pair $(w_{t:t+K-2}, \tilde{y}_{t:t+K-1})$ as fake.

Training Objective The total objective of the network consists of a mixture of four losses:

$$\mathcal{L}_G = \mathcal{L}_G^{\text{adv}} + \lambda_{\text{vgg}}\mathcal{L}_G^{\text{vgg}} + \lambda_{\text{feat}}\mathcal{L}_G^{\text{feat}} + \lambda_{\text{face}}\mathcal{L}_G^{\text{face}} \tag{4.3.2}$$

where $\lambda_{\text{vgg}} = \lambda_{\text{feat}} = \lambda_{\text{face}} = 10$ and each function is defined below.

The first loss is an adversarial GAN loss, which follows LSGAN [45] using the 0-1 binary coding scheme (labels b = c = 1 for real samples and label a = 0 for fake ones) which results in the following objective for the Generator:

$$\mathcal{L}_G^{\text{adv}} = \frac{1}{2}\mathbb{E}_t\left[\begin{array}{l}\Big(D_I(X_t, \tilde{Y}_t) - 1\Big)^2 + \Big(D_m(X_t^m, \tilde{Y}_t^m) - 1\Big)^2 \\ \quad + \Big(D_h(X_t^h, \tilde{Y}_t^h) - 1\Big)^2 + \Big(D_e(X_t^e, \tilde{Y}_t^e) - 1\Big)^2\end{array}\right] \tag{4.3.3}$$

The next two losses of the total function are a VGG loss and a feature matching loss. The VGG loss is computed by using the VGG network to extract visual features in different layers for both the ground-truth frame y_t and the synthesised frame $\tilde{y}_t$. The feature matching loss is computed by extracting features with the Image Discriminator D_I and computing the l_1 distance of these features for a fake frame $\tilde{y}_t$ and the corresponding ground truth y_t.

To further improve our results, additional to the VGG loss and face discriminators, we incorporated an identity loss similar to the SMIRK framework [50], implemented using a ResNet-50 model pre-trained on the VGG-Face2 dataset [11],[19]. This perceptual loss ensures that the synthesized frames retain identity consistency, preventing artifacts where facial features change unnaturally across frames.

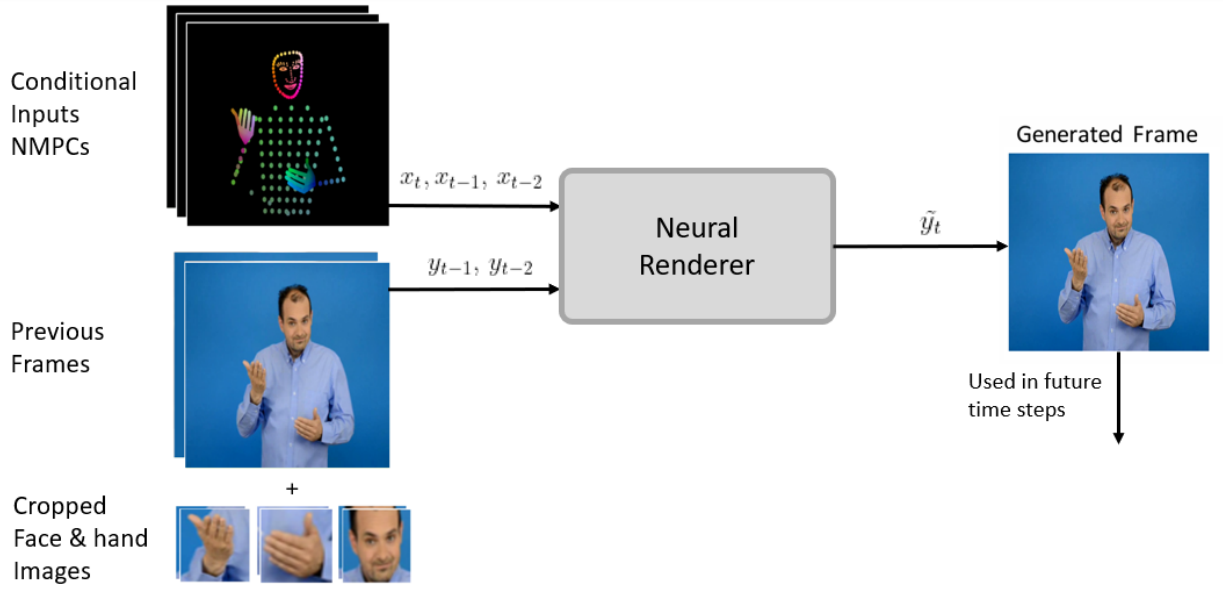


Figure 4.3.1: Generator architecture for the neural rendered

4.4 Summary

In this chapter, we presented our pipeline targeted to Sign Language Production from text, which combines Transformer text-to-pose generation with neural rendering to generate photo-realistic sign language videos. Our method introduces a novel approach to Greek Sign Language Production, leveraging a hybrid training approach that combines typical teacher forcing with auto-regressive decoding for pose generation. Additionally, we explored the use of data-driven gloss generation through large language models to reduce lexical diversity and enhance model accuracy. The inclusion of the video-to-vext Translation Loss further ensures that the model doesn't regress to the mean output pose.

The Photo-realistic Module employs a GAN architecture, inspired by the Head2head framework, to synthesize realistic videos of signers performing the generated poses. By incorporating multiple discriminators for different facial and hand regions, as well as a dynamic discriminator for temporal consistency, we ensure that the synthesized videos are both visually and temporally coherent. The use of perceptual losses, such as VGG and identity loss, further ensures that the synthesized signers retain natural-looking movements.

Overall, our method represents a significant step forward in the field of Sign Language Production, particularly for Greek Sign Language, which has not been extensively explored in prior research. By combining advanced Transformer architectures with neural rendering techniques, we have developed a system capable of generating realistic and accurate sign language videos from text, with potential

applications in education, accessibility, and communication for the deaf and hard-of-hearing community. Future work could focus on expanding the dataset, improving the robustness of the model across different signers, and exploring additional applications of this technology in real-world scenarios.

Chapter 5

Experiments

Contents

5.1 Datasets

While there are some large-scale and annotated datasets available for sign language translation (OpenASL [57], Youtube-ASL [66]), there are only a few publicly available large-scale datasets that have been used for continuous sign language production. After carefully considering our options for publicly available SL datasets we chose the following datasets:

- German Sign Language: **PHOENIX14T** [36] is a German Sign Language (GSL) dataset which contains weather forecast airings from 2009 to 2011, transcribed with gloss notation. It is the dataset that was used in Saunders and Stoll's previous works ([54], [56], [60], [61]) which were the first to tackle the Continuous Sign Language Production Task using Neural Machine Translation.
- American Sign Language: **How2Sign** [16] is a multimodal continuous American Sign Language (ASL) dataset. It constists of a parallel corpus of more than 80 hours of sign language videos and the corresponding modalities include speech, English transcripts, and depth. Its contents mainly consist of instructional videos with a wide vocabulary variety.
- Greek Sign Language: Regarding the Greek Sign Language, **Elementary23** [68] contains translation pairs based on the official syllabus of Greek Elementary School and has been used for transformer-based SLT. Its contents are separated based on the syllabus subject (i.e Maths, Greek Language) we enables us to perform subject- specific SLP training. This dataset has not yet been used on the sign Language Production Task and its one of the few Large scale datasets existing on Greek SL.

Dataset	Language	Year	Video	Text	Gloss
PHOENIX14T [36]	German SL	2014	✓	✓	✓
How2Sign [16]	ASL	2021	✓	✓	✗
Elementary23 [68]	Greek SL	23k	✓	✓	✗

Table 5.1: Sign Language Datasets Availiabily

Dataset	Vocab Size	# Signers	Lenghth (hours)	Frame Size
PHOENIX14T [36]	3k	9	11	210 x 260
How2Sign [16]	16.5k	11	79	1280 x 720
Elementary23 [68]	23k	9	71	1280 x 720

Table 5.2: Sign Language Datasets Size Details

Focusing on Greek SL, the Elementary23 dataset [68], extensively used in this thesis, contains annotations of the first three classes of Greek Elementary school books in all subjects, with a large vocabulary exceeding 30,000 words. In our work, we focus on The Greek Language subset which contains 9499 videos with a vocabulary of 14345 words, and the Math subset which contains 6583 videos with a vocabulary of 6457 words. Specifically for the Math subset, we begin our evaluation, by training the SLP pipeline on the two most prominent signers, referred to as Signer A and Signer B, who appear in 3,476 and 746 videos, respectively. Then, we generalize our training process across all signers to achieve a more holistic result. Table 5.3 visualizes the size and vocabulary of each subset used, while Figure 5.1.1

shows the original distribution of signers in Elementary23.

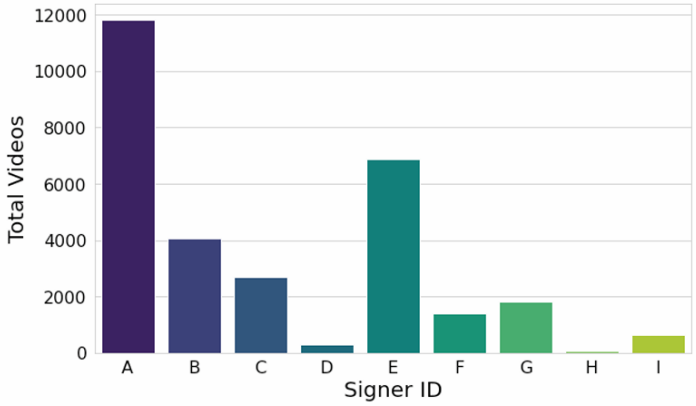


Figure 5.1.1: Signer Distribution inside the entirety of the Elementary23 Dataset. Since the Greek Language and Math subsets contain mostly videos of Signer A and Signer B, we select those for our future training[68]

		Videos	# Words
Math	**Signer A**	3473	3654
	Signer B	746	1059
Greek	**Signer A**	2467	4749
	Signer B	1927	3535

Table 5.3: Defined Elementary23 subsets used in this thesis

5.2 Feature Extraction and Data Preprocessing

We use **MediaPipe Holistic** [26] for landmark extraction on each SL video. As mentioned in chapter 2, MediaPipe Holistic employs a graph-based pipeline that processes different regions of interest (ROIs) within an image to estimate a total of up to 543 landmarks. These include 33 body pose landmarks, up to 468 facial landmarks, and 42 hand landmarks (21 per hand). Its also worth mentioning that MediaPipe Holistic runs on just the CPU, taking approximately 3 seconds of inference per frame.

In order to expedite the training process we sub-sample both the pose and face mesh landmarks. For the pose keypoints, we select the 8 points shown in figure 5.2.1 which include the body parts necessary for a SL video, such as the torso, elbows and wrists. For the face landmarks, we choose 141 instead of 468 face keypoints, which contain all the necessary face information, such as the mouth, eyes, nose and face perimeter. For each hand, we keep all 21 landmarks. This brings us to a total of 191 landmarks, instead of the original MP 468 landmarks, which is a substantial reduction to nearly one third. Finally, the total extracted landmark sequence for each frame is defined as follows:

$$\mathbf{P}_f = [\mathbf{a}_{left\ hand}||\mathbf{a}_{right\ hand}||\mathbf{a}_{face}||\mathbf{a}_{pose}||c_f] \tag{5.2.1}$$

where P_f is the landmark sequence for the f-th frame, c_f is the counter value ranging from 0 to 1 that indicates the relevant frame posisition and || the concatenation symbol.

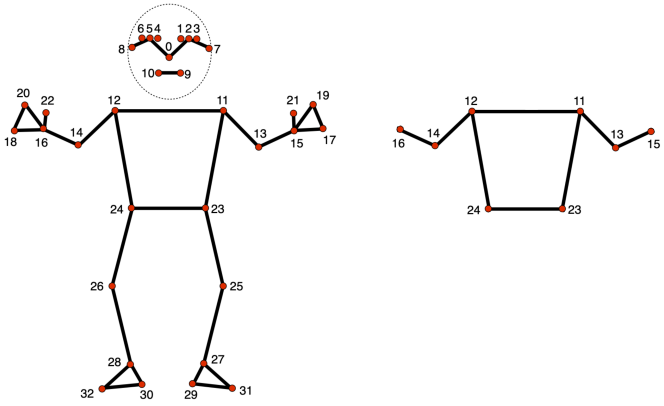


Figure 5.2.1: (left) Original 33 MP pose landmarks. (right) **Selected** 8 MP pose landmarks for SLP

Figure 5.2.2: (left) Original 478 MP face landmarks. (right) **Selected** 141 MP face landmarks for SLP

5.3 Pose Retargeting and CCBR extraction

After the initial step of pose estimation and the curation of the dataset for the Text-to-Pose models, we also have to prepare the data for the Neural Rendering process.

Since each SL dataset contains two or more distinct signers, when motion retargeting from one to another, we must consider possible differences in their bodily anatomy or a difference in camera-body placement in the original videos.

For this reason we adapt pose retargeting methods for the face and body landmarks seperetally, following the Procrustes Analysis methods used in Head2head [37] for face and Neural Sign Reanactor [64] for hands and body.

For the face retargeting, we focus on rigid facial regions less affected by deformations due to expressions. Using a subset of n rigid facial landmarks, we align the source and target faces to a mean facial template through Procrustes Analysis. At each frame, this alignment applies isotropic scaling, translation, and rotation transformations to minimize the disparity between the source and target landmarks. The aligned landmarks are then refined using a geometric median across all frames to compute median source and target face templates. Non-uniform scaling factors are subsequently calculated to match the spatial dimensions of the median source face to the median target face. These scaling factors are then applied frame-wise to adjust the source facial landmarks while ensuring proper alignment with the target's anatomy.

For the body retargeting, including torso and hands, we also adopt a similar Procrustes-based approach. Here, body landmarks are aligned with a mean skeletal template, and transformations are computed to preserve the relative spatial dynamics of the source pose while ensuring compatibility with the target's skeletal proportions.

After performing retargeting where necessary, we generate the corresponding color-coded frames, used to condition the neural rendering process. These are RGB images, where each keypoint is visualized as a fixed-radius disk with a unique color assigned through a predefined color-coding scheme. To define our color scheme, we use the normalized x,y 2D MediaPipe landmarks in the space [0,1] for the red and green values, and pre-define the blue coordinate. For the face, we utilize the triangular mesh representation from MediaPipe and assign colors to the mesh triangles based on the transformed vertex positions after retargeting. Each joint retains a fixed color across signers, ensuring consistent and semantic representation in the RGB space, which aids the renderer in learning the mapping to output images.

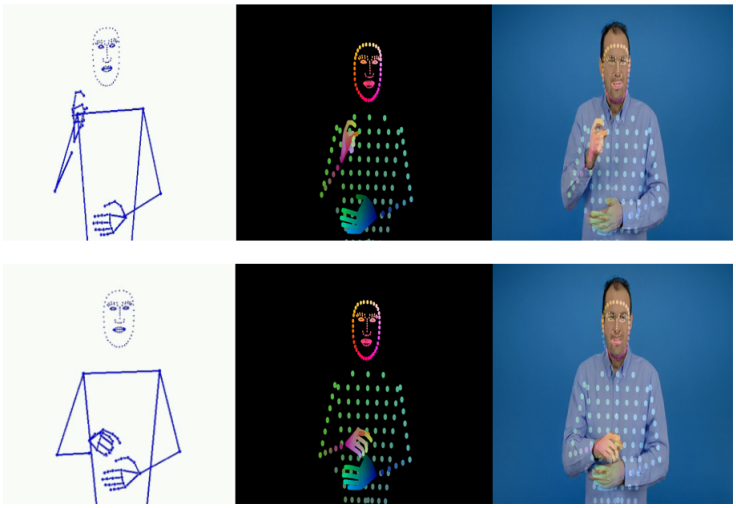


Figure 5.3.1: Example of MediaPipe extracted and sub-sampled keypoints (left), example of **retargeted** and **color-coded** extended skeletal pose used for renderer conditioning (center) and superimposition with the RGB reference frame (right).

5.4 Training and Implementation Details

The transformer models in the Text to Pose and backwards generation use the following configuration:

All models have been trained using 2-layer transformers with 4 attention heads, embedding dimension 512 and all weights are initialized with Xavier. All SL Production and SL Translation models used in a specific pipeline and data subset are trained on the same network specifications (i.e. vocab size, embedding dimensions) for model compatibility. Most SL transformer modules contain approximately 20 million parameters which ensures a balance between computational efficiency and modeling capability. This manageable parameter count enables the models to be effectively trained while maintaining high

correlation with the dependent data subsets.

The training process uses Adam optimization with an initial learning rate of 10^{-3}, which is dropped using a scheduler depending on the resulting BLEU-4 scores during evaluation, with a 50-epoch patience and decrease factor of 0.95, till reaching a learning minimum. Both Text-to-Pose and reverse generation models are trained with a batch size of 8, requiring 3 to 4 days depending on the alterations between auto-regressive decoding and teacher forcing.

Next, regarding the neural rendering process:

The neural rendering process focuses on generating realistic signer-specific video output. The video rendering network is trained on a signer-specific 8 minute concatenated video, chosen from the Elementary23 dataset. The network is trained on given signer for 100 epochs, and the training is complete in approximately 4 days on a NVIDIA GeForce GTX 1080 Ti GPU. The network is optimized using Adam wit an initial learning rate $\eta = 2 \cdot 10^{-4}$, $\beta_1 = 0.5$ and $\beta_2 = 0.999$.

5.5 Evaluation

5.5.1 Evaluation Methods

As mentioned in chapter 3, both SLT and SLP tasks use metrics from typical NLP seq to seq expirements in order to perform evaluation. Some of the **lingual** evaluation metrics used to measure the output quality include the BLUE [46] and ROUGE [40] metrics.

5.5.1.1 The BLEU Metric

In the BLEU@N metric [46], the matched N-grams between the machine-generated and the ground-truth answer are utilized to compute the precision score. BLEU@N metric is calculated for N = 1 to 4, where shorter N-grams are used to fulfill the adequacy and longer N-gram matching accounts for fluency. The BLEU@N score is calculated as the product of the Brevity Penalty and the translation Geometric Average Precision Score. Formally, the BLUE score can be calculated as follows:

$$BLUE = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{5.5.1}$$

where p_n is the precision for n-grams of length n, w_n are optional weights and N is the maximum n-gram length considered, in our case from BLUE1 to BLUE4.

5.5.1.2 The ROUGE Metric

ROUGE (Recall-Oriented Understudy for Gisting Evaluation), is a set of metrics and a software package specifically designed for evaluating automatic summarization, but that can be also used for machine translation [40]. ROUGE-L, which is used in our case, measures the overlap percentage of Longest Common subsequences between groundtruth and generated sentences. More formally ROUGE-L precision score is defined as:

$$ROUGEL_{precision} = \frac{Length\ of\ LCS}{Length\ of\ generated\ sequence\ (\#words)} \tag{5.5.2}$$

In both metrics higher scores indicate higher similarity between the produced sentence and the reference. BLEU focuses on precision (how much the n-grams in the model output appear in the ground-truth sequence) while ROUGE focuses on recall (how much the n-grams in the ground-truth sequence appear in the model output) and usually a precision-recall trade-off is observed.

5.5.1.3 Dynamic Time Warping

Dynamic Time Warping (DTW) is a time-series alignment technique used to measure the similarity between two sequences that may vary in time or speed. In the context of sign language production, DTW plays a crucial role in aligning motion trajectories of skeletal keypoints between the generated and ground-truth signs, specifically for evaluation purposes.

5.5.2 Compared Benchmarks

As mentioned, we aimed to evaluate the proposed pipeline as extensively as possible, conducting experiments on all three used datasets. We compare our method with at least three previously existing benchmarks on SLP, if those are available. The benchmarks shown where chosen based on their relevance and similarity to our method. It's important to also clarify which task (Translation or Production) each benchmark is referring to, since the differ in difficulty in implementation. Since PHOENIX14T is one of the most popular datasets for Sign Language Processing we considered it almost necessary to include it in our evaluation method.

5.5.2.1 Sign Language Translation Benchmarks

First, we structure our evaluation process by presenting results on the Sign Language Translation module. It's important to note the models resulting in this section are obtained from models solely trained on ground-truth skeleton sequences. This increases model performance and also follows accurately benchmarks presented in previous literature. Table 5.4 shows our SLT results on the three datasets, How2Sign (ASL), PHOENIX14T (German SL) and Elementary23 (Greek SL). The ground-truth keypoints for the PHOENIX14T set-up where retrieved from the Sign-Diff repository [18] for faster results during this thesis. For the How2Sign experiments, we extract data keypoints ourselves as described previously, only excluding the face landmarks (75 in total) for expedited training.

The evaluation results for Sign Language Translation (SLT) demonstrate that the proposed model performs competitively across multiple datasets. In the PHOENIX14T (German SL) dataset, the model achieves a BLEU-4 score of 21.22. This indicates strong translation quality comparable to previous literature benchmarks. However, in the How2Sign (ASL) dataset, the proposed model scores 8.53 BLEU-4, which is slightly lower than the reported 14.9 BLEU-4 in the most recent (sota) work. This discrepancy may be attributed to differences in pre-processing, such as the exclusion of face landmarks for expedited training. On the other hand, for the Elementary23 (Greek SL) dataset, the proposed model shows

Method	Dev		Test	
	BLEU-4↑	**ROUGE↑**	**BLEU-4↑**	**ROUGE↑**
How2Sign (SLT) [52]	-	-	14.9	36.0
ours	9.01	22.34	8.53	25.22
Progressive Transformers [54] (SLT)	20.23	55.41	19.10	54.55
Sign Language Transformers [10] (SLT)	22.38	-	21.32	-
ours	21.53	-	21.22	-
Elementary23 SLT (Voskou et al. [68])	6.67	-	5.69	-
Elementary23 SLT (ours)	8.34	32.36	**8.2**	32.16

Table 5.4: SLT Evaluation Banchmarks

a significant improvement over prior results, reaching 8.2 BLEU-4 compared to the previous original paper's 5.69 BLEU-4, suggesting a good adaptation to Greek Sign Language.

5.5.2.2 Sign Language Production Benchmarks

Naturally, when switching to SLP, we expect BLUE-4 scores to slightly drop as it's a more complex and challenging ML task. As previously mentioned in chapter 4 the SL translation loss is used during our SLP training along with the mean squared error differences, so a drop in BLUE-4 scores is justified. It's also important to note that the results shown in tables 5.4 and 5.5 are not directly comparable with previous literature, as the selected test set often differs across publications. However the scores remain clear indicators of how well are models perform.

Method	Dev		Test	
	BLEU-4↑	**ROUGE↑**	**BLEU-4↑**	**ROUGE↑**
Neural Sign Actors [2]	13.12	47.55	13.12	47.55
SignDiff [18]	16.92	49.74	15.92	46.57
MS2SL [42]	4.26	16.38	4.26	16.38
ours	4.5	12.36	**4.48**	12.16
Progressive Transformers [54]	11.82	33.18	10.51	32.46
There and Back again [61]	17.10	40.42	16.91	40.22
ours	12.34	33.98	12.51	34.21
Elementary23 SLT [68]	6.67	-	5.69	-
Elementary23 Math (ours)	7.58	15.11	**7.69**	15.26
Elementary23 Greek Lang (ours)	5.63	14.56	**5.52**	14.23

Table 5.5: SLP Evaluation Banchmarks

For Sign Language Production (SLP), as expected, the BLEU-4 scores generally decrease due to the complexity of generating sign language sequences. In the How2Sign dataset, the proposed model achieves 4.48 BLEU-4, which is slightly higher than MS2SL but still below the Neural Sign Actors and SignDiff models. Despite this, the proposed model performs well in other benchmarks, notably in the Elementary23 (Greek SL) dataset, where it achieves BLEU-4 scores of 7.69 (Greek Language subset) and 5.52 (Math subset), both well competing with the previous 5.69 SLT score.

5.5.3 Ablation Studies

In order to test the effect of each proposed component on the overall accuracy we perform several ablation studies on the Elementary23 dataset, presented below in detail.

To begin our SLP ablation analysis, we performed experiments on the Mathematics subset of the Elementary23 dataset, on the two most frequently appearing signers, Signer A and Signer B. To assess the model's ability to generalize across different signers, we performed two separate training sessions: one focused exclusively on Signer A and the other on Signer B. For evaluation, we alternated between their respective test sets to measure cross-signer performance. Results are shown in table 5.6. We clearly see that the model fails to produce accurate signs when the "wrong" test set is used. These findings suggest that the model struggles to generalize across signers, likely due to differences in signing styles or vocabulary correlations unique to individual signers. To address this limitation, we proceed to train our models on larger sections of the Elementary23 dataset, emphasizing the need for more generalized training approaches.

	Test - Signer A		**Test - Signer B**	
	BLEU-1↑	**BLEU-4**↑	BLEU-1↑	**BLEU-4**↑
Train - Signer A	17.05	**5.02**	5.93	0.00
Train - Signer B	6.29	1.18	21.87	**6.69**

Table 5.6: Ablation Study on the Elementary23 Greek Language SL Dataset. Best-performing results are highlighted in bold, while failure scores in the case of swapped signer test scores are shown in red.

Our next ablation study, shown in table 5.7, on the Elementary23 Greek Language Subset, shows that the inclusion of pose-to-text Loss and Gloss annotations possessively affects on performance. Although BLEU-4 scores improve independently (4.42 dev, 4.55 test), the combination with Gloss yields mixed results, slightly reducing BLEU-4 on the dev set (4.06) but improving on the test set (4.32). This interplay suggests that while gloss annotations simplify linguistic diversity, over-reliance on glosses can limit adaptability. This study includes the entirety of the Greek Language subset, addressing the dataset's signer and vocabulary dependencies.

		Dev	**Test**
$L_{Video2Text}$	*Gloss*	**BLEU-4**↑	**BLEU-4**↑
✗	✗	4.17	4.15
✗	✓	3.56	3.44
✓	✗	4.42	**4.55**
✓	✓	4.06	**4.32**

Table 5.7: Ablation Study on the Elementary23 Greek Language SL Dataset

We show a similar ablation study, in table 5.8, this time conducted on the Elementary23 Math Subset. Again, The results demonstrate that $L_{Video2Text}$ is a critical component for achieving high performance in our SLP pipeline, while The slight reduction in performance when combining $L_{Video2Text}$ and LLM retrieved gloss annotations suggests that there may be some redundancy or misalignment between the two components.

An example of the Video2Text Loss is show in figure 5.5.1. We ultimately observe that the use of the

$L_{Video2Text}$	$Gloss$	Dev BLEU-4↑	Test BLEU-4↑
✗	✗	3.17	3.15
✗	✓	4.36	4.44
✓	✗	5.42	**5.55**
✓	✓	5.12	**5.06**

Table 5.8: Ablation Study on the Elementary23 Math SL Dataset

CTC Loss added to the MSE training objective increases movement variability on the generated signs. While changes in body pose and head are minimal, the use of CTC Loss during training suggests the improvement of movement in the hands during specific representation of SL sentences from the dataset.

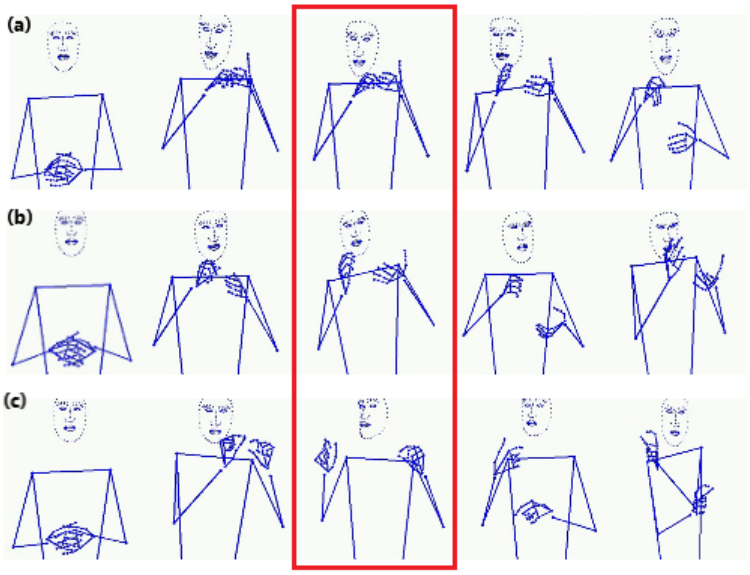


Figure 5.5.1: Sample visualization of the effect of the pose-to-text Loss. Top to bottom: **(a)** 2D Pose w/o pose-to-text Loss, **(b)** 2D Pose with pose-to-text Loss, **(c)** ground-truth sequence reference. When used, the generated poses show greater movement variability and regress less on mean pose.

Next, we focus on conducting experiments on entire sections of the Elementary23 dataset, disregarding the fact that videos are filmed with different signers. We specifically choose the entire Math subset and the Greek Language subset. Our first ablation study, shown in table 5.9, compares training with Teacher Forcing (TF), Auto-regressive Decoding (AD), and their combination (TF+AD), underlining the benefits of employing a hybrid approach.

While Auto-regressive Decoding achieves significantly higher BLEU-4 and ROUGE scores (5.4 and 14.5 on the dev set, respectively) compared to Teacher Forcing (1.69 and 8.52), the hybrid TF+AD model provides balance between computational efficiency and predictive accuracy. Notably, the hybrid model achieves the highest overall performance, both in the Greek and Math subsets, validating the importance of alternating decoding strategies during training.

Following the ablation study in table 5.9 we shown that the hybrid model of TF+AD successfully increases the quality of generated pose signs. Another way of performing quantitative evaluation is by

Subset	Method	Epochs	Time/ Epoch (s)	Dev BLEU-4↑	Test BLEU-4↑
Greek	Teacher Forcing, (PT [54])	2500	5	0.49	0.35
	Autoregressive Dec	2500	30	4.3	4.13
	TF + AD	1250 + 1250	5, 30	**4.67**	**4.46**
Math	Teacher Forcing, (PT [54])	2500	5	1.69	1.46
	Autoregressive Dec	2500	30	5.4	5.3
	TF + AD	1250 + 1250	5, 30	**5.69**	**5.59**
Signer A	Teacher Forcing, (PT [54])	3000	5	0.98	0.00
	Autoregressive Dec	3000	30	2.12	2.06
	TF + AD	2000 + 500	5, 30	**5.23**	**5.02**
Signer B	Teacher Forcing, (PT [54])	3000	5	0.18	0.18
	Autoregressive Dec	3000	30	2.5	2.2
	TF + AD	2000 + 500	5, 30	**6.7**	**6.69**

Table 5.9: Ablation Study on the Elementary23 Greek (Top) and Math (Bottom) SL Dataset. Best TF+AD results are highlighted in bold. Teacher Forcing (TF) method is equivalent to the Progressive Transformers (PT) work [54].

using Dynamic Time Wrapping to align ground-truth and produced sign sequence in an optimal way mathematically. Figure 5.5.2 compares DTW scores in late training both in the Math and Greek Language SL subsets.

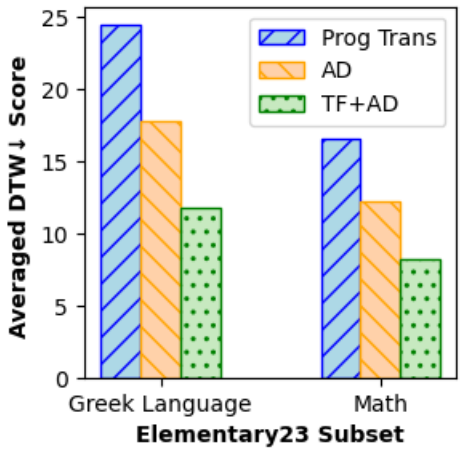


Figure 5.5.2: Comparison of DTW values for methods TF, AD and Hybrid TF,AD at late stage training

To further support the claim that the combination of Teacher Forcing combined with Autoregressive Decoding (TF+AD) works best in our training setups, we also provide below some of the plots retrieving during training. Figure 5.5.3 shows the averaged dynamic time wrapping values, calculated during evaluation (or inference), across the dev set (or the test set, respectively). Initially, the DTW values fluctuate within the range of 15-25, indicating a relatively random alignment between predictions and reference sequences. However, upon adopting AD training, there is a noticeable and significant drop in the DTW values, settling at a promising value of 8.8. This improvement in DTW similarity signals a substantial enhancement in the model's ability to generate more accurate and coherent sequences, further validating the effectiveness of TF+AD training in our specific application.

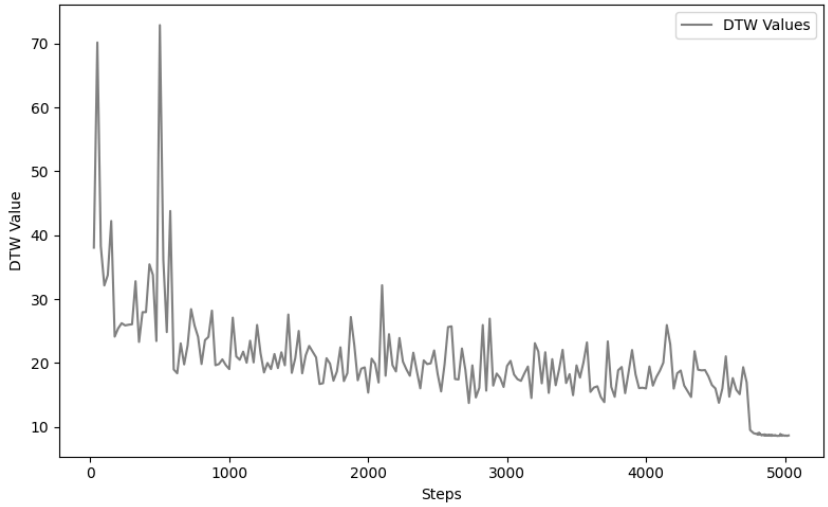


Figure 5.5.3: Visualization of the DTW values on the test set on an Elementary23 Math model. DTW values drop significantly after switching to training with autoregressive decoding, achieving a promising final similarity score of 8.8.

Along with the DTW plot shown above, we also provide a training loss diagram, to further analyze the model's learning dynamics. Figure 5.5.4 compares the progression of the training loss during two phases: first with Teacher Forcing (TF) and then with Autoregressive Decoding (AD) in the same hybrid TF+AD model. The plot reveals the following insights regarding our hybrid training process: During the TF phase, the model converges more quickly, achieving a lower Mean Squared Error (MSE) loss. Specifically, the MSE loss during TF training drops to approximately 10e-5, the AD phase shows converges at a slower rate, with the MSE decreasing to around 10e-3. However, despite the slower learning rate and higher loss values during AD training,this phase contributes to enhanced video quality and improved DTW scores which immensely benefits our case in SLP videos.

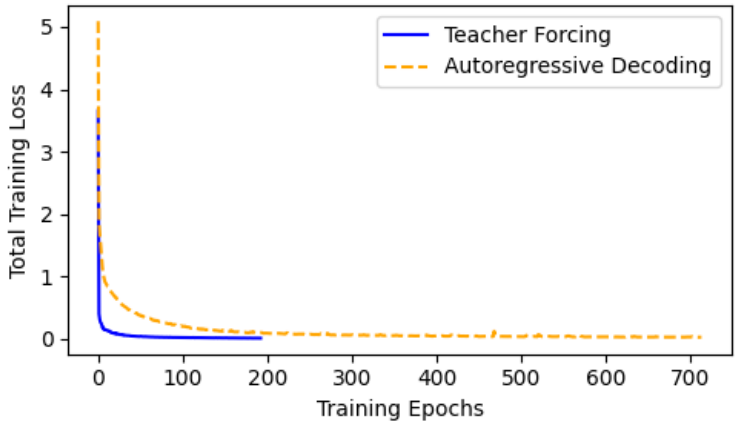


Figure 5.5.4: Visualization of the training MSE loss between TF and AD training on the same hybrid model, trained on Elementary23. As expected, AD converges slower but boost video quality.

5.5.4 User Study

While NLP-based metrics such as BLEU-4, ROUGE, and dynamic time warping provide a quantitative assessment of our model's performance, they do not fully capture the crucial aspect of human-like evaluation. To address this, we conducted a web-based study focusing on sign classification and sign realism. The study contains approximately 30 unseen SL videos, from the dev and test set of the Elementary23 Math dataset. The first half of our study required participants to select the correct corresponding sentence from a set of three options, evaluating the model's ability to generate semantically accurate signs. The second half, participants are asked to compare videos generated from **our pipeline** versus the progressive transformer network [54] and choose which of the two videos best describes the given sentence. Additionally, we included questions specifically measuring the perceived naturalness of the generated sign videos. Our pipeline was separately evaluated by 8 sign language experts, with the assistance of members of the Athena Research Center. Below we present all results retrieved from the study.

5.5.4.1 Sign Classification Study

In the first part of our web-based study we carefully selected 14 sentences (and their corresponding generated SL videos) from either the test or the dev set of the Mathematics Elementary23 subset. Both dev and test sentences are unseen from the transformers models and equally contributing to our user study, as dev sentences are only used during evaluation and DTW calculations.

For each of the 14 videos included in the first half of the study, we asked the participants to select the best matching text option over three sentences, based on which they believe was more accurate. one over the three text option that they think best matches the generated video. To account for cases where none of the provided text options accurately reflected the video, a "none of the above" option was also included. This initial question aimed to assess the semantic alignment between the generated signing videos and their corresponding textual descriptions. Following this, participants were asked two additional questions to evaluate the correctness and clarity of the signed sentences: Namely, we ask "How simple was it to understand the meaning in the signing video?" and "How well do you thing the person signed?". These questions are designed to best model the comprehensibility of the generated signing, focusing on both the linguistic accuracy and the naturalness of the signing performance. After these question we follow by adding three separate scales (1 through 5) rating the visual quality of the videos, namely for the face, hand and overall video quality. These ratings provided insights into the optical quality of the generated videos, highlighting areas for improvement in terms of visual detail. Together, these questions form a comprehensive evaluation framework, combining linguistic and visual elements needed for SLP.

Table 5.11 shows the results of the sign classification study based on correctly answered questions. We can see that our method outperforms PT by 35 percentage points suggesting that the proposed pipeline achieves better results on SLP for SL experts. The study's reliance on SL experts as participants ensures that the evaluation is grounded in real-world usability and comprehensibility, taking into account the needs of the DHH community. While the proposed pipeline achieves a strong accuracy of 71%, there

Table 5.10: Sentences (translated) from Elementary23 Math used in the user study, produced by different methods

Ours	PT
How many kids are in my class?	The green boxes are in total
Revision Lesson	Numbers and additions
Checking in the number line	I'm drawing geometrical shapes
Calculating the product	I'm checking the answer I wrote down
How can we find the result?	Approximately
Discuss the solutions in class	Examples
I'm explaining my thought process	Which kid got the most change?

Table 5.11: Sign Classification Study Comparative Results

Method	Answered Choices	Accuracy
Proposed Pipeline	40/56	**71.42%**
PT Baseline SLP Pipeline	20/56	**35.71%**

is still room for improvement. The remaining 29% of mismatches may be due to challenges such as ambiguous signing, insufficient visual quality, or limitations in the training data. On the other hand, Figure 5.5.5 depicts the answers on the comprehension and visual quality scales. The SL users rated both pipelines relatively low (1 through 3), possible due to low image analysis and sing ambiguity. Future work could focus on addressing these issues to further enhance our performance.

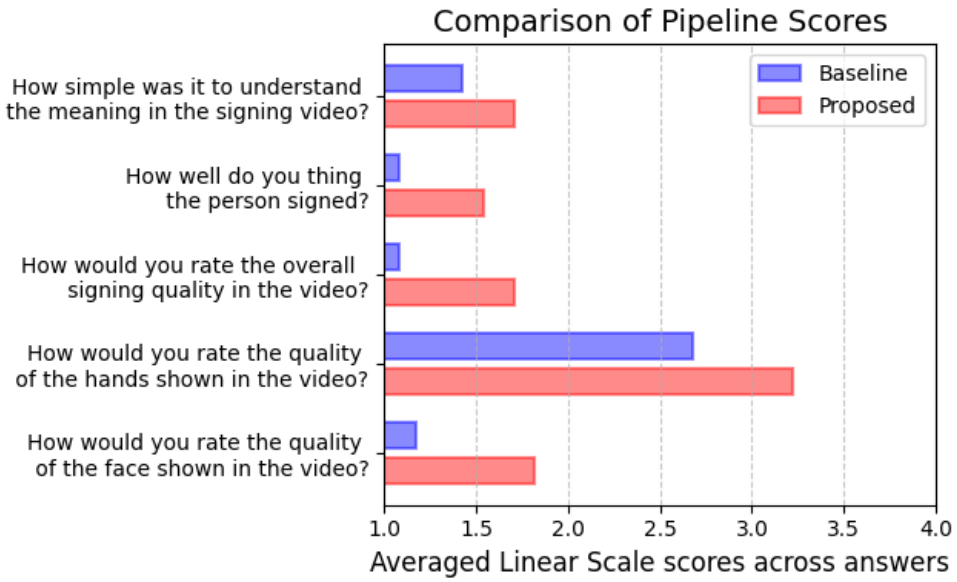


Figure 5.5.5: User study results: Picture Quality and Signer- related questions

5.5.4.2 Comparative Realism Study

In the second phase of our web-based study, we selected six sentences (along with their corresponding sign language videos) generated from both the proposed pipeline and the Progressive Transformer architecture, which serves as the baseline method. The sequences were again chosen from the Math Elementary23 subset and also previously appear in the first part of the questionnaire. This decision was

made to maintain consistency and avoid introducing additional complexity to the results. The findings demonstrate that the proposed SLP-realistic pipeline was overwhelmingly preferred by SL users in comparison to PT, highlighting the clear advantages and effectiveness of the proposed approach.

Table 5.12: Comparative Realism Study Results

Method	Answered Choices	Accuracy
Proposed Pipeline	44/48	**91.66%**
PT Baseline SLP Pipeline	4/48	**8.33%**

Table 5.12 shows the results obtained from the second half of the questionnaire. This Table indicates that 91.6% of participants preferred the videos generated by our method. This overwhelming majority highlights the pipeline's ability to produce sign language videos that are relatively more accurate and natural, in comparison to PT.

5.6 Qualitative Results and Additional Visualizations

Finally, this section contains various visualizations that capture each step of the end-to-end SLP pipeline in a explainable manner. First, we show figures solely from the text-to-video generation using MediaPipe as an intermediate representation step, and the we finalize by showing the results of the rendering process. Figure 5.6.1 depicts two sample sentences from the best performing model in the math subset, showing the (2D) generated skeleton and the original RGB video frame reference from the dataset.

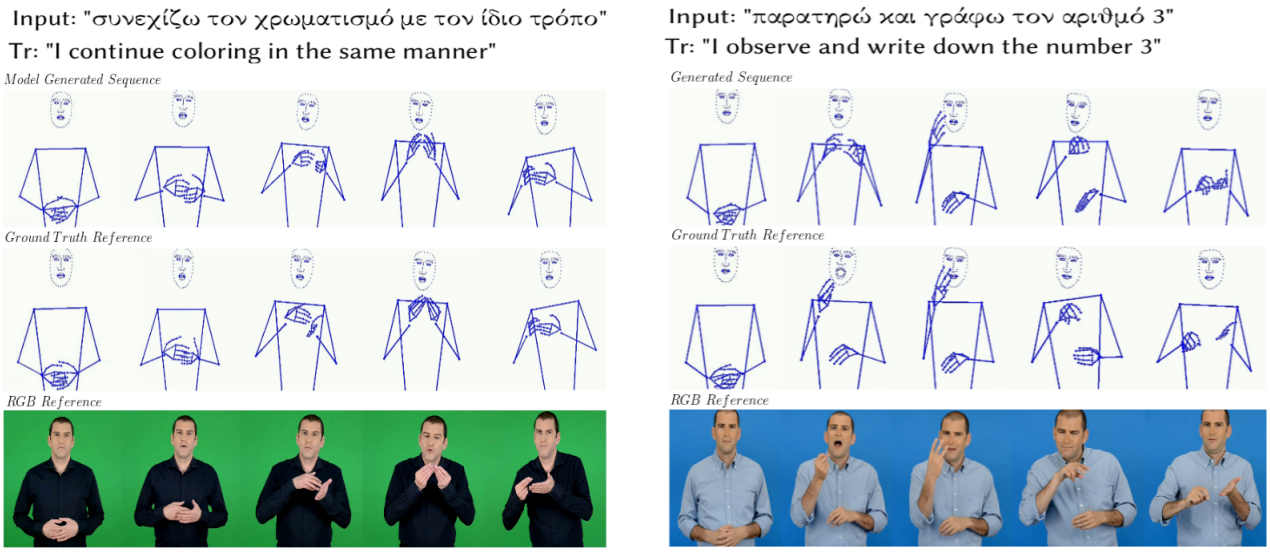


Figure 5.6.1: Sample (test set) visualizations of our SLP method. Top to bottom: Text inputs, 2D generated sign sequence from text embeddings, ground-truth sequence reference, RGB reference.

Figure 5.6.2 again shows the generated skeleton poses, this time comparing between out proposed pipeline and the PT framework. The results clearly demonstrate that our approach achieves superior hand quality in the generated sign.

Figure 5.6.3 shows two examples of the synthetic generated video output. Along with the synthetic output, we also provided visualizations of the generated 2D landmark seqeunce as well as the retargeted color-coded frames.

Finally, in table 5.13 we provide several examples of the video-to-text module that performs the Sign Language Translation task. We compare translated sequences from the ground-truth landmarks and the transformer landmarks.

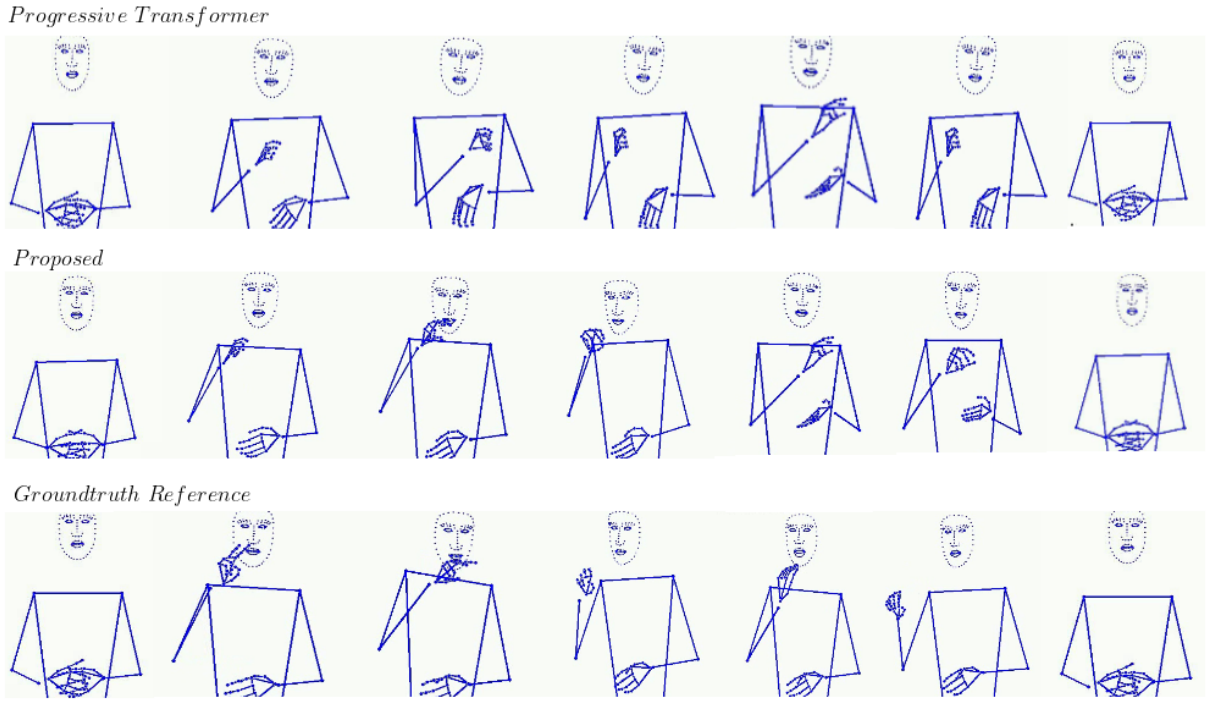


Figure 5.6.2: Sample visualization in order to compare Progressive Transformers [54] to the proposed generative pipeline. Top to Bottom: PT output, Proposed Pipeline Output (ours), Ground-truth Reference.

Figure 5.6.3: Sample total pipeline visualizations. (Top) Skeletal Pose generated from text, (Middle) Corresponding color-coded poses after retargeting and Procrustes analysis, (Bottom) Neural Renderer Result.

1	**Ref:** οι αριθμοί από το 6 μέχρι το 10 **Prod w/ GT:** οι αριθμοί από το 6 μέχρι το 10 **Prod w/ SLP:** οι αριθμοί από το 6 μέχρι το 10
	Ref (Tr): The numbers from 6 to 10 Prod w/ GT (Tr): The numbers from 6 to 10 Prod w/ SLP (Tr): The numbers from 6 to 10
2	**Ref:** φτιάχνω γεωμετρικά σχήματα **Prod w/ GT:** γεωμετρικά σχήματα **Prod w/ SLP:** φτιάχνω γεωμετρικά σχήματα
	Ref (Tr): I create geometrical shapes Prod w/ GT (Tr): geometrical shapes Prod w/ SLP (Tr): I create geometrical shapes
3	**Ref:** αν σε κάθε φύλλο του άλμπουμ βάλει 8 αυτοκόλλητα , πόσα φύλλα θα χρησιμοποιήσει; **Prod w/ GT:** αν σε κάθε φύλλο του άλμπουμ βάλει 10 αυτοκόλλητα, πόσα φύλλα θα χρησιμοποιήσει; **Prod w/ SLP:** αν σε κάθε φύλλο του άλμπουμ βάλει 10 αυτοκόλλητα, πόσα φύλλα θα χρησιμοποιήσει;
	Ref (Tr): if he puts 8 stickers on each sheet of the album, how many sheets will he use? Prod w/ GT (Tr): if he puts 10 stickers on each sheet of the album, how many sheets will he use? Prod w/ SLP (Tr): if he puts 10 stickers on each sheet of the album, how many sheets will he use?
4	**Ref:** κάνω τις διαιρέσεις και γράφω το αποτέλεσμα **Prod w/ GT:** κάνω τις παρακάτω πράξεις **Prod w/ SLP:** κάνω τις προσθέσεις και γράφω το αποτέλεσμα
	Ref (Tr): I do the divisions and write the result Prod w/ GT (Tr): I do the following operations Prod w/ SLP (Tr): I do the additions and write the result
5	**Ref:** ενώνω με το χάρακα τα σημεία που έχουν το ίδιο χρώμα **Prod w/ GT:** ενώνω με μια γραμμή τα σημεία που έχουν το ίδιο χρώμα **Prod w/ SLP:** ενώνω με το χάρακα τα σημεία και τα σύμβολα
	Ref (Tr): I join the points that have the same color with the ruler Prod w/ GT (Tr): join the points that have the same color with a line Prod w/ SLP (Tr): join the points and shapes with the ruler
6	**Ref:** συνδέω τα σχήματα με το όνομα τους **Prod w/ GT:** διηγούμαι ένα πρόβλημα **Prod w/ SLP:** συνδέω με μια γραμμή τα κομμάτια
	Ref (Tr): I connect the shapes with their names Prod w/ GT (Tr): I'm telling a problem Prod w/ SLP (Tr): I connect the pieces with a line

Table 5.13: Cumulative Sign Language Translation Examples. Notation meaning: Ref is Text Reference for Dataset , Prod w/ GT is the SLT text result using the ground-truth landmark sequences and Prod w/ SLP is the SLT text result using the landmark sequences **produced** from our SLP module. From top to bottom, with green are highlighted correctly translated sentences, with orange sentences that differ in words but not in meaning and in red the wrong translations. (Tr) denotes free translation in English.

Chapter 6

Conclusions and Future Work

Contents

6.1 Summary

In this thesis we explore the field of Sign Language Processing and specifically SL production and propose a deep learning system for SLP. We provide an extended review on exiting research regarding deep learning based SL translation and production to gain a better understanding of the field. We base our SLP pipeline on an Encoder - Decoder architecture that predicts pose - landmark sequences given just the text input. To our knowledge, this is the first sign language production system applied to Greek sign language datasets using deep learing architectures. The proposed method utilizes several architectural details that ultimately seem to improve the generated sign quality. More specifically, we incorporate components such as the use of a Pose-to-Text loss during training and SL gloss generation through text transcriptions, which help the quality of the generated sign poses. We also propose a scheduling algorithm that alternates between using teacher forcing and auto-regressive decoding during training. We evaluated the effectiveness of the proposed pipeline on three diverse datasets through an extensive series of comparative analyses and ablation studies, proving its effectiveness. Our work on SLP was accepted at the 18th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA 2025), titled "A Transformer-Based Framework for Greek Sign Language Production using Extended Skeletal Motion Representations".

Furthermore, we explore the photorealistic aspect of the problem, aiming to create a more complete and user-friendly pipeline that transforms text directly into realistic human SL videos. For the photorealistic module, we harness Generative Adversarial Networks (GANs) to perform neural rendering on the pose sequences generated by the transformer model. The rendering process produces a synthetic signer video that emulates the appearance of one of the original dataset's signers, this way addressing potential concerns regarding signer anonymization. To comprehensively assess the performance of the proposed model, we conducted a web-based user study answered by Greek Sign Language experts. The study included a series of questions assessing both the comprehensibility of the generated signs, as well as questions regarding the visual realism of the rendered output.

6.2 Computational Limitations and Future Work

This work, despite its competitive performance in several benchmarks, also posed some architectural limitations, highlighted as follows. As continuous sign language production is a relatively complex task with high computational cost, our model often struggles in unseen or longer sign language sentences. Additionally, the separate training of the neural renderer and the forward and backward Text2Pose models on different machines introduces inefficiencies, especially in the case of autoregressive decoding, where training can take up to 3-4 days to complete. Finally, our current rendering approach does not fully exploit the third coordinate of the MediaPipe skeleton signs, which encodes valuable information about image depth and camera positioning. After summarizing the work presented in this thesis and addressing seen limitations we focus our attention on expanding sign language production and synthesis research and several possible ideas arise to mind:

- **VQ-VAE, VQ-GAN architectures:** Recent advancements demonstrate that integrating variational autoencoders can significantly enhance the performance of sign language production systems. By mapping the encoded sign language sequences to their nearest candidate from a learnable codebook we could enhance the quality of the produced SL videos. Embedding such approaches into our Encoder-Decoder architecture in future implementations could further enhance the realism of produced signs while reducing the reliance on autoregressive decoding during training.
- **3D model reconstruction:** This work utilizes 2D representations to encode SL videos, as both the Transformer and neural rendering modules perform effectively within this framework. However, there still remains the capability of extracting 3D represenations with MediaPipe. 3D representations could contain useful information regarding the image depth and the camera position especially during the neural rendering process for better result quality. On the other hand, several recent works explore the creation of SL models which use realistic 3D figures, often employed using SMPL-X and other body representation frameworks.
- **Lexical Diversity:** Notable limitation of existing SL datasets is their constrained vocabularies. For example, Elementary23 focuses on elementary school textbooks, How2Sign on instructional and tutorial videos, and PHOENIX14T on weather forecasts. As a result, most SLP modules relay on reproducing words and sentences from the already limited SL training vocabulary, and thus may not covers the necessities of a complete Sign Language education system. The possible development of a larger and unified SL base would tremendously benefit the DHH community, particularly in the digital era.
- **End-to-end Training:** In this thesis, SLP is addressed as a twofold task: First, generating MP-extended skeletal poses from text and then performing neural rendering in a separate step using distinct models. Additionally, within the Text-to-Pose module, separate Transformer models are employed for forward and backward translation. Transitioning to a more holistic approach that doesn't require the training of diverse modules could be proven beneficial, by improving semantic coherence and enhancing computational efficiency.

6.3 Possible Applications and Social Impact

Sign Language Production is a way of bridging communication between the hard of hearing and non-DHH communities. Digital sign language systems can have a wide range of valuable applications in the daily life of people who might be in need of it. For example, lifelike generated sign language avatars could serve as active educational aids, interactive guides in museums, or presenters in news broadcasting.

Is of great importance to note that these technologies are not intended to replace human sign language interpreters; rather, they aim to enhance accessibility by supporting sign language education and dubbing, making these resources more accessible and widely available. Such artificial systems that often output a synthetic human should always be developed with a deep respect for the irreplaceable value

of human-to-human communication, while also prioritizing the genuine needs and preferences of the DHH community.

Chapter 7

Bibliography

[1] Ahmadian, S. and Khanteymoori, A. "Training back propagation neural networks using asexual reproduction optimization". In: May 2015. DOI: 10.1109/IKT.2015.7288738.

[2] Baltatzis, V., Potamias, R. A., Ververas, E., Sun, G., Deng, J., and Zafeiriou, S. "Neural Sign Actors: A Diffusion Model for 3D Sign Language Production from Text". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 1985–1995.

[3] Bantupalli, K. and Xie, Y. "American Sign Language Recognition using Deep Learning and Computer Vision". In: *2018 IEEE International Conference on Big Data (Big Data)*. 2018, pp. 4896–4899. DOI: 10.1109/BigData.2018.8622141.

[4] Bao, J., Chen, D., Wen, F., Li, H., and Hua, G. "CVAE-GAN: Fine-Grained Image Generation through Asymmetric Training". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2764–2773. DOI: 10.1109/ICCV.2017.299.

[5] Bazarevsky, V., Grishchenko, I., Raveendran, K., Zhu, T., Zhang, F., and Grundmann, M. *BlazePose: On-device Real-time Body Pose tracking*. 2020. arXiv: 2006.10204 [cs.CV].

[6] Bouzid, Y. and Jemni, M. "An Avatar Based Approach for Automatically Interpreting a Sign Language Notation". In: *2013 IEEE 13th International Conference on Advanced Learning Technologies* (2013), pp. 92–94.

[7] Brun, R., Turki, A., and Laville, A. "A 3D application to familiarize children with sign language and assess the potential of avatars and motion capture for learning movement". In: vol. 05-06-July-2016. Cited by: 4. 2016. DOI: 10.1145/2948910.2948917.

[8] Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. "Neural Sign Language Translation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[9] Camgoz, N. C., Koller, O., Hadfield, S., and Bowden, R. "Multi-channel Transformers for Multi-articulatory Sign Language Translation". In: *Computer Vision – ECCV 2020 Workshops*. Ed. by A. Bartoli and A. Fusiello. Cham: Springer International Publishing, 2020, pp. 301–319. ISBN: 978-3-030-66823-5.

[10] Camgöz, N. C., Koller, O., Hadfield, S., and Bowden, R. *Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation*. 2020.

[11] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. "VGGFace2: A Dataset for Recognising Faces across Pose and Age". In: *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. 2018, pp. 67–74. DOI: 10.1109/FG.2018.00020.

[12] Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: *CVPR*. 2017.

[13] Chan, C., Ginosar, S., Zhou, T., and Efros, A. A. "Everybody Dance Now". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019.

[14] Cihan Camgoz, N., Hadfield, S., Koller, O., and Bowden, R. "SubUNets: End-To-End Hand Shape and Continuous Sign Language Recognition". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017.

[15] Doukas, M. C., Koujan, M. R., Sharmanska, V., Roussos, A., and Zafeiriou, S. "Head2Head++: Deep Facial Attributes Re-Targeting". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3.1 (Jan. 2021), pp. 31–43. ISSN: 2637-6407. DOI: 10.1109/tbiom.2021.3049576.

[16] Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., and Giro-i-Nieto, X. "How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021.

[17] Elliott, R., Bueno, F., Kennaway, R., and Glauert, J. "Towards the Integration of Synthetic SL Animation with Avatars into Corpus Annotation Tools". In: (Jan. 2010).

[18] Fang, S., Sui, C., Zhang, X., and Tian, Y. *SignDiff: Learning Diffusion Models for American Sign Language Production*. 2023. arXiv: 2308.16082 [cs.CV].

[19] Feng, Y., Feng, H., Black, M. J., and Bolkart, T. "Learning an Animatable Detailed 3D Face Model from In-The-Wild Images". In: *CoRR* abs/2012.04012 (2020). arXiv: 2012.04012 [cs.CV].

[20] Filntisis, P. P., Retsinas, G., Paraperas-Papantoniou, F., Katsamanis, A., Roussos, A., and Maragos, P. "Visual Speech-Aware Perceptual 3D Facial Expression Reconstruction from Videos". In: *arXiv preprint arXiv:2207.11094* (2022).

[21] Forte, M.-P., Kulits, P., Huang, C.-H. P., Choutas, V., Tzionas, D., Kuchenbecker, K. J., and Black, M. J. "Reconstructing Signing Avatars From Video Using Linguistic Priors". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 12791–12801.

[22] Gibet, S., Lefebvre-Albaret, F., Hamon, L., Brun, R., and Turki, A. "Interactive editing in French Sign Language dedicated to virtual signers: requirements and challenges". In: *Universal Access in the Information Society* 15.4 (Nov. 2016), pp. 525–539. DOI: 10.1007/s10209-015-0411-6.

[23] Glauert, J. and Elliott, R. *Extending the SiGML Notation-a Progress Report*. 2011.

[24] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[25] Google. *MediaPipe Hands Solution Documentation*.

[26] Google. *MediaPipe Holistic Solution Documentation*.

[27] Google. *MediaPipe Pose Solution Documentation*.

[28] Grieve-Smith, A. B. "SignSynth: A Sign Language Synthesis Application Using Web3D and Perl". In: *Gesture and Sign Language in Human-Computer Interaction*. Ed. by I. Wachsmuth and T. Sowa. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 134–145. ISBN: 978-3-540-47873-7.

[29] Guimaraes, C., Guardezi, J. F., and Fátima Fernandes, S. de. "Sign Language Writing Acquisition – Technology for a Writing System". In: *2014 47th Hawaii International Conference on System Sciences* (2014), pp. 120–129.

[30] Hanke, T., Popescu, H., and Schmaling, C. "eSIGN: HPSG-assisted sign language composition". In: Jan. 2003.

[31] Hochreiter, S. and Schmidhuber, J. "Long Short-term Memory". In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[32] Isard, A. "Approaches to the Anonymisation of Sign Language Corpora". In: *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*. Ed. by E. Efthimiou, S.-E. Fotinea, T. Hanke, J. A. Hochgesang, J. Kristoffersen, and J. Mesch. Marseille, France: European Language Resources Association (ELRA), May 2020, pp. 95–100. ISBN: 979-10-95546-54-2.

[33] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. "Image-to-Image Translation with Conditional Adversarial Networks". In: *CVPR* (2017).

[34] Kennaway, R. "Experience with and Requirements for a Gesture Description Language for Synthetic Animation". In: Apr. 2003, pp. 300–311. ISBN: 978-3-540-21072-6. DOI: 10.1007/978-3-540-24598-8_28.

[35] Ko, S.-K., Kim, C. J., Jung, H., and Cho, C. "Neural Sign Language Translation Based on Human Keypoint Estimation". In: *Applied Sciences* 9.13 (2019). ISSN: 2076-3417. DOI: 10.3390/app9132683.

[36] Koller, O., Forster, J., and Ney, H. "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers". In: *Computer Vision and Image Understanding* 141 (Dec. 2015), pp. 108–125.

[37] Koujan, M., Doukas, M., Roussos, A., and Zafeiriou, S. "Head2Head: Video-Based Neural Head Synthesis". In: *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2020, pp. 319–326. DOI: 10.1109/FG47880.2020.00048.

[38] Lee, S., Glasser, A., Dingman, B., Xia, Z., Metaxas, D., Neidle, C., and Huenerfauth, M. "American Sign Language Video Anonymization to Support Online Participation of Deaf and Hard of Hearing Users". In: *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '21. <conf-loc>, <city>Virtual Event</city>, <country>USA</country>, </conf-loc>: Association for Computing Machinery, 2021. ISBN: 9781450383066. DOI: 10.1145/3441852.3471200.

[39] Lefebvre-Albaret, F., Gibet, S., Turki, A., Hamon, L., and Brun, R. "Overview of the Sign3D Project High-fidelity 3D recording, indexing and editing of French Sign Language content". In: *Third International Symposium on Sign Language Translation and Avatar Technology (SLTAT) 2013*. Chicago, United States, Oct. 2013.

[40] Lin, C.-Y. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81.

[41] Lin, J. F.-S., Karg, M., and Kulić, D. "Movement Primitive Segmentation for Human Motion Modeling: A Framework for Analysis". In: *IEEE Transactions on Human-Machine Systems* 46.3 (2016), pp. 325–339. DOI: 10.1109/THMS.2015.2493536.

[42] Ma, J., Wang, W., Yang, Y., and Zheng, F. "MS2SL: Multimodal Spoken Data-Driven Continuous Sign Language Production". In: *ACL*. 2024.

[43] Ma, X., Jin, R., Wang, J., and Chung, T.-S. "Attentional bias for hands: Cascade dual-decoder transformer for sign language production". In: *IET Computer Vision* (). DOI: https://doi.org/10.1049/cvi2.12273. eprint:

[44] Malaia, E. and Wilbur, R. "What sign languages show". In: Aug. 2012, pp. 263–276. ISBN: 9789027255778. DOI: 10.1075/la.194.12mal.

[45] Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., and Smolley, S. P. *Least Squares Generative Adversarial Networks*. 2017. arXiv: 1611.04076 [cs.CV].

[46] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. "BLEU: a method for automatic evaluation of machine translation". In: ACL '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 311–318. DOI: 10.3115/1073083.1073135.

[47] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D., and Black, M. J. *Expressive Body Capture: 3D Hands, Face, and Body from a Single Image*. 2019. arXiv: 1904.05866 [cs.CV].

[48] Pezeshkpour, F., Marshall, I., Elliott, R., and Bangham, J. "Development of a legible deaf-signing virtual human". In: vol. 1. Cited by: 14. 1999, pp. 333–338.

[49] Pratikaki, C., Filntisis, P., Katsamanis, A., Roussos, A., and Maragos, P. *A Transformer-Based Framework for Greek Sign Language Production using Extended Skeletal Motion Representations*. 2025. arXiv: 2503.02421.

[50] Retsinas, G., Filntisis, P. P., Danecek, R., Abrevaya, V. F., Roussos, A., Bolkart, T., and Maragos, P. "3D Facial Expressions through Analysis-by-Neural-Synthesis". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024.

[51] Roussos, A., Theodorakis, S., Pitsikalis, V., and Maragos, P. "Dynamic Affine-Invariant Shape-Appearance Handshape Features and Classification in Sign Language Videos". In: *Journal of Machine Learning Research* 14.51 (2013), pp. 1627–1663.

[52] Rust, P., Shi, B., Wang, S., Camgoz, N. C., and Maillard, J. "Towards Privacy-Aware Sign Language Translation at Scale". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, 2024, pp. 8624–8641.

[53] Sandler, W. and Lillo-Martin, D. *Sign Language and Linguistic Universals*. Cambridge University Press, 2006.

[54] Saunders, B., Camgoz, N. C., and Bowden, R. "Progressive Transformers for End-to-End Sign Language Production". In: (2020).

[55] Saunders, B., Camgoz, N. C., and Bowden, R. "Anonysign: Novel Human Appearance Synthesis for Sign Language Video Anonymisation". In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. 2021, pp. 1–8. DOI: 10.1109/FG52635.2021.9666984.

[56] Saunders, B., Camgoz, N. C., and Bowden, R. *Signing at Scale: Learning to Co-Articulate Signs for Large-Scale Photo-Realistic Sign Language Production*. 2022.

[57] Shi, B., Brentari, D., Shakhnarovich, G., and Livescu, K. *Open-Domain Sign Language Translation Learned from Online Video*. 2022. arXiv: 2205.12870 [cs.CV].

[58] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. "The Development and Evaluation of a Speech-to-Sign Translation System to Assist Transactions". In: *International Journal of Human–Computer Interaction* 16.2 (2003), pp. 141–161. DOI: 10.1207/S15327590IJHC1602_02.

[59] Stokoe, W. C. "Sign language structure: an outline of the visual communication systems of the American deaf. 1960." In: *Journal of deaf studies and deaf education* ().

[60] Stoll, S., Camgoz, N. C., Hadfield, S., and Bowden, R. "Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks". In: *International Journal of Computer Vision* 128.4 (Apr. 2020), pp. 891–908. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01281-2.

[61] Stoll, S., Mustafa, A., and Guillemaut, J. Y. "There and Back Again: 3D Sign Language Generation from Text Using Back-Translation". In: Institute of Electrical and Electronics Engineers Inc., 2022, pp. 187–196. ISBN: 9781665456708. DOI: 10.1109/3DV57658.2022.00031.

[62] Theodorakis, S., Pitsikalis, V., and Maragos, P. "Dynamic–static unsupervised sequentiality, statistical subunits and lexicon for sign language recognition". In: *Image and Vision Computing* 32.8 (2014), pp. 533–549. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2014.04.012.

[63] Tian, Y., Peng, X., Zhao, L., Zhang, S., and Metaxas, D. N. "CR-GAN: Learning Complete Representations for Multi-view Generation". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization, July 2018, pp. 942–948. DOI: 10.24963/ijcai.2018/131.

[64] Tze, C. O., Filntisis, P. P., Dimou, A.-L., Roussos, A., and Maragos, P. "Neural Sign Reenactor: Deep Photorealistic Sign Language Retargeting". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2023.

[65] Tze, C. O., Filntisis, P. P., Roussos, A., and Maragos, P. "Cartoonized Anonymization of Sign Language Videos". In: *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. 2022. DOI: 10.1109/IVMSP54334.2022.9816293.

[66] Uthus, D., Tanzer, G., and Georg, M. *YouTube-ASL: A Large-Scale, Open-Domain American Sign Language-English Parallel Corpus*. 2023. arXiv: 2306.15162 [cs.CL].

[67] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.

[68] Voskou, A., Panousis, K. P., Partaourides, H., Tolias, K., and Chatzis, S. *A New Dataset for End-to-End Sign Language Translation: The Greek Elementary School Dataset*. 2023. arXiv: 2310.04753 [cs.CL].

[69] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. *Video-to-Video Synthesis*. 2018. arXiv: 1808.06601 [cs.CV].

[70] Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. "High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[71] Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. "Convolutional pose machines". In: *CVPR*. 2016.

[72] Wolfe, R., McDonald, J. C., Schnepp, J. C., and Toro, J. "Synthetic and acquired corpora: meeting at the annotation". In: (2011).

[73] Wolfe, R., McDonald, J. C., Hanke, T., Ebling, S., Van Landuyt, D., Picron, F., Krausneker, V., Efthimiou, E., Fotinea, E., and Braffort, A. "Sign Language Avatars: A Question of Representation". In: *Information* 13.4 (2022). ISSN: 2078-2489. DOI: 10.3390/info13040206.

[74] Xia, Z., Chen, Y., Zhangli, Q., Huenerfauth, M., Neidle, C., and Metaxas, D. “Sign Language Video Anonymization”. In: ().

[75] Xia, Z., Neidle, C., and Metaxas, D. N. *DiffSLVA: Harnessing Diffusion Models for Sign Language Video Anonymization*. 2023. arXiv: 2311.16060 [cs.CV].

[76] Xie, P., Peng, T., Du, Y., and Zhang, Q. *Sign Language Production with Latent Motion Transformer*. 2023.

[77] Yanovich, P., Neidle, C., and Metaxas, D. “Detection of major ASL sign types in continuous signing for ASL recognition”. In: Cited by: 16. 2016, pp. 3067–3073.

[78] Zhang, L., Rao, A., and Agrawala, M. *Adding Conditional Control to Text-to-Image Diffusion Models*. 2023. arXiv: 2302.05543 [cs.CV].