National Technical University of Athens

School of Electrical and Computer Engineering
division of signals, control and robotics

Occlusion-Robust Audiovisual Face Reconstruction with Temporal Modeling

DIPLOMA THESIS

of

Angeliki Tsinouka

Supervisor: Petros Maragos
Professor NTUA

Co-Supervisor: Panagiotis Filntisis
Researcher Athena Research Center

Co-Supervisor: George Retsinas
Researcher Athena Research Center

COMPUTER VISION, SPEECH COMMUNICATION AND SIGNAL PROCESSING GROUP
Athens, October 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

Occlusion-Robust Audiovisual Face Reconstruction with Temporal Modeling

DIPLOMA THESIS

of

Angeliki Tsinouka

Supervisor: Petros Maragos
Professor NTUA

Co-Supervisor: Panagiotis Filntisis
Researcher Athena Research Center

Co-Supervisor: George Retsinas
Researcher Athena Research Center

Approved by the examination committee on 17th October, 2025.

.....	.....	.....
Petros Maragos	Athanasios Rontogiannis	Ioannis Kordonis
Professor NTUA	Associate Professor NTUA	Assistant Professor NTUA

Athens, October 2025

.....

Angeliki Tsinouka
Graduate of Electrical and
Computer Engineering NTUA

Στην οικογένειά μου

Abstract

Recent progress in deep learning for 3D face reconstruction and animation has enabled the creation of realistic digital humans, capable of reproducing subtle expressions and natural speech-driven motion. These advances open new opportunities for applications across communication, entertainment, AR/VR, and education. Nevertheless, significant challenges remain. Conventional approaches often fail to fully capture the expressive dynamics of human faces, they struggle with temporal consistency, and they are not robust to real-world conditions such as partial occlusions or interfering background voices. As the demand for detailed digital avatars grows, especially in interactive systems, developing methods that combine realism, expressiveness, and robustness becomes increasingly crucial.

This thesis addresses the challenges in 3D face reconstruction through the design of an audiovisual learning technique from input video, which we call FAVOR, extending the SMIRK framework. Our method applies synthetic occlusions to the training dataset to improve robustness and employs a lip-reading loss for supervision, guiding the model toward more accurate mouth movements. By combining multimodal signals from and training strategies that reflect real-world variability, the proposed approach generates talking avatars that remain coherent and natural even when visual information is missing or corrupted. The system also ensures proper synchronization between speech and facial motion, reducing common artifacts.

Extensive experimental analysis on videos with natural occlusions demonstrates that the proposed model achieves robust and temporally consistent results compared to single-modality methods, both in qualitative and quantitative evaluations. A user study further confirms the perceptual quality of the generated avatars, while an ablation study highlights the contribution of each component to the overall performance.

Keywords — Audiovisual Face Reconstruction, Multimodal Learning, Talking Avatars, Face Modeling, Synthetic Occlusions.

Περίληψη

Οι πρόσφατες εξελίξεις της βαθιάς μάθηση στην τρισδιάστατη ανακατασκευή και απεικόνιση ανθρώπινων προσώπων έχουν καταστήσει δυνατή τη δημιουργία ρεαλιστικών ψηφιακών ανθρώπινων άβαταρ, ικανών να αναπαράγουν λεπτομερείς εκφράσεις και φυσική κίνηση στόματος καθοδηγούμενη από την ομιλία. Οι εξελίξεις αυτές ανοίγουν νέες δυνατότητες για εφαρμογές στην επικοινωνία, την ψυχαγωγία, την εκπαίδευση και τα περιβάλλοντα επαυξημένης πραγματικότητας. Παρ' όλα αυτά, εξακολουθούν να υπάρχουν σημαντικές προκλήσεις. Οι συμβατικές προσεγγίσεις συχνά αδυνατούν να αποδώσουν πλήρως τη δυναμική των εκφράσεων του προσώπου, δυσκολεύονται να διατηρήσουν χρονική συνέπεια και δεν είναι εύρωστες σε πραγματικές συνθήκες. Καθώς αυξάνεται η ζήτηση για λεπτομερή ψηφιακά άβαταρ, ειδικά σε διαδραστικά συστήματα, η ανάπτυξη μεθόδων που συνδυάζουν ρεαλισμό, εκφραστικότητα και ανθεκτικότητα καθίσταται ολοένα και πιο σημαντική.

Η παρούσα διπλωματική εργασία αντιμετωπίζει τις προκλήσεις στην τρισδιάστατη ανακατασκευή ανθρώπινου προσώπου μέσω μιας οπτικοακουστικής προσέγγισης που ονομάζουμε FAVOR. Η μέθοδος παίρνει ως είσοδο βίντεο δεδομένα και επεκτείνει την αρχιτεκτονική του SMIRK. Παράλληλα, κατά τη διαδικασία εκπαίδευσης, εφαρμόζονται συνθετικές αποκρύψεις προσώπου στο σύνολο δεδομένων, ενώ αξιοποιείται μια συνάρτηση απώλειας βασισμένη στην αναγνώριση των χειλιών, η οποία κατευθύνει το μοντέλο προς την παραγωγή ακριβέστερων κινήσεων του στόματος. Μέσω του συνδυασμού πολυτροπικών σημάτων εισόδου και ενισχυμένων στρατηγικών εκπαίδευσης, η προτεινόμενη προσέγγιση παράγει ψηφιακά άβαταρ που παραμένουν συνεπή και λεπτομερή, ακόμη και όταν η οπτική πληροφορία απουσιάζει. Το σύστημα διασφαλίζει επιπλέον τον ακριβή συγχρονισμό μεταξύ της ομιλίας και της κίνησης των χειλιών, περιορίζοντας συνήθη σφάλματα που εντοπίζονται στη περιοχή του στόματος.

Τα πειράματά μας σε οπτικοακουστικά δεδομένα με φυσικές αποκρύψεις προσώπου, δείχνουν την εξαιρετική ποιότητα ανακατασκευής, όπως αποτυπώνεται τόσο σε αντικειμενικές μετρικές όσο και σε ανθρώπινες αξιολογήσεις. Τα αποτελέσματα αυτά αναδεικνύουν την ευρωστία και την αξιοπιστία της προτεινόμενης μεθόδου σε ρεαλιστικές συνθήκες, επιβεβαιώνοντας τη δυνατότητά της να αποδίδει συνεπή και ποιοτικά αποτελέσματα ακόμη και σε απαιτητικά σενάρια.

Λέξεις Κλειδιά — Οπτικοακουστική Ανακατασκευή Προσώπου, Πολυτροπική Μάθηση, Ομιλούντα Άβαταρ, Μοντελοποίηση Προσώπου, Συνθετικές Αποκρύψεις Προσώπου.

Ευχαριστίες

Η ολοκλήρωση της παρούσας διπλωματικής εργασίας σηματοδοτεί και την ολοκλήρωση των προπτυχιακών μου σπουδών στη Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών του Εθνικού Μετσόβιου Πολυτεχνείου. Στο σημείο αυτό θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη προς όλους όσους συνέβαλαν με τον δικό τους τρόπο σε αυτή την πορεία.

Πρώτα απ' όλα, θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου, κ. Πέτρο Μαραγκό, για την εμπιστοσύνη που μου έδειξε και την ευκαιρία που μου έδωσε να πραγματοποιήσω τη διπλωματική μου εργασία στο Εργαστήριο Όρασης Υπολογιστών και Επεξεργασίας Σημάτων.

Θα ήθελα επίσης να ευχαριστήσω εγκάρδια τους συνεπιβλέποντές μου, Δρ. Παναγιώτη Φιλντίση και Δρ. Γεώργιο Ρετσινά, για την ακούραστη και ουσιαστική υποστήριξή τους. Η καθοδήγησή τους, οι γνώσεις που μοιράστηκαν μαζί μου, καθώς και η συνεχής διαθεσιμότητά τους για συζήτηση και βοήθεια, έπαιξαν καθοριστικό ρόλο στην επιτυχή ολοκλήρωση της διπλωματικής μου εργασίας.

Ιδιαίτερες ευχαριστίες οφείλω στους φίλους μου που έδωσαν χρώμα στα φοιτητικά μου χρόνια και μου χάρισαν όμορφες στιγμές. Τέλος, το πιο μεγάλο ευχαριστώ ανήκει στην οικογένειά μου, για την αδιάκοπη στήριξη και την πίστη τους σε εμένα όλα αυτά τα χρόνια.

Αγγελική Τσινούκα
Οκτώβριος 2025

Table of contents

List of Acronyms

CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
FAN	Face Alignment Network
3DMM	3D Morphable Models
SPECTRE	Speech-Informed Perceptual 3D Facial Expression Reconstruction
SMIRK	Spatial Modeling for Image-based Reconstruction of Kinesics
FAVOR	Face AudioVisual Occlusion-robust Reconstruction

List of figures

List of tables

Chapter 1

Εκτεταμένη Περίληψη στα ελληνικά

Contents

1.1 Εισαγωγή

Η μελέτη και η ανάλυση του ανθρώπινου προσώπου αποτελούν κρίσιμο πεδίο για την επιστήμη της όρασης υπολογιστών και της μηχανικής μάθησης, καθώς το πρόσωπο μεταφέρει βασικές πληροφορίες για τη ταυτότητα, το συναίσθημα και τη συμπεριφορά. Η πρόοδος της τεχνητής νοημοσύνης επιτρέπει στους υπολογιστές να αναγνωρίζουν πρόσωπα, να ερμηνεύουν εκφράσεις και να κατανοούν κινήσεις και ομιλία, ανοίγοντας τον δρόμο για μια πιο φυσική αλληλεπίδραση ανθρώπου-μηχανής. Στο πλαίσιο αυτό, η τρισδιάστατη ανακατασκευή προσώπων έχει αναδειχθεί σε βασικό ερευνητικό αντικείμενο, με ποικίλες εφαρμογές στην ασφάλεια, στην ιατρική, στην ψυχαγωγία, στα κοινωνικά δίκτυα και στην εικονική/επαυξημένη πραγματικότητα.

Η διαδικασία της ανακατασκευής προσώπου είναι ιδιαίτερα απαιτητική, καθώς το ανθρώπινο πρόσωπο είναι δυναμικό, παραμορφώνεται συνεχώς με την ομιλία, τις εκφράσεις και την πάροδο του χρόνου, ενώ η ποιότητα των δεδομένων επηρεάζεται από το φωτισμό, την απόκρυψη ορισμένων χαρακτηριστικών του και τη μεταβολή της γωνίας λήψης. Έτσι, οι μονοτροπικές μέθοδοι (που βασίζονται για παράδειγμα μόνο στην εικόνα ή μόνο στον ήχο) παρουσιάζουν σημαντικούς περιορισμούς. Για τον λόγο αυτό, η παρούσα διπλωματική εργασία προτείνει μια πολυτροπική, οπτικοακουστική προσέγγιση, η οποία συνδυάζει τα οπτικά χαρακτηριστικά (γεωμετρία, εκφράσεις) με τις ακουστικές πληροφορίες (κινήση του στόματος, άρθρωση). Ο συνδυασμός αυτών των δεδομένων βελτιώνει την ακρίβεια και την ευρωστία του συστήματος σε απαιτητικές συνθήκες εφαρμογής.

Πολυτροπική Αναπαράσταση και Βαθιά Μάθηση

Οι αρχές της πολυτροπικής αναπαράστασης μιμούνται τον τρόπο με τον οποίο ο άνθρωπος συνδυάζει τις αισθητηριακές πληροφορίες (όραση, ακοή, γλώσσα). Η πολυτροπικότητα οδηγεί σε πιο πλούσιες και συνεκτικές αναπαραστάσεις σε σχέση με μονοτροπικά μοντέλα και βρίσκει εφαρμογή σε τομείς όπως η αναγνώριση λόγου με οπτική υποβοήθηση [66], η αυτόματη περιγραφή εικόνας [29] και η ανάλυση συναισθήματος [76].

Για την υλοποίηση της προτεινόμενης μεθοδολογίας, αξιοποιούνται τεχνικές βαθιάς μάθησης [36], οι οποίες βασίζονται σε νευρωνικά δίκτυα πολλαπλών επιπέδων [39]. Τα δίκτυα αυτά εκπαιδεύονται μέσω μιας επαναληπτικής διαδικασίας πρόβλεψης και διόρθωσης: κάθε δίκτυο παράγει μια πρόβλεψη που συγκρίνεται με την πραγματική τιμή, και στη συνέχεια, μέσω του αλγορίθμου αντίστροφης διάδοσης, βελτιώνει την απόδοσή του. Ιδιαίτερη έμφαση δίνεται στις παρακάτω βασικές κατηγορίες αρχιτεκτονικών:

- Τα **Συνελικτικά Νευρωνικά Δίκτυα** (CNNs) [38], [48] αποτελούν τον πυρήνα των περισσότερων μεθόδων επεξεργασίας εικόνας. Χρησιμοποιούν συνελικτικά φίλτρα, τα οποία εξάγουν ιεραρχικά χαρακτηριστικά διαφορετικού επιπέδου αφαίρεσης ανά επίπεδο, από απλές ακμές έως σύνθετα μορφολογικά χαρακτηριστικά. Μια επέκταση που χρησιμοποιούμε στην μέθοδό μας είναι οι χρονικές συνελίξεις (TCNs), οι οποίες εφαρμόζονται στη διάσταση του χρόνου και επιτρέπουν τη μοντελοποίηση των συσχετίσεων μεταξύ διαδοχικών καρέ σε μια ακολουθία βίντεο.

- Τα **Αναδρομικά Νευρωνικά Δίκτυα** (RNNs) [45] επεξεργάζονται δεδομένα με χρονική διάσταση, όπως η ανθρώπινη ομιλία. Αξιοποιούν πληροφορίες από προηγούμενες χρονικές στιγμές, αλλά αδυνατούν να διατηρήσουν μακροχρόνιες εξαρτήσεις. Αυτό το πρόβλημα καλούνται να αντιμετωπήσουν **Δίκτυα Μακράς-Βραχύχρονης Μνήμης** [26], τα οποία ελέγχουν τη ροή πληροφορίας στο χρόνο.

- Οι **μετασχηματιστές** (Transformers) [69] αποτελούν μια καινοτόμα προσέγγιση, οι οποίοι εισήγαγαν τον μηχανισμό προσοχής. Είναι σχεδιασμένοι για την αποδοτική επεξεργασία ακολουθιακών δεδομένων και επιτρέπουν τη μοντελοποίηση μακροχρόνιων εξαρτήσεων. Ο μηχανισμός προσοχής συσχετίζει το κάθε στοιχείο με όλα τα υπόλοιπα της ακολουθίας, δίνοντας έμφαση σε εκείνα που είναι πιο σημαντικά για την τελική αναπαράσταση. Η ιδιότητά τους να επεξεργάζονται παράλληλα μεγάλες ποσότητες δεδομένων έχει καταστήσει τους μετασχηματιστές κυρίαρχους σε πολλές εφαρμογές μηχανικής μάθησης.

Εφαρμογές της Τρισδιάστασης Ανακατασκευής Προσώπων

Η δυναμική αυτών των αρχιτεκτονικών αποτελεί τη βάση για τον σχεδιασμό συστημάτων ικανών να ανακατασκευάζουν το ανθρώπινο πρόσωπο με υψηλή ακρίβεια και ρεαλισμό. Τα συστήματα αυτά βρίσκουν εφαρμογή σε πολλούς τομείς. Στην ψυχαγωγία και στην επαυξημένη πραγματικότητα [28] χρησιμοποιούνται για τη δημιουργία εξατομικευμένων ψηφιακών άβαταρ που αναπαράγουν την ταυτότητα και τις εκφράσεις του χρήστη. Στη βιομηχανία παραγωγής ταινιών επιτρέπουν τόσο την τροποποίηση συναισθημάτων σε σκηνές [50] όσο και την αυτόματη μεταγλώττιση μέσω συγχρονισμού των χειλιών [21]. Η χρήση τους επεκτείνεται επίσης στα μέσα κοινωνικής δικτύωσης, προσφέροντας εργαλεία δημιουργικής επεξεργασίας και παραγωγής περιεχομένου. Ειδικές εφαρμογές εντοπίζονται στην ιατρική [23], [58], με έμφαση στη μοντελοποίηση του προσώπου για διαγνωστικούς και προεγχειρητικούς σκοπούς, ενώ σημαντικός είναι και ο ρόλος τους σε συστήματα αλληλεπίδρασης ανθρώπου–μηχανής και στη ρομποτική.

Προκλήσεις και Συνεισφορά της Διπλωματικής Εργασίας

Παρά την σημαντική εξέλιξη που έχει σημειωθεί στη κατεύθυνση της τρισδιάστατής ανακατασκευής προσώπου εξακολουθούν να παρουσιάζονται σημαντικές προκλήσεις, ιδιαίτερα σε ρεαλιστικές συνθήκες όπου το πρόσωπο δεν εμφανίζεται πλήρως, αλλά συχνά καλύπτεται από αντικείμενα, κινήσεις των χεριών ή προστατευτικές μάσκες. Η αποκλειστική χρήση οπτικών δεδομένων αποδεικνύεται περιοριστική, καθώς όταν μέρος του προσώπου κρύβεται χάνουμε σημαντική πληροφορία. Αντίθετα, η χρήση μόνο ακουστικών δεδομένων προσφέρει πλεονεκτήματα στη μοντελοποίηση της κίνησης των χειλιών και του συγχρονισμού με την ομιλία, στερείται όμως της χωρικής πληροφορίας που απαιτείται για την ακριβή αποτύπωση της γεωμετρίας και της εκφραστικότητας του προσώπου. Επιπλέον, η περιορισμένη διαθεσιμότητα εκτενών και υψηλής ποιότητας συνόλων δεδομένων που περιλαμβάνουν τέτοιου είδους σενάρια καθιστά δυσχερή την ανάπτυξη εύρωστων μοντέλων ικανών να λειτουργούν αξιόπιστα σε πραγματικές εφαρμογές. Συνεπώς, η συνδυαστική αξιοποίηση οπτικών και ακουστικών σημάτων, σε συνδυασμό με τη μοντελοποίηση των χρονικών εξαρτήσεων της ακολουθίας ενός βίντεο, αποτελεί κρίσιμο βήμα για την εξέλιξη και βελτίωση της τρισδιάστατης ανακατασκευής προσώπων.

Η παρούσα διπλωματική εργασία διερευνά την ανάπτυξη ενός νέου μοντέλου για την τρισδιάστατη ανακατασκευή προσώπου που λαμβάνει ως είσοδο βιντέο δεδομένα και αξιοποιεί τη οπτική και την ακουστική πληροφορία. Στόχος είναι η δημιουργία ενός εύρωστου συστήματος ικανού να παράγει ρεαλιστικά και ακριβή αποτελέσματα ακόμη και σε απαιτητικές συνθήκες, όπως αποκρύψεις τμημάτων του προσώπου ή μεταβολές φωτισμού και πόζας του προσώπου. Μέσω εκτεταμένων πειραματικών αξιολογήσεων, αποδεικνύεται η αποδοτικότητα του προτεινόμενου πλαισίου και αναδεικνύονται οι δυνατότητές του σε σχέση με τις υπάρχουσες μεθόδους.

Οι κύριες συνεισφορές της εργασίας είναι:

1. Ανάπτυξη ενός οπτικοακουστικού μοντέλου ανακατασκευής προσώπου, ισχυρό σε πραγματικές συνθήκες με αποκρύψεις προσώπου.
2. Μοντελοποίηση των χρονικών εξαρτήσεων στην ακολουθία βίντεο, ώστε οι εκφράσεις να αποτυπώνονται με συνέπεια και συνέχεια στο χρόνο.
3. Σχεδίαση μιας τριμορφικής αρχιτεκτονικής που υποστηρίζει τόσο οπτικοακουστική όσο και μόνο-ακουστική ή μόνο-οπτική είσοδο, επιτυγχάνοντας υψηλή απόδοση και για τα τρία συστήματα.
4. Ενσωμάτωση μιας ενισχυμένης συνάρτησης κόστους για τη μείωση της εσφαλμένης κίνησης των χειλιών.
5. Επαύξηση των δεδομένων εκπαίδευσης με συνθετικές αποκρύψεις του προσώπου για ενίσχυση της γενίκευσης του μοντέλου σε ρεαλιστικά σενάρια.
6. Εκτενής ποιοτική και ποσοτική αξιολόγηση, συμπεριλαμβανομένης μελέτης χρηστών και μιας μελέτης αφαίρεσης, που αποδεικνύει την υπεροχή της προτεινόμενης μεθόδου έναντι προηγούμενων προσεγγίσεων.

1.2 Μοντελοποίηση Προσώπου

Η μοντελοποίηση τρισδιάστατων προσώπων αποτελεί ένας ενδιαφέρον τομέας στην όραση υπολογιστών, με εφαρμογές στην αναγνώριση προσώπου, την ανάλυση εκφράσεων, την εικονική πραγματικότητα και την ιατρική. Υπάρχουν διάφοροι τρόποι αναπαράστασης της γεωμετρίας και της εμφάνισης του προσώπου που αφορούν είτε τη χρήση ογκομετρικών μεθόδων είτε τη χρήση 3Δ Μορφοποιήσιμων Μοντέλων όπως θα αναλυθούν παρακάτω.

Πλέγματα

Τα πλέγματα αποτελούν τον πιο διαδεδομένο τρόπο αναπαράστασης τρισδιάστατων επιφανειών. Ένα πλέγμα ορίζεται από σύνολα κορυφών και επιφανειών, συνήθως τριγώνων, τα οποία συνδέουν τις κορυφές για να σχηματίσουν μια συνεχή επιφάνεια. Η δομή αυτή είναι ιδιαίτερα ευέλικτη, καθώς υποστηρίζει παραμορφώσεις, μορφοποίηση και χαρτογράφηση υφής. Εκτός από τριγωνικά πλέγματα, υπάρχουν και πιο σύνθετες εκδοχές με τετράπλευρα ή πολύγωνα, ωστόσο σε αυτή τη διπλωματική χρησιμποιούμε τη πρώτη μορφή.

Σημεία Αναφοράς

Τα σημεία αναφοράς είναι χαρακτηριστικά γεωμετρικά σημεία του προσώπου, όπως οι γωνίες των ματιών, η κορυφή της μύτης και το περίγραμμα των χειλιών που μεταφέρουν χρήσιμη πληροφορία για τη γεωμετρία του. Στη διπλωματική αυτή χρησιμοποιούνται δύο σύνολα σημείων αναφοράς, Το πρώτο το εξάγουμε από το Face Alignment Network (FAN) [8], που εντοπίζει 68 2Δ αραιά σημεία μέσω ενός συνελικτικού νευρωνικού δικτύου το οποίο είναι εύρωστο σε διαφορετικές πόζες, εκφράσεις και φωτισμό. Το δεύτερο το αντλούμε από το MediaPipe Face Landmarker [33], το οποίο εντοπίζει 468 πυκνά σημεία με 3Δ συντεταγμένες, καλύπτοντας ολόκληρη την επιφάνεια του προσώπου. Ο συνδυασμός των δύο εξασφαλίζει συνέπεια στη γενική μορφή του προσώπου και λεπτομέρεια στις εκφράσεις.

Ογκομετρικές Μέθοδοι

Οι ογκομετρικές επιχειρούν να αναδομήσουν την τρισδιάστατη γεωμετρία ενός προσώπου με βάση πολλαπλές εικόνες ή μετρήσεις βάθους. Ο στόχος είναι να συνδυαστούν παρατηρήσεις από διαφορετικές οπτικές γωνίες σε μια συνεκτική, πυκνή και ακριβή τρισδιάστατη αναπαράσταση.

Το **Structure-from-Motion** (SfM) [64] είναι η βάση για πολλές μεταγενέστερες τεχνικές. Μέσω της ανίχνευσης και αντιστοίχισης σημείων-κλειδιών σε διαφορετικές εικόνες, υπολογίζει με ακρίβεια τις θέσεις των καμερών και δημιουργεί ένα αραιό νέφος 3D σημείων. Παράγει αξιόπιστες εκτιμήσεις για τη γεωμετρία, αλλά το αποτέλεσμα περιορίζεται μόνο σε λίγα σημεία. Το **Multi-View-Stereo** (MVS) [65] μπορεί να χτιστεί πάνω στο SfM. Συνήθως αξιοποιεί τις εκτιμήσεις για τη θέση της κάμερας από το SfM και υπολογίζει πυκνά νέφη σημείων ή πλέγματα εξετάζοντας την αντιστοιχία εικονοστοιχείων μεταξύ πολλαπλών όψεων.

Οι **Volumetric Fusion** μέθοδοι όπως το KinectFusion[47], εκμεταλλεύονται αισθητήρες βάθους για να συνθέσουν πολλαπλούς χάρτες βάθους σε ένα ενιαίο 3D μοντέλο. Το αντικείμενο αναπαρίσταται ως πλέγμα voxel, όπου κάθε voxel αποθηκεύει την απόσταση από την προσημασμένη επιφάνεια (Signed Distance Function – SDF). Καθώς νέες λήψεις προστίθενται, το μοντέλο γίνεται πιο πλήρες και λείο. Μια πρόσφατη προσέγγιστη είναι η χρήση **Neural Radiance Fields** (NeRFs) [20]. Αντί για ρητά γεωμετρικά πλέγματα, μαθαίνουν συνεχείς όγκους μέσω νευρωνικών δικτύων, τα οποία αντιστοιχούν 3Δ συντεταγμένες και κατευθύνσεις θέασης σε χρώμα και πυκνότητα. Έτσι, επιτυγχάνεται ρεαλιστική απόδοση νέων οπτικών γωνιών και καταγραφή λεπτομερειών όπως υφή δέρματος ή ακόμη και τρίχες.

Τρισδιάστατα Μορφοποιήσιμα Μοντέλα

Οι ογκομετρικές μέθοδοι παρέχουν υψηλής ακρίβειας αναπαραστάσεις, αλλά συχνά παράγουν μοντέλα χωρίς εννοιολογική παραμετροποίηση (π.χ. διάκριση των παραμέτρων ταυτότητας, έκφρασης). Για να ξεπεραστούν οι περιορισμοί των καθαρά γεωμετρικών μεθόδων, αναπτύχθηκαν τα 3Δ Μορφοποιήσιμα Μοντέλα (3ΔΜΜ) [62]. Εισήχθησαν από τους Blanz και Vetter [6], οι οποίοι πρότειναν μια στατιστική, παραμετρική αναπαράσταση του προσώπου. Η διαδικασία ξεκινά με συλλογή σαρώσεων υψηλής ανάλυσης, οι οποίες ευθυγραμμίζονται ώστε όλα τα πρόσωπα να έχουν κοινή τοπολογία (ίδιο αριθμό κορυφών και κοινά σημεία αναφοράς). Έπειτα εφαρμόζεται Ανάλυση Κυρίων Συνιστωσών (ΑΚΣ), ώστε να εξαχθούν οι κύριες διαστάσεις μεταβλητότητας σχήματος και υφής. Έτσι, κάθε νέο πρόσωπο περιγράφεται με λίγες παραμέτρους (συντελεστές ΑΝΚ), που είναι ερμηνεύσιμες και συμπαγείς. Το βασικό πλεονέκτημα είναι ότι τα 3ΔΜΜ προσφέρουν σημασιολογικό έλεγχο. Μπορούμε, δηλαδή, να αλλάξουμε την ταυτότητα ή την έκφραση του προσώπου μεταβάλλοντας λίγες μόνο παραμέτρους, χωρίς να χρειάζεται να επεξεργαστούμε όλο το πλέγμα.

Ένα ευρέως διαδεδομένο μοντέλο στην ανακατασκευή του ανθρώπινου προσώπου, το οποίο χρησιμοποιούμε και στη παρούσα διπλωματική είναι το **FLAME** (Faces Learned with an Articulated Model and Expressions) [40]. Αποτελεί ένα στατιστικό μοντέλο κεφαλιού που επιτυγχάνει τον διαχωρισμό των τριών βασικών συνιστωσών του προσώπου: σχήμα, πόζα και έκφραση. Για τη μοντελοποίηση του χρησιμοποιεί τριγωνικό πλέγμα, με 5023 κορυφές, και ενσωματώνει 4 αρθρωτές συνδέσεις για τα μάτια, το λαιμο και το πηγούνι. Τα δεδομένα εκπαίδευσης προήλθαν από ακολουθίες τρισδιάστατων σαρώσεων τα οποία έχουν την ίδια τοπολογία, επιτρέποντας στο μοντέλο να μάθει ρεαλιστικά σχήματα παραμόρφωσης. Συγκεκριμένα, κάθε πρόσωπο στο FLAME χρησιμοποιεί τρια διακρητά σχήματα παραμόρφωσης που κωδικοποιούν τις μεταβολές στο σχήμα, τη πόζα και την έκφραση διαφορετικών ατόμων. Το τελικό πλέγμα προσώπου προκύπτει από την πρόσθεση αυτών των συνιστωσών στο μέσο πρότυπο πλέγμα, και εφαρμόζεται η διαδικασία Γραμμικής Παραμόρφωσης

Σκελετού για να δημιουργήσει φυσικές και ρεαλιστικές κινήσεις, με κάθε κορυφή του πλέγματος να επηρεάζεται σε διαφορετικό βαθμό από τις κοντινές αρθρώσεις.

Το FLAME αποτελεί εξέλιξη του μοντέλου σώματος SMPL [41], το οποίο έχει επεκταθεί ειδικά για την αναπαράσταση του κεφαλιού. Παράλληλα η παραμετροποίηση που περιγράφηκε επιτρέπει στο μοντέλο την παραγωγή ρεαλιστικών, συνεπών και ελεγχόμενων τρισδιάστατων προσώπων, καθιστώντας το FLAME ένα από τα πιο αξιόπιστα και ευρέως χρησιμοποιούμενα μοντέλα στην κατεύθυνση της τρισδιάστατης μοντελοποίησης προσώπου.

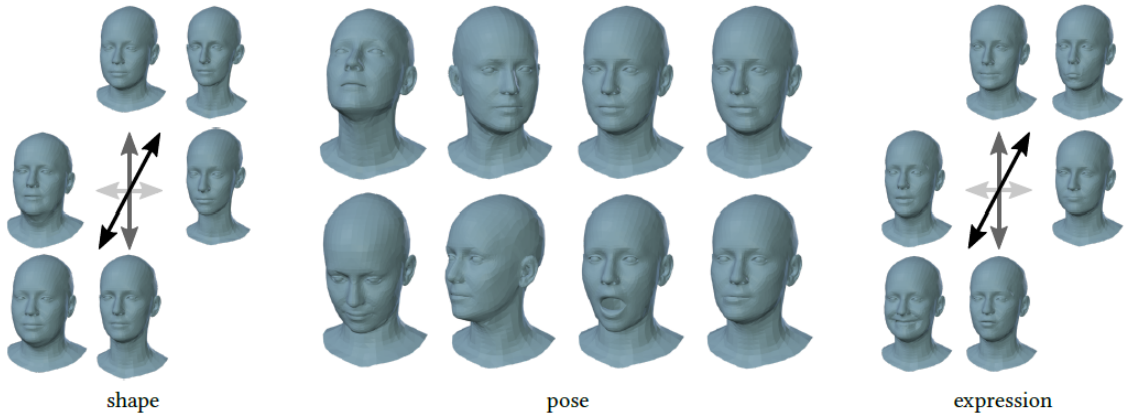

Figure 1.2.1: Παραμετροποίηση του FLAME μοντέλου. Αριστερά: Ενεργοποίηση των τριών πρώτων συνιστωσών σχήματος. Κέντρο: Παράμετροι πόζας που κινούν τέσσερις από τις έξι αρθρώσεις του λαιμού και της γνάθου. Δεξιά: Ενεργοποίηση των τριών πρώτων συνιστωσών έκφρασης. [40]

Άλλα γνωστά 3ΔΜΜ μοντέλα που προηγήθηκαν του FLAME αποτελούν το Basel Face Model (BFM) [53] και το Large Scale Face Model (LSFM) [7]. Και τα δύο μοντέλα βασίζονται στην Ανάλυση Κυρίων Συνιστωσών για τη μοντελοποίηση του σχήματος και της υφής. Η βασική τους διαφορά είναι ότι το δεύτερο κατασκευάστηκε από πολύ μεγαλύτερο αριθμό σαρώσεων, με σημαντικά πλουσιότερη ποικιλομορφία σκαναρισμένων προσώπων όσον αφορά την ηλικία, το φύλο και την εθνικότητα, προσφέροντας έτσι μεγαλύτερη γενίκευση και ακρίβεια σε σχέση με το BFM.

1.3 Βιβλιογραφία μεθόδων για Τρισδιάστατη Ανακατασκευή Ανθρώπινου Προσώπου

Η τρισδιάστατη ανακατασκευή ανθρώπινου προσώπου έχει γνωρίσει εντυπωσιακή πρόοδο την τελευταία δεκαετία, κυρίως χάρη στην αξιοποίηση μεγάλων συνόλων δεδομένων και στις εξελίξεις της βαθιάς μάθησης. Οι υπάρχουσες προσεγγίσεις μπορούν να κατηγοριοποιηθούν σε τρεις βασικές ομάδες, ανάλογα με τον τύπο δεδομένων εισόδου που χρησιμοποιούν: (α) μεθόδους βασισμένες στην εικόνα ή στο βίντεο, (β) μεθόδους βασισμένες στον ήχο και (γ) πολυτροπικές μεθόδους που συνδυάζουν ταυτόχρονα τις δύο προηγούμενες. Κάθε κατηγορία έχει πλεονεκτήματα και περιορισμούς, γεγονός που καθιστά την επιλογή της κατάλληλης προσέγγισης κρίσιμη για την εκάστοτε εφαρμογή. Στην παρούσα εργασία, βασιζόμαστε στην τρίτη κατηγορία, αξιοποιώντας τη συνδυαστική υπερόχη εικόνας και ήχου για την ανακατασκευή 4Δ προσώπων με μεγαλύτερη ακρίβεια και χρονική συνέπεια.

Μέθοδοι βασισμένοι στην οπτική πληροφορία

Οι μέθοδοι που βασίζονται στην εικόνα προσπαθούν να αναπαραστήσουν τη γεωμετρία του προσώπου από μία ή περισσότερες εικόνες, συνήθως μέσω της προσαρμογής παραμέτρων σε ένα 3Δ Μορφοποιήσιμο Μοντέλο. Στην περίπτωση όπου δέχονται ως είσοδο βίντεο, επιχειρούν όχι μόνο να εκτιμήσουν τη γεωμετρία σε κάθε καρέ, αλλά και να διατηρήσουν τη χρονική συνοχή τους στις εκφράσεις και τις κινήσεις του προσώπου.

Το **DECA** (Detailed Expression Capture and Animation) [17] αποτελεί σημείο αναφοράς στο πεδίο της τρισδιάστατης ανακατασκευής προσώπου, καθώς επικεντρώνεται στην ανάδειξη λεπτομερειών της γεωμετρίας του προσώπου. Σε ένα αρχικό στάδιο εκτιμά μία χονδροειδή ανακατασκευή του προσώπου στον χώρο αναπαράστασης του FLAME μέσω της στρατηγικής ανάλυση-μέσω-σύνθεσης. Σε δεύτερο στάδιο εμπλουτίζει τη γεωμετρία του με λεπτομέρειες όπως ρυτίδες και υφές δέρματος χρησιμοποιώντας χάρτες παραμόρφωσης σε χώρο UV. Με αυτόν τον τρόπο, το DECA μοντέλο καταφέρνει να αναπαριστά δυναμικές λεπτομέρειες με υψηλό ρεαλισμό ακόμη και σε σύνολο δεδομένων από πραγματικές συνθήκες.

Το **EMOCA** (EMOtion Capture and Animation) μοντέλο [13] βασίζεται πάνω στο DECA, εστιάζοντας ωστόσο στην ακριβή αποτύπωση συναισθηματικών εκφράσεων. Ενσωματώνει μια καινοτόμα συνάρτηση κόστους βασισμένη σε αντιληπτικά χαρακτηριστικά συναισθήματος, ώστε το ανακατασκευασμένο πρόσωπο να διατηρεί συναισθηματική συνέπεια με το αρχικό.

Μία προσέγγιση που στοχεύει στην αποτύπωση ακραίων και ασύμμετρων εκφράσεων είναι το **SMIRK** [59]. Το μοντέλο αξιοποιεί τον χώρο αναπαράστασης του FLAME και περιλαμβάνει δύο βήματα. Στο πρώτο βήμα, η εικόνα εισόδου κωδικοποιείται σε παραμέτρους του FLAME και περνά από έναν renderer. Το αποτέλεσμα αυτό μαζί με μια μάσκα του προσώπου συνοδευόμενη από λίγα εικονοστοιχεία της αρχικής εικόνας, τροφοδοτούν ένα διαφορικό νευρωνικό renderer που παράγει μια φωτορεαλιστική ανακατασκευή. Στο δεύτερο βήμα, εφαρμόζεται μια επαυξημένη έκφραση στην αναπαράσταση FLAME, και ακουληθεί ένα δεύτερο βήμα εκπαίδευσης. Ο στόχος είναι οι παράμετροι έκφρασης που τελικά προβλέπει ο κωδικοποιητής να ταυτίζονται με τις αρχικές τροποποιημένες παραμέτρους, διασφαλίζοντας έτσι ότι το μοντέλο μπορεί να αναπαράγει λεπτομερή ανακατασκευές με πολύπλοκες εκφράσεις.

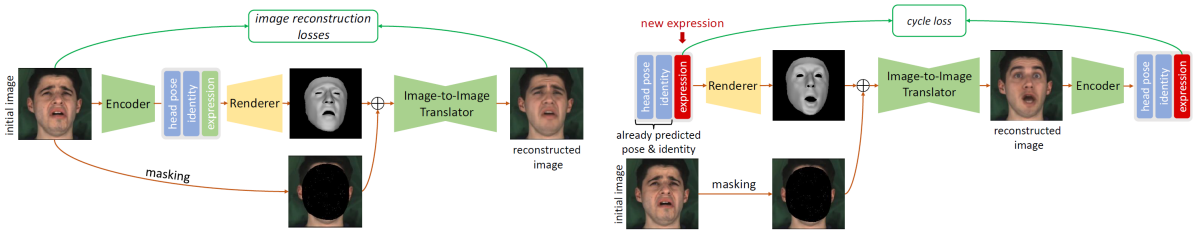

Figure 1.3.1: Η εκπαίδευση του SMIRK πραγματοποιείται σε δύο βήματα [59].

Μια μέθοδος που επικεντρώνεται στην ακριβή αναπαράσταση των κινήσεων της στοματικής περιοχής κατά την ομιλία είναι το **SPECTRE** [18]. Η προσέγγιση αυτή αξιοποιεί ένα προεκπαιδευμένο μοντέλο αναγνώρισης χειλιών τόσο στις αρχικές όσο και στις ανακατασκευασμένες εικόνες, εξάγοντας χαρακτηριστικά διανύσματα που περιγράφουν την άρθρωση των χειλιών. Ο στόχος είναι η ελαχιστοποίηση της διαφοράς ανάμεσα στα δύο σύνολα χαρακτηριστικών, ώστε να διασφαλιστεί η πιστότητα της κίνησης του στόματος κατά την ομιλία.

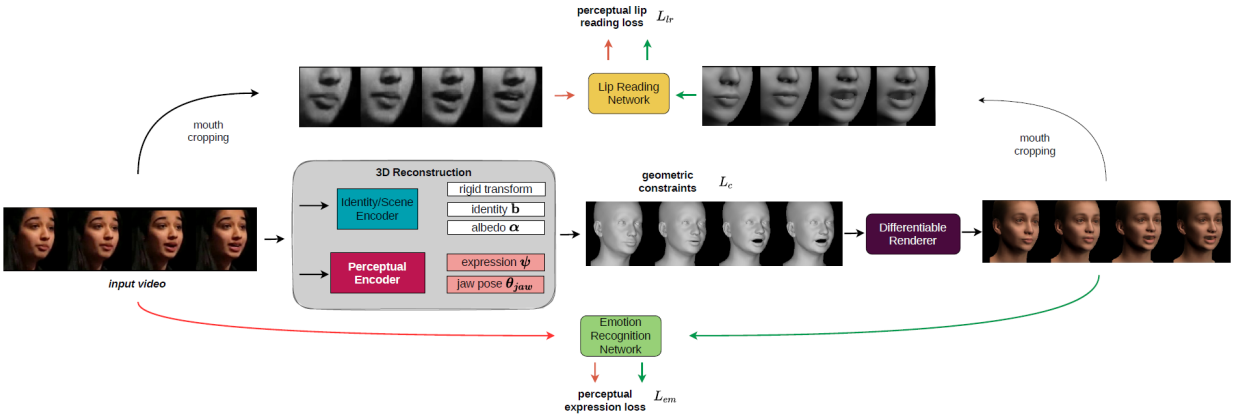

Figure 1.3.2: Επισκόπηση της αρχιτεκτονικής SPECTRE. Η περιοχή του στόματος αποκόπτεται τόσο από το αρχικό όσο και από το ανακατασκευασμένο βίντεο, και εφαρμόζεται ένα δίκτυο ανάγνωσης χειλιών για τον υπολογισμό της αντιληπτικής συνάρτησης κόστου αναγνώρησης χειλιών. [18].

Παρά την υψηλή απόδοσή τους, οι προαναφερθείσες μέθοδοι εμφανίζουν κοινούς περιορισμούς που απορρέουν από την αποκλειστική τους εξάρτηση στην οπτική πληροφορία. η οποία συχνά μπορεί να περιλαμβάνει αποκρύψεις ή να είναι ανεπαρκής.

Μέθοδοι βασισμένοι στην Ακουστική Πληροφορία

Οι μέθοδοι που λαμβάνουν ως είσοδο ηχητικά σήματα βασίζονται στη συσχέτιση της ακουστικής πληροφορίας με την άρθρωση του προσώπου και παράγουν άβαταρ των οποίων οι κινήσεις των χειλιών είναι απόλυτα συγχρονισμένες με την ομιλία.

VOCA (Voice Operated Character Animation) [12] αποτελεί μία μέθοδο δημιουργίας ομιλούντων τρισδιάστατων προσώπων. Ως είσοδο λαμβάνει ένα πρότυπο πλέγμα στον χώρο του FLAME, μαζί με μια παράμετρο που καθορίζει την ταυτότητα του υποκειμένου, καθώς και το αντίστοιχο ηχητικό σήμα. Μέσω αρχιτεκτονικής κωδικοποιητή–αποκωδικοποιητή, το σύστημα προβλέπει τις τρισδιάστατες μετατοπίσεις των κορυφών του προτύπου πλέγματος, οι οποίες εφαρμόζονται για να παραχθεί το τελικό πρόσωπο σε ουδέτερη πόζα, συγχρονισμένο με το σήμα της ομιλίας.

MeshTalk [60] εισάγει έναν διακριτό λανθάνοντα χώρο εκφράσεων, όπου η πληροφορία από το σήμα ομιλίας συνδυάζεται με πληροφορία γεωμετρίας, αποδίδοντας συνεπείς ακολουθίες πλεγμάτων. Η διακριτοποίηση των εκφράσεων σε κατηγορίες καθιστά το μοντέλο ικανό να παράγει χρονικά συνεπείς και εκφραστικές ακολουθίες που αντικατοπτρίζουν την άρθρωση του λόγου.

FaceFormer [16] αξιοποιεί αρχιτεκτονική μετασχηματιστών για να αντιμετωπίσει τον περιορισμό μακροχρόνιων εξαρτήσεων. Έτσι καταφέρνει να αποδώσει χρονικά συνεπείς και φυσικές κινήσεις που ανταποκρίνονται στο σήμα εισόδου.

CodeTalker [73] αντί να προβλέπει συνεχείς παραμέτρους, μαθαίνει ένα διακριτό λεξιλόγιο κινήσεων μέσω διακρητών διανυσμάτων αυτοκωδικοποιητών. Με αυτόν τον τρόπο περιορίζει τον χώρο εξόδου σε ρεαλιστικά μοτίβα κίνησης και βελτιώνει τον συγχρονισμό χειλιών.

Οι μέθοδοι αυτές είναι ιδιαίτερα χρήσιμες, αλλά συχνά αποτυγχάνουν να αποδώσουν πλήρη εκφραστικότητα όταν απουσιάζει οπτική πληροφορία.

Μέθοδοι βασισμένοι στην Οπτικο-Ακουστική Πληροφορία

Η τρίτη κατηγορία, και η πιο πρόσφατη, συνδυάζει ήχο και εικόνα ώστε να αξιοποιήσει τα συμπληρωματικά πλεονεκτήματα και των δύο. Η εικόνα προσφέρει γεωμετρική ακρίβεια και λεπτομερή απεικόνιση των εκφράσεων, ενώ ο ήχος προσφέρει συμπληρωματική πληροφορία σε περιπτώσεις ανεπαρκούς ή ελλιπούς οπτικής εισόδου. Δεν υπάρχει εκτεταμένη βιβλιογραφία σε αυτή τη κατεύθυνση, ενώ σύμφωνα με όσα γνωρίζουμε το **AVFace** [11] αποτελεί το πρώτο ολοκληρωμένο μοντέλο για πολυτροπική ανακατασκευή 4D προσώπου. Η αρχιτεκτονική του συνδυάζει ένα δίκτυο ResNet-50 και χρήση μετασχηματιστών για τα χωροχρονικά χαρακτηριστικά. Η εκπαίδευση ακολουθεί δύο στάδια: αρχικά πραγματοποιείται μια χονδροειδής ανακατασκευή της γεωμετρίας και στη συνέχεια αυτή εμπλουτίζεται με λεπτομέρειες στην υφή και στις εκφράσεις. Επιπλέον, το σύστημα εμπλουτίζει το σύνολο δεδομένων εκπαίδευσης με συνθετικές αποκρύψεις προσώπου, ώστε να ενισχύσει την ευρωστία του μοντέλου σε πραγματικές συνθήκες.

1.4 Προτεινόμενη Μεθοδολογία

Η παρούσα εργασία προτείνει μια ολοκληρωμένη μεθοδολογία για τρισδιάστατη ανακατασκευή προσώπων από βίντεο, με στόχο τη ρεαλιστική και ακριβή αποτύπωση τόσο της γεωμετρίας του προσώπου όσο και των κινήσεων του στόματος σε δύσκολες ρεαλιστικες συνθήκες. Σε αντίθεση με τις περισσότερες υπάρχουσες προσεγγίσεις που βασίζονται αποκλειστικά στην οπτική πληροφορία, το προτεινόμενο σύστημα συνδυάζει οπτικά και ακουστικά χαρακτηριστικά, με ιδιαίτερη έμφαση στην αντιμετώπιση δύο βασικών προβλημάτων: (α) τις αποκρύψεις προσώπου και (β) το θορυβώδες περιβάλλον.

Συνθετικές Αποκρύψεις Προσώπου

Δεδομένου του περιορισμού των διαθέσιμων συνόλων δεδομένων να περιέχουν βίντεο με φυσικές αποκρύψεις, ενισχύουμε το σύνολο εκπαίδευσης δημιουργώντας συνθετικές αποκρύψεις χρησιμοποιώντας δύο διαφορετικούς τύπους.

Η Εικόνα 1.4.1 παρουσιάζει αποκρύψεις από διαφορετικά αντικείμενα/χέρια: Για τυχαίο κάθε φορά μήκος ακολουθίας βίντεο από τα δεδομένα μας, συνθέτουμε γεγονότα απόκρυψης όπου ένας αποκρυπτικός παράγοντας (χέρι ή αντικείμενο) ακολουθεί καμπύλη τροχιά μπροστά από το πρόσωπα. Εισέρχεται, δηλαδή, από το όριο ενός πρώτου καρέ, κινείται προς την περιοχή του στόματος, και εξέρχεται από την απέναντι πλευρά σε μεταγενέστερη χρονική στιγμή του βίντεο. Προσθέτουμε, επίσης ομαλές περιστροφές του αντικειμένου, δημιουργώντας δυναμικά γεγονότα που κάθε χρονική στιγμή αποκρύβεται διαφορετικό μέρος του προσώπου.

Figure 1.4.1: Εφαρμογή συνθετικών αποκρύψεων προσώπου που απεικονίζουν χέρια ή αντικείμενα.

Στην Εικόνα 1.4.2 βλέπουμε τον δεύτερο τύπο αποκρύψεων όπου εφαρμόζουμε μία μάσκα στην περιοχή του στόματος. Χρησιμοποιούμε τα Mediapipe σημεία αναφοράς για τον ορισμό πολυγώνων

γύρω από τη γνάθο και τη μύτη, τα οποία γεμίζουμε με διαφορετικές υφές από ένα εκτενές σύνολο δεδομένων. Με αυτό τον τρόπο παράγουμε δεδομένα στα οποία ολόκληρη η περιοχή του στόματος αποκρύπτεται.

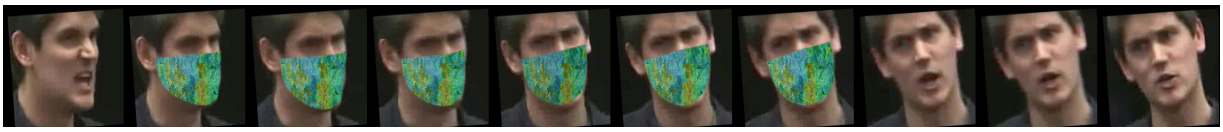

Figure 1.4.2: Εφαρμοφή συνθετικής χειρουργικής μάσκας στο εικονιζόμενο πρόσωπο.

Επεξεργασία Δεδομένων

Τα βίντεο χωρίζονται σε τμήματα που αποτελούνται από K καρέ. Για κάθε ένα εξάγονται τα σημεία ανάφορας του προσώπου από το MediaPipe (468 σημεία) και από το FAN (68 σημεία). Για λόγους αποδοτικότητας, διατηρούμε ένα συμπαγές σύνολο από 126 σημεία αναφοράς, συνδυάζοντας αυτά της γνάθου από το FAN με τα σημεία του στόματος/ματιών/μύτης από το MediaPipe. Κάθε δείγμα περιλαμβάνει το πρωτότυπο και το βίντεο με τις συνθετικές αποκρύψεις, τα σημεία αναφοράς, εξάγεται η μάσκα περιγράμματος του προσώπου (hull μάσκα) και το αντίστοιχο ηχητικό σήμα.

Θεωρητικό Υπόβαθρο

Για τη μοντελοποίηση του προσώπου υιοθετούμε το FLAME [40], ένα στατιστικό 3Δ μορφοποιήσιμο μοντέλο προσώπου με παραμέτρους ταυτότητας, έκφρασης και πόζας. Για το ακουστικό κομμάτι χρησιμοποιούμε το **wav2vec** [4], ένα προεκπαιδευμένο μοντέλο αναπαράστασης ομιλίας, το οποίο βασίζεται στην αρχιτεκτονική των μετασχηματιστών. Το wav2vec έχει εκπαιδευτεί σε μεγάλο όγκο μη επισημασμένων δεδομένων ομιλίας και στη συνέχεια έχει προσαρμοστεί για την εφαρμογή της αναγνώρισης φωνής, παρέχοντας έτσι πλούσιες αναπαραστάσεις φωνημάτων. Η αρχιτεκτονική του περιλαμβάνει έναν συνελικτικό κωδικοποιητή χαρακτηριστικών που μετατρέπει το ηχητικό σήμα σε αναπαριστάσεις διανυσμάτων, καθώς και ένα δίκτυο μετασχηματιστή που μοντελοποιεί τις χρονικές εξαρτήσεις στην ακολουθία.

Αρχιτεκτονική

Η προτεινόμενη αρχιτεκτονική βασίζεται στο μοντέλο **SMIRK** [59], ένα καινοτόμο μοντέλο για την αποτύπωση ποικίλων εκφράσεων, το οποίο επεκτείνουμε ώστε να υποστηρίζει πολυτροπική είσοδο (εικόνα και ήχο) και να είναι εύρωστο σε αποκρύψεις. Χρησιμοποιούμε έναν κωδικοποιητή $E(\cdot)$ που δέχεται ως είσοδο μια ακολουθία μεγέθους K με συνθετικές αποκρύψεις $I^{\text{occl}}_{1:K} = \{I^{\text{occl}}_1, I^{\text{occl}}_2, \ldots, I^{\text{occl}}_K\}$ και προβλέπει τις παραμέτρους του FLAME. Ακολουθώντας τον σχεδιασμό του SMIRK, ο κωδικοποιητής χωρίζεται σε τρεις επιμέρους κλάδους:

Ο **Κωδικοποιητής Πόζας** E_θ υπολογίζει τις παραμέτρους πόζας $\theta_{1:K} = \{\theta_1, \ldots, \theta_K\}$ (θέση και περιστροφή προσώπου) και βασίζεται σε ένα δίκτυο MobileNetV3:

$$\theta_{1:K} = E_\theta(I^{\text{occl}}_{1:K})$$

Ο **Κωδικοποιητής Σχήματος** E_β εκτιμά τις παραμέτρους ταυτότητας $\beta_{1:K} = \{\beta_1, \ldots, \beta_K\}$, που περιγράφουν το σταθερό σχήμα του προσώπου. Βασίζεται στο MobileNetV3 δίκτυο και χρησιμοποιεί επιπλέον 1Δ συνελικτικά στρώματα κατά μήκος της χρονικής διάστασης, ώστε να μοντελοποιήσει τις χρονικές εξαρτήσεις της ακολουθίας βίντεο:

$$\beta_{1:K} = E_\beta(I^{\text{occl}}_{1:T})$$

Ο **Κωδικοποιητής Εκφράσεων** E_ψ δέχεται ως είσοδο τόσο οπτική πληροφορία όσο και ηχητικά σήματα. Η οπτική πληροφορία κωδικοποιείται και πάλι με χρήση του MobileNetV3 δικτύου, ενώ για το ακουστικό σήμα εξάγονται χαρακτηριστικά αξιοποιώντας το **wav2vec**.

Τα ηχητικά χαρακτηριστικά προβάλλονται στην ίδια διάσταση με τα οπτικά ώστε να μοιράζονται έναν κοινό ενδιάμεσο χώρο. Οι δύο τύποι δεδομένων ενώνονται και εισάγονται σε ένα χρονικό συνελικτικό δίκτυο, το οποίο μαθαίνει τις χρονικές εξαρτήσεις. Το τελικό αποτέλεσμα υπολογίζει τις παραμέτρους έκφρασης $\psi_{1:K} = \{\psi_1, \ldots, \psi_K\}$ και μοντελοποιείται ως εξής:

$$\psi_{1:K} = E_\psi(I^{\text{occl}}_{1:K})$$

Κατά τη διάρκεια της εκπαίδευσης, οι κωδικοποιητές πόζας και σχήματος είναι προεκπαιδευμένοι και παραμένουν σταθεροί, ενώ ο κωδικοποιητής εκφράσεων εκπαιδεύεται πλήρως.

Για την προβολή του ανακατασκευασμένου πλέγματος του FLAME χρησιμοποιούμε **διαφορικό renderer** του οποίου το σφάλμα ρέει πίσω στις παραμέτρους FLAME και στον κωδικοποιητή. Τυπικά:

$$S_{1:K} = R(\theta_{1:K}, \beta_{1:K}, \psi_{1:K})$$

όπου S αντιστοιχεί στον renderer του ανακατασκευασμένου πλέγματος.

Το τελευταίο βήμα είναι ένας **γεννήτορας τύπου U-Net**, ο οποίος λαμβάνει ως είσοδο το αποτέλεσμα του renderer και το αποτέλεσμα καθοδηγείται από το αρχικό βίντεο στο οποίο έχουμε εφαρμόσει μια ειδικά σχεδιασμένη μάσκα σε όλο το πρόσωπο. Συγκεκριμένα μία συνάρτηση μασκαρίσματος $M(\cdot)$ εφαρμόζεται στο βίντεο εισόδου $I_{1:K}$ χρησιμοποιώντας τη hull μάσκα ώστε να καλύπτει το πρόσωπο και να κρατά μόνο λίγα τυχαία εικονοστοιχεία. Ο συνδυασμός $S_{1:K} \oplus M(I_{1:K})$ περνά από τον γεννήτορα T για να παραχθεί η τελική ανακατασκευή:

$$I'_{1:K} = T(S_{1:K} \oplus M(I_{1:K})$$

Αυτό το βήμα είναι κρίσιμο διότι το δίκτυο καλείται να ανακατασκευάσει την μη-αποκρυμμένη εικόνα, ώστε να μάθει να αγνοεί τις αποκρύψεις και να παράγει καθαρές, ρεαλιστικές προβλέψεις προσώπων. Η αρχιτεκτονική του μοντέλου παρουσιάζεται στο σχήμα 1.4.3

Για την αποτελεσματική κύρια εκπαίδευση του μοντέλου, η συνάρτηση κόστους που χρησιμοποιείται αποτελείται από:

Φωτομετρικά Σφάλματα. Υπολογίζουμε το σφάλμα L1 μεταξύ των αρχικών καρέ I του βίντεο και των ανακατασκευασμένων I':

$$\mathcal{L}_{\text{photo}} = \|I' - I\|_1.$$

Αυτή η απώλεια εξασφαλίζει συνέπεια σε επίπεδο εικονοστοιχείων και διατηρεί την οπτική εγγύτητα με το βίντεο αναφοράς.

Σφάλματα Αντιληπτικής Ομοιότητας με χρήση του VGG. Για την ενίσχυση της αντιληπτικής ομοιότητας και την ταχύτερη σύγκλιση στα αρχικά στάδια εκπαίδευσης, χρησιμοποιούμε το VGG μοντέλο [32]:

$$\mathcal{L}_{\text{vgg}} = \|\Gamma(I') - \Gamma(I)\|_1,$$

όπου $\Gamma(\cdot)$ συμβολίζει τα χαρακτηριστικά που εξάγονται από έναν προεκπαιδευμένο κωδικοποιητή VGG. Η απώλεια αυτή δίνει έμφαση σε δομικά στοιχεία και στοιχεία υφής στοιχεία που δεν αποτυπώνονται από τις διαφορές σε επίπεδο εικονοστοιχείων.

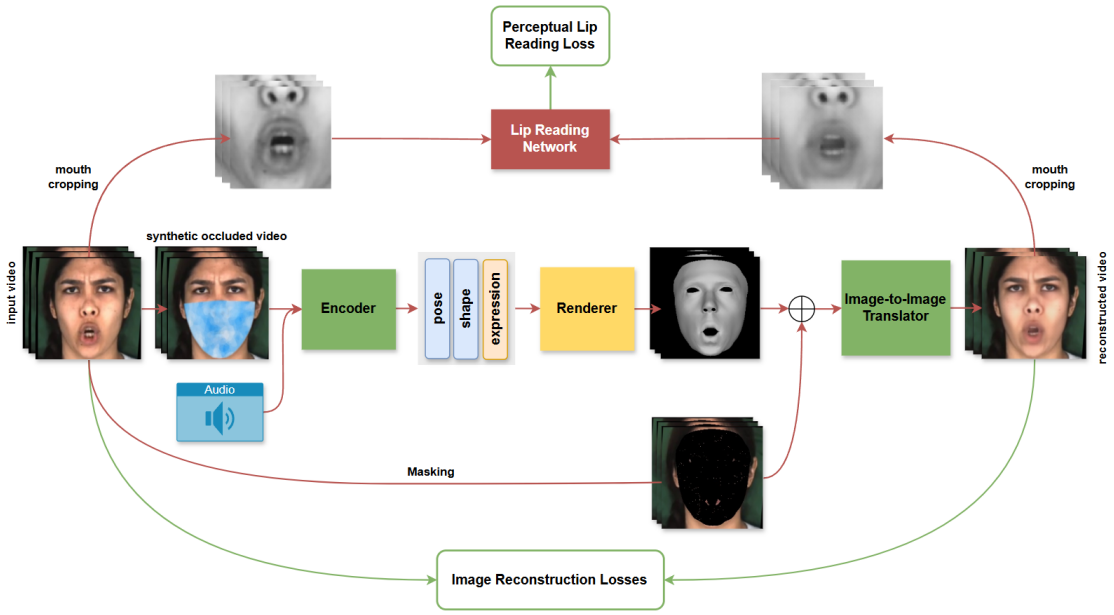

Figure 1.4.3: Συνολική αρχιτεκτονική του προτεινόμενου μοντέλου κατά την εκπαίδευση. Η είσοδος περνά από τον encoder που προβλέπει τις FLAME παραμέτρους. Ακολουθεί η κατασκευή του άβαταρ και τελικά παράγεται το ανακατασκευασμένο βίντεο μέσω του γενήτορα.

Σφάλματα Σημείων Αναφοράς. Για τη διατήρηση της γεωμετρικής συνέπειας στη δομή του προσώπου, ελαχιστοποιούμε την απόσταση L_2 μεταξύ των πραγματικών δισδιάστατων σημείων αναφοράς k και των προβλεπόμενων σημείων k':

$$\mathcal{L}_{\text{lmk}} = \sum_{i=1}^{N} \|k_i - k'_i\|_2^2,$$

όπου N είναι ο συνολικός αριθμός σημείων. Έτσι διασφαλίζεται ακρίβεια στη γεωμετρία του προσώπου, ιδίως στα μάτια, στη μύτη και στο στόμα.

Σφάλματα Αναγνώρισης Χειλιών. Εμπνευσμένοι από την προσέγγιση του SPECTRE [18], εξάγουμε τα χαρακτηριστικά του ανακατασκευασμένου βίντεο ϵ_I χρησιμοποιώντας ένα μοντέλο αναγνώρισης χειλιών, και υπολογίζουμε την απόστασή τους από τα αντίστοιχα χαρακτηριστικά του πραγματικού ϵ_R:

$$\mathcal{L}_{\text{lr}} = \frac{1}{K} \sum_{k=1}^{K} d(\epsilon_{I_k}, \epsilon_{R_k}),$$

όπου $d(\cdot)$ δηλώνει τη συνημιτονική απόσταση και K τον αριθμό καρέ στη χρονική ακολουθία.

Κανονικοποίηση εκφράσεων. Για την αποφυγή υπερβολικών ή μη ρεαλιστικών παραμορφώσεων, επιβάλλουμε ποινή L_2 στις παραμέτρους έκφρασης:

$$\mathcal{L}_{\text{reg}} = \|\psi\|_2^2,$$

ώστε να ενισχύονται πιο ομαλές και φυσικές εκφράσεις.

Η εκπαίδευση γίνεται σε δύο φάσεις:

1. **Προεκπαίδευση με μόνο οπτικά δεδομένα:** Η συνάρτηση κόστους περιλαμβάνει τα σφάλματα των σημείων αναφοράς, ενώ ο κωδικοποιητής της πόζας εκπαιδεύεται επιπλέον με

επίβλεψη βασισμένη στις προβλέψεις του μοντέλου MICA. Χρησιμοποιούμε επιπλέον σφάλματα για κανονικοποίηση της πόζας και της έκφρασης, ώστε να αποφεύγονται ακραίες παραμορφώσεις.

2. **Κύρια εκπαίδευση πολυτροπικού μοντέλου με χρήση και των δύο εισόδων (εικόνας και ήχου):**, Χρησιμοποιούμε τη συνολική συνάρτηση κόστους που περιγράφηκε παραπάνω για την εκπαίδευση μόνο του κωδικοποιητή έκφρασης.

Η βελτιστοποίηση πραγματοποιείται με Adam βελτιστοποιητή, με ρυθμό μάθησης 10^{-3}, μέγεθος παρτίδας 6 και χρήση 20 καρέ άνα βίντεο εισόδου. Η εκπαίδευση διαρκεί 10 εποχές και ολοκληρώνεται σε περίπου 1.5 ημέρα σε GPU NVIDIA L40S.

1.5 Πειράματα - Αξιολόγηση

Σύνολα Δεδομένων

Για την εκπαίδευση και αξιολόγηση του μοντέλου χρησιμοποιήθηκαν τα σύνολα MEAD, LRS3, CelebV-Text και ViCo Listeners, καθώς και πρόσθετες συνθετικές αποκρύψεις. Το MEAD παρείχε δεδομένα υψηλής ποιότητας σε εργαστηριακές συνθήκες με ποικιλία εκφράσεων. Το LRS3 κάλυψε σκηνές ρεαλιστικών συνθηκών με μεγάλο όγκο ομιλιών TED/TEDx. Το CelebV-Text προσέφερε παραδείγματα μερικής ή πλήρους απόκρυψης προσώπου, ενώ το ViCo, σε συνδυασμό με το ESC-50, χρησιμοποιήθηκε για σενάρια με ακροατές αλλά χωρίς ομιλία. Τέλος, δημιουργήθηκαν συνθετικές αποκρύψεις προσώπου ώστε να ενισχυθεί η ανθεκτικότητα του συστήματος σε ρεαλιστικές συνθήκες καταγραφής του προσώπου.

Η αξιολόγηση πραγματοποιήθηκε σε επιλεγμένα δείγματα από τα σύνολα δεδομένων CelebV-Text και CelebV-HQ, καθώς και σε δύο βίντεο που καταγράψαμε στο εργαστήριο. Για την απομαγνητοφώνηση χρησιμοποιήθηκε το σύστημα Whisper, προκειμένου να ελεγχθεί η αντιστοίχιση λόγου–κινήσεων χειλιών. Επειδή δεν υπάρχει καθιερωμένο benchmark για πρόσωπα με φυσικές αποκρύψεις, η αξιολόγηση βασίστηκε σε έναν συνδυασμό ποιοτικής ανάλυσης, μετρικών ανάγνωσης χειλιών και μία μελέτη χρηστών.

Ποιοτική Αξιολόγηση

Το μοντέλο απέδωσε με συνέπεια σε διαφορετικούς τύπους απόκρυψης όπως παρουσιάζεται στο σχήμα 1.5.1. Στις περιπτώσεις μερικής απόκρυψης, όπως όταν ένα μικρόφωνο καλύπτει το στόμα, το σύστημα συνδύασε τα περιορισμένα οπτικά στοιχεία με το ακουστικό σήμα και κατάφερε να διατηρήσει την ομαλή άρθρωση. Αντίστοιχα, σε δυναμικές αποκρύψεις, όπως όταν τα χέρια κινούνται μπροστά από το πρόσωπο, το μοντέλο δημιούγησε ομαλές ανακατασκευές χωρίς απότομες παραμορφώσεις. Η ικανότητα αυτή είναι σημαντική, καθώς τέτοιες αυθόρμητες κινήσεις εμφανίζονται συχνά σε φυσικούς διαλόγους.

Μία απαιτητική περίπτωση που παρουσιάζεται και στο σχήμα 1.5.1 (δεξιά) αφορά την πλήρη απόκρυψη του στόματος. Εδώ το σύστημα αναγκάζεται να βασιστεί σημαντικά στον ήχο και παρόλο που δυσκολεύεται να απεικονίσει απόλυτα λεπτομερή άρθρωση, παράγει ρεαλιστικά αποτελέσματα. Επιπλέον, σε σενάρια όπου το πρόσωπο είναι ορατό αλλά δεν υπάρχει ομιλία που να προέρχεται από το απεικονιζόμενο πρόσωπο, το μοντέλο δεν παράγει ψευδείς κινήσεις στόματος. Συνολικά, τα ποιοτικά αποτελέσματα δείχνουν ότι το μοντέλο μπορεί να αντιμετωπίσει ένα ευρύ φάσμα συνθηκών, από μερικές μέχρι πλήρεις αποκρύψεις, και να αποδώσει εκφράσεις που παραμένουν ρεαλιστικές και συνεπείς με το ακουστικό περιεχόμενο.

Figure 1.5.1: Ποιοτικά παραδείγματα του μοντέλου μας υπό διαφορετικές συνθήκες. Από αριστερά προς τα δεξιά, το εισερχόμενο ηχητικό σήμα είναι: ένα πορτογαλικό τραγούδι, γαλλικός λόγος και μία αγγλική έκφραση.

Στη συνέχεια, συγκρίνουμε, τη μέθοδός μας με δύο σύγχρονες προσεγγίσεις που βασίζονται στο ίδιο μοντέλο παραμετρικής αναπαράστασης FLAME, τα SPECTRE και SMIRK. Η οπτικοποίηση των αποτελεσμάτων παρουσιάζεται στην Εικόνα 1.5.2. Η ανάλυση ανέδειξε σαφείς διαφορές στην ποιότητα και τη ρεαλιστικότητα των αποτελεσμάτων.

Το **SPECTRE** διατηρεί σταθερά την πόζα και το γενικό σχήμα του κεφαλιού, αλλά αποτυγχάνει να συλλάβει λεπτές εκφράσεις στην περιοχή του στόματος δημιουργώντας σημαντικά σφάλματα. Αυτό οδηγεί σε λιγότερο εκφραστικές αναπαραστάσεις, ιδιαίτερα σε συνθήκες όπου η άρθρωση των χειλιών είναι κρίσιμη για την κατανόηση του λόγου. Το **SMIRK**, αντιθέτως, παράγει πιο ζωντανές και έντονες εκφράσεις, όμως παρουσιάζει συχνά γεωμετρικές ανακρίβειες και παραμορφώσεις, οι οποίες μειώνουν τον ρεαλισμό και δημιουργούν αντιληπτικά σφάλματα. Η προτεινόμενη μέθοδος **FAVOR** καταφέρνει να ισορροπήσει ανάμεσα σε αυτές τις δύο αδυναμίες. Διατηρεί με ακρίβεια την ταυτότητα του ατόμου, αποδίδει τις εκφράσεις με φυσικό τρόπο, και κυρίως προσφέρει σταθερά ικανοποιητικά αποτελέσματα ακόμη και σε δύσκολες περιπτώσεις μερικής ή πλήρους απόκρυψης του στόματος.

Ποσοτική Αξιολόγηση

Η ποσοτική αξιολόγηση της ανακατασκευής εκφράσεων προσώπου είναι ιδιαίτερα απαιτητική, καθώς τα καθαρά γεωμετρικά σφάλματα δεν αντικατοπτρίζουν πάντα την αντιληπτική ποιότητα των αποτελεσμάτων. Για τον λόγο αυτό χρησιμοποιήθηκαν μετρικές ανάγνωσης χειλιών, οι οποίες αξιολογούν την αντιστοίχιση λόγου και κινήσεων στόματος. Συγκεκριμένα, υπολογίστηκε ο Ρυθμός Σφάλματος Χαρακτήρων (CER), ο Ρυθμός Σφάλματος Λέξεων (WER), ο Ρυθμός Οπτικών Φωνημάτων (VER) και ο Ρυθμός Σφάλματος Οπτικών Λέξεων (VWER). Τα αποτελέσματα παρουσιάζονται στον πίνακα 1.1 και έδειξαν ότι η μέθοδός μας υπερέχει σταθερά σε σχέση με των SPECTRE και SMIRK, ιδιαίτερα στο σύνολο CelebV-HQ [78], το οποίο περιλαμβάνει δείγματα με έντονες αποκρύψεις προσώπου. Στο σύνολο LRS3 οι διαφορές ήταν μικρότερες, κάτι που εξηγείται από το γεγονός ότι περιέχει λιγότερα περιστατικά απόκρυψης και είναι πιο κοντά στον χώρο εκπαίδευσης των ανταγωνιστικών μοντέλων. Συνολικά, οι μετρικές επιβεβαιώνουν ότι το FAVOR γενικεύει καλύτερα

Figure 1.5.2: Οπτικά αποτελέσματα από δεξιά προς τα αριστερά του SPECTRE, του SMIRK και της μεθόδου μας.

και παράγει πιο αξιόπιστες κινήσεις χειλιών.

	CELEBV-HQ				LRS3			
	CER	WER	VER	VWER	CER	WER	VER	VWER
SMIRK [59]	108.9	149.6	112.2	140.6	126.7	157.4	121.8	153.5
SPECTRE [18]	115.8	156.5	125.4	150.2	**116.9**	147.3	**112.2**	145.5
Ours	**89.0**	**118.3**	**82.6**	**116.9**	118.1	**146.5**	114.0	**143.9**

Table 1.1: Αποτελέσματα της μετρικής αναγνώρισης χειλιών στο LRS3 και σε 30 βίντεο του CELEBV-HQ συνόλου δεδομένων.

Μελέτη Χρηστών

Επειδή οι αριθμητικές μετρικές δεν αποτυπώνουν πλήρως την αντιληπτική ποιότητα, διεξήχθη και μία μελέτη χρηστών με 46 συμμετέχοντες. Στο πείραμα αυτό, οι χρήστες κλήθηκαν να συγκρίνουν βίντεο που παρήχθησαν με τη μέθοδό μας και με τα SPECTRE και SMIRK, επιλέγοντας τη πιο ρεαλιστική ανακατασκευή. Στα σενάρια με αποκρύψεις προσώπου, οι χρήστες προτίμησαν με μεγάλη συνέπεια το FAVOR, γεγονός που επιβεβαιώνει την ανωτερότητα της μεθόδου. Σε δείγματα χωρίς αποκρύψεις, το FAVOR ξεπέρασε το SMIRK, ενώ η απόδοσή του ήταν συγκρίσιμη με του SPECTRE, το οποίο αποδίδουμε στο γεγονός ότι το δεύτερο δημιουργεί πιο έντονες εκφράσεις, κάτι που στον παρατηρητή ίσως φαίνεται πιο ζωντανό. Τα αποτελέσματα που αφορούν το σύνολο των βίντεο που χρησιμοποιήθηκαν (με και χωρίς αποκρύψεις προσώπου) συγκεντρώνονται στον πίνακα 1.2, τα οποία πιστοποιούν ότι το προτεινόμενο μοντέλο επιτυγχάνει υψηλής ποιότητας ανακατασκευές, επιβεβαιώνοντας τη χρησιμότητα της πολυτροπικής προσέγγισης.

	SMIRK	SPECTRE
Ours	**288**/106	**247**/157

Table 1.2: Προτιμήσεις χρηστών ανά μέθοδο.

Μελέτη Αφαίρεσης Συνιστωσών

Για να αξιολογηθεί η συμβολή κάθε επιμέρους στοιχείου του προτεινόμενου συστήματος, πραγματοποιήθηκε μία εκτεταμένη μελέτη αφαίρεσης συνιστωσών. Ακολουθήθηκε η εξής βηματική διαδικασία: ξεκινώντας από το SMIRK ως βασικό μοντέλο, προστίθενται σταδιακά τα προτεινόμενα υποσυστήματα, ώστε να απομονωθεί η επίδραση του καθενός στην τελική απόδοση.

Με vs χωρίς συνθετικές αποκρύψεις. Αρχικά εξετάστηκε η σημασία της εκπαίδευσης με συνθετικές αποκρύψεις προσώπου. Όπως φαίνεται στην Εικόνα 1.5.3 το SMIRK, όταν εκπαιδεύεται χωρίς αποκρύψεις στα δεδομένα εκπαίδευσης, τείνει να παράγει εκφράσεις που δεν ευθυγραμμίζονται με το πραγματικό βίντεο και παρουσιάζει εμφανή σφάλματα. Αντίθετα, η χρήση του επαυξημένου συνόλου δεδομένων αποτρέπει τέτοιες αστοχίες, οδηγώντας σε πιο ομαλές και σταθερές ανακατασκευές. Παρ' όλα αυτά, η αποκλειστική αξιοποίηση σημείων αναφοράς δεν αποδεικνύεται επαρκής, καθώς σε περιπτώσεις όπου αυτά απουσιάζουν το μοντέλο δυσκολεύεται να διατηρήσει φυσικές εκφράσεις, όπως το ανοιγόκλεισμα των ματιών.

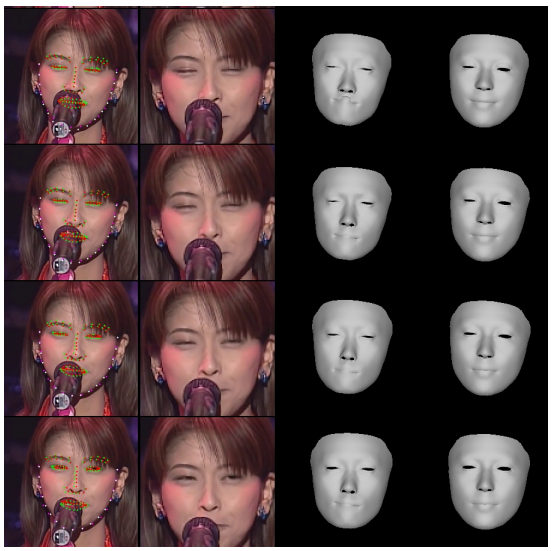

Figure 1.5.3: Ποιοτική σύγκριση του SMIRK με το προεκπαιδευμένο μοντέλο μας. Από αριστερά προς τα δεξιά: βίντεο εισόδου με τα αντίστοιχα σημεία αναφοράς, μια περικομμένη εικόνα για πιο προσεκτική παρατήρηση, η έξοδος του SMIRK και η έξοδος του προεκπαιδευμένου μας μοντέλου.

Με vs χωρίς Lipreading Loss. Στη συνέχεια ενσωματώθηκε ακουστική πληροφορία και δοκιμάστηκε η χρήση του σφάλματος αναγνώρισης χειλιών. Αν και αναμενόταν να βελτιώσει την αναπαράσταση των χειλιών, στην πράξη προκάλεσε σφάλματα, κυρίως όταν το στόμα ήταν κλειστό, γεγονός που αναδεικνύεται στην Εικόνα 1.5.4. Το φαινόμενο αυτό αποδόθηκε στη διαφορά μεταξύ του χώρου εισόδου και του χώρου των rendered εικόνων.

Lipreading Loss σε rendered vs fused εικόνα. Αρχικά, το σφάλμα αναγνώρισης χειλιών υπολογιζόταν στα rendered καρέ. Στην τρέχουσα προσέγγιση, το εφαρμόζουμε στο ανακατασκευασμένο

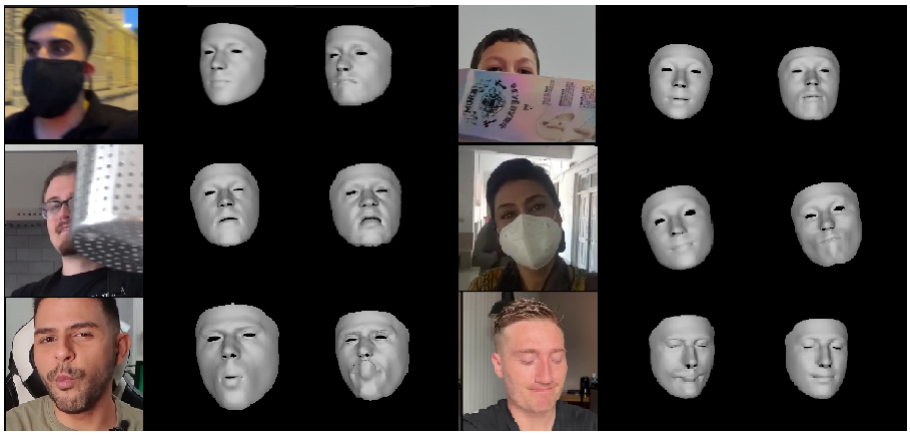

Figure 1.5.4: Σφάλματα που προκύπτουν από τη χρήση της απώλειας ανάγνωσης χειλιών. Τα αποτελέσματα παρουσιάζονται για το οπτικοακουστικό μας μοντέλο χωρίς και με την χρήση της συνάρτησης κόστου για την αναγνώση των χειλιών, από αριστερά προς τα δεξιά.

βίντεο που παράγει ο γεννήτορας, γεγονός που οδηγεί σε πιο συνεπή και σταθερά αποτελέσματα όπως παρουσιάζεται στην Εικόνα 1.5.5.

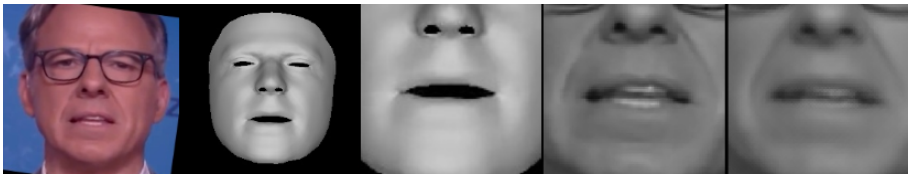

Figure 1.5.5: Από αριστερά προς τα δεξιά: Είσοδος, FAVOR, περικομμένο στόμα από το βίντεο εισόδου, περικομμένο στόμα rendered αποτέλεσμα, περικομμένο στόμα από το πραγματικό βίντεο σε κλίμακα του γκρι, περικομμένο στόμα από την ανακατασκευασμένη έξοδο.

Με τον τρόπο αυτό μειώνονται τα σφάλματα και παράγονται πιο ρεαλιστικές κινήσεις στόματος. Επιπλέον, η εισαγωγή ενός κατωφλίου στο σφάλμα αποτρέπει την υπερδιόρθωση, η οποία οδηγούσε σε λανθασμένες ανακατασκευές του στόματος.

Οπτικό vs Ακουστικό vs Οπτικοακουστικό Μοντέλο Τέλος, συγκρίθηκαν τα αποτελέσματα από τις τρεις εκδοχές του μοντέλου μας με διαφορετικούς τύπους εισόδου: μόνο ήχος, μόνο εικόνα και συνδυασμός τους. Το μοντέλο που βασίζεται αποκλειστικά στον ήχο απέτυχε να αποδώσει σωστά τις εκφράσεις, παράγοντας σχεδόν στατικές κινήσεις που δεν ανταποκρίνονταν στην είσοδο. Αντίθετα, το οπτικό μοντέλο παρήγαγε πιο ρεαλιστικές κινήσεις στόματος και εκφράσεις, με μικρές μόνο αδυναμίες σε περιπτώσεις αποκρύψεων, το οποίο αποδίδεται στο γεγονός ότι το προεκπαιδευμένο μοντέλο που χρησιμοποιεί έχει εκπαιδευτεί σε συνθετικές αποκρίψεις. Η οπτικοακουστική εκδοχή του μοντέλου εμφάνισε την καλύτερη απόδοση, με συνεπείς και ακριβείς ανακατασκευές, όπως φαίνεται στο σχήμα 1.5.6.

1.6 Συμπέρασμα

Η εργασία εστίασε στην πρόκληση της τρισδιάστατης ανακατασκευής προσώπου σε πραγματικές συνθήκες, όπου συχνά υπάρχουν αποκρύψεις και θορύβοι στα δεδομένα εισόδου. Παρουσιάστηκε αναλυτική βιβλιογραφική ανασκόπηση, η οποία ανέδειξε ότι οι περισσότερες υπάρχουσες μέθοδοι επικεντρώνονται αποκλειστικά σε μονοτροπικές εισόδους. Σε αυτό το πλαίσιο αναπτύχθηκε ένα

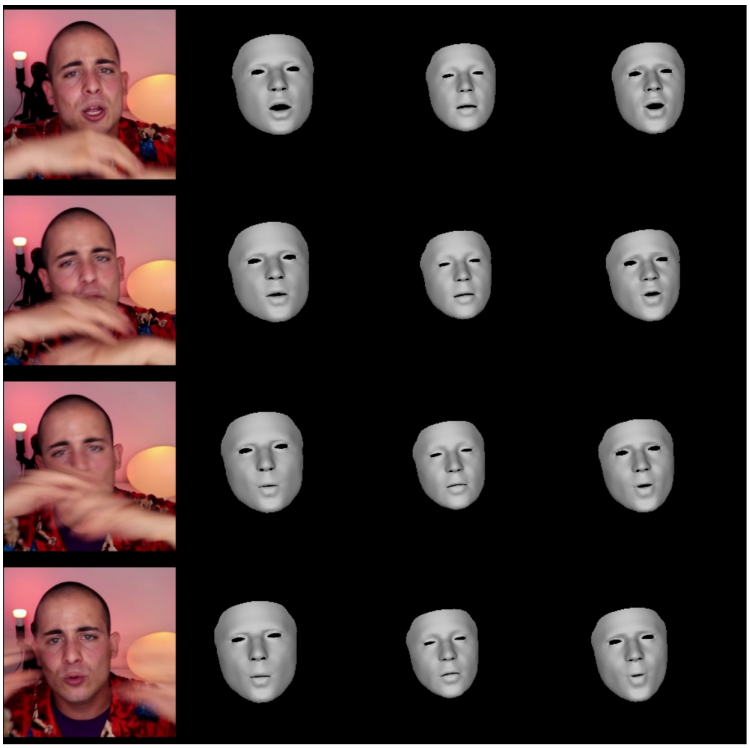

Figure 1.5.6: Οπτική σύγκριση της τρισδιάστατης ανακατασκευής προσώπου για οπτική, ηχητική και οπτικοακουστική είσοδο.

νέο οπτικοακουστικό μοντέλο που εκμεταλλεύεται και τις δύο πηγές πληροφορίας, επιτυγχάνοντας ρεαλιστική και συνεπή ανακατασκευή προσώπων. Η εκπαίδευση με συνθετικές αποκρύψεις και η χρήση αντιληπτικών συναρτήσεων κόστους ενίσχυσαν την ευρωστία του συστήματος. Η ποιοτική και ποσοτική αξιολόγηση των αποτελεσμάτων σε συνδυασμό με τη μελέτη χρηστών έδειξε σαφή υπεροχή έναντι άλλων σύγχρονων μεθόδων, ιδιαίτερα σε συνθήκες αποκρύψεων.

Περιορισμοί και Μελλοντικές Κατευθύνσεις

Παρά τα θετικά αποτελέσματα της μεθόδου μας, εντοπίστηκαν ορισμένοι περιορισμοί που αποτελούν προοπτικές για μελλοντική έρευνα. Παρατηρούνται σφάλματα ασυμφωνίας στις κινήσεις του στόματος στο μοντέλο με είσοδο μόνο τον ήχο, τα οποία μειώνουν την αντιληπτική ποιότητα των αποτελεσμάτων. Επιπλέον, η μέθοδος παρουσιάζει δυσκολίες σε περιπτώσεις ακραίων αποκρύψεων (π.χ. πλήρης κάλυψη του στόματος). Προς αυτή τη κατεύθυνση, θα μπορούσαν να διερευνηθούν τεχνικές όπως το modality dropout ή μηχανισμοί προσοχής, ώστε το δίκτυο να προσαρμόζει δυναμικά τη βαρύτητα μεταξύ οπτικών και ακουστικών ενδείξεων.

Ηθικές Διαστάσεις και Κοινωνικός Αντίκτυπος

Οι τεχνολογίες τρισδιάστατης ανακατασκευής προσώπων παρουσιάζουν σημαντικές προοπτικές σε τομείς όπως η επαυξημένη πραγματικότητα, η ιατρική και η αλληλεπίδραση ανθρώπου–μηχανής. Ωστόσο, συνοδεύονται και από σοβαρούς κινδύνους, όπως η χρήση τους σε deepfakes και σε παραπλανητικό περιεχόμενο χωρίς συναίνεση. Η πρόοδος στον τομέα αυτό οφείλει να διέπεται από αρχές υπευθυνότητας και διαφάνειας, ώστε να αξιοποιούνται αποκλειστικά οι θετικές εφαρμογές του.

Chapter 2

Introduction

Contents

2.1 Prefaces

Human faces are central to perception and communication, as they convey essential information about identity, expression, and emotion. With the rapid advances in machine learning, human–computer interaction has become an important research area in the last decades. Teaching computers to recognize people and interpret their behavior requires understanding not only how humans look but also how they move and express themselves. As faces and their expressions provide such a rich source of information, researchers started early to analyze real-world faces and to develop methods for the generation of realistic digital avatars. These technologies are increasingly applied across many domains, and their impact continues to expand as reconstruction methods become more accurate and accessible.

Facial analysis supports a variety of applications. In security and interaction systems, it enables tasks such as identity verification, visual speech recognition, and expression-based interfaces. In creative and medical domains, it is applied to personalized avatars, prosthetics, and 3D printing. At the same time, capturing both the geometry and the texture of the face is essential for producing realistic digital humans in film, games, and social media. Building on this, computer vision community has increasingly focused on developing 3D face reconstruction methods. This refers to recovering a detailed three-dimensional representation of a human face, capturing both its geometry and appearance, from recorded data. Depending on the method, the input can be a single monocular image [13], [17], [59], a video sequence [9], [18], [35], audio signals [12], [60], [67], a text [46], [72] or a combination of these modalities [11].

Reconstructing a human face in three dimensions though is a challenging problem. Unlike rigid objects, faces are deformable and change continuously with expressions, speech, and aging. Factors such as changes in lighting, occlusions from hand movements or facial masks, and differences in viewpoint further complicate the task. Capturing both the global geometry and the fine details of a face is essential for realistic reconstruction, yet it requires methods that can generalize well to diverse conditions. So far, most existing methods rely on a single modality, which limits their robustness and accuracy in real-world scenarios.

Driven by these challenges in the current thesis we design and develop an audiovisual 3D face reconstruction approach that exploits both audio and visual information from input monocular videos to generate realistic and accurate avatars. Visual data provides rich spatial cues about facial geometry and expressions, while audio signals offer complementary temporal information that helps capture lip movements and articulation. By fusing these modalities, the method becomes more robust to occlusions, lighting changes, and pose variations, while also improving the temporal consistency of the reconstructed faces. This integration leads to more natural and coherent animations compared to single-modality approaches as showcased by our experimental results and user study.

Building such multimodal systems has become feasible largely thanks to the rapid advances in deep learning over the past decade. The strength of deep learning lies in the ability to learn complex patterns directly from large amounts of data and to generate accurate representations of facial diversity. By training deep learning models on extensive datasets, we can develop systems that generalize well to unseen faces and conditions. In this first chapter, we present the theoretical foundations of multimodal representation and deep learning, which form the basis of our work. We also provide an overview of the main applications, discuss the key challenges, and highlight the contributions of this thesis.

2.2 Multimodality Representation Learning

Humans naturally perceive and process information through multiple sensory channels such as vision, hearing, and language, integrating them to form a coherent understanding of the world. Inspired by this, recent advancements in machine learning have focused on developing models that can effectively leverage information from multiple modalities. A modality corresponds to a specific type or form of data, each capturing a distinct aspect of the same phenomenon.

Traditional machine learning approaches have primarily relied on unimodal models, each designed to process a single type of input data. In contrast, more recent research has shifted towards multimodal learning, which integrates and exploits complementary information from different modalities. Multimodality refers to a model's ability to jointly process and understand diverse types of inputs, resulting in richer and more robust representations. This approach, also, is especially valuable in cases where a single modality may be corrupted across data samples or where unimodal methods fail to capture the complexity of the dataset.

Applications of multimodal learning have wide applications in various domains, including Audio-Visual Speech Recognition [66] and Visual Question Answering [3]. Other prominent applications include image captioning, where models generate descriptions of visual content [29], sentiment analysis from text and images in social media [76], and medical image analysis combining imaging data with clinical reports [30].

2.3 Background on Deep Learnning

Machine learning is a subset of artificial intelligence that focuses on the development of algorithms and statistical models that enable computers to perform specific tasks without explicit instructions. It emphasizes on systems to learn patterns from data and automatically build models to solve specific tasks. In recent years, the increased availability of large datasets and advancements in computational power have led to the rise of deep learning [31], [36].

A deep learning algorithm utilizes artificial neural networks with multiple layers to model complex patterns in data. Unlike traditional machine learning algorithms that often require manual feature engineering, deep learning models automatically learn hierarchical representations of data through multiple layers of abstraction. This capability allows them to capture intricate relationships and patterns, making them particularly effective for tasks such as image processing, natural language processing, and speech recognition.

Such models are typically trained using large datasets and optimization techniques such as stochastic gradient descent [63]. The training process involves adjusting the weights and biases of the network to minimize a predefined loss function, which quantifies the difference between the model's predictions and the actual target values. This iterative process continues until the model converges to an optimal set of parameters that generalize well to unseen data [39].

In the field of 3D face reconstruction, deep learning has introduced new approaches through end-to-end frameworks that directly map input data such as 2D scans or audio signals, to detailed 3D face models. These methods exploit the representational power of neural networks to automatically learn complex relationships from data, reducing the reliance on manual feature engineering and enabling more accurate and robust reconstructions results. The key network architectures will be described in the following section.

2.3.1 Convolutional Neural Networks

In the context of face reconstruction, which is inherently a visual task, CNNs [37], [38], [48] have proven highly effective for detecting key elements such as head pose and facial expressions. The convolutional layer is the core building block, consisting of a set of learnable filters (also called kernels) that slide across the spatial dimensions of the input. At each position, the filter performs a dot product between its weights and the local input region, producing an activation map that highlights the presence of relevant features. Deeper layers in the network introduce non-linearities and progressively reduce the dimensionality of the data, allowing the network to capture increasingly abstract patterns while reducing computational complexity. The final layers of a CNN are usually fully connected, transforming the extracted features into a 1D representation suitable for classification or regression tasks. Figure 2.3.1 illustrates the overall architecture of a typical CNN.

In addition, Temporal Convolutions Networks (TCNs), which we use in this work, extends the concept of convolution into temporal domain. These networks are well-suited for modeling dependencies across sequential data, enabling the model to learn how features evolve over time.

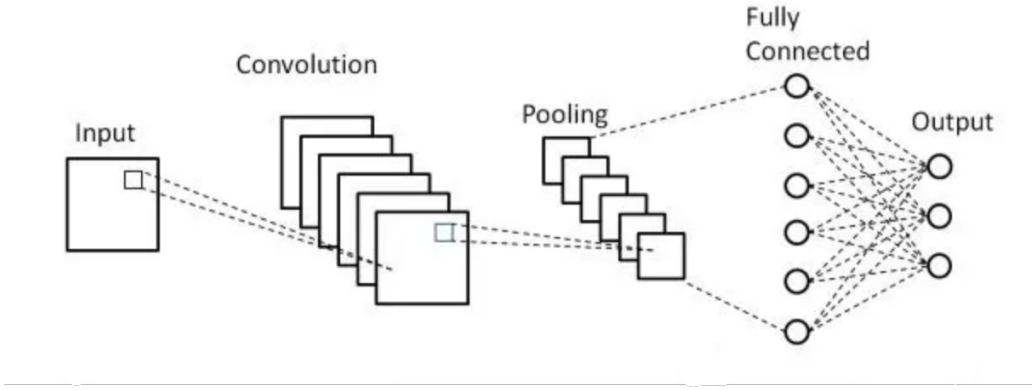

Figure 2.3.1: Typical CNN architecture.

2.3.2 Recurrent Neural Networks

Unlike CNNs that operate on independent data instances, such as images, certain types of data, like audio signals or natural language, require models that can capture their sequential dependencies. Traditional feedforward networks process information in a single direction: from the input layer, through one or more hidden layers, to the output layer. In contrast, Recurrent Neural Networks (RNNs) [45] incorporate feedback connections, allowing the output from previous time steps to be used as input for the current one, as illustrated in Figure 2.3.2. This enables RNNs to retain contextual information through their hidden states, which are updated over time based on prior outputs.

However, standard RNNs struggle with learning long-term dependencies because they tend to lose information over extended sequences, a limitation commonly known as the long-term context problem. To address this issue, Long Short-Term Memory (LSTM) networks [26] introduce a gating mechanism that regulates how information is stored, updated, and forgotten across time steps, thereby improving the model's ability to capture long-range temporal relationships.

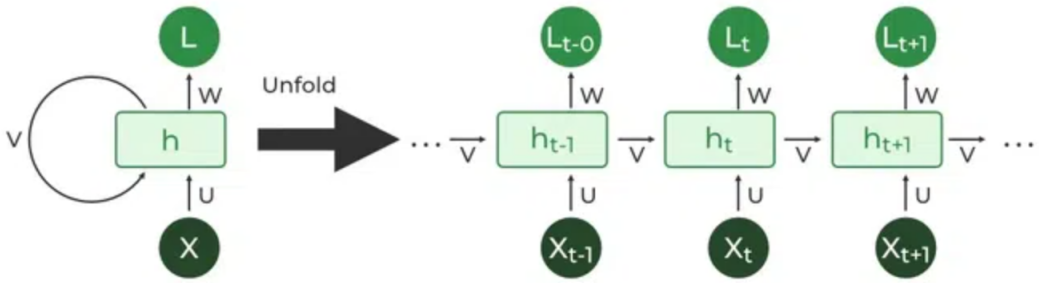

Figure 2.3.2: Unrolled Recurrent Neural Network Architecture.

2.3.3 Transformers

Transformers were first introduced by Vaswani et al. in the seminal work *"Attention Is All You Need"* [69], which proposed a novel architecture for sequence modeling that eliminates recurrence and instead relies entirely on the mechanism of self-attention. The Transformer addresses the limitations of Recurrent Neural Networks (RNNs) by modeling global dependencies within a sequence, allowing it to capture long-range contextual relationships more effectively. Through the use of self-attention, each input element can attend to all others in the sequence, weighting their influence based on learned similarity scores to produce the final output. The complete Transformer architecture follows an encoder–decoder structure, with both components composed of stacked layers containing self-attention. Transformers scale effectively to large datasets, and their encoder–decoder structure is illustrated in Figure 2.3.3.

2.4 Applications

Significant progress in 3D face reconstruction techniques have led to a wide range of applications across various fields. The ability to generate accurate and realistic 3D models of human faces has enabled new possibilities in entertainment, communication, healthcare, and security.

In the context of virtual and augmented reality (VR/AR), 3D facial reconstruction provides the foundation for the creation of personalized digital avatars that can reproduce a user's identity, expressions, and emotions in immersive environments [28]. Such avatars facilitate natural telepresence and enhance user experience in gaming, social interaction, and virtual meetings.

In the entertainment industry, 3D face reconstruction plays a central role in visual effects, performance capture, and film post-production. The ability to accurately reconstruct and manipulate an actor's facial expressions enables filmmakers to refine performances and achieve greater emotional realism. Recent systems such as NED [50] allow photo-realistic modification of an actor's emotions directly in in-the-wild video footage, while VDub [21] enables realistic visual dubbing by synchronizing lip motion to new audio tracks without altering the actor's original performance.

With the rapid growth of social media as a platform for sharing and promoting digital content,

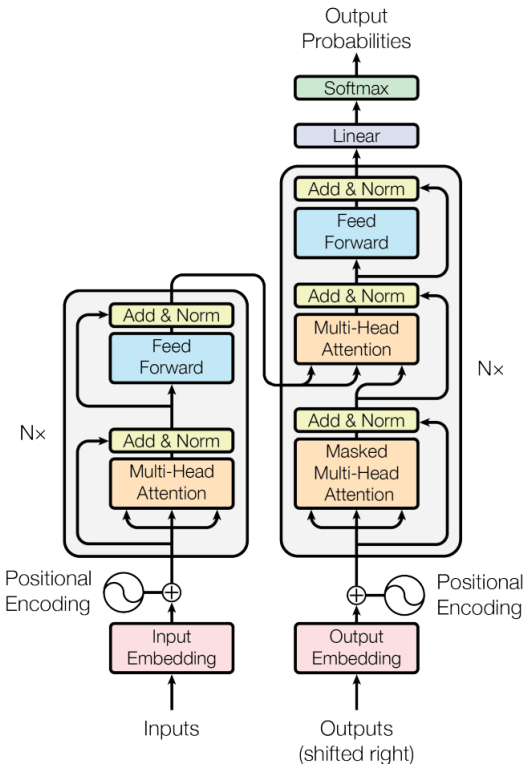

Figure 2.3.3: The Transformer - model architecture.

realistic facial editing has gained significant attention. A reliable method for changing facial emotions or expressions can provide a creative tool for face manipulation. Deepfake [67] and neural rendering techniques make this process easier and more accessible, reducing the need for expensive and time-consuming visual effects while maintaining high visual quality.

Finally, 3D face reconstruction has important applications in both biometrics and medicine. In biometrics, 3D facial models provide robust identity representations that are invariant to pose, illumination, and expression variations. In medical contexts, 3D facial geometry supports the automated diagnosis of genetic disorders that affect facial morphology [23] and enables precise quantitative assessment of craniofacial structures for diagnostic and surgical planning purposes [58].

Many other applications stand to benefit from 3D face reconstruction, and ongoing technological advancements in the field are expected to enhance existing ones while also opening new opportunities in areas where it has not yet been explored.

2.5 Challenges

Audiovisual 3D face reconstruction remains an ill-posed problem, especially under real-world conditions where the face is often partially occluded by hands, surgical masks or other objects. Using only visual input for reconstruction is inherently limited due to the lack of visual information in these occluded regions, leading to insufficient results. Conversely, relying solely on audio information provides advantages in modeling lip motion and maintaining temporal consistency, but it lacks the spatial information necessary to recover identity-specific facial geometry. More-

over, the limited availability of large-scale, high-quality datasets containing occlusions limits the development of robust models capable of handling real-world scenarios. In addition to leveraging complementary audio and visual modalities, modeling the temporal information across sequences of frames is crucial, as it enforces consistency in dynamic facial expressions.

2.6 Contributions

As already mentioned, the main topic of this Diploma Thesis is audiovisual 3D face reconstruction. Our work aims to leverage recent advancements in deep learning and multimodal representation learning to reconstruct accurate and expressive 3D facial geometry from both visual and audio inputs. The contributions of this thesis can be briefly summarized as follows:

1. We introduce a novel audiovisual model for 3D face reconstruction that takes input video and is robust under real-world conditions, including challenging scenarios with occlusions.
2. We model the temporal information across sequence of frames.
3. We propose a trimodal architecture that supports audiovisual, visual-only and audio-only input.
4. We augment the training data with synthetical occlusions to improve robustness and generalization to real-world scenarios where parts of the face are hidden.
5. By leveraging both modalities, our model reduces erroneous mouth movement generation when the speech does not correspond to the visible speaker in the video.
6. We conduct extensive qualitative and quantitative experiments, along with a user study and an ablation study, to evaluate the performance of our method and compare it with recent state-of-the-art approaches. The results demonstrate that our method achieves superior reconstruction accuracy especially in occluded data samples.

2.7 Thesis Outline

The remainder of this thesis is organized as follows.

- Chapter 3 provides an overview of mesh representations, facial landmark extraction, face modeling.
- Chapter 4 reviews the most relevant literature on 3D face reconstruction, highlighting prior work in visual-driven, audio-driven and audiovisual-driven methods.
- Chapter 5 presents the proposed audiovisual 3D face reconstruction framework, detailing the network architecture, training objectives, and data processing pipeline.
- Chapter 6 describes both qualitative and quantitative analyses and a user study. Next, we presents the ablation study, which investigates the contribution of individual components and design choices within our model.
- Finally, Chapter 7 concludes a summary of the results, discusses limitations, and outlines potential directions for future research.

Chapter 3

Face Modeling

Contents

3.1 Meshes

Meshes are widely used in computer graphics and vision for various applications, including 3D modeling, animation, and rendering. They provide a flexible and efficient way to represent complex surfaces and can be easily manipulated for tasks such as deformation, morphing, and texture mapping. A 3D mesh defines a surface via a collection of vertices and usually triangular faces. It is represented by two matrices:

- A vertex matrix $V \in \mathbb{R}^{N\times 3}$, where each row corresponds to the 3D coordinates (x, y, z) of a vertex in the mesh.
- A face matrix $F \in \mathbb{N}^{M\times 3}$, where each row contains indices into the vertex matrix that define a triangular face. Each face is represented by three vertex indices.

While in this thesis we only deal with meshes consisting of triangles, other types of meshes can include quadrilaterals (quads), or other simple convex polygons (n-gons). Formally, each mesh can be transformed into a graph $M = (V, F)$ which is a purely geometric representation, meaning that it does not involve any texture. Similarly, a textured mesh is represented by $M = (V, F, C)$, with the texture $C \in \mathbb{R}^{N\times 3}$ encoded as a per-vertex color vector.

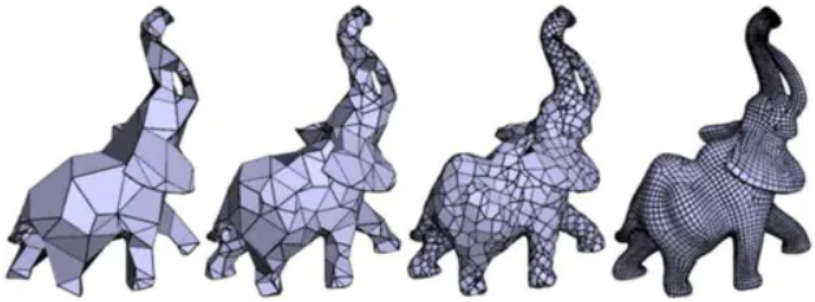

Figure 3.1.1: Polygonal 3D meshes of the same object, illustrated at multiple vertex densities [44].

3.2 Texture Mapping

Texture mapping is a technique used in computer graphics to enhance the visual detail of 3D models by applying 2D images (textures) onto their surfaces. This process involves mapping points on the 3D surface to corresponding points on the 2D texture image, allowing for the simulation of complex surface details such as colors, patterns, and material properties. Texture mapping is widely used in applications such as video games and virtual reality to create more realistic and visually appealing representations of objects and environments.

Figure 3.2.1: Examples of texture mapping [56].

3.3 Landmarks

Facial landmarks are semantically meaningful points on the human face, such as the corners of the eyes, the tip of the nose, and the outline of the lips. They provide a compact geometric representation of facial structure and are widely used in tasks such as face recognition, expression analysis, and 3D face reconstruction. In this thesis, we make use of two landmark extraction methods: the Face Alignment Network (FAN), which produces a sparse keypoint set, and MediaPipe Face Landmarker, which provides a dense set of landmarks covering the entire facial surface. Together, these representations allow us to capture both coarse structural cues and fine-grained details of the face.

3.3.1 Facial Landmark Extraction using FAN

The Face Alignment Network (FAN) [8] is a convolutional neural network designed for facial landmark localization. It is built on stacked HourGlass networks and outputs heatmaps for each landmark, where the peak indicates the estimated coordinate. FAN has been widely adopted in face analysis research due to its robustness across different poses, expressions, and illumination conditions. In this work, we employ FAN through the iBUG Face Alignment framework, an open-source Python library developed by the Intelligent Behaviour Understanding Group at Imperial College London. The framework, implemented in PyTorch, provides a unified interface that allows pretrained detectors to be applied directly to input images. Within this framework, we use the FANPredictor, a 2D facial landmark detector that encapsulates a pretrained FAN model.

Our facial landmark extraction pipeline follows a two-stage process. First, face regions are detected using the RetinaFace detector [14], which produces robust bounding boxes for all visible faces in the input images. These detected face regions are then passed to FAN for landmark localization which predicts a 68-point landmark set, covering key regions of the face such as the eyes, eyebrows, nose, lips, and jawline. Each landmark ℓ_i is defined as a two-dimensional coordinate:

$$\ell_i = (x_i, y_i), \quad i = 1, 2, \ldots, 68,$$

with (x_i, y_i) expressed in pixel coordinates relative to the input image.

Figure 3.3.1: Facial landmark topology extracted by the Face Alignment Network [8].

3.3.2 Facial Landmark Extraction using Mediapipe

MediaPipe (MP) is an open-source, cross-platform framework developed by Google for constructing multimodal perception pipelines [22]. It is designed around the concept of calculators, modu-

lar components that can be linked together in a directed graph to form complex data-processing pipelines. This architecture allows MediaPipe to handle the full workflow of machine learning inference, including data pre-processing, model execution, and post-processing, while maintaining real-time performance across desktop, mobile, and web platforms. In addition to its role as a general-purpose framework, MediaPipe provides a suite of pre-trained solutions for vision and audio tasks, including face detection, hand tracking, and holistic body pose estimation [42].

One of the most widely used solutions within the MediaPipe ecosystem is the Face Landmarker, which performs dense facial landmark detection in real time. This capability is based on the work of Kartynnik et al. [33], who introduced the Face Mesh model for estimating 3D facial geometry from monocular video on mobile GPUs. The Face Mesh model is integrated into the MediaPipe framework as a complete pipeline: a lightweight face detector [5] first localizes the face region of interest, after which a regression network predicts a dense set of facial landmarks.

The MediaPipe Face Landmarker estimates 468 landmarks per face, covering the entire surface of the face. Each landmark is represented as a triplet

$$\ell_i = (x_i, y_i, z_i), \quad i = 1, 2, \ldots, 468,$$

where x_i and y_i denote normalized 2D coordinates within the image, later rescaled to pixel units, and z_i represents the relative depth in the same scale as the image width. This dense representation enables reconstruction of a three-dimensional facial mesh, capturing fine-grained features.

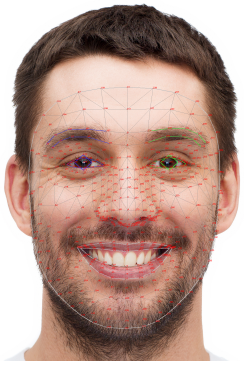

Figure 3.3.2: Facial landmark topology from the MediaPipe Face Mesh model [22].

3.4 3D Face Modelling

Modeling human faces has long been a challenge in computer graphics and computer vision. Since Parke's early contributions [51], [52], many approaches have been introduced for both representing facial geometry [6], [10], [40] and animating expressions [68].

In this thesis, we focus specifically on the task of audiovisual 3D face reconstruction from input videos. This involves estimating the time-varying 3D geometry and appearance of a human face by leveraging both visual and auditory modality. Currently, a plethora of methods for single-image or video-based 3D face reconstruction are primarily built upon 3D Morphable Models (3DMMs). In these approaches, the parameters of a statistical face model are typically determined either through iterative optimization procedures or by direct regression using deep learning techniques.

However, to provide a comprehensive overview, we will first explore some powerful 3D face modeling methods that do not rely on 3DMMs. Following this, we will present a detailed analysis of 3D Morphable Models, specifically focusing on the FLAME model, which serves as the core face representation model utilized in this thesis.

3.5 Volumetric reconstruction methods for Face Modeling

Volumetric reconstruction methods aim to recover 3D facial geometry by combining information from multiple images or depth observations. These approaches may include both traditional geometric techniques and also more recent neural methods.

3.5.1 Structure-from-Motion

Structure-from-Motion (SfM) [64] primary role is to determine the intrinsic properties of cameras such as focal length and lens distortion and, crucially, the precise 3D position and orientation (pose) from which each input image was captured. It operates by detecting and matching distinctive keypoints within each image, robustly filtering erroneous matches, and then performing a global optimization known as bundle adjustment. The output of SfM is an accurate set of camera poses for all input images and a sparse 3D point cloud, representing only the locations of the detected keypoints. While geometrically precise, this sparse output alone is generally insufficient for a complete surface reconstruction.

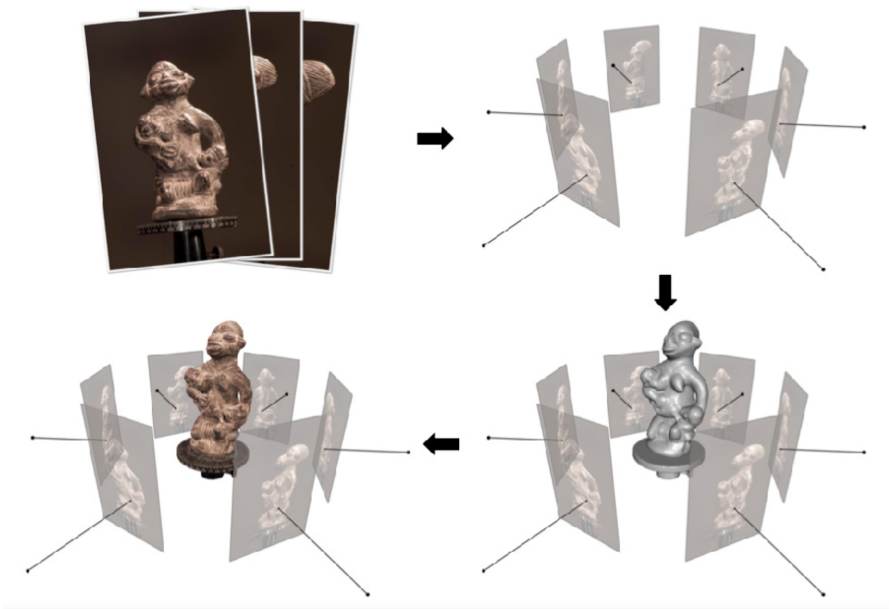

Figure 3.5.1: Structure from Motion (in the first row) estimates camera poses from multiple input images, while Multi-View Stereo (in the second row) reconstructs a dense 3D model [19].

3.5.2 Multi-View Stereo

Multi-View Stereo (MVS) [65] utilizes multiple images taken from different viewpoints around an object to reconstruct its 3D geometry as shown in Figure 3.5.1. The core idea is to identify corresponding points across different images and then triangulate their 3D positions. Traditional MVS

pipelines involve feature extraction, feature matching, bundle adjustment for camera pose estimation, and dense reconstruction. While computationally intensive and sensitive to textureless regions, new MVS methods are increasingly leveraging deep learning. Neural networks enhance feature matching robustness and improve depth map [74], making MVS an effective approach for generating high-fidelity 3D face models. Furtheremore, MVS can be build directly upon the output provided by SfM utilizing the camera poses from SfM to densify the sparse 3D structure into a continuous surface model. A leading implementation of this powerful SfM-MVS paradigm is COLMAP [64]. COLMAP is an open-source framework that uses state-of-the-art algorithms for both SfM and MVS into a unified and optimized pipeline.

3.5.3 Volumetric fusion techniques

Volumetric fusion methods, such as KinectFusion [47], reconstruct 3D surfaces by integrating multiple depth maps into a unified signed distance function volume. The scene is discretized into volumetric grid (voxels), each storing a truncated signed distance to the nearest surface: positive values indicate free space, negative values lie inside the object, and zero marks the surface itself. As new depth frames are aligned and fused, the SDF representation averages observations, filling gaps and smoothing noise. This approach creates complete 3D models and is widely used for real-time scanning with depth cameras.

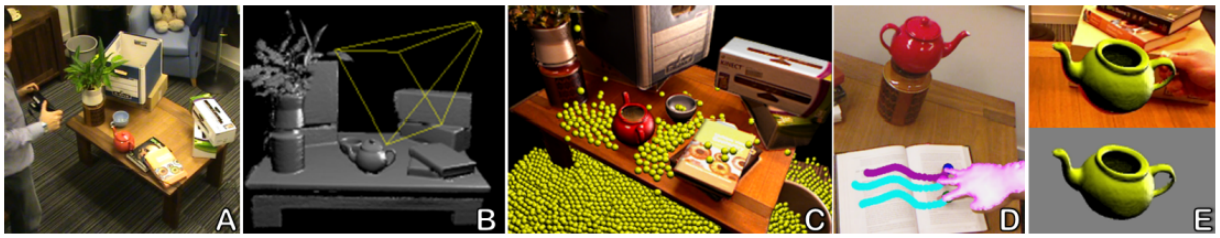

Figure 3.5.2: Demonstration of KinectFusion for real-time 3D reconstruction and interaction using a Kinect depth camera. (A) User scanning an indoor scene with Kinect. (B) Phong-shaded 3D reconstruction with wireframe frustum showing the tracked pose. (C) Texture-mapped 3D model reconstructed from Kinect RGB-D data. (D) Multi-touch interaction on the reconstructed surface. (E) Real-time segmentation and tracking of a physical object. [47]

3.5.4 Neural Radiance Fields

Neural Radiance Fields (NeRFs) [20] represent direction for 3D reconstruction. Originally introduced for novel view synthesis from a sparse set of input images, NeRFs learn a continuous volumetric scene representation that is also rich in geometric detail. A NeRF model uses a multilayer perceptron to map 3D coordinates and viewing directions to both RGB color and volume density. The density field defines geometry, with high-density regions indicating surfaces, and this can later be converted into a 3D mesh using techniques such as marching cubes. These neural methods can model subtle aspects of facial shape, skin, and even hair, making them a powerful tool. Though their volumetric nature can make editing and export to mesh pipelines more challenging.

3.6 3D Morphable Face Models

While volumetric methods provide detailed reconstructions, they also present certain challenges. They often produce dense, high-resolution models that can be difficult to animate or manipulate. More importantly, they lack a semantic understanding of identity and facial expression. This is where 3D Morphable Models have historically proven effective. Unlike raw geometric outputs, 3DMMs [62] provide a statistically learned representation of face variations. They offer a parametric framework in which a wide range of shapes and textures can be generated and controlled using only a small set of interpretable parameters.

Blanz and Vetter introduced the concept of the 3D Morphable Models [6] as a general face representation and a principled approach to image-based facial analysis. In their seminal work, they proposed a novel method for modeling textured 3D faces by transforming both shape and texture information from example scans into a vector space representation. They proposed modelling shape and texture variations using three-dimensional vertices rather than image coordinates. This formulation allows for the generation of new, realistic faces while constraining the model to avoid producing facial shapes or appearances that are statistically unlikely.

Constructing a 3DMM begins with building a 3D face database using high-resolution scans. A crucial step is solving the correspondence problem: ensuring that semantically consistent landmarks (e.g., nose tip, eye corners) are aligned across all scans so that each face shares the same topology and vertex ordering. This alignment process, known as registration, ensures that each 3D face in the dataset can be represented as a shape vector (capturing the 3D coordinates of its vertices) and a texture vector (representing RGB color values at each vertex). The shape vector of its n vertex coordinates takes the form $(X_1, Y_1, Z_1, \ldots, X_n, Y_n, Z_n)^{\mathrm{T}} \in \mathbb{R}^{3n}$, and similarly, the texture vector encodes $(R_1, G_1, B_1, \ldots, R_n, G_n, B_n)^{\mathrm{T}} \in \mathbb{R}^{3n}$.

Since the shape and texture vectors are not orthogonal and are high-dimensional, they cannot be directly used as basis vectors for modeling. Therefore, Principal Component Analysis (PCA) is employed to perform dimensionality reduction. First, the average shape and texture vectors across all training samples are computed. Then, each individual face vector is centered by subtracting the mean $\Delta S_i = S_i - \overline{S}$ and $\Delta T_i = T_i - \overline{T}$.

Covariance matrices for shape and texture C_S and C_T are calculated from these centered vectors, and PCA is applied to extract their eigenvalues and eigenvectors. The resulting eigenvectors form the principal components (or basis vectors) for shape and texture, denoted as s_i and t_i respectively. The 3D shape and texture of any face can now be represented as a weighted sum of these components:

$$S_{\text{model}} = \bar{S} + \sum_{i=1}^{k} \alpha_i S_i, \quad T_{\text{model}} = \bar{T} + \sum_{i=1}^{k} \beta_i T_i$$

where: $\bar{S}, \bar{T}$ are the mean shape and texture, S_i, T_i are the eigenvectors (principal components), α_i, β_i are the PCA coefficients

To reconstruct a 3D face from a single 2D image, the task is to estimate the optimal shape and texture coefficients α and β, along with additional rendering parametest. Theng parameters, combined with the 3DMM coefficients, allow the synthesized 3D face to be projected onto a 2D plane and compared with the input image. The fitting process then involves minimizing the difference between the rendered projection and the actual 2D image using an optimization loop.

3.6.1 FLAME

In this Thesis, we make use of FLAME [40]. FLAME is a statistical head model that separates **identity**, **pose**, and **facial expression** into separate controllable components, enabling flexible and accurate facial modeling. FLAME is designed to be efficient by using a relatively low-polygon mesh, while still maintaining realism through articulated joints and blend skinning. It is trained on sequences of 3D scans and is able to capture realistic blendshapes, which are approximate semantic parameterizations of facial expression.

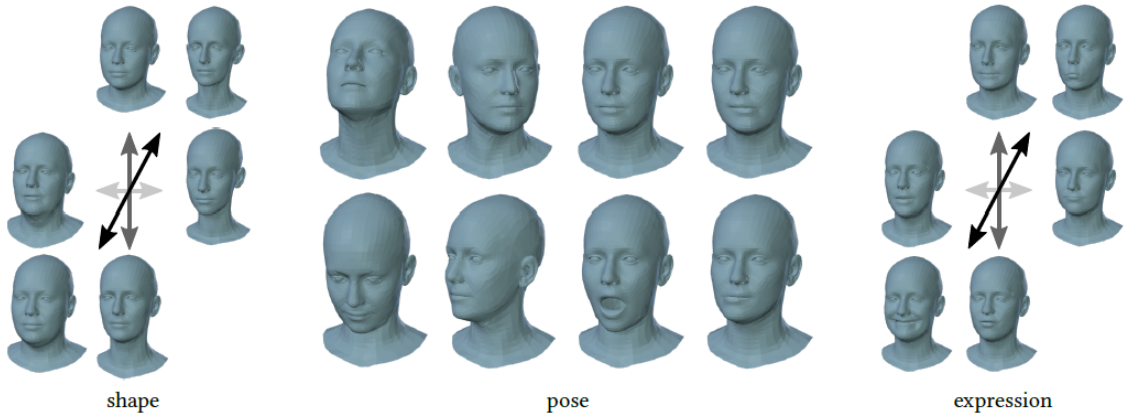

Figure 3.6.1: Parametrization of the FLAME model. Left: Activation of the first three shape components. Middle: Pose parameters actuating four of the six neck and jaw joints. Right: Activation of the first three expression components [40].

FLAME adapts the SMPL body model formulation [41] to heads. The model represents the human face as a triangular mesh consisting of N = 5023 vertices connected by edges, which define the surface geometry. Each vertex is associated with skinning weights W (values between 0 and 1 that sum to 1), describing how much the vertex is influenced by nearby joints (K = 4 joints: neck, jaw, and eyeballs). These joints, in turn, define the kinematic skeleton that allows the mesh to be articulated.

The model is defined by a mean template shape represented by a vector of N concatenated vertices $\bar{T} \in \mathbb{R}^{3N}$ in the zero pose θ^* and a set of blend weights $W \in \mathbb{R}^{N \times K}$. To account for variability, they use blendshapes. Each blendshape is essentially a displacement function that shifts vertex positions relative to a neutral template. Three sets of learned offsets are added:

- Shape blendshapes, which capture person-specific shape variation:

$$B_S(\beta; S) = \sum_{n=1}^{|\beta|} \beta_n S_n,$$

  where β is the vector of identity parameters and $S_n \in \mathbb{R}^{3N}$ are orthonormal principal components of shape displacements.

- Expression blendshapes, which capture deformations due to facial expressions:

$$B_E(\psi; E) = \sum_{n=1}^{|\psi|} \psi_n E_n,$$

  where ψ is the vector of expression parameters and $E_n \in \mathbb{R}^{3N}$ are expression basis vectors.

- Pose blendshapes, which capture deformations caused by joint rotations, such as jaw opening:

$$B_P(\theta; P) = \sum_{n=1}^{9K} \big(R_n(\theta) - R_n(\theta^*)\big) P_n,$$

where θ are the pose parameters, R_n are entries of the joint rotation matrices, and $P_n \in \mathbb{R}^{3N}$ are pose-corrective shapes.

The three components are added to the mean template to form the posed mesh:

$$T_P(\beta, \theta, \psi) = \bar{T} + B_S(\beta) + B_P(\theta) + B_E(\psi).$$

To produce natural articulation, FLAME applies Linear Blend Skinning, which blends vertex transformations from nearby joints according to the skinning weights. The complete model is written as:

$$M(\beta, \theta, \psi) = W\big(T_P(\beta, \theta, \psi), J(\beta), \theta, W\big),$$

where W is the skinning function, W are the skinning weights, and $J(\beta)$ is a sparse matrix defining how to compute joint locations from mesh vertices.

In summary, FLAME represents faces as the mean template mesh plus identity, expression, and pose blendshapes, articulated through linear blend skinning. This decomposition provides a compact and controllable model for realistic face generation and animation.

3.6.2 Other 3DMMs

Although FLAME provides a more advanced parametric representation by jointly modeling identity, pose, and expression in a unified framework, it is important to also present earlier notable 3D Morphable Models that primarily focused on facial identity and texture. The Basel Face Model (BFM) [53] is a seminal and widely adopted 3D Morphable Model build from 100 high-quality 3D scans. It uses PCA to create separate statistical models for face shape and texture , ensuring high fidelity through a refined registration pipeline. BFM's strength lies in its analysis-by-synthesis fitting, which disentangles identity from external factors enable generation of robust face models.

While BFM laid the foundation for 3DMM-based face modeling, its limited dataset restricted demographic diversity. LSFM was developed to address this limitation by scaling the model to thousands of scans, offering broader coverage and stronger generalization. The Large Scale Facial Model (LSFM) [7] is a 3D Morphable Model automatically constructed from 9,663 distinct facial identities, making it one of the largest-scale morphable models available. LSFM employs a PCA-based statistical framework to model facial shape and texture variations, relying on an advanced pipeline to ensure dense point-to-point correspondence across all scans. An example of the LSFM model and its principal components is shown in Figure 3.6.2. By leveraging its rich demographic data, LSFM also enables the creation of tailored submodels for specific age, gender, and ethnicity groups, significantly enhancing its accuracy and generalization compared to smaller, less diverse models in generating realistic 3D faces.

.

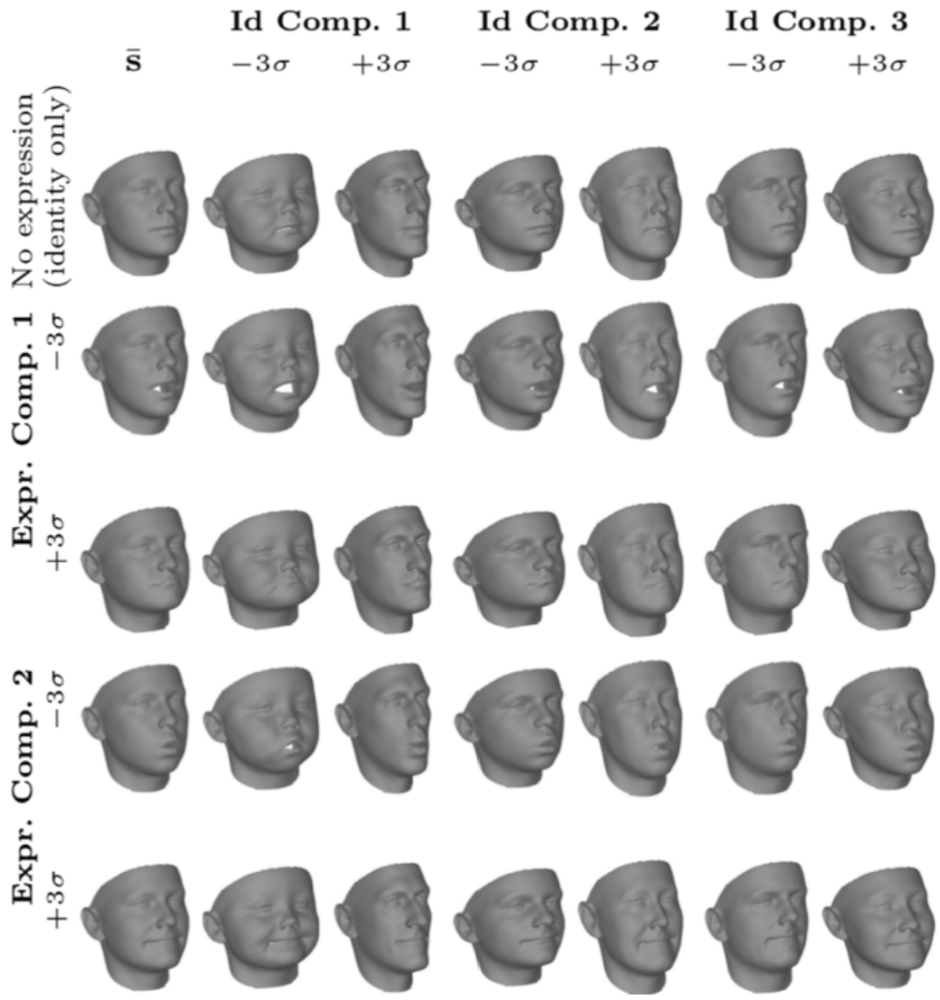

Figure 3.6.2: Principal component analysis of the LSFM , illustrating facial identity variation using the first three components and expression variation using the first two components. [15].

Chapter 4

Literature

Contents

4.1 3D Face Reconstruction Literature

Existing approaches in face reconstruction can be broadly categorized into three groups based on the input data: image/video-driven, audio-driven, and audiovisual methods.

- **Image/Video-driven methods** rely on monocular visual input to reconstruct the facial surface. The task is to infer the parameters of a 3D Morphable Model or directly regress the 3D geometry that best explain the face depicted in a given image. Predicting the 3D facial performance from a video refers to the dynamic version of the reconstruction problem. Here, the objective is not only to estimate the 3D facial geometry for each frame, but also to capture the temporal information of expressions and motions, resulting in a coherent sequence that represents the performance of the face over time.
- **Audio-driven methods** use speech signals as the primary input for predicting facial motion. These models typically focus on synchronizing mouth and lip movements with speech content generating talking head avatars.
- **Audiovisual methods** combine both image and audio modalities to leverage the complementary strengths of each modality. By fusing visual input with speech-driven dynamics, these methods can produce reconstructions that capture both fine-grained geometry mesh and temporally coherent expressions. Such approaches have become an active research area in recent years with many open challenges remaining.

In this thesis, we explore the third category and we propose a novel audiovisual method for 4D face reconstruction. In the following sections, we review some of the most relevant works in each category.

4.2 Facial Reconstruction Methods: Image/Video-Driven

In this section, we present some recent models that take a single monocular image as input and reconstruct the corresponding 3D face geometry.

4.2.1 DECA

Detailed Expression Capture and Animation (DECA) [17] is a 3D face reconstruction model that emphasizes on capturing fine-grained details. DECA addresses limitations of previous approaches, which struggles to model animatable faces with dynamic details like wrinkles, while trained on in-the-wild images. The model consists of two main components: a coarse reconstruction and a detail reconstruction.

In the first step, the model learns a coarse 3D face reconstruction in FLAME model space using an analysis-by-synthesis strategy. Given a 2D image as input, the model encodes the image into a latent representation, decode it to render a synthetic 2D image, and minimizes the difference between the rendered and input images. The encoder consists of a ResNet50 network [25] followed by a fully connected layer that regress FLAME and enviroment parameters. Formally, given an image I, the coarse encoder $\mathcal{E}_c$

$$E_c(I) \rightarrow (\ \beta, \theta, \psi, l, c, A)$$

outputs FLAME parameters β, θ, ψ, lighting l and camera parameters c and and albedo A. Training uses a dataset of 2D face images with multiple samples per subject, together with identity labels and ground-truth landmarks. Supervision combines several objectives: landmark

alignment, eye-closure consistency, photometric error, identity preservation, shape consistency, and regularization.

In the second step, the model augments the coarse FLAME geometry with a UV displacement map. The detail encoder shares the same architecture as the coarse encoder and predicts a subject-specific latent code δ.

$$E_d(I) \rightarrow \delta$$

This code is concatenated with FLAME expression and jaw-pose parameters and decoded into a UV displacement map D.

$$F_d(\psi, \delta, \theta_{jaw}) \rightarrow D$$

The displacement map is then converted to a detailed normal map, which augments the differential renderer with dynamic details. An overview of the architecture is illustrated in Figure 4.2.1. Training minimizes a combination of photometric loss, ID-MRF loss, symmetry loss, regularization, and a consistency loss. The consistency loss ensures that subject-specific details are disentangled from expression-dependent ones. Specifically, exchanging detail codes between two images of the same subject should not alter the rendered output.

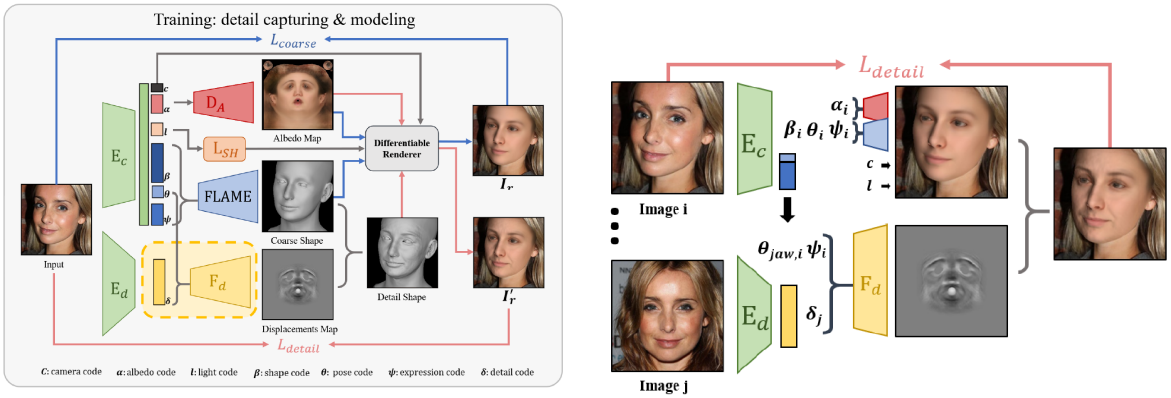

Figure 4.2.1: DECA Training and Detail Consistency Loss. In the training stage (left box), DECA enforces shape consistency and learns an expression-conditioned displacement model from detail consistency across multiple images of the same person. On the right scheme, extracting the detail code from image j and combining it with the expression of image i should have no effect on the rendered image [17].

4.2.2 EMOCA

Emotion Capture and Animation (EMOCA) [13] is a neural network that reconstructs an animatable 3D face from in-the-wild images, capturing expression details that convey the emotional content of the input. It is build on top of DECA by leveraging its accurate identity shape reconstruction accuracy. EMOCA discards the expression parameters predicted by DECA and introduces a separate encoder to estimate facial expressions, while keeping the other components fixed.

$$E_e(I) \rightarrow \psi_e$$

This architecture reduces the number of trainable parameters, leading to lower resource requirements, faster training, and reduced memory consumption. The model can be trained directly

on emotion-rich datasets without requiring multiple images per subject and it uses a state-of-the-art emotion recognition model for expression supervision. A novel perceptual emotion loss is introduced to enforce similarity between the emotion features of the input and rendered images:

$$L_{emo} = \|\epsilon_1 - \epsilon_2\|_2$$

where ϵ_1 and ϵ_2 are the emotion feature vectors of the input and the synthesized image respectevely predicted by the pretrained emotion recognition model.

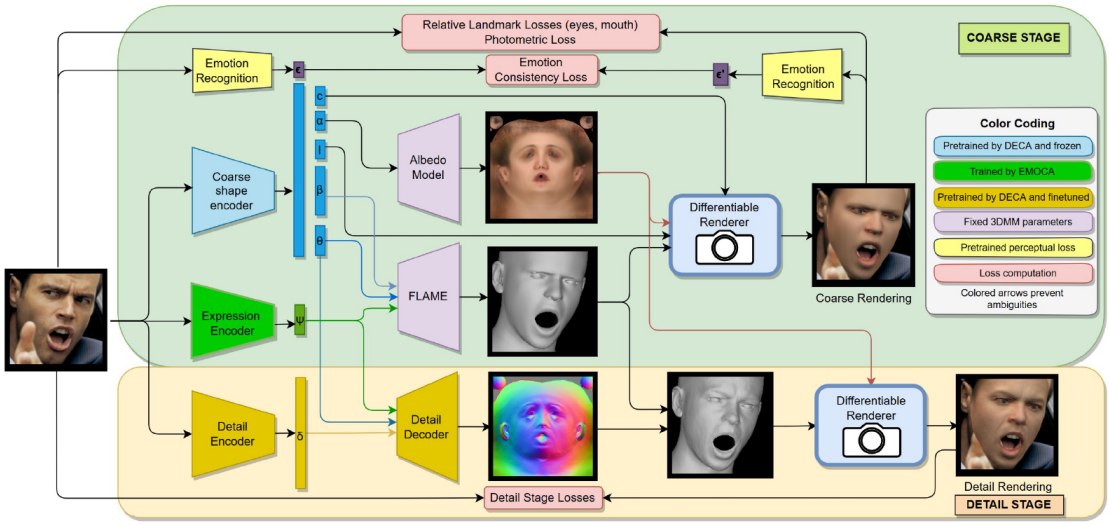

Figure 4.2.2: EMOCA: extension of DECA for emotional face capture. Given an input image, the coarse shape encoder (initialized from DECA and kept fixed) predicts the coarse facial shape, while EMOCA's trainable encoder estimates the expressions. During training of the detail encoder, the EMOCA's expression encoder is fixed. To estimate the emotion consistency loss, both the original and the coarse rendered images are passed through a pretrained emotion recognition network [13].

4.2.3 SMIRK

In this subsection we present Spatial Modeling for Image-based Reconstruction of Kinesics (SMIRK) [59], a method for accurate expressive 3D face reconstruction from images. It adresses shortcomings in self-supervised formulation and pure expression diversity data in previous approaches, presenting a novel image-to-image translator model based on U-Net [61]. The key idea is to reconstruct the image while relying only on the rendered predicted geometry and a small number of sampled pixels. SMIRK uses separate encoders, which consist of a MobilenetV3 backbone [27] followed by a fully connected layer to regress the FLAME parameters.

$$\theta = E_\theta(I), \quad \beta = E_\beta(I), \quad \psi = E_\psi(I).$$

Encoder for shape E_θ and pose E_β are pretrained and remain freezed, while expression encoder E_ψ is trainable. Supervision is enabled with two alternative separate passes during each training iteration.

Reconstruction path: Given an image I the model encodes the FLAME parameters to render the face geometry S. Then, the input image is masked out $M(I)$ and along with S, is fed into the neural renderer T to obtain the reconstruction image I'

$$I' = T\big(S \oplus M(I)\big)$$

where $\oplus$ denotes concatenation. The reconstruction path is supervised using a combination of standard self-supervised landmark, photometric and perceptual losses.

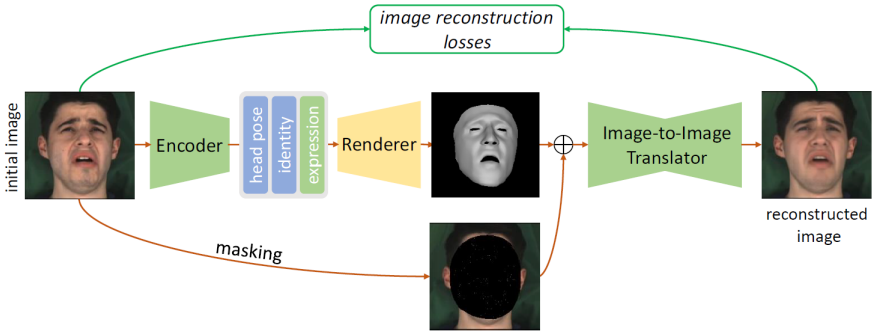

Figure 4.2.3: Reconstruction path. The encoder predicts head pose, identity, and expression parameters, which are rendered into a 3D geometry. This rendering is concatenated with the masked input image and passed through an image-to-image translator to produce the reconstructed image.[59].

Augmented expression cycle path: In this path, the predicted expression parameters ψ are replaced with an augmented expression ψ_{aug}, while keeping the predicted shape and pose fixed. The translator network T generates a new image I'_{aug} corresponding to ψ_{aug}. This image is then passed through the expression encoder E_ψ to recover the expression parameters. A cycle consistency loss is applied to enforce that the predicted expression matches the original augmented expression:

$$\mathcal{L}_{\text{exp}} = \left\| E_\psi\big(T\big(R(\theta, \beta, \psi_{\text{aug}}) \oplus M(I)\big)\big) - \psi_{\text{aug}} \right\|_2^2$$

During training, the expression encoder and the image-to-image translator are alternately frozen to prevent the translator from compensating for encoder errors, ensuring robust disentanglement of expression parameters.

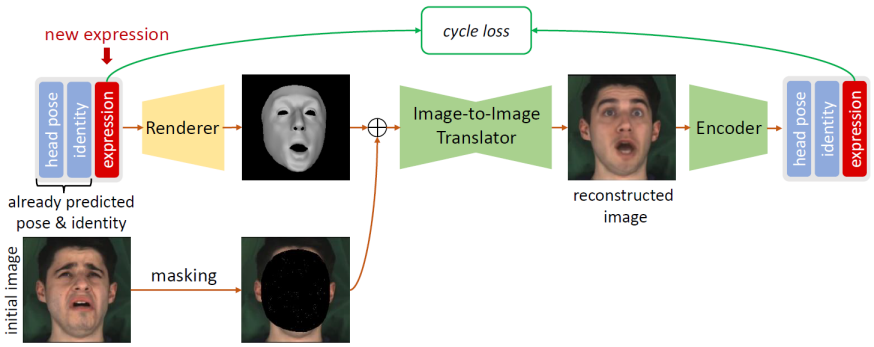

Figure 4.2.4: Augmented Cycle Path. Identity and pose are fixed while new expressions are used to generate augmented faces. The cycle loss enforces consistency, allowing the model to learn from varied expression inputs [59].

4.2.4 SPECTRE

We continue the literature review with SPECTRE [18], a method for perceptual 3D face from reconstruction from videos which focuses on accurately modeling the mouth area. To achieve this, it introduces a lip-read loss that enforces strong correlation between lip movements and speech articulation.

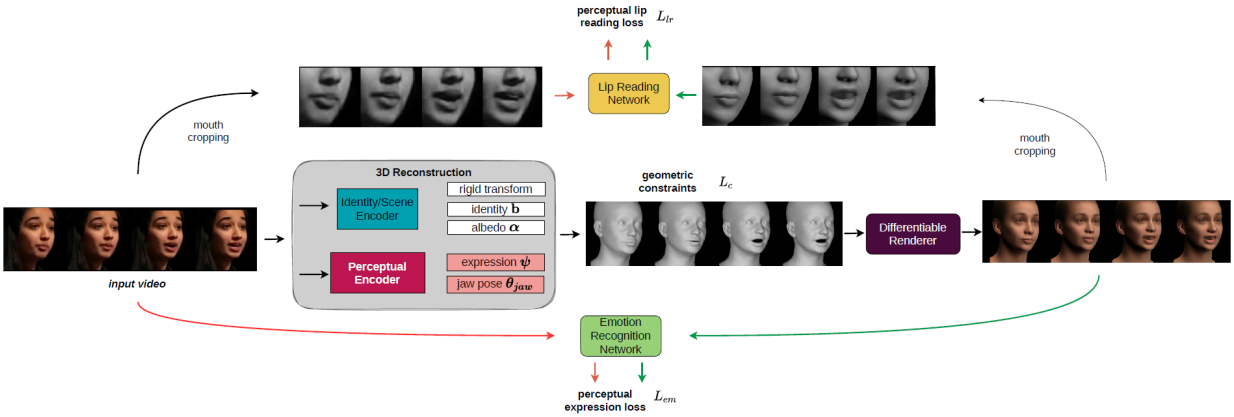

Figure 4.2.5: Overview of the SPECTRE architecture. The input video is fed into a fixed encoder which estimates scene parameters and identity parameters, and an coarse prediction of jaw and expression parameters. A mouth/expression encoder then augments the expression and jaw pose, and a differentiable renderer produces the corresponding 3D face. The mouth region is cropped from both the input and the rendered sequences, and a lip-reading network is applied to compute the perceptual lipreading loss. In parallel, a facial expression recognizer is used in the same way to compute the perceptual expression loss. [18].

SPECTRE borrows EMOCA network architecture but replaces the ResNet50 encoder with a lightweight MobileNetV2 followed by a temporal convolution layer to reduce the computational overhead of the system. For training supervision, a perceptual expression loss is applied between the emotional feature vectors of the input video and those of the reconstructed 3D mesh.

To improve the reconstruction of the mouth (expression and jaw parameters in FLAME space), SPECTRE introduces a perceptual mouth-oriented loss. Since 2D landmarks often suffer from inaccuracies, a network pretrained on the LRS3 dataset [43] is used. It takes as input grayscale sequences of both the original and differentiably rendered images cropped around the mouth, and estimates the cosine distance between them.

Geometric constraints are added to address the domain gap between rendered and original images. These penalize significant deviations of the predicted expression and jaw parameters from the corresponding DECA predictions, which serve as a reliable initialization. In addition, landmark losses are included as an alternative geometric constraint by enforcing consistency between the 2D landmarks of the reconstructed and the original face images.

4.3 Facial Reconstruction Methods: Audio-Driven

Audio-driven reconstrauction methods, also refered to as talking head generation, aim to generate a talking head avatar whose facial movements is accurately synced with the input audio.

4.3.1 VOCA

Voice Operated Character Animation (VOCA) [12] is a speaker-independent animation framework for talking faces. It is trained on VOCASET, a dataset of 4D facial scans aligned with the FLAME model. By conditioning on subject labels, the model can reproduce a wide range of speaking styles. The input consists of a subject-specific template mesh and the raw audio signal. The audio is processed by DeepSpeech [24], a well establish speech recognition RNN model, to extract speech features that are then passed through an encoder. The encoder is composed of four temporal convolution layers and two fully connected layers while subject identity is represented by an one-hot vector. The decoder maps the low-dimensional embeddings to 3D vertex displacements, which are added to the input template. The output is an animated 3D face in a neutral pose, represented in FLAME model space.

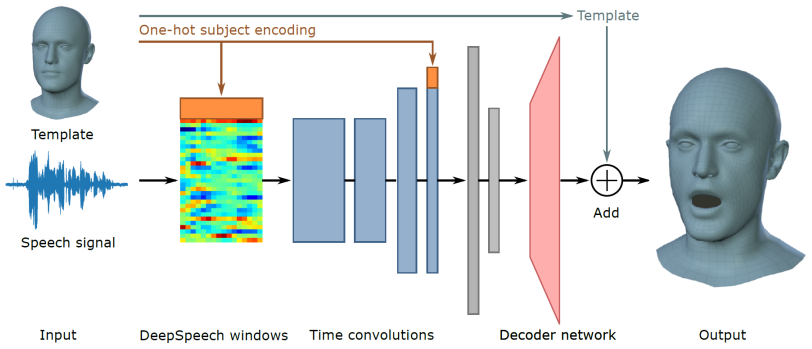

Figure 4.3.1: VOCA architecture. The model is trained to predict expression coefficients from audio features using a loss function with two terms: a position loss that enforces vertex alignment with ground truth, and a velocity loss that promotes temporal stability across frames.[12].

4.3.2 MeshTalk

MeshTalk [60] is a cross-modal framework for speech-driven 3D face animation that disentangles identity-specific geometry from speech-correlated motion. Given a sequence of face meshes $x_{1:T} = (x_1, \ldots, x_T)$, where each mesh $x_t \in \mathbb{R}^{V\times 3}$ represents V vertices in 3D space, and a corresponding sequence of aligned speech snippets $a_{1:T} = (a_1, \ldots, a_T)$, with each audio feature $a_t \in \mathbb{R}^D$ be a sequence of T speech snippets, each with D samples, aligned to the corresponding visual frame t. MeshTalk learns a shared categorical latent expression space that captures speech-related facial motion.

The system comprises an encoder and a decoder. The encoder contains two parallel streams: an audio encoder that extracts temporal features from the speech signal and an expression encoder that encodes mesh-based motion dynamics. Their outputs are fused and mapped to a categorical latent expression space. These categorical codes serve as an interpretable and compact representation of facial motion.

The decoder, implemented as a U-Net–style architecture, takes a neutral face template mesh h and the latent expression sequence to predict an animated sequence of meshes. This design enables the model to reconstruct temporally consistent and expressive facial motion. MeshTalk is trained with objectives promoting cross-modal consistency, motion realism, and temporal smoothness, ensuring that the generated animations accurately reflect speech content while remaining independent of speaker identity.

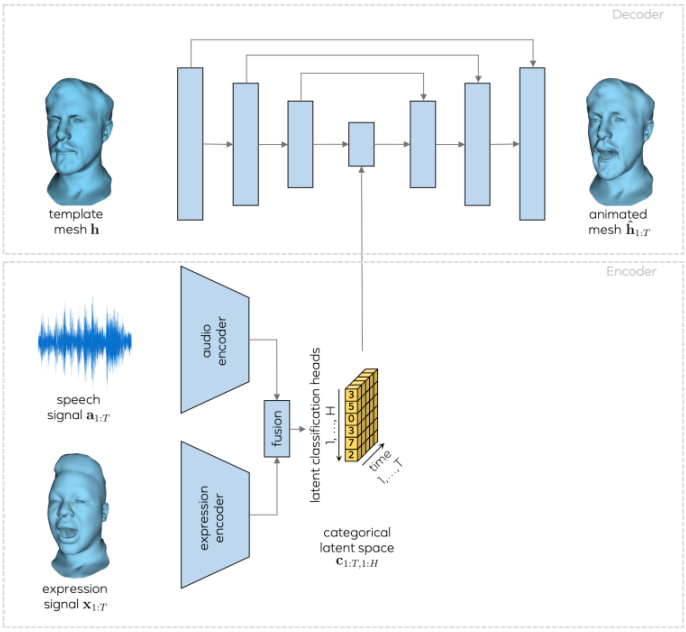

Figure 4.3.2: Overview of MeshTalk architecture.[60].

4.3.3 FaceFormer

FaceFormer [16] is an autoregressive transformer-based model for speech-driven 3D facial animation. It is designed to overcome two main limitations of earlier approaches: the difficulty of modeling long-term audio context and the lack of large-scale 3D audio–visual training data. Instead of using convolutional layers for temporal modeling, FaceFormer adopts a transformer architecture, which can attend to all tokens in the input sequence in parallel. This design enables the effective capture of long-range contextual information and supports more realistic full-face animation.

The model follows an encoder–decoder structure. The encoder is inspired by wav2vec and begins with several temporal convolution layers that extract features directly from raw audio. A linear interpolation layer is applied for resampling, and a multi-layer transformer encoder with a linear projection maps the audio features to speech representations.

The decoder contains two core modules. The first is a biased multi-head self-attention with periodic positional encoding, which allows the model to generalize to longer input sequences. The second is a biased cross-modal multi-head attention module, which aligns the audio representations with motion features to produce temporally coherent 3D facial animations. In this way, FaceFormer effectively integrates speech information with motion dynamics to generate smooth and realistic talking-face sequences.

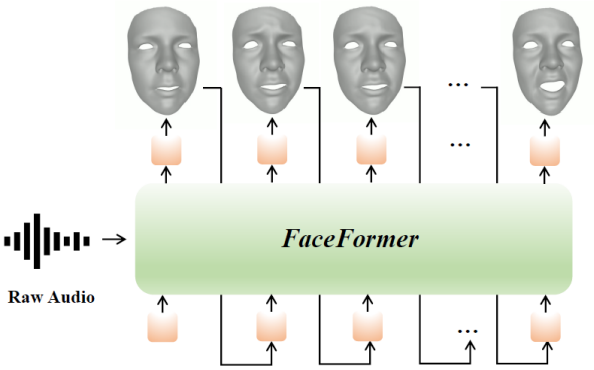

Figure 4.3.3: High-level overview of FaceFormer. Given a raw audio signal and a neutral 3D face mesh, FaceFormer is composed of an end-to-end transformer-based architecture that generates 3D facial motion sequences with accurate synced lip.[16].

4.3.4 CodeTalker

CodeTalker [73] introduces a speech-driven 3D facial animation framework designed to address the over-smoothing and ambiguity issues common in traditional regression-based methods. Rather than directly regressing continuous facial motion parameters from speech, CodeTalker learns a discrete motion prior that constrains the generation process to realistic motion patterns. This prior is established through a vector-quantized autoencoder, which encodes real facial motion sequences into a set of discrete motion primitives stored in a learned motion codebook. During inference, a cross-modal transformer decoder takes as input the speech features, a style vector encoding identity information, and past facial motions to predict the corresponding sequence of motion codes by querying this codebook. This discretization constrains the output space to realistic facial motions improving lip synchronization.

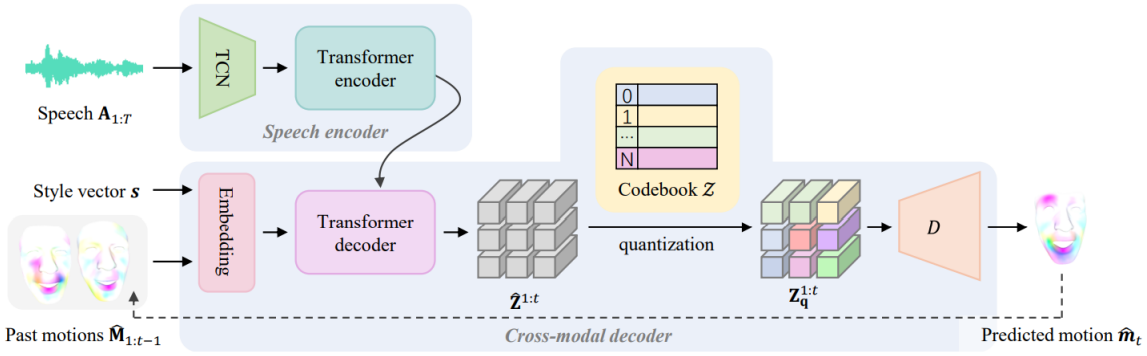

Figure 4.3.4: Overview of the CodeTalker speech-driven 3D facial motion synthesis.[73].

4.4 Facial Reconstruction Methods: Audio and Image-Driven

Audiovisual talking face generation aims to synthesize realistic videos of talking humans by jointly using audio and visual information. Combining both modalities enables the model to generate lip movements that are accurately synchronized with speech while maintaining the visual consistency of the speaker's face. Limited work has be done toward this direction and to the best of our knowledge AVFace [11] is the only method for spatio-temporal (4D) face reconstruction.

4.4.1 AVFace

Audio-Visual 4D Face Reconstruction (AVFace) [11] is a method that leverage a combination of these modalities for 4D face reconstruction. Its architecture consists of two stages.

The encoder combines a ResNet-50 with a transformer using multi-head self-attention. Expression and jaw parameters are obtained from the concatenation of audio features extracted with DeepSpeech and processed by a 1D convolution layer, and visual embeddings from ResNet-50. These fused features are passed through a transformer encoder to capture the temporal structure of the input sequence. During training, drop-modality is applied to ensure effective use of both audio and visual modalities.

In addition, a FiLM-conditioned SIREN MLP [54] predicts lip vertex offsets from audio, which are added to the coarse lip vertices to refine the mouth shape and position.

In the second stage, high-frequency facial details such as wrinkles and folds are recovered. A UNet-ResNet, initialized with pretrained weights, predicts normal displacements D from the coarse output. Audio features are fused in the encoder to improve prediction around the mouth.

AVFace exploits the temporal modeling and audio signals to remain robust under face occlusions, such as hand motions. While the training set already includes some occluded frames, robustness is further improved by fine-tuning with synthetic data using the method proposed by [70].

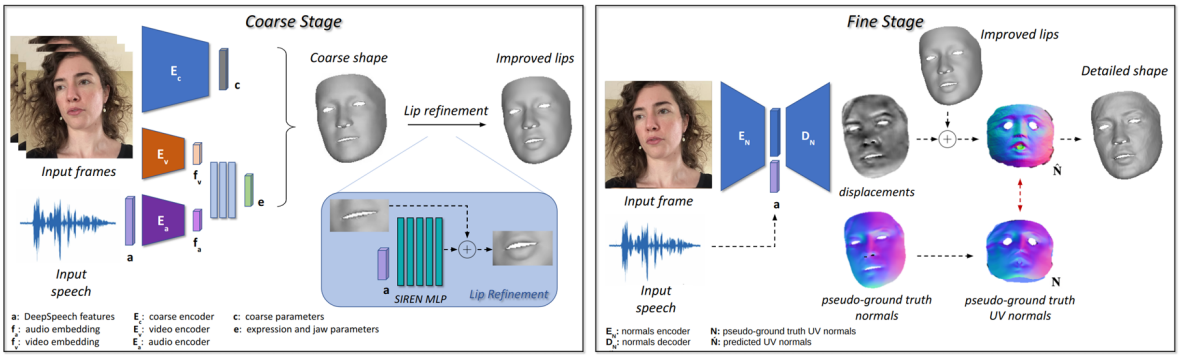

Figure 4.4.1: AVFace Network Architecture. Given a video of a talking face and its corresponding speech segment, the method applies a coarse-to-fine optimization process to reconstruct detailed 4D facial geometry. [11].

Chapter 5

Proposed Method

Contents

5.1 Preface

This chapter describes the proposed pipeline for end-to-end 3D face reconstruction. The aim of the method is to reconstruct faces with accurate geometry and realistic mouth movements under challenging real-world conditions. Our dataset consist of in-the-wild videos containing noise, occlusions, and other unpredictable variations. These issues limit the reliability of existing approaches and motivate the development of more robust reconstruction systems.

We focus on two main difficulties. The first is occlusion, where parts of the face, such as the mouth, are covered (e.g., by a hand in front of the speaker). Such occlusions are frequent in talking face videos, and they limit the performance of purely image-based methods. To address this, the proposed model combines temporal information with audio features so that the face and especially the mouth area can be reconstructed even when it is not visible. The second difficulty is background noise, which can be either speech from other speakers or environmental sounds. In this case, the model is designed to rely more heavily on the visual modality to reconstruct a closed mouth when the speech signal does not originate from the main character depicted in the video.

In addition, our framework has a dual-mode design. It can operate as an audio-visual model, exploiting the complementary strengths of both modalities, or as an audio-only model, which itself achieves high overall performance. This flexibility enables the system to adapt to a wide range of application scenarios, depending on the availability and reliability of the input data.

5.2 Preliminaries

As we established in Chapter 4 modeling the human face as a three-dimensional object instead of an two-dimensional image can better capture its variability, which is crucial for the perception of realism. Hence, in our method we use FLAME [40] 3D morphable model. FLAME is a statistical 3D head model with seperate parameters for identity shape $\beta \in \mathbb{R}^{|\beta|}$, facial expression $\psi \in \mathbb{R}^{|\psi|}$, and pose parameters $\theta \in \mathbb{R}^{3k+3}$ for rotations around $k = 4$ joints (neck, jaw, and eyeballs) and the global rotation. Given all parameters, FLAME outputs a mesh with $n_v = 5023$ vertices. Formally, FLAME is:

$$M(\beta, \theta, \psi) \rightarrow (V, F),$$

with vertices $V \in \mathbb{R}^{n_v \times 3}$ and $n_f = 9976$ faces $F \in \mathbb{R}^{n_f \times 3}$.

Most existing approaches rely exclusively on visual input, whereas the proposed model is further enhanced by incorporating audio information. For this purpose, wav2vec [4] is employed as a pretrained speech representation model to extract audio embeddings. The wav2vec architecture consists of three components. A multi-layer convolutional feature encoder $f : X \rightarrow Z$ maps raw audio X into latent speech representations $z_1, \ldots, z_T$. These latent vectors are provided to a Transformer network $g : Z \rightarrow C$, which produces contextualized representations $c_1, \ldots, c_T$ by modeling long-range dependencies across the entire sequence. In parallel, the latent representations are discretized by a quantization module $Z \rightarrow Q$, creating codebook entries q_t that serve as prediction targets for the self-supervised training objective. The overall architecture is illustrated in Figure 5.2.1.

During pre-training, a proportion of the feature encoder outputs is masked before being passed into the Transformer. Masking is performed by randomly sampling a subset of time steps as

starting indices and replacing spans of M consecutive latent vectors with a single trainable mask embedding, which is shared across all masked positions.

The model is supervised by a contrastive loss $\mathcal{L}_m$, which requires to identify the true quantized latent speech representation q_{true} among a set of distractors. In addition, a diversity loss $\mathcal{L}_d$ regularizes the quantization module by encouraging uniform usage of all codebook entries.

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d,$$

where α is a tunable hyperparameter.

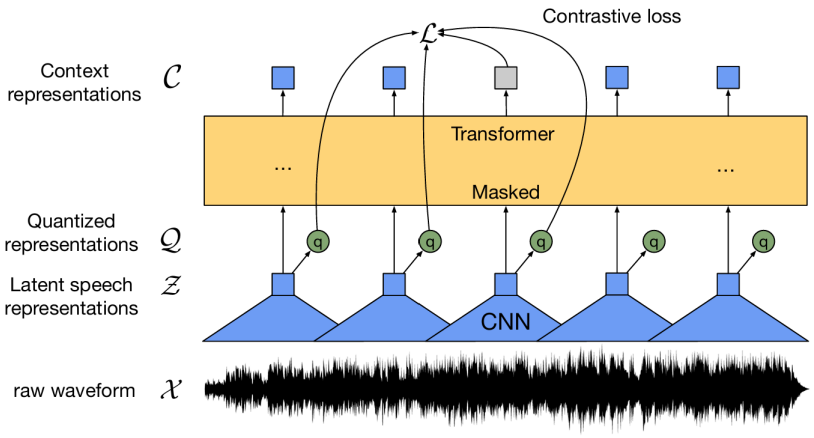

Figure 5.2.1: Wav2vec framework overfiew. The model encodes raw audio into latent representations, applies masking and contextual modeling with a Transformer, and learns discretized speech units through quantization. [4].

5.3 Synthetic Occlusions

Our dataset consists of in-the-wild videos which provide challenging input sequences for our model. However, two difficulties arise: (i) there are relatively few videos containing natural facial occlusions, and (ii) landmark extraction often fails under heavy occlusions, which can mislead the model.

To address these issues, we augment the dataset with two categories of synthetic occlusions. The first category consists of random objects or hands using the face occlusion generation method proposed by [70]. For a random sequence of consecutive frames from our real data, we then synthesize an occlusion event. The occluder is first placed at an initial position in the starting frame. From this point onward, we introduce two forms of temporal transformation: translation and rotation. The occluder's motion across frames is defined by three key points: a random starting point outside the image boundary, a target point near the mouth region, and an end point outside the boundary on the opposite side of the frame. Intermediate positions are interpolated smoothly between these points, creating a curved trajectory. This guarantees that the occluder enters from outside the frame, moves smoothly across the face, and exits again outside the frame. In addition, a total rotation angle is randomly sampled in the range $[10°, 120°]$ (positive or negative), which is applied progressively to the occluder across frames using a cosine-based easing function.

The second category consists of medical masks, which are synthesized using MediaPipe landmarks

Figure 5.3.1: Synthetic hand and object occlusions applied to a training video sample, visualized every second frame.

[34]. We define polygons around the lower face using predefined landmark chains along the left and right jawlines and the nose. We fill this polygon with a randomly selected texture from a large dataset. For each video, a random subsequence of consecutive frames is selected, and the mask is applied consistently across those frames.

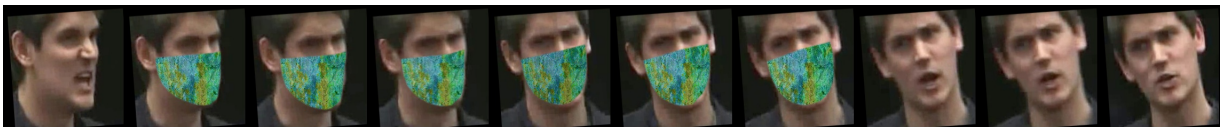

Figure 5.3.2: Synthetic textured surgical mask applied on video frames. The synthetic mask successfully follows the face motion across frames.

5.4 Data Processing

To process the videos for training, we divide them into fixed -length segments and we use a variable K to denote the number of frames extracted from each video. Depending on the segment length, we sample a random continuous subsequence of K frames. For each frame, we load pre-computed facial landmarks obtained from FAN and MediaPipe. We use the MediaPipe Face Landmarker for landmark extraction on each video frame. As described in Chapter 3, the model detects a single face and estimates 468 three-dimensional keypoints that describe the overall face geometry. In addition, we extract landmarks using the FAN model which predict 68 face keypoints.

In order to expedite the training process, we sub-sample both the MediaPipe and FAN landmarks and we keep the combination of them. For the jawline, we select the 17 FAN landmarks, while for the remaining facial regions such as the mouth, eyes and nose, we retain 109 MediaPipe keypoints instead of the full 468. This brings us to a total of 126 landmarks which results in a compact landmark set that preserves the most relevant facial information.

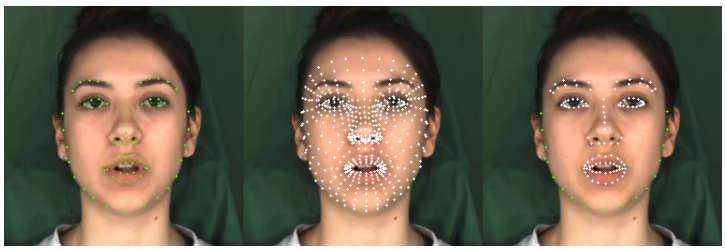

Figure 5.4.1: First image depicted the FAN landmarks while the second visualize the Mediapipe keypoints. For out method we use a combined subset.

Also, our method relies of masking faces. To that end, we use the mediapipe landmarks (or the fan landmarks if the former are missing) and calculate the convex hull of the landmarks in order to generate the binary mask that covers the full face.

Finally, each dataset sample contains a sequence of original frames, the synthetic occluded frames, the corresponding FAN and MediaPipe landmarks together with flags indicating whether they were valid, the convex hull mask, and the audio waveform.

5.5 Architecture

The proposed architecture, **Face AudioVisual Occluded-robust Reconstruction** (FAVOR), built on top of SMIRK [59] which is a state-of-the-art model for capturing extreme expressions. We model the human face as a three-dimensional object in FLAME space. An encoder $E(\cdot)$ takes as input a sequence K of synthetic occluded video frames $I^{\text{occl}}_{1:K} = \{I^{\text{occl}}_1, I^{\text{occl}}_2, \dots, I^{\text{occl}}_K\}$ and regresses the FLAME parameters that correspond to each frame. Following the design of SMIRK, we separate E into three branches:

Pose encoder $\mathbf{E}_\theta$ predicts the pose parameters $\theta_{1:K} = \{\theta_1, \dots, \theta_K\}$. It consists of a MobileNetV3 backbone [27] applied to $I^{\text{occl}}_{1:K}$.

$$\theta_{1:K} = E_\theta(I^{\text{occl}}_{1:K})$$

Shape encoder $\mathbf{E}_\beta$ regresses the shape parameters $\beta_{1:K} = \{\beta_1, \dots, \beta_K\}$. It also uses a MobileNetV3 backbone followed by 1D convolution layers applied across the temporal dimension, which allows the model to capture short-term dynamics from a sequence of video frames.

$$\beta_{1:K} = E_\beta(I^{\text{occl}}_{1:T})$$

Expression encoder $\mathbf{E}_\psi$ takes as input both frames and audio. The frames are processed through a MobileNetV3 backbone. For the audio input, we extract features using wav2vec2.0, a transformer-based model pre-trained on 960 hours of unlabeled audio from LibriSpeech dataset [49] and fine-tuned for ASR on the same audio with the corresponding transcripts. Given an input audio segment of length K, wav2vec outputs features $a_u \in \mathbb{R}^{49\times 768}$ where 49 is the number of audio frames and 768 is the embedding dimension. To align the frequency of the wav2vec output with the video frame rate, we use a 1D convolution with stride 2, resampling features to 25 fps. This step forms the first layer of the audio head, which is followed by four additional 1D convolution layers. Then, the audio features are projected to the same dimension as the visual features so that both share a common latent space. The two modalities are concatenated, and a temporal CNN is applied to capture cross-modal dependencies. This design enables smooth and accurate prediction of facial expressions $\psi_{1:K} = \{\psi_1, \dots, \psi_K\}$ across frames. For training Pose and Shape Encoders were pre-trained and remain frozen.

$$\psi_{1:K} = E_\psi(I^{\text{occl}}_{1:K})$$

We follow the same architecture from SMIRK for the renderer and the generator. We use a **differentiable renderer** based on orthographic projection and a mesh rasterization step. The predicted 3D vertices are projected with scale and translation parameters, and per-vertex attributes are interpolated across pixels using barycentric coordinates. Because the rasterization is differentiable, gradients can flow from the rendered images back to the FLAME parameters and the encoder. Formally,

$$S_{1:K} = R(\theta_{1:K}, \beta_{1:K}, \psi_{1:K})$$

where $S_{1:K}$ denote the sequence of outputs of the differentiable rasterization step, where $S_{1:K}$ are the monochrome renderings of the reconstructed face mesh.

Generator consists of a U-Net archineture and takes as input the sequence of rendered predicted mesh $S_{1:K}$ and sparsely sampled pixels of the input. A masking function $M(\cdot)$ is applied to the original non-occluded input frames masking out the face and only retaining a small amount of randomlly selected pixels. $M(I_{1:K})$ is concatenated with $S_{1:K}$ and the resulting tensor is passed through the neural renderer T to produce a reconstruction of the original image

$$I'_{1:K} = T(S_{1:K} \oplus M(I_{1:K})$$

where $\oplus$ denotes concatenation. It is important that the model learns to reconstruct the non-occluded image so that it will not be affected by the occlusions in the dataset.

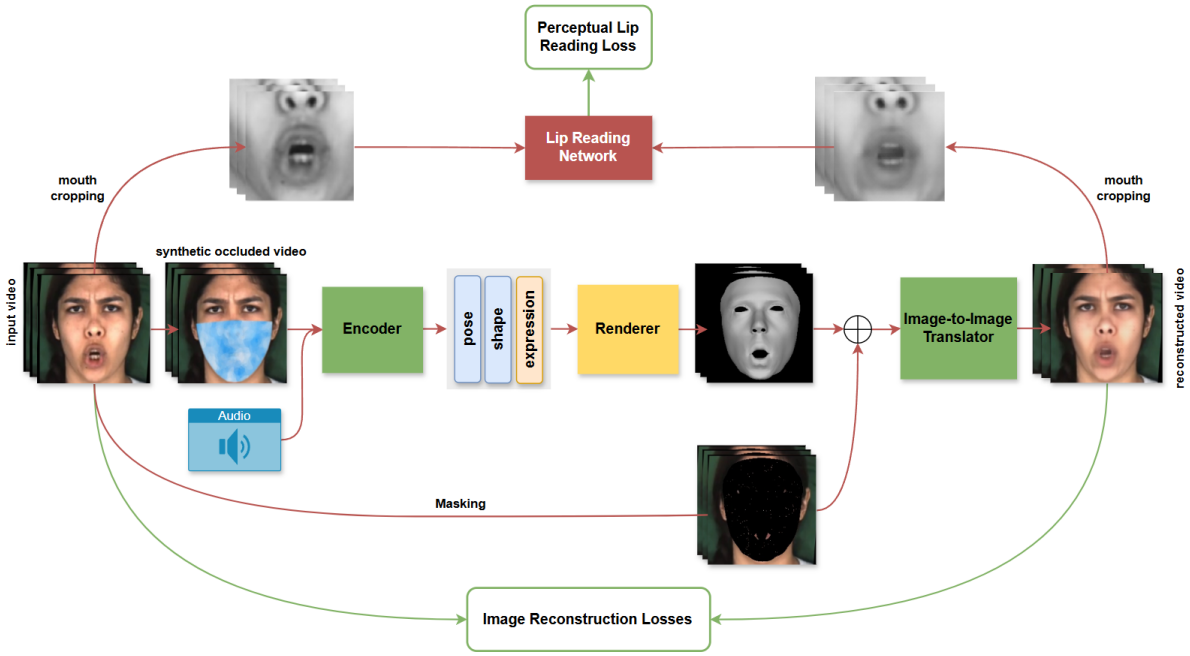

Figure 5.5.1: Overall Architecture of our model during training. An input image is passed to the encoder which regresses FLAME and camera parameters. A 3D shape is reconstructed, rendered with a differentiable rasterizer and finally translated into the output domain with the image translation network. Then, standard self-supervised landmark, photometric and perceptual losses are computed.

5.6 Loss Functions

To train our model effectively, we combine several loss functions, each addressing a different aspect of the reconstruction process. Below, we describe each component in detail.

Photometric loss. We compute the L1 reconstruction error between the original frame I and the fused frame I':

$$\mathcal{L}_{\text{photo}} = \|I' - I\|_1.$$

This loss enforces pixel-level consistency and ensures that the fused frame remains close to the ground truth in terms of low-level appearance.

VGG perceptual loss. To encourage high-level perceptual similarity and accelerate convergence during the initial training phases, we adopt a perceptual loss using the VGG network [32]:

$$\mathcal{L}_{\text{vgg}} = \|\Gamma(I') - \Gamma(I)\|_1,$$

where $\Gamma(\cdot)$ denotes the feature embeddings extracted from a pretrained VGG encoder. This loss emphasizes structural and textural details that are perceptually important but may not be captured by pixel-wise differences.

Landmark loss. To ensure geometric consistency in facial structure, we minimize the squared L_2 distance between the ground-truth 2D facial landmarks k and the projected landmarks k' from the fused frame:

$$\mathcal{L}_{\text{lmk}} = \sum_{i=1}^{N} \|k_i - k'_i\|_2^2,$$

where N denotes the total number of landmarks. This loss guides the model to preserve accurate facial geometry, especially around key regions such as the eyes, nose, and mouth.

Lip-reading loss. Inspired by the speech-informed training strategy of SPECTRE [18], we integrate a lip-reading loss to capture the temporal dynamics of lip motion during speech. Both the ground-truth and fused videos are cropped around the mouth region, and the corresponding feature vectors ϵ_I and ϵ_R are extracted from a pretrained lip-reading model for each frame. The similarity between the two feature sequences is quantified using the cosine distance:

$$\mathcal{L}_{\text{lr}} = \frac{1}{K} \sum_{k=1}^{K} d(\epsilon_{I_k}, \epsilon_{R_k}),$$

where $d(\cdot)$ denotes the cosine distance and K is the number of frames in the sequence. This loss ensures that lip movements in the fused video remain synchronized with the spoken content.

Expression regularization. To avoid exaggerated or unrealistic facial deformations, we impose an L_2 penalty on the expression parameters:

$$\mathcal{L}_{\text{reg}} = \|\psi\|_2^2,$$

which encourages smoother and more natural expressions.

5.7 Experiment Setup

Before training the full audiovisual model, we pretrain all three encoders (pose, shape, and expression) using only the visual modality. This stage is supervised primarily with landmark-based reconstruction losses for pose and expression, combined with expression regularization to prevent extreme deviations from neutral expressions. Additionally, we introduce a temporal smoothing loss on pose and expression parameters to enforce consistency across consecutive frames. The shape encoder is further supervised with predictions from MICA [79], after which the pose and shape encoders are kept frozen.

During main training, both visual and audio modalities are used. The model is optimized with a weighted combination of all aforementioned losses in the previous section. For simplicity, we incorporate the weighting coefficients directly into the loss terms, and thus express the overall loss as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{photo}} + \mathcal{L}_{\text{vgg}} + \mathcal{L}_{\text{lmk}} + \mathcal{L}_{\text{lr}} + \mathcal{L}_{\text{reg}}.$$

The training framework, including inputs, outputs, and the loss functions, is illustrated in Figure 5.5.1.

We adopt the Adam optimizer with an initial learning rate of 10^{-3}, a batch size of 6, and video clips of 20 frames. All experiments are conducted on an NVIDIA L40S GPU with 46 GB of memory. The model is trained for 10 epochs, which takes approximately 1.5 days to complete.

Chapter 6

Experiments

Contents

6.1 Datasets

For training our model, we use four diverse datasets: MEAD, LRS3, CelebV-Text, and ViCo Listeners. Each dataset provides complementary attributes, covering a wide range of speaking styles, visual conditions, and linguistic contexts. However, a common limitation across all of them is the scarcity of natural occlusions, which are important for evaluating model robustness in realistic scenarios. To mitigate this issue, we introduce synthetic occlusions across our training samples. The statistics of each dataset are summarized in Table 6.1.

6.1.1 MEAD

MEAD [71] is a large-scale emotional audio-visual dataset that we use to enhance the diversity of our data in terms of facial expressions. It comprises 60 actors and actresses from multiple racial backgrounds, recorded under controlled conditions with different camera angles. Each subject speaks with eight distinct emotions, each rendered at three levels of intensity, across seven view angles. The dataset consists of 281,400 video clips of high audio-visual fidelity, providing clear facial detail and consistent lighting, which is useful for learning 3D facial geometry and expression modeling.

6.1.2 LRS3

While MEAD enables controlled evaluation, we also incorporate LRS3 [1] to ensure that our experiments are not limited to laboratory settings but also cover in-the-wild conditions. LRS3 is a large-scale dataset for visual speech recognition, constructed from TED and TEDx talks collected from YouTube. It contains over 400 hours of video extracted from 5,594 talks in English, providing 151,819 video clips. The dataset offers cropped face tracks (224×224, 25 fps), single-channel 16-bit audio at 16 kHz, and aligned text transcripts with word-level boundaries. It is divided into three subsets: Pre-train, TrainVal, and Test. In our experiments, we use the TrainVal split, which contains 31,982 video clips, for model training and validation, and the Test set, comprising 1,321 clips, to evaluate the performance of our approach.

6.1.3 CelebV-text

To improve our model's robustness to natural occlusions, we incorporated a dataset containing a plethora of videos in which the face is partially or fully obscured. For this purpose, we use CelebV-Text [75]. This large-scale, in-the-wild dataset comprises approximately 70,000 face video clips with diverse visual content. Its scale and diversity make it suitable for training and evaluating 3D face reconstruction models, as it provides a wide range of subjects, poses, and expressions under unconstrained conditions.

6.1.4 ViCo Listeners

A combination of ViCo [77] and ESC-50 dataset is also used for training. ViCo is so far the first video conversation dataset containing face-to-face dialogue video clips in various scenarios. It involves 92 unique identities, comprising 67 speakers and 76 listeners across 483 paired speaking–listening clips. Each listener is recorded displaying non-verbal feedback behaviors such as nodding and smiling, reflecting three distinct listening attitudes: positive, neutral, and negative. In this work, we use only the listener recordings, as they directly provide the visual and behavioral information necessary for our task. We further replace their original audio with randomly selected

samples from the ESC-50 dataset [55], which contains 2,000 environmental audio clips across 50 semantic categories, including urban, domestic, natural, animal, and ambient noises. By pairing listener videos with non-speech audio, we augment our dataset with scenarios where human faces are visible but no speech signal is present. This design choice prevents the model from learning spurious correlations and discourages the generation of lip movements in the absence of speech.

6.1.5 Synthetic Occlusions Dataset

Finally, we apply synthetic occlusions only to the training and validation datasets, as shown in Fig.5.3.1 and Fig.5.3.2. These occlusions include dynamic events, where objects or hands move smoothly across the face, and static events, where surgical masks are synthesized using facial landmarks. This augmentation exposes the model to realistic occlusion scenarios and improves its robustness under challenging visual conditions.

Figure 6.1.1: Example frames from the datasets employed in this thesis. The first row presents subjects from the MEAD dataset recorded in lab conditions. The second row contains in-the-wild samples from LRS3. The third row depicts identities under occlusions from the CelebV-Text dataset, while the fourth row consists of frames from the ViCo listeners dataset.

Dataset	Videos Clips	Hours	Resolution	Environment
MEAD [71]	24,436	∼40	1920 × 1080	lad-conditioned
LRS3 (train-val) [1]	31,982	∼400	224 × 224	In-the-wild
CelebV-Text [75]	65,261	∼279	512 × 512	In-the-wild
ViCo [77]	483	∼95	1920 × 1080	In-the-wild

Table 6.1: Dataset statistics for training.

6.2 Evaluation

For evaluation, we curated 98 representative samples from the CelebV-Text dataset together with two self-recorded videos. This process involved a careful selection from the available data to ensure that the constructed evaluation set accurately represents the challenges targeted in our work. To assess the generalization ability of our model to unseen data, we further conducted experiments on the recently introduced CelebV-HQ dataset [78]. CelebV-HQ is a large-scale, high-quality, and diverse video dataset consisting of 35,666 videos, including 15,653 identities which was not included during training. From it, we curated a subset of samples that specifically reflect our problem setting, with a particular focus on both extreme and mild occlusions.

For our evaluation, we use transcriptions of the spoken content in the videos. To obtain these, we employ OpenAI Whisper [57], a large-scale automatic speech recognition (ASR) model trained on 680,000 hours of multilingual and multitask supervised data collected from the web. Whisper has demonstrated strong robustness across diverse acoustic and recording conditions, effectively handling background noise, speaker accents, and varying environments. In our experiments, Whisper is used to generate text transcriptions of the evaluation videos, which allows us to verify the alignment between spoken content and corresponding lip movements.

While standard benchmarks exist for quantitatively evaluating 3D face shape reconstruction, no such benchmark is available for assessing the accuracy of reconstructed facial expressions. Quantitatively measuring the difference between a reconstructed 3D expression and a ground-truth scan is less meaningful, as the resulting errors are often dominated by inaccuracies in the reconstructed identity shape. Furthermore, a low geometric error does not necessarily reflect perceptual accuracy, as subtle expression differences can strongly affect human interpretation. Therefore, we rely primarily on qualitative and perceptual evaluations, complemented by a user study, to better assess the realism and expressiveness of the reconstructed faces. Finally, we compare the results of our method with recent state-of-the-art approaches to evaluate its overall effectiveness and robustness.

6.2.1 Compared Methods

This thesis addresses 3D face generation driven by audio and image input, a field of study where only a few works exist and, to the best of our knowledge, no recent methods with publicly available implementations are available. Approaches such as AVFace [11] explore audiovisual input but no implementation has been released, making direct comparison challenging.

For this reason, we evaluate our method against state-of-the-art models that provide publicly available implementations and reconstruct 3D face geometry from videos. In particular, we consider **SPECTRE** [18] and **SMIRK** [59], both of which employ the FLAME model as our work. We focus on visual-based models because our goal is to address challenging real in-the-wild senarios while still preserving the subject's key facial attributes. This comparison highlights the contribution of combining audio and visual information, leading to more robustface reconstruction results. Both the above models are analyzed in Chapter 4.

6.2.2 Qualitative Evaluation

We first present qualitative results from our model and provide comparisons to demonstrate its effectiveness in challenging conditions. Given the nature of the task, visual evaluation provides the most direct and intuitive mean of assessment, as the advantages of different methods can be directly observed in the generated outputs.

In Figure 6.2.1, we illustrate three representative occlusion scenarios: object, hand, and total mouth mask occlusions (from left to right). These are natural occlusions from real-world data, matching the types we synthetically generated during training, thereby demonstrating the model's generalization ability.

- In the first scenario of Figure 6.2.1, we evaluate cases where the mouth region is partially occluded by an object, such as a microphone. This type of occlusion produces a persistent but incomplete obstruction: certain portions of the lips remain visible, while other regions are hidden throughout the sequence. Such conditions are highly representative of everyday recording environments, including live performances, podcast recordings, and television interviews, where microphones frequently cover part of the speaker's face. Under these settings, the model demonstrates its to reconstruct realistic articulations by effectively fusing limited visual cues with audio input. The results show that even when the lower lip is intermittently invisible, the generated reconstructions preserve the timing and dynamics of speech articulation.

- The second scenario involves dynamic hand occlusions, where a speaker's hand temporarily passes in front of the mouth. In contrast to the first case, the occlusion change location across frames, with different parts of the face becoming hidden at different times. Such occlusions are highly relevant for natural human communication, where gestures and spontaneous movements frequently obscure the face during conversation. Our model remains robust under these conditions, producing plausible lip synchronization even when visual cues are absent. Importantly, the outputs show a smooth temporal transition as the occlusion appears and disappears, without introducing flickering or abrupt deformations. This demonstrates that the model has learned to rely on audio when visual information does not give information for the face.

- The third and most challenging scenario considers full surgical mask occlusions, where the entire mouth region is covered for the duration of the sequence. Unlike the previous two cases, no visual cues are available at any point, forcing the model to rely entirely on the audio stream to drive articulation. This scenario has gained particular importance in recent years due to the widespread use of masks in public spaces, making it both realistic and socially relevant. The results demonstrate that our model is capable of producing plausible and temporally consistent lip movements even under such extreme occlusions. For instance, we observe accurate reconstruction of key phonetic events, such as the wide mouth opening required for the articulation of vowels like "I". Although more subtle articulations, such as closed-mouth consonants, remain difficult to reproduce, the outputs remain coherent, demonstrating that the model is more robust.

Overall, this qualitative evaluation provides strong evidence that our method achieves realistic, expressive, and robust reconstructions in challenging in-the-wild settings. While limitations remain in reproducing fine-grained articulations under complete occlusion, the results suggest that multimodal approaches offer a substantial advantage over visual-only baselines.

Figure 6.2.1: Qualitative examples of our model under different conditions. From left to right, the input audio is: a Portuguese song, French speech, and the English phrase *"I woke up"*.For visualization diversity, we display every third frame instead of continuous sequences.

A key contribution of our model is its robustness to surrounding sounds. In Figure 6.2.2, we present random selected frames from a video clip in which the main character remains silent while background speech is present. In this example, Whisper produces the transcription: "Ada, he ordered your favorite. He got Hawaiian. I saw the order." The model correctly relies on the visual information and does not generate mouth movements for the silent character.

Figure 6.2.2: Visualization of Listener sample frames from ViCo dataset. In this video, the main character does not speak, while background noise is present. Our model correctly generates an avatar that does not move its mouth, since the background noise does not correspond to the main character's voice.

After highlighting the benefits of our model, it is important to evaluate its performance in comparison with competing methods. To this end, we visually compare reconstructions obtained by our approach against the competitors on samples from an unseen dataset. Figure 6.2.3 illustrates the input frames alongside the corresponding reconstructions from each model. Our observations are summarized below:

- **SPECTRE:** Captures the overall head shape and pose consistently, but tends to smooth out fine details of facial expressions. In particular, the lip region is often not accurately reproduced, which reduces the naturalness of the reconstructed mesh. The results are stable but sometimes lack expressiveness compared to the input.
- **SMIRK:** Reproduces certain dynamic expressions better than SPECTRE, but often fails to maintain faithful geometry in the mouth region. Since lips are central to perceived articulation, this limitation leads to noticeable mismatches between input and reconstruction. These artifacts, especially around the mouth area, further reduce the realism of the generated meshes and overall performance.
- **FAVOR (our):** Produces reconstructions that remain more faithful to the input frames, preserving both identity and expressions. The correspondence in the mouth and lip regions is particularly stronger, leading to more realistic and intuitive facial geometry. Compared to SPECTRE and SMIRK, our approach aligns identity and expression more effectively while avoiding artifacts.

We further present examples across time to demonstrate the temporal consistency of reconstructed expressions. In particular, Figure 6.2.4 illustrates the weaknesses of competing methods in handling challenging occlusions. SPECTRE struggles to correctly interpret the occluded mouth

Figure 6.2.3: Visual comparison of 3D face reconstruction on unseen data for selected frames of video sequences. From left to right: Input, SPECTRE, SMIRK, ours.

region: instead of articulating realistic lip movements, it misinterprets the round shape of the microphone as part of the mouth and generates an incorrect, "sad" expression. SMIRK, while more expressive overall, fails to accurately capture the fine-grained details of lip motion. In contrast, FAVOR produces results that are closest to the ground truth, showing plausible mouth opening in alignment with the spoken content.

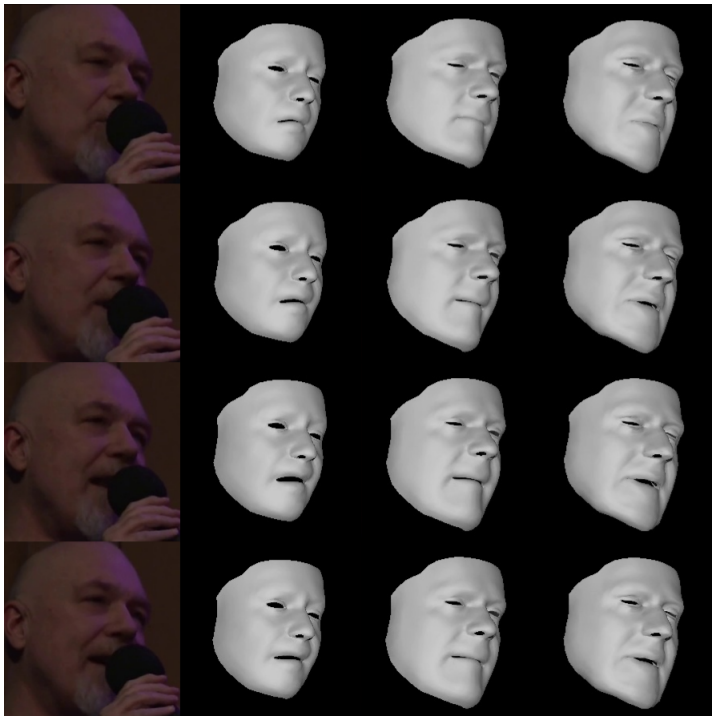

Figure 6.2.4: Visual comparison of 3D face reconstruction using SPECTRE, SMIRK and ours from left to right.

Furthermore, Figure 6.2.5 presents a sequence of frames in which the speaker's hand partially occludes the mouth region. Both SMIRK and SPECTRE exhibit noticeable artifacts when the occlusion reaches the peak of the event, producing facial expressions that deviate from the ground truth. In particular, SPECTRE tends to exaggerate the deformation around the mouth, generating expressions that appear unnatural and inconsistent with the input video. By contrast, FAVOR demonstrates greater robustness to this event, maintaining consistent and realistic facial reconstructions that align more closely with the correspoding speech content.

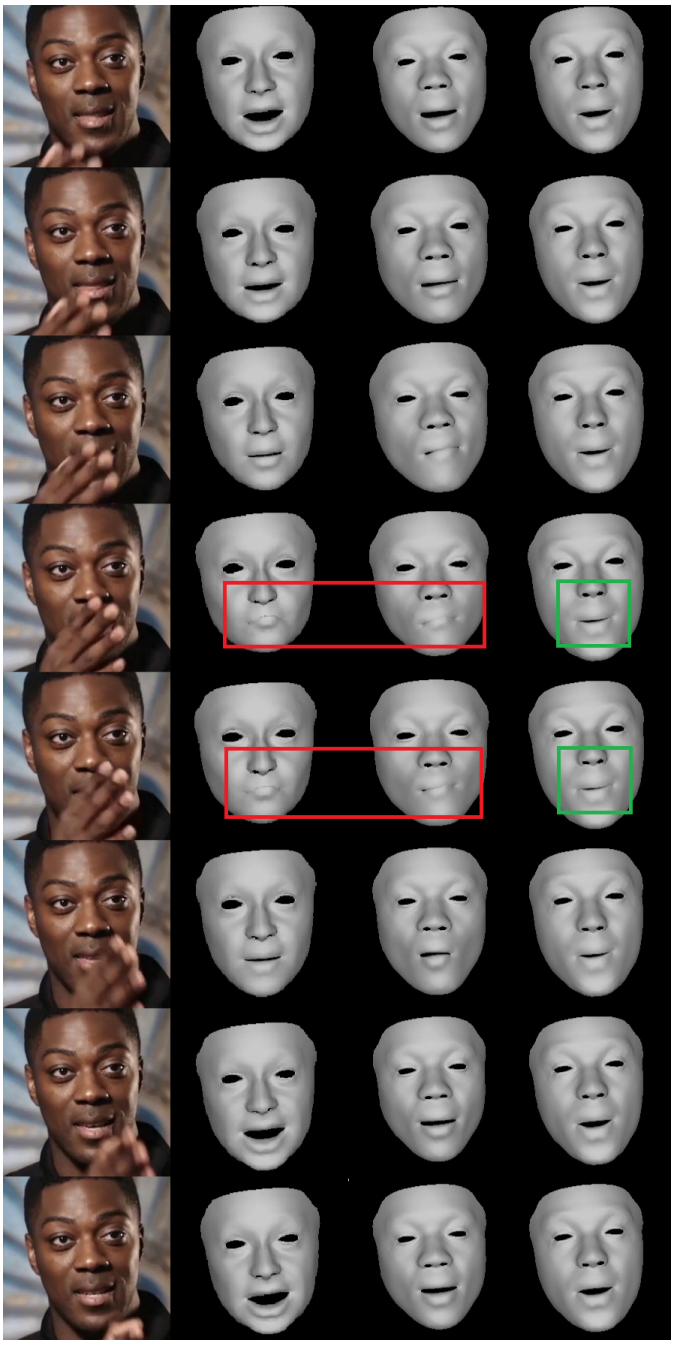

Figure 6.2.5: Visual results from SPECTRE, SMIRK and ours method.

6.2.3 Quantitative Evaluation

Quantitative evaluation of audiovisual 3D face reconstruction under occlusion is challenging due to the absence of large-scale publicly available datasets providing paired ground-truth 3D geometry and occluded visual data. Moreover, geometric errors are often dominated by identity-related inaccuracies and do not necessarily correspond to perceptual quality. To obtain a more meaningful measure of reconstruction accuracy, we incorporate a lip-aware perceptual loss that better captures the perceptual quality and expressiveness of the reconstructed faces.

6.2.3.1 Lip-Aware Perspective Loss

In this section, we focus on quantitative metrics to assess the performance of our model. Evaluating reconstructed face meshes requires reliable measures of facial reconstruction quality. Landmark-based evaluation, while commonly used, is not suitable in our case, as occluded videos often lack valid ground-truth landmarks for assessing the generated face mesh. To overcome this limitation and to capture the contribution of the audio modality, particularly in the mouth region, we follow SPECTRE [18] evaluation process and adopt a perceptual evaluation based on lipreading metrics. Particularly, we use

- Character Error Rate (CER)
- Word Error Rate (WER)
- Viseme Error Rate (VER)
- Viseme-Word Error Rate (VWER),

after mapping the predicted and ground-truth transcriptions into visemes using the Amazon Polly phoneme-to-viseme mapping [2]. We present the performance of these metrics for all compared methods on both the LRS3 test set and a selected subset of CelebV-HQ in Table 6.2. We observe that our method achieves lower error rates across most metrics, with particularly strong improvements on CelebV-HQ which contains many in-the-wild cases with frequent occlusions. This indicates that our approach generalizes better to challenging in-the-wild cases and produces more faithful articulations. On LRS3, the differences are smaller, which can be explained by the fact that this dataset is closer to the training domain of the competing models and contains fewer occlusion scenarios. We point out that the objective evaluation results on CER and WER, remain much higher compared to the original footage. This can be attributed to the different domains of the rendered images compared to the ground truth, as well as the absence of teeth and tongue, which are important for detecting specific types of phonemes/visemes.

	CELEBV-HQ				LRS3			
	CER	WER	VER	VWER	CER	WER	VER	VWER
SMIRK [59]	108.9	149.6	112.2	140.6	126.7	157.4	121.8	153.5
SPECTRE [18]	115.8	156.5	125.4	150.2	**116.9**	147.3	**112.2**	145.5
Ours	**89.0**	**118.3**	**82.6**	**116.9**	118.1	**146.5**	114.0	**143.9**

Table 6.2: Lipreading results on the LRS3 dataset and on 30 sample videos from CELEBV-HQ.

6.2.3.2 User Study

Perceptual evaluation have an important role in 3D face reconstruction, since the goal is to generate faces that humans perceive as natural and faithful. To complement the quantitative evaluation, we therefore designed a user study to assess the perceived quality of the reconstructed faces from human participants. For this purpose, 20 videos were randomly selected from the CelebV-HQ evaluation dataset and reconstructed using our method as well as SMIRK and SPECTRE. In each trial, participants were shown two pairs of video clips: one consisting of the original clip alongside our reconstruction, and the other consisting of the original video alongside a reconstruction from a competing method as shown in Figure 6.2.6. They were asked to select the one they considered to be the closest to the ground-truth appearance. A total of 46 users took part in the study and each of them answered 20 questions. The collected responses, summarized in Table 6.3, indicate a consistent preference for FAVOR, confirming the advantage of our method.

	SMIRK	SPECTRE
Ours	**288**/106	**247**/157

Table 6.3: User study preferences for each method. The values correspond to the number of times our method was selected over the competitor method. In both comparisons, participants consistently selected our method, demonstrating its superior perceptual quality.

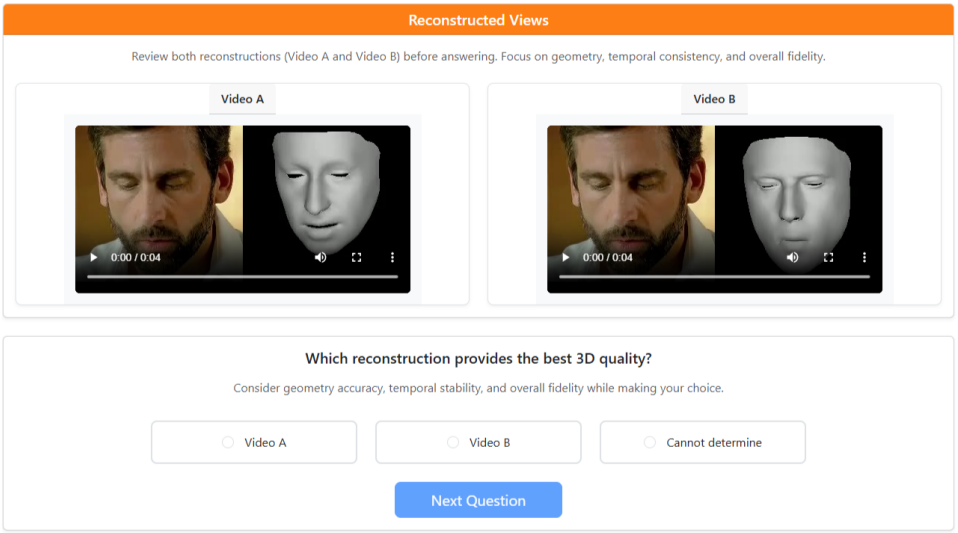

Figure 6.2.6: Instance of the user study.

We selected a subsample of five evaluation videos without any occlusions and reported the user preferences in Table 6.4. Our method outperformed SMIRK, indicating that incorporating audio features increased the overall accuracy of the model, while its performance was very close to that of SPECTRE. We attribute this result to the tendency of SPECTRE to exaggerate facial expressions, which often appear more plausible to human perception.

	SMIRK	SPECTRE
Ours	**33**/19	33/**37**

Table 6.4: User study preferences for the non-occluded samples.

6.3 Ablation Study

To present the contribution of each proposed component, we conduct an extended ablation study. We follow a step-by-step protocol: starting from SMIRK as the baseline model, we progressively add modules to build up to the full system. This design isolates the effect of each componet and clarifies its contribution to the overall performance.

6.3.1 With vs. Without occlusions

We first evaluate the impact of synthetic occlusion augmentation. Starting from SMIRK, which takes as input video frames and we train it with synthetic occlusions. We then compare the performance of SMIRK and a pretrained version of our model to examine how the augmented dataset affects reconstruction performance after 10 epochs. At this stage we are not concerned with audio-visual synchronization, since no audio information is used.

Figure 6.3.1 illustrates an example where SMIRK produces tweaked expressions, resulting in noticeable misalignments between the expected expressions and the reconstructed images. In contrast, training with the proposed occlusion-augmented dataset prevents the model from generating artifacts.

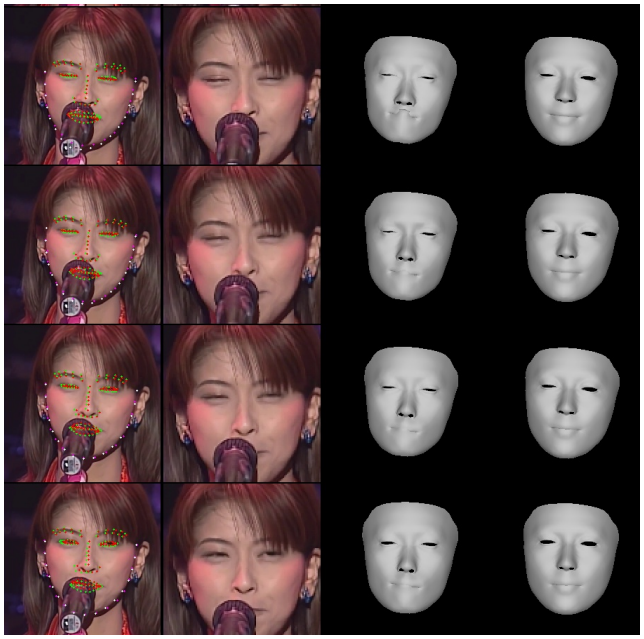

Figure 6.3.1: Qualitative comparison of SMIRK and our pretrain model. From left to right: the input frame with its corresponding landmarks, a cropped image for closer observation, the output of SMIRK, and the output of our pretrained model.

A more representative example confirming the sensitivity of SMIRK to occlusions is shown in Figure 6.3.2, where it fails to capture the overall facial structure and produces erroneous deformations in several frames. By contrast, our occlusion-augmented dataset surpass such cases, predicting smoother expression parameters and generating more robust reconstructions.

Another important observation here, is the lack of mediapipe landmarks in the last input frame. Particularly, the first column presents the ground-truth MediaPipe landmarks in red and the predicted landmarks from our method in green, while ground-truth jawline FAN landmarks are

shown in white and produced FAN landmarks in magenta. In the final input frame we can see that ground-truth MediaPipe landmarks are missing, confirming that landmarks alone are not a sufficiently reliable loss for stable training.

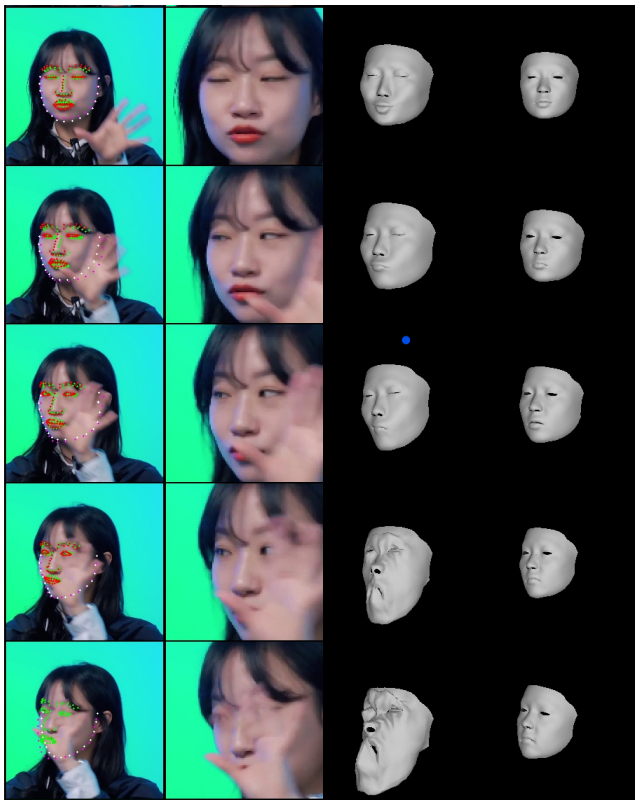

Figure 6.3.2: Examples of crucial artifacts generated by SMIRK. From left to right: input with landmarks, cropped close-up, SMIRK output, and our pretrained output.

This also highlights a limitation of our pretrained model. Since this version was trained using only landmark-based supervision, object or hand occlusions that partially cover the face reduce the available visual information. As a result, the model struggles in certain cases to reproduce natural eye closures when necessary as illustrated in Figure 6.3.3.

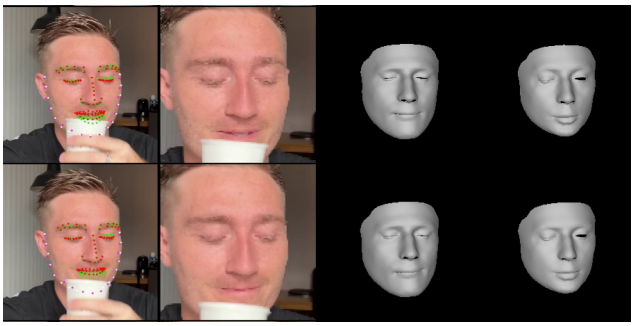

Figure 6.3.3: Visualization of the limitations of our pretrained model. From left to right: the input frame with its corresponding landmarks, a cropped image, SMIRK, our pretrained model.

6.3.2 With vs. Without Lipreading Loss on rendered image

Next, we incorporated audio information and introduced stronger supervision through the reconstruction path. In this setting, we conducted experiments to examine the impact of the lip-reading loss. Intuitively, one would expect that adding more supervision focused on the lips would lead to improved results. However, as shown in Figure 6.3.4, the model trained with the lip-reading loss instead generates artifacts which was also observed. The artifacts were found to occur primarily in cases where the mouth is closed. We attribute this behavior to the domain gap between the input space and the rendered image space.

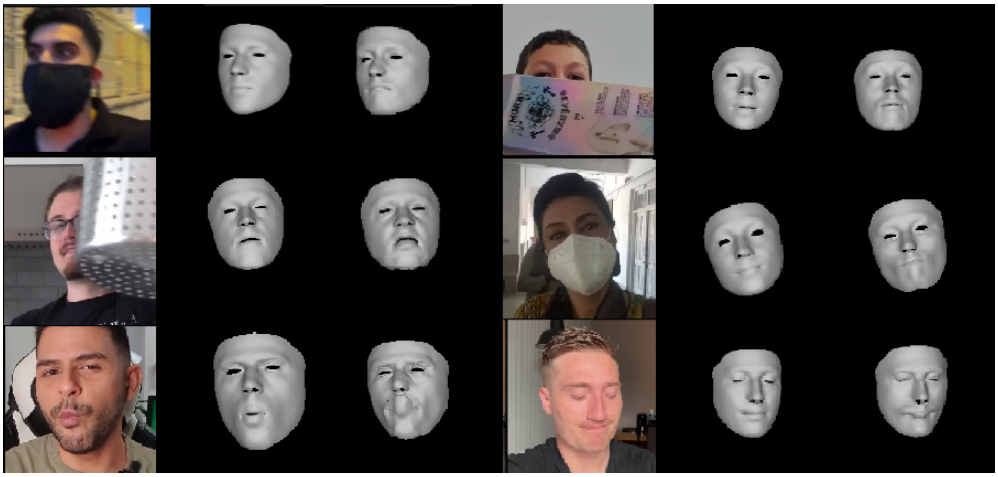

Figure 6.3.4: Artifacts arising from the use of lip-reading loss. Results are shown for our audio-visual model without and with the lip-reading loss from left to right.

6.3.3 Lipreading Loss on rendered vs. on fused image

To address the domain gap between the rendered frame and the input frame observed in the previous section, we apply the lip-reading loss between the input and the fused frame.

Figure 6.3.5: From left to right: Input, FAVOR, cropped mouth from input, cropped mouth from rendered frame, cropped mouth from input in greyscale, cropped mouth from fused frame.

To further mitigate artifacts, we introduce a threshold below which the model is not penalized. While moderate lip-reading loss helps reconstruct the mouth geometry accurately, excessively high values push the model to over-correct not only the geometry but also other aspects of the mouth region, leading to erroneous results. We also observed that training the model using only audio input tends to generate a large number of artifacts. Finally, it is important to note that our training dataset contains natural occlusions. In such cases, applying lip-reading loss can be disruptive, since the mouth is sometimes occluded. Figure 6.3.6 shows that this new pipeline produces artifact-free results in practice, generating plausible and consistent reconstructions.

Figure 6.3.6: Visual Comparison of reconstructions using lipreading loss on rendered vs. fused frames.

6.3.4 Visual vs Audio vs Audio-Visual model

We compare our model using as input different modalities audio-only, image-only, both. In Figure 6.3.7, on the left video frames, we observe that the audio-only setting fails to accurately capture the avatar's expressions. This behavior is expected, since the model receives no visual information in this case. In contrast, both the image-only and the audio-visual settings generate realistic mouth movements and accurate facial expressions. The image-only model shows a slight tendency to close the mouth on the occluded side by the microphone, but overall both image-only and audio-visual models produce plausible facial reconstructions.

Figure 6.3.7: Visualization of models driven by visual, audio, and audio-visual inputs from left to right.

Figure 6.3.8 further illustrates the effectiveness of the visual model in the presence of occlusions, such as when a hand partially hides facial areas. We attribute this robustness to the fact that the visual model was also trained with occlusion-augmented data and incorporates temporal information, allowing it to maintain consistency and avoid failure under such conditions.

We also present observations regarding the audio-driven model. In Figure 6.3.7 (right), we observe that its outputs remain relatively static over time. Even when the audio signal corresponds to wide mouth openings, the generated mesh fails to reproduce the expected variations. This behavior highlights a limitation of our audio-driven setting compared to the other modalities.

In particular, this case illustrates a fundamental difference in how the two modalities handle temporal alignment. Audio-driven synthesis allows temporal flexibility, since the mapping between phonemes and mouth shapes does not require strict frame-by-frame correspondence. For example, an open-mouth phoneme may be captured across neighboring frames and still be valid. In contrast, image-based reconstruction enforces rigid one-to-one alignment, where each output frame must precisely match the mouth shape in the corresponding input frame. This distinction explains why audio-only methods may experience temporal drift, leading to misleading supervision signals for the encoder. However, when both audio and visual modalities are used jointly,

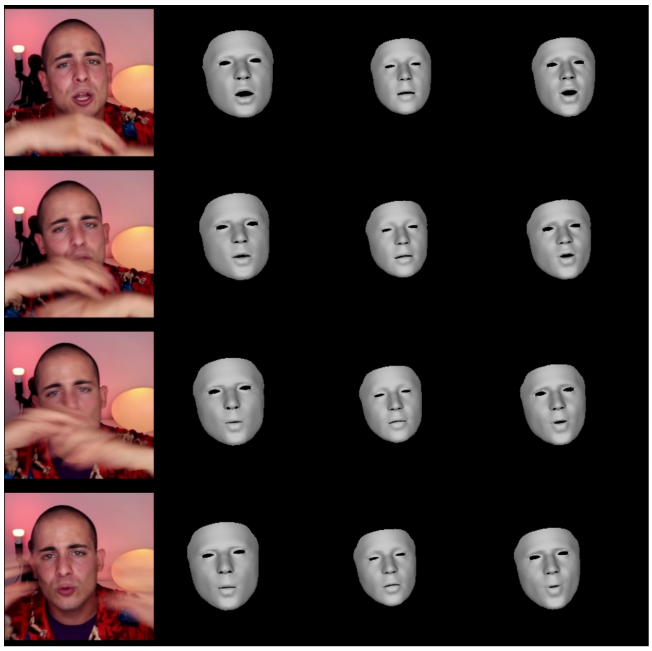

Figure 6.3.8: Visual comparison of 3D face reconstruction for visual, audio, audio-visual input. The audio-driven model (middle) tends to produce more static avatars.

the rendered avatar remains more faithful to the input sequence, as illustrated in Figure 6.3.9 (right).

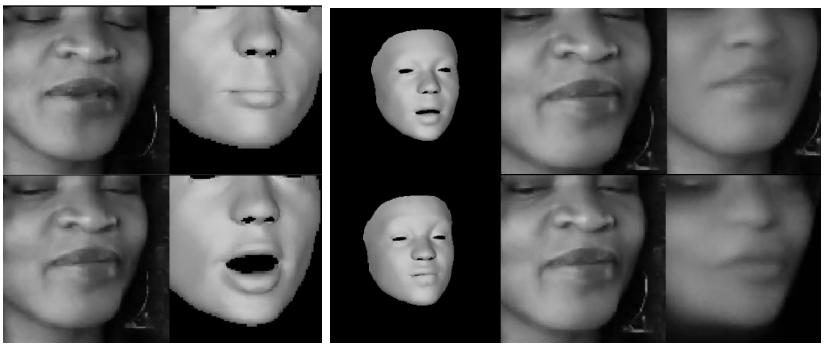

Figure 6.3.9: On the left, misaligned mouth movements between the input and the rendered frame, when using only audio input. On the right, exploiting both audio and visual input the model outputs more accurate mouth reconstruction.

Chapter 7

Conclusions and Future Work

Contents

7.1 Summary

In this thesis, we explored the field of 3D face reconstruction, with a focus on real-world scenarios and particularly on cases involving facial occlusions. We proposed a robust and flexible model that leverages audiovisual input to reconstruct realistic and temporally consistent 3D faces. This is a challenging task, as it requires achieving an effective balance between accurately reconstructing facial geometry under occlusions, while preserving the identity of the speaker.

We first conducted an extensive review of the related literature, which revealed that most existing methods are primarily focused on either image-driven or audio-driven models. To the best of our knowledge, very limited work has been carried out on audiovisual reconstruction architectures. Audiovisual face reconstruction, as the name suggests, combines information from both audio and vision modalities.

After presenting the relevant bibliography in Chapter 4, we introduced our proposed method, whose main contributions are summarized as follows:

- We propose FAVOR, a novel audiovisual model that reconstructs 3D facial geometry directly from video input. By jointly leveraging both visual and audio features, the model generates realistic and expressive talking faces while maintaining robustness under real-world conditions.
- FAVOR captures temporal dependencies across frame sequences to ensure coherent motion and stable facial geometry. This leads to smoother transitions and prevents frame-by-frame inconsistencies commonly observed in 3D reconstruction methods.
- The proposed framework supports audiovisual, visual-only and audio-only inputs making it suitable for a wide range of scenarios.
- To improve robustness, the training data were augmented with synthetic occlusions, simulating real-world conditions such as hands, objects, or masks covering parts of the face.
- The fusion of audio and visual cues enhances lip synchronization and reduces erroneous mouth movements, especially when the speech does not correspond to the visible speaker in the video.

We then conducted extensive experiments and compared FAVOR with recent state-of-the-art methods. The qualitative and quantitative evaluations demonstrated that our model achieves superior performance, particularly in terms of lip articulation on datasets with facial occlusions. Furthermore, the ablation study provided insight into the development pipeline and included visualizations that highlighted the contribution of each component to the final model. Finally, the user study further validated the performance of the proposed approach, confirming its ability to produce more natural and expressive reconstructions compared to existing methods.

7.2 Limitations and Future Work

While our method demonstrates strong performance across diverse scenarios, there are still some limitations in our approach that can guide future research. We briefly outline the key challenges identified in our experiments together with potential directions for improvement.

- **Misalignment on mouth movements using the audio-driven model:** As shown in Figure 6.3.9, our audio-driven model occasionally suffers from misalignment between the

original and the rendered frame. This reduces the effectiveness of the audio modality, particularly in the mouth region. A potential solution is to leverage the lipreading network outputs across consecutive rendered frames and enforce correspondence with the closest original frame, thereby achieving more consistent alignment.

- **Incorrect penalization under natural occlusions in the reconstruction path:** A key limitation of the reconstruction path arises when input frames contain natural occlusions. Since the fused generator relies on both geometry and sampled input pixels, it produces a photorealistic image that may attempt to reconstruct regions hidden by occlusions. So, when the loss is computed against the natural occluded input frame, the model is penalized even though it has reasonably reconstructed the missing areas. As illustrated in Figure 7.2.1, the fused output attempts to recover the mouth region, while the original frame is partially occluded by a mask. Such mismatch introduces misleading supervision during training. This can be handled with incorporating an occlusion-aware mechanism, such as a pretrained occlusion detector, to exclude heavily hidden regions from the reconstruction loss.

Figure 7.2.1: From left to right: Input frame, our model's result, cropped frame in mouth area, cropped fused image.

- **Robustness under extreme occlusions:** We also observed that when the entire mouth region is occluded, our method struggles to generate well-articulated mouth movements. Improving robustness under such extreme cases remains an important direction for future research. A promising approach is the use of modality dropout during training, which could encourage the model to generalize better and reconstruct effectively the face even when one modality is missing. Furthermore, integrating attention mechanisms may allow the network to dynamically adjust the relative contribution of audio and visual input, focusing on the most reliable modality for reconstruction.

7.3 Ethical Issues and Social Impact

In 3D face reconstruction research, it is crucial to account not only for the technical contributions but also for the wider ethical and societal implications. On the positive side, advances in 3D face reconstruction and neural rendering have the potential to benefit areas such as virtual reality, medical domain, and human-computer interaction. More robust and realistic reconstructions can enable new applications in content creation, assistive technologies, and personalized digital avatars.

At the same time, however, the ability to generate photo-realistic videos araises important ethical concerns. They could be misused to create manipulated content of individuals without their consent, for example in the form of deepfakes of public figures. Such misuse has implications for misinformation, privacy, and trust in digital media. So even though our method is designed with the goal of reconstructing detailed avatars under challenging conditions, it is important to recognize that the same technical advances could be repurposed in harmful ways.

In conclusion, while the work presented in this thesis contributes to creating a state-of-art audiovisual 3D face reconstruction model, its ethical and social dimensions are equally significant. Progress in this field should continue to be guided by principles of responsibility and transparency, ensuring that the positive applications of this technology can be realized while minimizing its potential for misuse.

Bibliography

[1] Afouras, T., Chung, J. S., and Zisserman, A., *LRS3-TED: A Large-Scale Dataset for Visual Speech Recognition*, 2018. arXiv: 1809.00496.

[2] *Amazon Polly. Developer Guide.* 2015.

[3] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D., "Vqa: Visual Question Answering," in *Proceedings the IEEE International Conference on Computer vision*, 2015, pp. 2425–2433.

[4] Baevski, A., Zhou, H., Mohamed, A., and Auli, M., "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proceedings the International Conference on Neural Information Processing Systems*, 2020, 1044:1–12.

[5] Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., and Grundmann, M., *Blazeface: Submillisecond Neural Face Detection on Mobile GPUs*, 2019. arXiv: 1907.05047.

[6] Blanz, V. and Vetter, T., "A Morphable Model For The Synthesis of 3D Faces," in *Proceedings the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1999, pp. 187–194.

[7] Booth, J., Roussos, A., Ponniah, A., Dunaway, D., and Zafeiriou, S., "Large Scale 3D Morphable Models," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 233–254, 2018.

[8] Bulat, A. and Tzimiropoulos, G., "How Far are We from Solving the 2D & 3D Face Alignment Problem? (and a dataset of 230,000 3d facial landmarks)," in *Proceedings the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1021–1030.

[9] Cao, C., Bradley, D., Zhou, K., and Beeler, T., "Real-Time High-Fidelity Facial Performance Capture," *ACM Transactions on Graphics (ToG)*, vol. 34, no. 4, pp. 1–9, 2015.

[10] Cao, C., Weng, Y., Zhou, S., Tong, Y., and Zhou, K., "Facewarehouse: A 3D Facial Expression Database for Visual Computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013.

[11] Chatziagapi, A. and Samaras, D., " AVFace: Towards Detailed Audio-Visual 4D Face Reconstruction," in *Proceedings the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16 878–16 889.

[12] Cudeiro, D., Bolkart, T., Laidlaw, C., Ranjan, A., and Black, M., "Capture, Learning, and Synthesis of 3D Speaking Styles," in *Proceedings the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 101–10 111.

[13] Danecek, R., Black, M. J., and Bolkart, T., "EMOCA: Emotion Driven Monocular Face Capture and Animation," in *Proceedings the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 20 311–20 322.

[14] Deng, J., Guo, J., Ververas, E., Kotsia, I., and Zafeiriou, S., "Retinaface: Single-Shot Multi-Level Face Localisation in the Wild," in *Proceedings the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5202–5211.

[15] Deng, J., Roussos, A., Chrysos, G., Ververas, E., Kotsia, I., Shen, J., and Zafeiriou, S., "The Menpo Benchmark for Multi-Pose 2D and 3D Facial Landmark Localisation and Tracking," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 599–624, 2019.

[16] Fan, Y., Lin, Z., Saito, J., Wang, W., and Komura, T., "Faceformer: Speech-Driven 3D Facial Animation with Transformers," in *Proceedings the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 749–18 758.

[17] Feng, Y., Feng, H., Black, M. J., and Bolkart, T., "Learning an Animatable Detailed 3D Face Model from In-The-Wild Images," *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, vol. 40, no. 8, pp. 1–13, 2021.

[18] 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

Proceedings the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023, pp. 5745–5755.

[19] Furukawa, Y. and Hernández, C., "Multi-view stereo: A tutorial," *Foundations and Trends in Computer Graphics and Vision*, vol. 9, no. 1-2, pp. 1–148, 2015.

[20] Gao, K., Gao, Y., He, H., Lu, D., Xu, L., and Li, J., *Nerf: Neural Radiance Field in 3D Vision, a Comprehensive Review*, 2022. arXiv: 2210.00379.

[21] Garrido, P., Valgaerts, L., Sarmadi, H., Steiner, I., Varanasi, K., Perez, P., and Theobalt, C., "Vdub: Modifying Face Video of Actors for Plausible Visual Alignment to a Dubbed Audio Track," in *Proceedings the Computer graphics forum*, vol. 34, 2015, pp. 193–204.

[22] Google, *Mediapipe Face Mesh*, Accessed: 2025-08-31.

[23] Hallgrímsson, B. et al., "Automated Syndrome Diagnosis by Three-Dimensional Facial Imaging," *Genetics in medicine*, vol. 22, no. 10, pp. 1682–1693, 2020.

[24] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al., *Deep Speech: Scaling Up End-to-End Speech Recognition*, 2014. arXiv: 1412.5567.

[25] He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition," in *Proceedings the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[26] Hochreiter, S. and Schmidhuber, J., "Long Short-Term Memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] Howard, A., Sandler, M., Chen, B., Wang, W., Chen, L.-C., Tan, M., Chu, G., Vasudevan, V., Zhu, Y., Pang, R., Adam, H., and Le, Q., "Searching for mobilenetv3," in *Proceedings the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1314–1324.

[28] Hu, L., Saito, S., Wei, L., Nagano, K., Seo, J., Fursund, J., Sadeghi, I., Sun, C., Chen, Y.-C., and Li, H., "Avatar Digitization from a Single Image for Real-Time Rendering," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 6, pp. 1–14, 2017.

[29] Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., and Wang, L., "Scaling Up Vision-Language Pre-Training for Image Captioning," in *Proceedings the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 980–17 989.

[30] Huang, S.-C., Pareek, A., Seyyedi, S., Banerjee, I., and Lungren, M. P., "Fusion of Medical Imaging and Electronic Health Records using Deep Learning: A Systematic Review and Implementation Guidelines," *NPJ digital medicine*, vol. 3, no. 1, 2020.

[31] Janiesch, C., Zschech, P., and Heinrich, K., "Machine Learning and Deep Learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021.

[32] Johnson, J., Alahi, A., and Fei-Fei, L., "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in *Proceedings the European Conference on Computer Vision*, 2016, pp. 694–711.

[33] Kartynnik, Y., Ablavatski, A., Grishchenko, I., and Grundmann, M., *Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs*, 2019.

[34] Kegkeroglou, N., Filntisis, P. P., and Maragos, P., "Medical Face Masks and Emotion Recognition from the Body: Insights from a Deep Learning Perspective," in *Proceedings the International Conference on PErvasive Technologies Related to Assistive Environments*, 2023, pp. 69–76.

[35] Koujan, M. R. and Roussos, A., "Combining Dense Nonrigid Structure from Motion and 3D Morphable Models for Monocular 4D Face Reconstruction," in *Proceedings the ACM SIGGRAPH European Conference on Visual Media Production*, 2018, pp. 1–9.

[36] Koutras, P., Retsinas, G., and Maragos, P. (2025). *Deep Learning for Computer Vision Applications* [Chapter]. In Maragos, P. 2025. *Topics in Computer Vision and Machine Learning* [Postgraduate textbook]. Kallipos, Open Academic Editions. https://hdl.handle.net/11419/15132

[37] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[38] LeCun, Y. and Bengio, Y., "Convolutional Networks for Images, Speech, and Time Series," *The Handbook of Brain Theory and Neural Networks*, 1998.

[39] LeCun, Y., Bengio, Y., and Hinton, G., "Deep Learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[40] Li, T., Bolkart, T., Black, M. J., Li, H., and Romero, J., "Learning a Model of Facial Shape and Expression from 4D Scans," *ACM Transactions on Graphics*, vol. 36, no. 6, 194:1–194:17, 2017.

[41] Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J., "SMPL: A Skinned Multi-person Linear Model," *ACM Transactions on Graphics (SIGGRAPH Asia)*, vol. 34, no. 6, 248:1–248:16, 2015.

[42] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., and Grundmann, M., "Mediapipe: A Framework for Building Perception Pipelines," in *Proceedings the Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, 2019.

[43] Ma, P., Petridis, S., and Pantic, M., "Visual Speech Recognition for Multiple Languages in the Wild," *Nature Machine Intelligence*, vol. 4, no. 11, pp. 930–939, 2022.

[44] Maglo, A., Courbet, C., Alliez, P., and Hudelot, C., "Progressive Compression of Manifold Polygon Meshes," *Computers & Graphics*, vol. 36, no. 5, pp. 349–359, 2012.

[45] Medsker, L. R. and Jain, L., "Recurrent Neural Networks," *Design and applications*, vol. 5, no. 64-67, p. 2, 2001.

[46] Mitsui, K., Hono, Y., and Sawada, K., *UniFLG: Unified Facial Landmark Generator from Text or Speech*, 2023. arXiv: 2302.14337.

[47] Newcombe, R. A., Izadi, S., Hilliges, O., Molyneaux, D., Kim, D., Davison, A. J., Kohi, P., Shotton, J., Hodges, S., and Fitzgibbon, A., "Kinectfusion: Real-Time Dense Surface Mapping and Tracking," in *IEEE International symposium on mixed and augmented reality*, 2011, pp. 127–136.

[48] O'Shea, K. and Nash, R., *An Introduction to Convolutional Neural Networks*, 2015. arXiv: 1511.08458.

[49] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S., "Librispeech: An asr corpus based on public domain audio books," in *Proceedings the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[50] Papantoniou, F. P., Filntisis, P. P., Maragos, P., and Roussos, A., "Neural Emotion Director: Speech-Preserving Semantic Control of Facial Expressions in" in-the-wild" Videos," in *Proceedings the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 781–18 790.

[51] Parke, F. I., "Computer generated animation of faces," *ACM Annual Conference*, vol. 1, pp. 451–457, 1972.

[52] Parke, F. I., "Parameterized models for facial animation," *IEEE computer graphics and applications*, vol. 2, no. 9, pp. 61–68, 1982.

[53] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T., "A 3D Face Model for Pose and Illumination Invariant Face Recognition," in *Proceedings the IEEE International conference on advanced video and signal based surveillance*, 2009, pp. 296–301.

[54] Perez, E., Strub, F., Vries, H. de, Dumoulin, V., and Courville, A., "Film: Visual Reasoning with a General Conditioning Layer," in *Proceedings the AAAI conference on artificial intelligence*, 2018, 483:1–10.

[55] Piczak, K. J., "ESC: Dataset for Environmental Sound Classification," in *Proceedings Annual ACM Conference on Multimedia*, 2015, pp. 1015–1018.

[56] Pietroni, N., Cignoni, P., Otaduy, M. A., and Scopigno, R., "A Survey on Solid Texture Synthesis," *IEEE Computer Graphics and Applications*, vol. 30, no. 4, pp. 74–89, 2010.

[57] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I., "Robust speech recognition via large-scale weak supervision," in *Proceedings the International Conference on Machine Learning*, 2023, 1182:1–27.

[58] Ren, Q., Yang, Z., Lu, Y., Pan, J., Li, Y., Guo, Y., Bi, M., Zhou, Y., Yang, H., Zhou, L., and Ji, F., "3D X-ray Microscope Acts as an Accurate and Effective Equipment of Pathological Diagnosis in Craniofacial Imaging," *Scientific Reports*, vol. 14, no. 1, 2024.

[59] Retsinas, G., Filntisis, P. P., Daněček, R., Abrevaya, V. F., Roussos, A., Bolkarr, T., and Maragos, P., "3D Facial Expressions through Analysis-by-Neural-Synthesis," in *Proceedings the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 2490–2501.

[60] Richard, A., Zollhöfer, M., Wen, Y., Torre, F. de la, and Sheikh, Y., "MeshTalk: 3D Face Animation From Speech Using Cross-Modality Disentanglement," in *Proceedings the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1173–1182.

[61] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional Networks for Biomedical Image Segmentation," in *Proceedings the International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

[62] Roussos, A., and Maragos, P. (2025). *Three-Dimensional Modeling of Deformable Objects* [Chapter]. In Maragos, P. 2025, *Topics in Computer Vision and Machine Learning* [Postgraduate textbook]. Kallipos, Open Academic Editions. https://hdl.handle.net/11419/15132

[63] Schmidhuber, J., "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85–117, 2015.

[64] Schonberger, J. L. and Frahm, J.-M., "Structure-from-motion revisited," in *Proceedings the IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2016, pp. 4104–4113.

[65] Seitz, S. M., Curless, B., Diebel, J., Scharstein, D., and Szeliski, R., "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," in *Proceedings the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2006, pp. 519–528.

[66] Shi, B., Hsu, W.-N., Lakhotia, K., and Mohamed, A., "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction," 2022.

[67] Thies, J., Elgharib, M., Tewari, A., Theobalt, C., and Nießner, M., "Neural Voice puppetry: Audio-Driven Facial Reenactment," in *Proceedings the European Conference on Computer Vision*, 2020, pp. 716–731.

[68] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M., "Face2face: Real-Time Face Capture and Reenactment of RGB Videos," in *Proceedings the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 2387–2395.

[69] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention is All you Need," vol. 30, 2017.

[70] Voo, K. T. R., Jiang, L., and Loy, C. C., "Delving into High-Quality Synthetic Face Occlusion Segmentation Datasets," in *Proceedings the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 4710–4719.

[71] Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., and Loy, C. C., "MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation," in *Proceedings the European Conference on Computer Vision (ECCV)*, 2020, pp. 700–717.

[72] Wang, X., Xie, Q., Zhu, J., Xie, L., and Scharenborg, O., "Anyonenet: Synchronized Speech and Talking Head Generation for Arbitrary Persons," *IEEE Transactions on Multimedia*, vol. 25, pp. 6717–6728, 2022.

[73] Xing, J., Xia, M., Zhang, Y., Cun, X., Wang, J., and Wong, T.-T., "CodeTalker: Speech-Driven 3D Facial Animation with Discrete Motion Prior," in *Proceedings the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 12 780–12 790.

[74] Yao, Y., Luo, Z., Li, S., Fang, T., and Quan, L., "MVSNet: Depth Inference for Unstructured Multi-view Stereo," in *Proceedings the European Conference on Computer Vision (ECCV)*, 2018, pp. 785–801.

[75] Yu, J., Zhu, H., Jiang, L., Loy, C. C., Cai, W., and Wu, W., "CelebV-Text: A Large-Scale Facial Text-Video Dataset," in *Proceedings the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[76] Zadeh, A., Chen, M., Poria, S., Cambria, E., and Morency, L.-P., "Tensor Fusion Network for Multimodal Sentiment Analysis," in *Proceedings the Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[77] Zhou, M., Bai, Y., Zhang, W., Yao, T., Zhao, T., and Mei, T., "Responsive Listening Head Generation: A Benchmark Dataset and Baseline," in *Proceedings the European Conference on Computer Vision (ECCV)*, 2022, pp. 124–142.

[78] Zhu, H., Wu, W., Zhu, W., Jiang, L., Tang, S., Zhang, L., Liu, Z., and Loy, C. C., "CelebV-HQ: A Large-Scale Video Facial Attributes Dataset," in *Proceedings the European Conference on Computer Vision (ECCV)*, 2022, pp. 650–667.

[79] Zielonka, W., Bolkart, T., and Thies, J., "Towards Metrical Reconstruction of Human Faces," in *Proceedings the European Conference on Computer Vision (ECCV)*, 2022, pp. 250–269.