



NATIONAL TECHNICAL UNIVERSITY OF  
ATHENS

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING  
DIVISION OF SIGNALS, CONTROL AND ROBOTICS

Deep Facial Expression Recognition Exploiting  
Categorical and Continuous Emotional Dependencies

DIPLOMA THESIS

of

**Panagiotis Antoniadis**

**Supervisor:** Petros Maragos  
Professor NTUA

COMPUTER VISION, SPEECH COMMUNICATION AND SIGNAL PROCESSING GROUP  
Athens, November 2021





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

# Deep Facial Expression Recognition Exploiting Categorical and Continuous Emotional Dependencies

## DIPLOMA THESIS

of

**Panagiotis Antoniadis**

**Supervisor:** Petros Maragos  
Professor NTUA

Approved by the examination committee on 5<sup>th</sup> November, 2021.

.....  
Petros Maragos  
Professor NTUA

.....  
Costas Tzafestas  
Associate Professor NTUA

.....  
Gerasimos Potamianos  
Associate Professor NTUA

Athens, November 2021

.....  
**PANAGIOTIS ANTONIADIS**  
Graduate of Electrical and  
Computer Engineering NTUA

Copyright © – All rights reserved Panagiotis Antoniadis, 2021.

The copying, storage and distribution of this diploma thesis, all or part of it, is prohibited for commercial purposes. Reprinting, storage and distribution for non-profit, educational or of a research nature is allowed, provided that the source is indicated and that this message is retained.

The content of this thesis does not necessarily reflect the views of the Department, the Supervisor, or the committee that approved it.

*to my family for their endless support*



# Abstract

Facial Expression Recognition (FER) has been a topic of study in the field of computer vision and machine learning for decades. Despite huge efforts to improve the accuracy of FER systems, existing methods still are not generalizable and accurate enough for use in real-world applications. Most traditional methods use hand-crafted features for representation of facial images that often require rigorous hyper-parameter tuning to achieve favorable results. Over the past few years, deep learning methods have shown remarkable results in FER managing to achieve almost human performance in lab-controlled environments. However, recognizing facial expressions in real-world settings is still very challenging due to large variations, occlusions and the ambiguity of human emotion. Meanwhile, we have no clear evidence as to which emotion representation is more appropriate for FER. Numerous models describing the human emotional states have been proposed by the psychology community and the majority of FER systems use either the categorical or the dimensional model of affect.

The goal of this diploma thesis is to explore the challenges that are present in the task and present novel deep learning techniques for recognizing facial expressions in-the-wild. The main contributions of our work can be summarized as follows:

1. We investigate the types of variations that are connected to the task and employ metric learning techniques to reduce their impact.
2. We explore the relation between the categorical and the dimensional emotion representation and train multi-task learning networks to exploit their dependencies.
3. Inspired by recent work in multi-label image recognition, we propose Emotion-GCN; a novel multi-task learning framework that uses a Graph Convolution Network (GCN) to capture the emotional dependencies.
4. We train and evaluate our proposed methods under real-world settings using AffectNet dataset, the largest in-the-wild database of facial expressions.

A part of our work has been accepted to the IEEE International Conference on Automatic Face and Gesture Recognition 2021 (FG 2021) with the authors being Panagiotis Antoniadis, Panagiotis Paraskevas Filntisis and Petros Maragos [AFM21]. Also, the same authors along with Ioannis Pikoulis participated in ICCV ABAW2 competition where we leveraged facial, bodily and context information to recognize emotion in real-world videos [Ant+21].

**Keywords** — Emotion Analysis, Facial Expression Recognition, Deep Neural Networks, Multi-task learning, Metric learning, Models of Affect





# Περίληψη

Η Αναγνώριση των Εκφράσεων του Προσώπου αποτελεί θέμα μελέτης στον τομέα της όρασης υπολογιστών και της μηχανικής μάθησης εδώ και δεκαετίες. Παρά την συνεχή προσπάθεια για τη βελτίωση της απόδοσης συστημάτων αναγνώρισης, οι υπάρχουσες μέθοδοι εξακολουθούν να μην γενικεύουν σε διαφορετικές συνθήκες και να μην είναι ακριβείς για χρήση σε πραγματικές εφαρμογές. Οι περισσότερες παραδοσιακές μέθοδοι χρησιμοποιούν χειροποίητα χαρακτηριστικά για την αναπαράσταση της εικόνας του προσώπου το οποίο συχνά απαιτεί διαρκή ρύθμιση των υπερπαραμέτρων για να επιτευχθούν ευνοϊκά αποτελέσματα. Τα τελευταία χρόνια, οι μέθοδοι βαθιάς μάθησης έχουν δείξει αξιοσημείωτα αποτελέσματα στην αναγνώριση των εκφράσεων του προσώπου πετυχαίνοντας επιδόσεις συγκρίσιμες με αυτές του ανθρώπου σε εργαστηριακά ελεγχόμενα περιβάλλοντα. Ωστόσο, η αναγνώριση των εκφράσεων του προσώπου σε πραγματικό περιβάλλον εξακολουθεί να είναι πολύ δύσκολη λόγω των μεγάλων διακυμάνσεων, των αποφράξεων και της ασάφειας των ανθρώπινων συναισθημάτων. Την ίδια στιγμή, δεν έχουμε σαφή στοιχεία για το ποια αναπαράσταση συναισθημάτων είναι πιο κατάλληλη για την αναγνώριση. Πολλά μοντέλα που περιγράφουν τις ανθρώπινες συναισθηματικές καταστάσεις έχουν προταθεί από την ψυχολογική κοινότητα και η πλειοψηφία των συστημάτων χρησιμοποιούν είτε το κατηγορικό είτε το διαστασιακό μοντέλο του συναισθήματος.

Ο σκοπός της παρούσας διπλωματικής εργασίας είναι να διερευνήσει τις προκλήσεις που υπάρχουν στον τομέα και να παρουσιάσει νέες τεχνικές βαθιάς μάθησης για την αναγνώριση των εκφράσεων του προσώπου σε πραγματικές συνθήκες. Οι κύριες συνεισφορές της εργασίας μας συνοψίζονται ως εξής:

1. Διερευνούμε τους τύπους διακυμάνσεων που συνδέονται με τον τομέα και χρησιμοποιούμε μετρική μάθηση για να μειώσουμε την επίδρασή τους.
2. Εξερευνούμε τη σχέση μεταξύ της κατηγορικής και της διαστασιακής αναπαράστασης συναισθημάτων και εκπαιδεύουμε δίκτυα μάθησης πολλαπλών εργασιών για να εκμεταλλευτούν τις εξαρτήσεις τους.
3. Εμπνευσμένοι από μία πρόσφατη εργασία στην αναγνώριση εικόνας με πολλές ετικέτες, προτείνουμε το Emotion-GCN, ένα νέο πλαίσιο εκμάθησης πολλαπλών εργασιών που χρησιμοποιεί συνελκτικά δίκτυα σε γράφους για να εκμεταλλευτεί τις συναισθηματικές εξαρτήσεις.
4. Εκπαιδεύουμε και αξιολογούμε όλες τις μεθόδους μας σε πραγματικές συνθήκες χρησιμοποιώντας το σύνολο δεδομένων AffectNet, τη μεγαλύτερη βάση δεδομένων εκφράσεων του προσώπου.

Ένα μέρος της εργασίας μας έχει δημοσιευθεί στο συνέδριο IEEE International Conference on Automatic Face and Gesture Recognition 2021 (FG 2021) με συγγραφείς τους Παναγιώτη Αντωνιάδη, Παναγιώτη Παρασκευά Φιλντίση και Πέτρο Μαραγκό [AFM21]. Επίσης, οι ίδιοι συγγραφείς μαζί με τον Ιωάννη Πίκουλη συμμετείχαν στον διαγωνισμό ICCV ABAW2 όπου αξιοποιήσαμε τις πληροφορίες προσώπου, σώματος και περιβάλλοντος για την αναγνώριση συναισθήματος σε βίντεο [Ant+21].

**Λέξεις κλειδιά** — Ανάλυση Συναισθήματος, Αναγνώριση Εκφράσεων του Προσώπου, Βαθιά Νευρωνικά Δίκτυα, Μάθηση Πολλαπλών Εργασιών, Μετρική Μάθηση, Μοντέλα Συναισθήματος



# Acknowledgements

First of all, I would like to thank Professor Petros Maragos for giving me the opportunity to conduct my diploma thesis in the Computer Vision, Speech Communication and Signal Processing Laboratory (CVSP). The courses of Computer Vision and Pattern Recognition taught by him acted as an early inspiration to me and increased my research interest in the domain of computer vision and machine learning. During my thesis he offered me valuable advice and guided me through my first steps as a researcher. Also, I would like to wholeheartedly thank the Ph.D. candidate Panagiotis Filntisis for our constructive collaboration in my thesis. He was constantly present to discuss my thoughts, answer any question and guide me to the right research directions. I am very excited that I started my first research steps and published my first scientific publication with him.

In addition, I thank my friends for our unforgettable experiences during our studies. Their support and encouragement was invaluable and without them my studies would not have been as memorable. Finally, I would like to thank my parents for the support and understanding they have shown me throughout my studies.

Panagiotis Antoniadis  
November 2021



# Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον καθηγητή Πέτρο Μαραγκό για την ευκαιρία που μου έδωσε να πραγματοποιήσω τη διπλωματική μου εργασία στο Εργαστήριο Όρασης Υπολογιστών, Επικοινωνίας Λόγου και Επεξεργασίας Σημάτων. Τα μαθήματα Όραση Υπολογιστών και Αναγνώριση Προτύπων που μου δίδαξε αποτέλεσαν την πρώτη πηγή έμπνευσης για μένα και αύξησαν το ερευνητικό μου ενδιαφέρον στον τομέα της όρασης υπολογιστών και της μηχανικής μάθησης. Κατά τη διάρκεια της διπλωματικής εργασίας μου προσέφερε πολύτιμες συμβουλές και με καθοδήγησε στα πρώτα μου βήματα ως ερευνητής. Επίσης, θα ήθελα να ευχαριστήσω ολόψυχα τον Διδακτορικό φοιτητή Παναγιώτη Φιλντίση για την εποικοδομητική συνεργασία μας στη διπλωματική μου. Ήταν συνεχώς παρών για να συζητήσει τις σκέψεις μου, να απαντήσει σε οποιαδήποτε ερώτηση και να με καθοδηγήσει στις σωστές ερευνητικές κατευθύνσεις. Είμαι πολύ ενθουσιασμένος που ξεκίνησα τα πρώτα μου ερευνητικά βήματα και δημοσίευσα την πρώτη ερευνητική μου εργασία μαζί του.

Επιπλέον, ευχαριστώ τους φίλους μου για τις αξέχαστες εμπειρίες μας κατά τη διάρκεια των σπουδών μας. Η υποστήριξη και η ενθάρρυνσή τους είναι ανεκτίμητη και χωρίς αυτούς οι σπουδές μου δεν θα ήταν τόσο αξέχαστες. Τέλος, ευχαριστώ τους γονείς μου για την υποστήριξη και την κατανόηση που μου έδειξαν σε όλη τη διάρκεια των σπουδών μου.

Παναγιώτης Αντωνιάδης  
Νοέμβριος 2021



# Εκτεταμένη Περίληψη

## Εισαγωγή

### Συναισθηματική Υπολογιστική

Τα συναισθήματα μπορούν να εκφραστούν μέσω λεκτικών ή μη λεκτικών σημάτων όπως οι εκφράσεις του προσώπου, ο τονισμός της φωνής και οι χειρονομίες. Τα περισσότερα συστήματα Αλληλεπίδρασης Ανθρώπου Υπολογιστή είναι ανεπαρκή στην σωστή ερμηνεία των συναισθηματικών σημάτων παρουσιάζοντας έλλειψη συναισθηματικής νοημοσύνης. Ο στόχος της Συναισθηματικής Υπολογιστικής είναι να αναπτύξει συστήματα που να μπορούν να αναγνωρίσουν, να ερμηνεύσουν, να επεξεργαστούν και να προσομοιώσουν τα συναισθήματα του ανθρώπου. Είναι ένας διεπιστημονικός τομέας που καλύπτει την επιστήμη των υπολογιστών, την ψυχολογία και τη γνωσιακή επιστήμη.

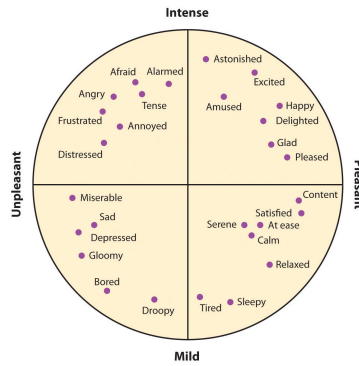
### Τρόποι Έκφρασης Συναισθήματος

Ο άνθρωπος εκφράζει την συναισθηματική του κατάσταση με τους εξής τρόπους:

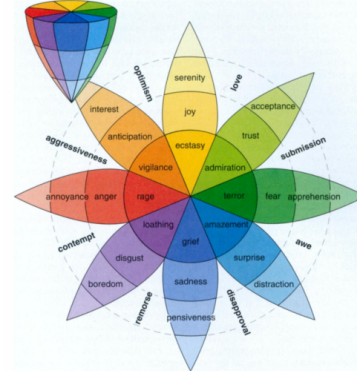
- 1. Εκφράσεις του προσώπου:** Ένας από τους πιο άμεσους τρόπους έκφρασης της συναισθηματικής κατάστασης του ανθρώπου. Ο Paul Ekman έδειξε ότι οι άνθρωποι αντιλαμβάνονται ορισμένα βασικά συναισθήματα με τον ίδιο τρόπο ανεξάρτητα από την κουλτούρα τους [EF71], [Ekm94]. Αυτές οι έξι καθολικές εκφράσεις του προσώπου (χαρά, θλίψη, έκπληξη, φόβος, αηδία και θυμός) αποτελούν το κατηγορικό μοντέλο.
- 2. Κινήσεις του σώματος:** Αρχικά, τα χέρια είναι η βασική πηγή πληροφοριών για τη γλώσσα του σώματος. Για παράδειγμα, η θέση των χεριών μπορεί κανείς να κρίνει αν ένα άτομο είναι ειλικρινές (θα γυρίσει τα χέρια προς τα μέσα) ή όχι (κρύβοντας τα χέρια πίσω από την πλάτη). Η κατεύθυνση του κεφαλιού αποκαλύπτει επίσης πολλές πληροφορίες σχετικά με τη συναισθηματική κατάσταση. Πολλές μελέτες δείχνουν ότι οι άνθρωποι είναι επιρρεπείς στο να μιλούν περισσότερο εάν ο ακροατής τους ενθαρρύνει με ένα νεύμα. Ο ρυθμός του νεύματος μπορεί να σηματοδοτήσει την υπομονή ή την έλλειψή αυτής. Τέλος, ο κορμός είναι ίσως το λιγότερο επικοινωνιακό μέρος του σώματος. Για παράδειγμα, η τοποθέτηση του κορμού μετωπικά στον συνομιλητή μπορεί να θεωρηθεί ως επίδειξη επιθετικότητας. Γυρίζοντας τον κορμό σε μια μικρή γωνία μπορεί κανείς να θεωρηθεί σίγουρος για τον εαυτό του και χωρίς επιθετικότητα. Η κλίση προς τα εμπρός, ειδικά όταν συνδυάζεται με ένα νεύμα και χαμόγελο, είναι ένας τρόπος για να δείξει κανείς περιέργεια.
- 3. Ομιλία:** Μία από τις ταχύτερες και πιο φυσικές μεθόδους επικοινωνίας μεταξύ των ανθρώπων. Η ομιλία μεταφέρει όχι μόνο γλωσσικά μηνύματα, αλλά και συναισθηματική πληροφορία. Για παράδειγμα, μια κατάσταση θυμού προκαλεί συχνά αλλαγές στην αναπνοή και αύξηση της μυϊκής έντασης, οι οποίες επηρεάζουν τη δόνηση των φωνητικών πτυχών και το σχήμα της φωνητικής οδού, επηρεάζοντας τα ακουστικά χαρακτηριστικά της ομιλίας, τα οποία με τη σειρά τους μπορούν να χρησιμοποιηθούν από τον ακροατή για να συμπεράνει την αντίστοιχη κατάσταση.
- 4. Κείμενο:** Σήμερα μεγάλο μέρος της επικοινωνίας και της ανθρώπινης αλληλεπίδρασης πραγματοποιείται μέσω κοινωνικών ιστότοπων και εφαρμογών. Το συναίσθημα μπορεί να εκφραστεί εκεί μέσω του κειμένου και αναλύεται με τεχνικές επεξεργασίας φυσικής γλώσσας.



(a) Κατηγορικό μοντέλο



(b) Διαστατικό μοντέλο



(c) Συνδυαστικό μοντέλο

Τρόποι αναπαράστασης του συναισθήματος.

## Μοντέλα Συναισθήματος

Τα προηγούμενα χρόνια πολλά υπολογιστικά μοντέλα που περιγράφουν το συναίσθημα έχουν προταθεί από την ψυχολογική κοινότητα. Ωστόσο, δεν έχουμε σαφή εικόνα για το ποια αναπαράσταση είναι η κατάλληλη για την ανάλυση των εκφράσεων του προσώπου. Ως εκ τούτου, είναι σημαντικό να αναλύσουμε τα οφέλη και τα μειονεκτήματα κάθε αναπαράστασης συναισθήματος.

1. **Κατηγορικό:** Η ταξινόμηση των συναισθημάτων σε διακριτές κατηγορίες που μπορούν να αναγνωριστούν και να περιγραφούν εύκολα στην καθημερινή γλώσσα ήταν κοινή από την εποχή του Δαρβίνου. Πιο πρόσφατα, ξεκινώντας από την έρευνα του Paul Ekman [Ekm94] [EF71] η κυρίαρχη άποψη για το συναίσθημα βασίζεται στην υπόθεση ότι οι άνθρωποι εκφράζουν και αναγνωρίζουν καθολικά ένα σύνολο διακριτών βασικών συναισθημάτων τα οποία είναι η χαρά, η λύπη, ο φόβος, ο θυμός, η αγάπη και η έκπληξη. Κυρίως λόγω της απλότητας και της καθολικότητάς του, η υπόθεση των καθολικών συναισθημάτων είναι συνήθως η πρώτη επιλογή στα συστήματα αναγνώρισης. Ωστόσο, ο περιορισμός των ανθρώπινων συναισθημάτων σε ένα προκαθορισμένο σύνολο κατηγοριών έχει τα μειονεκτητά του. Γνωρίζοντας ότι το ανθρώπινο συναίσθημα είναι διαφορούμενο και υποκειμενικό, το κατηγορικό μοντέλο δημιουργεί μεγάλες ενδοταξικές διακυμάνσεις και μικρές διαταξικές διακυμάνσεις.
2. **Διαστατικό:** Μια άλλη δημοφιλής προσέγγιση είναι η μοντελοποίηση των συναισθημάτων σε ένα συνεχές χώρο 2 ή 3 διαστάσεων [Rus80]. Αυτές οι διαστάσεις περιλαμβάνουν το valence (πόσο ευχάριστο ή δυσάρεστο είναι ένα συναίσθημα), arousal (πόσο πιθανό είναι το άτομο να αναλάβει δράση υπό αυτή τη συναισθηματική κατάσταση) και dominance (την αίσθηση ελέγχου του συναισθήματος). Λόγω της συνεχούς φύσης τους, τέτοια μοντέλα μπορούν θεωρητικά να περιγράψουν πιο περίπλοκα συναισθήματα.
3. **Συνδυαστικό:** Ανάμεσα στο κατηγορικό και το διαστατικό μοντέλο όσον αφορά την περιγραφική ικανότητα, τα συνδυαστικά μοντέλα οργανώνουν τα συναισθήματα με ιεραρχικό τρόπο όπου κάθε ανώτερο στρώμα περιέχει πιο πολύπλοκα συναισθήματα που αποτελούνται από συναισθήματα προηγούμενων στρώματων. Το πιο γνωστό παράδειγμα συνδυαστικών μοντέλων προτάθηκε από τους Plutchik et al. [Plu01]. Σύμφωνα με τη θεωρία του, τα πιο σύνθετα συναισθήματα είναι συνδυασμοί ζευγών πιο βασικών συναισθημάτων, που ονομάζονται δυάδες. Για παράδειγμα, η αγάπη θεωρείται ένας συνδυασμός χαράς και εμπιστοσύνης. Οι πρωτογενείς δυάδες, π.χ. αισιοδοξία = προσμονή+χαρά, γίνονται συχνά αισθητές, δευτερεύουσες δυάδες, π.χ. ενοχή = χαρά+φόβος, μερικές φορές γίνονται αισθητές και σπάνια γίνονται αισθητές τριτογενείς δυάδες, π.χ. απόλαυση = χαρά+έκπληξη.

## Εφαρμογές

Ένα σύστημα αναγνώρισης των εκφράσεων του προσώπου έχει χρήσιμες εφαρμογές σε πολλούς τομείς όπως:

- **Αλληλεπίδραση Ανθρώπου Μηχανής:** Μια σημαντική πρόκληση στον τομέα αυτόν είναι η δυνατότητα να αποκτήσουν οι μηχανές συναισθηματική νοημοσύνη για να γίνει η αλληλεπίδρασή τους με τον άνθρωπο πιο διαισθητική και φυσική.



- **Ψηφιακή Ψυχαγωγία:** Η αναγνώριση συναισθημάτων επιτρέπει τη δημιουργία εξατομικευμένων και πιο διαδραστικών μορφών ψηφιακής ψυχαγωγίας. Στον τομέα της ανάπτυξης παιχνιδιών, η αναγνώριση εκφράσεων του προσώπου μπορεί να χρησιμοποιηθεί για τη δημιουργία επιπέδων του παιχνιδιού, έτσι ώστε οι χώροι να βελτιστοποιούν την δυσκολία για τον κάθε παίκτη χωρίς να ζητούν ρητά σχόλια κατά τη διάρκεια του παιχνιδιού.
- **Υγεία:** Δεδομένου ότι οι εκφράσεις του προσώπου αλλάζουν ανάλογα με την κατάσταση της υγείας, ένα σύστημα αναγνώρισης μπορεί να είναι επωφελές σε ένα πλαίσιο υγειονομικής περίθαλψης.
- **Διαφήμιση:** Οι εκφράσεις του προσώπου έχουν αποδειχθεί χρήσιμες στον τομέα της διαφήμισης. Μεγαλύτερη μυϊκή δραστηριότητα στα ζυγωματικά (χαμόγελο) παρατηρείται κατά τη διάρκεια διαφημίσεων με θετικό συναισθηματικό τόνο και μεγαλύτερη δραστηριότητα των φρυδιών παρατηρείται κατά τη διάρκεια διαφημίσεων με αρνητικό συναισθηματικό τόνο. Επομένως, η αναγνώριση των εκφράσεων του προσώπου θα ήταν χρήσιμη για την κατανόηση της σχέσης μεταξύ των συναισθηματικών αποκρίσεων στο περιεχόμενο και των μέτρων αποτελεσματικότητας της διαφήμισης.
- **Εκπαίδευση:** Σε περιβάλλοντα ηλεκτρονικής μάθησης, ένα σύστημα αναγνώρισης μπορεί να βοηθήσει τους εκπαιδευτικούς να προσδιορίσουν εάν οι μαθητές κατανοούν το διδακτικό περιεχόμενο σύμφωνα με τις διαφορετικές εκφράσεις των μαθητών. Επίσης, η χρήση συναισθημάτων σε συστήματα λογισμικού για ηλεκτρονική μάθηση αυξάνει σημαντικά την απόδοση εάν το λογισμικό μπορεί να προσαρμοστεί στη συναισθηματική κατάσταση του μαθητή.

## Κύριες Προκλήσεις

Προκειμένου να σχεδιάσουμε συστήματα που μπορούν να αναγνωρίσουν τις εκφράσεις του προσώπου υπό πραγματικές συνθήκες, πρέπει πρώτα να εξετάσουμε τις προκλήσεις που παρουσιάζει ο συγκεκριμένος τομέας.

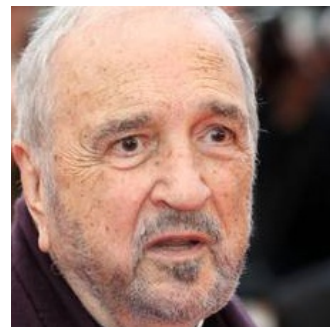
1. **Διακυμάνσεις:** Μια μεγάλη πρόκληση είναι η αποσύνδεση των διαφόρων παραγόντων διακύμανσης που υπάρχουν σε μια εικόνα. Συγκεκριμένα, μια εικόνα προσώπου περιέχει πολλές πληροφορίες που δεν σχετίζονται με την έκφραση αλλά με τα προσωπικά χαρακτηριστικά του ατόμου. Για παράδειγμα, τα αζεσουάρ ή τα μαλλιά θεωρούνται ως θόρυβος από το σύστημα και μπορεί να επηρεάσουν τη συνολική του απόδοση. Επίσης, η πόζα του κεφαλιού και ο φωτισμός μπορεί να ποικίλουν δραστικά, ενώ τα περισσότερα διαθέσιμα σύνολα δεδομένων αναγνώρισης αποτελούνται κυρίως από μετωπικές εικόνες σε ιδανικές συνθήκες φωτισμού. Τέλος, ένα συναίσθημα μπορεί να εκφραστεί με διάφορους τρόπους. Αυτό το εγγενές χαρακτηριστικό των εκφράσεων προκαλεί μεγάλη ενδοταξική και μικρή διαταξική διακύμανση των χαρακτηριστικών.
2. **Occlusions:** Η περιοχή του προσώπου μπορεί να κρύβεται από μαλλιά, γυαλιά ή ρούχα δυσκολεύοντας την αναγνώριση.
3. **Υποκειμενικότητα:** Η επισήμειση ενός συνόλου δεδομένων εκφράσεων προσώπου επηρεάζεται από την υποκειμενικότητα, καθώς υπάρχουν εκφράσεις που δεν αντιστοιχούν σε έναν μόνο τύπο συναισθήματος. Για παράδειγμα, μια εικόνα ενός άνδρα που ανοίγει ευρέως το στόμα του και τα μάτια του θα



Διακυμάνσεις



Occlusions



Υποκειμενικότητα

Κύριες προκλήσεις στην αναγνώριση των εκφράσεων του προσώπου.

μπορούσε να ταξινομηθεί είτε ως έκφραση έκπληξης είτε φόβου, καθώς και τα δύο συναισθήματα μπορούν να εκφραστούν μέσω αυτών των χαρακτηριστικών.

## Βαθιά Μάθηση

Ένας αλγόριθμος μηχανικής μάθησης είναι ένας αλγόριθμος που μπορεί να μάθει από δεδομένα. Σύμφωνα με τους Mitchell et al. [Mit+97] «Ένα πρόγραμμα υπολογιστή λέμε ότι μαθαίνει από την εμπειρία  $E$  σε σχέση με κάποια κατηγορία εργασιών  $T$  και μέτρο απόδοσης  $P$ , εάν η απόδοσή του σε εργασίες στο  $T$ , όπως μετράται από το  $P$ , βελτιώνεται με εμπειρία  $E$ ». Πιο τεχνικά, ο στόχος ενός μοντέλου μηχανικής μάθησης είναι να μάθει ένα σύνολο παραμέτρων χρησιμοποιώντας ένα σύνολο σημείων εισόδου δεδομένων. Ενώ υπάρχουν πολλά είδη εργασιών που μπορούν να επιλυθούν με τη μηχανική εκμάθηση, παρουσιάζουμε αυτά που θα χρησιμοποιηθούν στη διπλωματική:

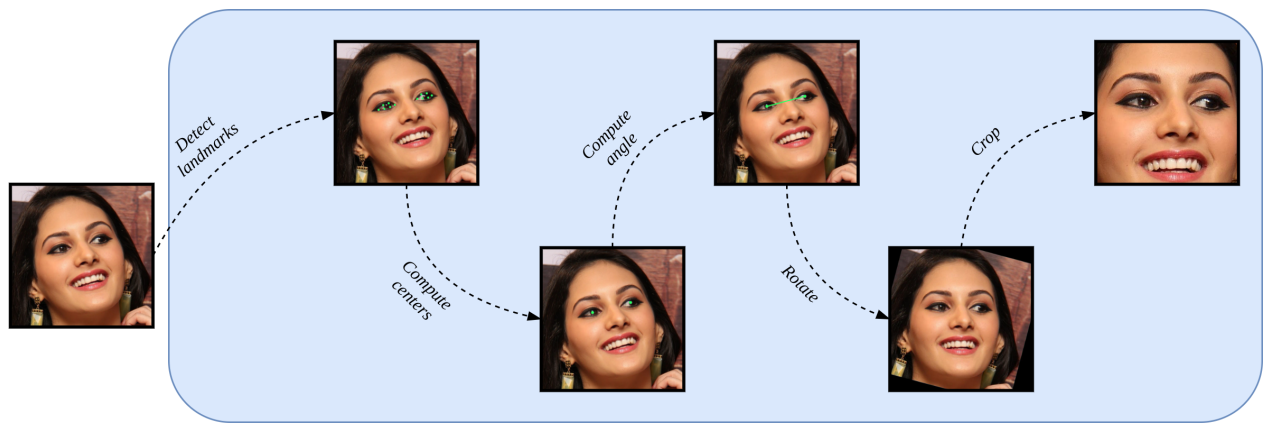
- **Ταξινόμηση:** Εδώ το μοντέλο καλείται να καθορίσει σε ποια από τις  $k$  κατηγορίες ανήκει η είσοδος. Για την επίλυση αυτού του προβλήματος, ο αλγόριθμος μάθησης συνήθως καλείται να μάθει μια συνάρτηση  $f : R^n \rightarrow \{1, \dots, k\}$ . Όταν  $y = f(x)$ , το μοντέλο αντιστοιχεί μια είσοδο που περιγράφεται από το διάνυσμα  $x$  σε μια κατηγορία που προσδιορίζεται από τον αριθμό  $y$ . Υπάρχουν πολλές παραλλαγές της ταξινόμησης, για παράδειγμα, όπου η συνάρτηση  $f$  εξάγει μια κατανομή πιθανότητας πάνω στις κλάσεις. Ένα παράδειγμα ταξινόμησης είναι η αναγνώριση αντικειμένου, όπου η είσοδος είναι μια εικόνα και η έξοδος είναι ένας αριθμός που προσδιορίζει το αντικείμενο στην εικόνα.
- **Παλινδρόμηση:** Σε αυτόν τον τύπο εργασιών, το μοντέλο καλείται να προβλέψει μια αριθμητική συνεχή τιμή δεδομένης κάποιας εισόδου. Για την επίλυση αυτού του προβλήματος, ο αλγόριθμος μάθησης καλείται να εξάγει μια συνάρτηση  $f : R^n \rightarrow R$ . Ένα παράδειγμα παλινδρόμησης είναι η πρόβλεψη του αναμενόμενου ποσού απαίτησης που θα κάνει ένας ασφαλισμένος (χρησιμοποιείται για τον καθορισμό ασφαλίστρων) ή η πρόβλεψη μελλοντικών τιμών των κινητών αξιών.

Οι παραδοσιακές τεχνικές μηχανικής μάθησης είναι περιορισμένες στην ικανότητά τους να επεξεργάζονται φυσικά δεδομένα στην ακατέργαστη μορφή τους. Για δεκαετίες, η κατασκευή ενός συστήματος αναγνώρισης προτύπων ή μηχανικής μάθησης απαιτούσε προσεκτικό σχεδιασμό και αρκετή γνώση του τομέα για να σχεδιάσει κανείς έναν εξαγωγέα χαρακτηριστικών που μετατρέπει τα ακατέργαστα δεδομένα (όπως οι τιμές των εικονοστοιχείων μιας εικόνας) σε μια κατάλληλη εσωτερική αναπαράσταση ή διάνυσμα χαρακτηριστικών από την οποία το υποσύστημα μάθησης, συχνά ταξινομητής, θα μπορούσε να ανιχνεύσει ή να ταξινομήσει μοτίβα στην είσοδο.

Η μάθηση αναπαραστάσεων αποτελείται από ένα σύνολο μεθόδων που επιτρέπουν σε ένα μηχανήμα να τροφοδοτείται με ακατέργαστα δεδομένα και να ανακαλύπτει αυτόματα τις αναπαραστάσεις που απαιτούνται για τον εντοπισμό ή την ταξινόμηση. Οι μέθοδοι βαθιάς μάθησης είναι μέθοδοι μάθησης αναπαράστασης με πολλαπλά επίπεδα αναπαράστασης, που λαμβάνονται με τη σύνθεση απλών αλλά μη γραμμικών ενοτήτων που καθένα μετατρέπει την αναπαράσταση σε ένα επίπεδο (ξεκινώντας από την ακατέργαστη εισαγωγή) σε αναπαράσταση σε υψηλότερο, ελαφρώς πιο αφηρημένο επίπεδο. Με τη σύνθεση αρκετών τέτοιων μετασχηματισμών, μπορούν τα μοντέλα αυτά να μάθουν πολύ σύνθετες λειτουργίες. Για τις εργασίες ταξινόμησης, υψηλότερα επίπεδα αναπαράστασης ενισχύουν πτυχές της εισόδου που είναι σημαντικές για τις διακρίσεις και καταστέλλουν άσχετες παραλλαγές. Μια εικόνα, για παράδειγμα, έρχεται με τη μορφή μιας συστοιχίας τιμών εικονοστοιχείων και τα χαρακτηριστικά που μαθαίνονται στο πρώτο επίπεδο αναπαράστασης αντιπροσωπεύουν τυπικά την παρουσία ή την απουσία ακμών σε συγκεκριμένους προσανατολισμούς και τοποθεσίες της εικόνας. Το δεύτερο στρώμα τυπικά ανιχνεύει μοτίβα εντοπίζοντας συγκεκριμένες διατάξεις των άκρων, ανεξάρτητα από τις μικρές παραλλαγές στις θέσεις των άκρων. Το τρίτο στρώμα μπορεί να συγκεντρώνει μοτίβα σε μεγαλύτερους συνδυασμούς που αντιστοιχούν σε μέρη οικείων αντικειμένων και τα επόμενα στρώματα θα ανιχνεύουν αντικείμενα ως συνδυασμούς αυτών των τμημάτων. Η βασική πτυχή της βαθιάς μάθησης είναι ότι αυτά τα επίπεδα χαρακτηριστικών δεν έχουν σχεδιαστεί από ανθρώπους: μαθαίνονται από δεδομένα χρησιμοποιώντας μια διαδικασία μάθησης γενικού σκοπού.

## Αρχικό Μοντέλο

Το πρώτο βήμα στην εκπαίδευση οποιουδήποτε μοντέλου μηχανικής μάθησης είναι η προεπεξεργασία των δεδομένων εκπαίδευσης. Δεδομένου ότι μας ενδιαφέρει η έκφραση του προσώπου, αρχικά πρέπει είναι να περικόψ-



Τα βήματα προεπεξεργασίας σε μία εικόνα από το σύνολο δεδομένων AffectNet.

ουμε την περιοχή του προσώπου στην εικόνα. Το επόμενο στάδιο είναι η ευθυγράμμιση του προσώπου η οποία συνήθως αυξάνει την απόδοση επειδή μειώνει την υψηλή διακύμανση του συνόλου δεδομένων. Το τελευταίο στάδιο στην προεπεξεργασία δεδομένων είναι η επαύξηση των δεδομένων εκπαίδευσης με χρήση augmentation το οποίο συμβάλλει στη μείωση του overfitting κατά την εκπαίδευση ενός μοντέλου μηχανικής μάθησης.

Ως βασική αρχιτεκτονική χρησιμοποιούμε το μοντέλο DenseNet [Hua+17a] στο οποίο η είσοδος κάθε επιπέδου αποτελείται από τους χάρτες χαρακτηριστικών όλων των προηγούμενων επιπέδων και η έξοδος του περνά σε κάθε επόμενο επίπεδο. Πέρα από την αντιμετώπιση του προβλήματος vanishing gradient, η αρχιτεκτονική αυτή ενθαρρύνει την επαναχρησιμοποίηση χαρακτηριστικών, καθιστώντας το δίκτυο εξαιρετικά αποδοτικό. Σχετικά με τις συναρτήσεις κόστους, στο κατηγορικό μοντέλο εφαρμόζουμε μια σταθμισμένη έκδοση της παραδοσιακής συνάρτησης διασταυρούμενης εντροπίας καθώς στις περισσότερες βάσεις δεδομένων, η κατανομή των κλάσεων είναι μη ισορροπημένη. Έτσι, το δίκτυο "τιμωρεί" περισσότερο την λανθασμένη ταξινόμηση δειγμάτων από υποεκπροσωπούμενες κατηγορίες παρά από καλά εκπροσωπούμενες κατηγορίες ως εξής:

$$L_{CE_w} = - \sum_{i=1}^7 \frac{f_i}{f_{min}} y_i \log(\hat{y}_i)$$

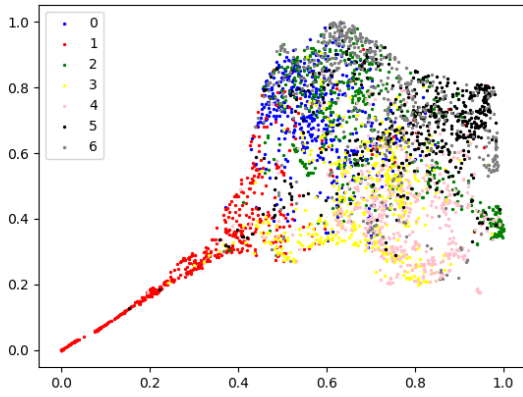
όπου  $y_i = 1$  εάν η κλάση  $i$  είναι η κλάση του δείγματος,  $f_i$  είναι ο αριθμός των δειγμάτων της κλάσης  $i$  και  $f_{min}$  είναι ο αριθμός των δειγμάτων στην λιγότερο εκπροσωπούμενη κλάση.

## Μετρική Μάθηση

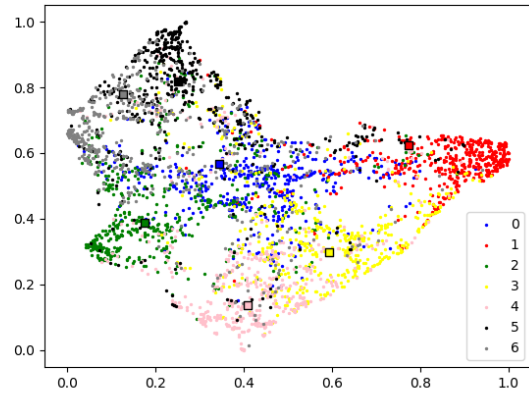
Η εικόνα ενός προσώπου περιέχει πολλές πληροφορίες που δεν σχετίζονται με το συναίσθημα και λειτουργούν ως θόρυβος για το σύστημα αναγνώρισης μειώνοντας τη συνολική του ακρίβεια. Αυτές οι διακυμάνσεις συνήθως ονομάζονται διακυμάνσεις εμφάνισης. Ωστόσο, στο πρόβλημά μας υπάρχουν δύο άλλοι τύποι διακυμάνσεων που προέρχονται από τη φύση της ανθρώπινης έκφρασης. Συγκεκριμένα, οι ανθρώπινες εκφράσεις αλλάζουν με συνεχή τρόπο, ενώ το κατηγορικό μοντέλο περιορίζει τις εκφράσεις σε διακριτές κατηγορίες συναισθημάτων. Γνωρίζουμε ότι ένα συναίσθημα μπορεί να εκφραστεί με διάφορους τρόπους. Για παράδειγμα, ο θυμός μπορεί να εκφραστεί με τρόπους που είναι πολύ διαφορετικοί μεταξύ τους και πιο κοντά στην έκφραση άλλων συναισθημάτων όπως η αγδία. Αυτό το εγγενές χαρακτηριστικό του κατηγορικού μοντέλου οδηγεί σε μια μεγάλη ενδοταξική και μικρή διαταξική διακύμανση των χαρακτηριστικών. Για να μειώσουμε τις διακυμάνσεις αυτές, προτείνουμε διάφορες τεχνικές μετρικής μάθησης.

Αρχικά, εφαρμόζουμε το Center loss που έχει σκοπό να φέρει κοντά τα χαρακτηριστικά της ίδιας κλάσης προς ένα κέντρο της κλάσης. Για να επιτευχθεί αυτό, το center loss τιμωρεί τις αποστάσεις μεταξύ των χαρακτηριστικών και των αντίστοιχων κέντρων τους:

$$L_c = \frac{1}{2} \sum_{i=1}^m \|x_i - c_{y_i}\|_2^2$$



Αρχικό μοντέλο.



Center loss.

Προβολή των χαρακτηριστικών στο 2-διάστατο επίπεδο πριν και μετά την εφαρμογή του Center loss.

όπου  $x_i$  είναι το διάνυσμα χαρακτηριστικών του δείγματος  $i$ ,  $y_i$  είναι η κλάση του δείγματος  $i$  και  $c_{y_i}$  είναι το κέντρο της κλάσης  $y_i$ . Προκειμένου να διερευνήσουμε ποιοτικά την επίδραση του Center loss στο σύστημά μας, εφαρμόζουμε την τεχνική μείωσης διαστάσεων UMAP [MHM18] για την προβολή των μαθημένων χαρακτηριστικών από το  $R^{1024}$  στο  $R^2$ . Παρατηρούμε ότι πράγματι το Center loss φέρνει πιο κοντά τα χαρακτηριστικά της ίδιας κλάσης μειώνοντας την ενδοταξική διακύμανση. Ωστόσο, το center loss παρουσιάζει δύο προβλήματα:

1. Αντιμετωπίζει μόνο το πρόβλημα των μεγάλων ενδοταξικών διακυμάνσεων και αγνοεί τις μικρές διαταξικές διακυμάνσεις. Τα δείγματα μιας συγκεκριμένης κλάσης πλησιάζουν στο κέντρο της. Ωστόσο, το center loss δεν εγγυάται ότι τα κέντρα διαφορετικών κατηγοριών θα είναι μακριά.
2. Δεν σέβεται την εγγενή ενδοταξική διακύμανση του προβλήματος καθώς το πρόβλημά μας απαιτεί να υπάρχει κάποια διακύμανση σε κάθε κλάση. Για παράδειγμα, το συναίσθημα της χαράς εκφράζεται με περισσότερους τρόπους από το συναίσθημα του φόβου. Επομένως, η κλάση Χαράς θα πρέπει να είναι σε θέση να μάθει χαρακτηριστικά με περισσότερη διακύμανση από την κλάση Φόβος.

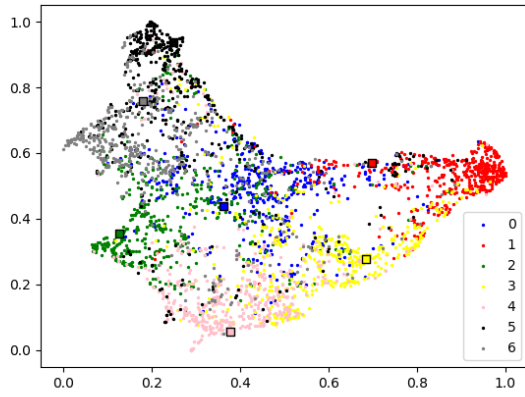
Για να αντιμετωπίσουμε το πρώτο πρόβλημα, χρησιμοποιούμε το island loss που ορίζεται ως το άθροισμα του center loss και των αποστάσεων μεταξύ των κέντρων των κλάσεων στο χώρο χαρακτηριστικών:

$$L_{island} = L_c + \lambda_1 \sum_{c_j \in N} \sum_{\substack{c_k \in N \\ c_k \neq c_j}} \left( \frac{c_k \cdot c_j}{\|c_k\|_2 \|c_j\|_2} + 1 \right)$$

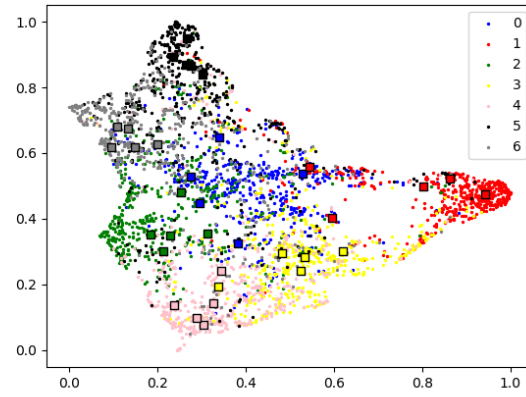
όπου  $N$  είναι το σύνολο των κλάσεων.  $c_k$  και  $c_j$  δηλώνουν το κέντρο  $k$  και  $j$  με  $L_2$  νόρμα  $\|c_k\|_2$  και  $\|c_j\|_2$ , αντίστοιχα. Συγκεκριμένα, ο πρώτος όρος τιμωρεί την απόσταση μεταξύ του δείγματος και του αντίστοιχου κέντρου του και ο δεύτερος όρος τιμωρεί την ομοιότητα μεταξύ των κέντρων. Για να αντιμετωπίσουμε το δεύτερο πρόβλημα, χρησιμοποιούμε το local subclass loss που χρησιμοποιεί παραπάνω από ένα κέντρα ανά κλάση για να επιτρέψει την ύπαρξη διακύμανσης μέσα σε κάθε κλάση. Ορίζεται ως:

$$L_{subclass} = \frac{1}{2} \sum_{i=1}^m \|x_i - c_{y_i}^{min}\|_2^2$$

όπου  $c_{y_i}^{min}$  είναι το κοντινότερο υπο-κέντρο της κλάσης  $y_i$  στο διάνυσμα  $x_i$ . Η απόσταση μεταξύ του διανύσματος χαρακτηριστικών και του υπο-κέντρου της κλάσης του υπολογίζεται στον Ευκλείδειο χώρο. Αυτή η μέθοδος στοχεύει στην εκμάθηση ενός τοπικά συμπαγούς χώρου χαρακτηριστικών ελαχιστοποιώντας την απόσταση μεταξύ των διανυσμάτων χαρακτηριστικών και του πλησιέστερου υπο-κέντρου. Προφανώς, η ακριβής τιμή των υπο-κέντρων της κάθε κλάσης είναι δύσκολο να οριστεί ακριβώς και πρέπει να ενημερώνεται καθώς μαθαίνονται τα χαρακτηριστικά.



Island loss.



Local Subclass loss.

Προβολή των χαρακτηριστικών στο 2-διάστατο επίπεδο μετά από την εφαρμογή του Island loss και του Local Subclass loss.

## Μάθηση Πολλαπλών Εργασιών

Η μάθηση πολλαπλών εργασιών είναι ιδιαίτερα σημαντική στην όραση υπολογιστών, καθώς έχει αποδειχθεί ότι αυξάνει την απόδοση μιας εργασίας με τη συμπερίληψη άλλων σχετικών εργασιών στη διαδικασία μάθησης. Έτσι, η κύρια εργασία μπορεί να επωφεληθεί από άλλες εργασίες μοιράζοντας μια κοινή αναπαράσταση χαρακτηριστικών και μεταφέροντας τη γνώση μεταξύ διαφορετικών τομέων. Στην αναγνώριση της έκφρασης του προσώπου, ένα μοντέλο μπορεί να ωφεληθεί από την ταυτόχρονη εκπαίδευση άλλων εργασιών που σχετίζονται με το πρόσωπο. Για τον λόγο αυτό υλοποιούμε διάφορα μοντέλα μάθησης πολλαπλών εργασιών με το αποδοτικότερο να είναι αυτό που χρησιμοποιεί ως επιπλέον εργασία την πρόβλεψη των τιμών valence/arousal του διαστατικού μοντέλου. Για το διαστατικό μοντέλο χρησιμοποιείται το CCC loss το οποίο μετρά τη συμφωνία μεταξύ της πραγματικής διάστασης του συναισθήματος με τον βαθμό συναισθήματος που προβλέψαμε και ορίζεται ως:

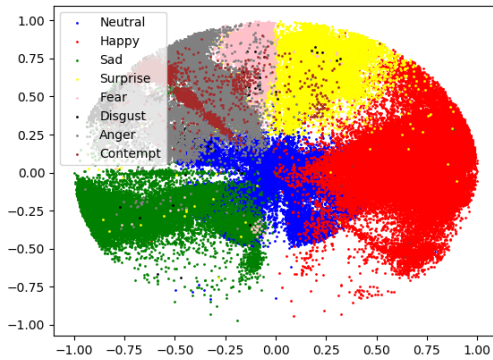
$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2}$$

όπου  $s_x$  και  $s_y$  δηλώνουν τη διακύμανση των προβλεπόμενων και αληθινών τιμών αντίστοιχα,  $\bar{x}$  και  $\bar{y}$  είναι οι αντίστοιχες μέσες τιμές και  $s_{xy}$  είναι η αντίστοιχη τιμή συνδιακύμανσης. Το εύρος του CCC είναι από -1 (τέλεια διαφωνία) έως 1 (τέλεια συμφωνία). Ως εκ τούτου, στην περίπτωση μας ορίζουμε τη συνάρτηση κόστους του διαστατικού μοντέλου ως:

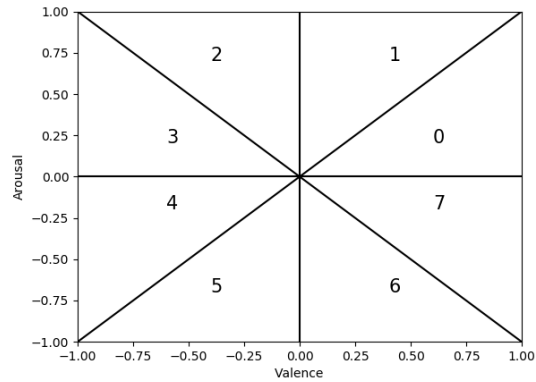
$$L = 1 - \frac{\rho_v + \rho_a}{2} \quad (0.0.1)$$

όπου  $\rho_v$  και  $\rho_a$  είναι η τιμή CCC του valence και arousal αντίστοιχα.

Στην εκμάθηση πολλαπλών εργασιών οι επιμέρους εργασίες πρέπει να είναι ισορροπημένες, δηλαδή η δυσκολία τους πρέπει να είναι ανάλογη. Ωστόσο, η ταυτόχρονη αναγνώριση στο κατηγορικό και διαστατικό μοντέλο δεν είναι ισορροπημένη και βλάπτει την τελική απόδοση. Επομένως, μετατρέπουμε την εργασία παλινδρόμησης στο VA χώρο σε μία εργασία ταξινόμησης διαιρώντας τον χώρο VA σε επιμέρους περιοχές. Το πρόβλημα που εμφανίζεται εδώ είναι πώς να διαιρέσουμε αποτελεσματικά τον χώρο VA για να εκμεταλλευτούμε πλήρως τα οφέλη που μπορούν να προσφέρουν οι τιμές VA. Ο ιδανικός αριθμός τμημάτων που πρέπει να διαιρέσουμε τον χώρο VA είναι 7 όσες και οι κλάσεις του κατηγορικού μοντέλου. Για πρακτικούς και υπολογιστικούς λόγους, διαιρούμε το χώρο σε 8 μέρη. Αν παρατηρήσουμε την κατανομή των δειγμάτων στο χώρο VA, τα βασικά συναισθήματα εντοπίζονται γύρω από το ουδέτερο συναισθήμα που εμφανίζεται όταν οι τιμές valence και arousal είναι κοντά στο μηδέν. Επομένως, μια γωνιακή διαίρεση ταιριάζει καλύτερα. Εκπαιδεύοντας το δίκτυο να προβλέπει την περιοχή κάθε δείγματος στον χώρο VA μαζί με την αναγνώριση των εκφράσεων στο κατηγορικό μοντέλο, το δίκτυο επωφελείται από το πρώτο και η ακρίβεια αυξάνεται.



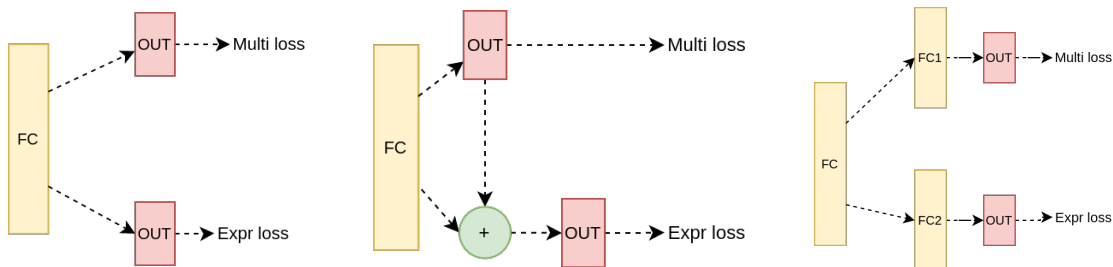
Κατανομή των δειγμάτων εκπαίδευσης.



Στρατηγική χωρισμού.

Γωνιακός χωρισμός του VA χώρου σε 8 ίσα μέρη.

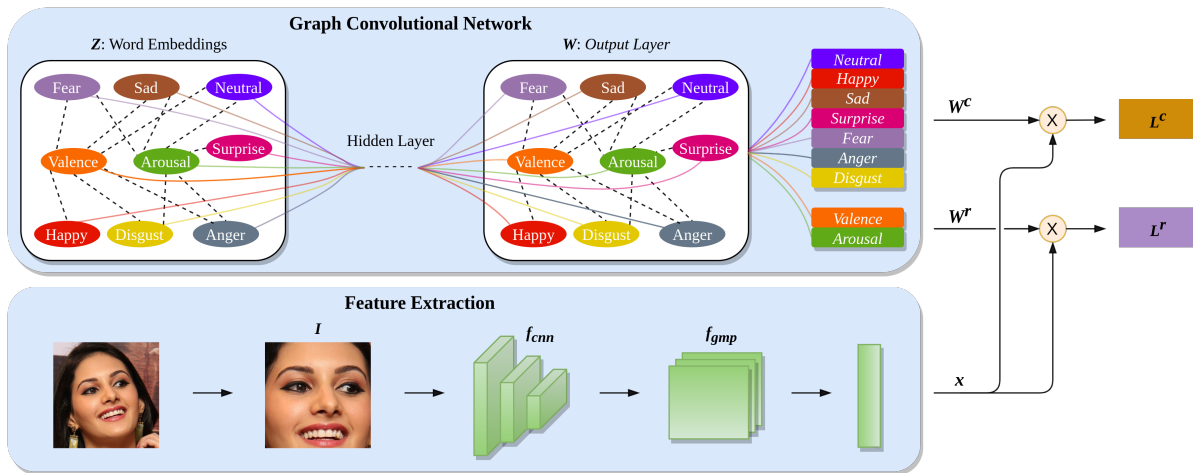
Τέλος, μια κρίσιμη παράμετρος στην εκμάθηση πολλαπλών εργασιών είναι ο τρόπος συνδυασμού των διαφορετικών εργασιών σε επίπεδο χαρακτηριστικών. Όταν οι εργασίες μοιράζονται όλα τα επίπεδα (εξαιρουμένου του τελευταίου επιπέδου πρόβλεψης), μοιράζονται πολλή πληροφορία μεταξύ τους. Στην περίπτωση μας, και οι δύο εργασίες ασχολούνται με ένα πρόβλημα αναγνώρισης συναισθήματος αλλά χρησιμοποιώντας διαφορετική αναπαράσταση του συναισθήματος. Επομένως, υπάρχει πολλή πληροφορία που πρέπει να μοιραστεί και η συνένωση γίνεται στο τελευταίο επίπεδο. Στην περίπτωση που μας ενδιαφέρει περισσότερο η απόδοση στο κατηγορικό μοντέλο, μπορούμε να αλλάξουμε τη διαμόρφωση έτσι ώστε να δώσουμε περισσότερη προσοχή εκεί. Προτείνουμε μια αρχιτεκτονική όπου το τελευταίο επίπεδο ταξινόμησης λαμβάνει ως είσοδο τόσο την έξοδο του τελευταίου επιπέδου όσο και τις προβλέψεις στον χώρο VA. Με αυτόν τον τρόπο, δίνουμε μια επιπλέον ώθηση στο πρόβλημα της ταξινόμησης. Τέλος, μια άλλη διαμόρφωση είναι να υπάρχει διαφορετικό τελευταίο επίπεδο για κάθε εργασία. Σε αυτήν τη διαμόρφωση, επιτρέπουμε στις δύο εργασίες να μάθουν πιο συγκεκριμένες γνώσεις και να μοιραστούν λιγότερες πληροφορίες.



Τρόποι συνδυασμού των διαφορετικών εργασιών στην εκμάθηση πολλαπλών εργασιών.

## Emotion-GCN

Ο επόμενος στόχος μας είναι να εκμεταλλευτούμε ακόμα περισσότερο την ισχυρή εξάρτηση μεταξύ του κατηγορικού και του διαστατικού μοντέλου. Για να το πετύχουμε αυτό βασιζόμαστε στο μοντέλο ML-GCN και γενικότερα στα Graph Convolutional Networks. Πολλά σημαντικά σύνολα δεδομένων έρχονται με τη μορφή γραφημάτων ή δικτύων: κοινωνικά δίκτυα, ο Παγκόσμιος Ιστός κ.λπ. Πριν από τα GCN, ελάχιστη προσοχή είχε αφιερωθεί στη γενίκευση των μοντέλων νευρωνικών δικτύων σε τέτοια δομημένα σύνολα δεδομένων. Η γενίκευση καθιερωμένων νευρωνικών μοντέλων όπως τα RNN ή τα CNN για να λειτουργούν σε αυθαίρετα δομημένα γραφήματα είναι ένα δύσκολο πρόβλημα. Οι Kipf et al. [KW17] ξεκίνησαν από το πλαίσιο των φασματικών συστάσεων γραφημάτων και εισήγαγαν απλουστεύσεις που σε πολλές περιπτώσεις επιτρέπουν τόσο ταχύτερους χρόνους εκπαίδευσης όσο και υψηλότερη ακρίβεια, επιτυγχάνοντας κορυφαία αποτελέσματα ταξινόμησης σε πολλά σύνολα δεδομένων με γράφους. Για τα GCN, ο στόχος είναι να μάθουν μια συνάρτηση



Αρχιτεκτονική του προτεινόμενου μοντέλου Emotion-GCN.

σημάτων/χαρακτηριστικών σε ένα γράφο  $G = (V, E)$  που λαμβάνει ως είσοδο

- ένα χαρακτηριστικό  $x_i$  για κάθε κόμβο  $i$  που συνοψίζεται σε μια  $N \times D$  μήτρα χαρακτηριστικών  $X$  ( $N$ : αριθμός κόμβων,  $D$ : αριθμός χαρακτηριστικών εισόδου).
- μία περιγραφή της δομής του γράφου σε μορφή πίνακα, τυπικά με τη μορφή ενός πίνακα γειτνίασης  $A$  (ή κάποιας συνάρτησης αυτού).

και παράγει μια έξοδο σε επίπεδο κόμβου  $Z$  (μια μήτρα χαρακτηριστικών  $N \times F$ , όπου  $F$  είναι το μέγεθος των χαρακτηριστικών εξόδου ανά κόμβο). Κάθε επίπεδο νευρωνικού δικτύου μπορεί στη συνέχεια να γραφτεί ως μη γραμμική συνάρτηση:

$$H^{l+1} = f(H^l, A)$$

με  $H^{(0)} = X$  και  $H^{(L)} = Z$  (το  $L$  είναι ο αριθμός των επιπέδων). Τα συγκεκριμένα μοντέλα διαφέρουν στη συνέχεια μόνο στον τρόπο επιλογής και παραμετροποίησης της συνάρτησης  $f$ .

Βασιζόμενοι στο ML-GCN που είναι μια αρχιτεκτονική για ταξινόμηση πολλαπλών ετικετών, προτείνουμε το Emotion-GCN ένα μοντέλο μάθησης πολλαπλών εργασιών που βασίζεται σε GCN για την αναγνώριση των εκφράσεων του προσώπου. Η κύρια ιδέα στο Emotion-GCN είναι η δημιουργία εξαρτημένων ταξινομητών συναισθήματος και προβλεπτών των τιμών VA μέσω μιας συνάρτησης που βασίζεται σε GCN. Τα διανύσματα που δημιουργούνται στη συνέχεια εφαρμόζονται σε μια εξαγόμενη αναπαράσταση εικόνας. Ως εκ τούτου, η εξάρτηση μεταξύ των δύο μοντέλων συναισθημάτων αποτυπώνεται τόσο μέσω μιας κοινής αναπαράστασης χαρακτηριστικών όσο και των εξαρτημένων ταξινομητών και προβλεπτών.

Σημαντικό κομμάτι στο Emotion-GCN είναι ο σχεδιασμός του πίνακα γειτνίασης του γράφου. Σύμφωνα με τον κανόνα ενημέρωσης ενός GCN, το διάνυσμα ενός κόμβου στο γράφημα είναι το σταθμισμένο άθροισμα του δικού του και των γειτονικών διανυσμάτων. Δεδομένου ότι ο σκοπός πίσω από τη χρήση ενός GCN είναι η εκμετάλλευση των εξαρτήσεων μεταξύ του κατηγορικού και του διαστασιακού μοντέλου, θα πρέπει να σχεδιάσουμε τον πίνακα γειτνίασης  $A$  προς αυτήν την κατεύθυνση. Υποθέτουμε ότι οι πρώτες επτά σειρές του  $A$  αντιστοιχούν στις βασικές εκφράσεις και οι δύο τελευταίες είναι οι συνεχείς διαστάσεις valence και arousal. Δεδομένου ότι ασχολούμαστε με ένα πρόβλημα αναγνώρισης πολλαπλών εργασιών και όχι πολλαπλών ετικετών, μας ενδιαφέρει μόνο ο συσχετισμός μεταξύ του κατηγορικού και του διαστασιακού μοντέλου. Ως εκ τούτου, θέτουμε τα άλλα ζεύγη στον πίνακα  $A$  στο μηδέν, εκτός από την διαγώνιο. Ο πίνακας συσχέτισης  $A$  μπορεί να γραφτεί ως:

$$A_{ij} = \begin{cases} 1, & \text{if } i = j \\ |c_{ij}|, & \text{if } i \in \text{Cat} \wedge j \in \text{Dim} \\ |c_{ij}|, & \text{if } j \in \text{Cat} \wedge i \in \text{Dim} \\ 0, & \text{else} \end{cases}$$

όπου *Cat* και *Dim* είναι το σύνολο δεικτών του κατηγορικού και διαστατικού μοντέλου αντίστοιχα. Ως μέτρηση συσχέτισης, χρησιμοποιούμε τον συντελεστή συσχέτισης Spearman [Spe61]. Ακολουθώντας τις ιδέες στο ML-GCN [Che+19b], χρησιμοποιούμε ένα κατώφλι  $\tau$  για να φιλτράρουμε τις θορυβώδεις άκρες ως εξής:

$$A'_{ij} = \begin{cases} 1, & \text{if } A_{ij} \geq \tau \\ 0, & \text{if } A_{ij} < \tau \end{cases}$$

όπου  $\tau = 0.1$  για να επιτρέψουμε τη διάδοση πληροφοριών μεταξύ ασθενώς συσχετισμένων κόμβων. Επίσης, για να λύσουμε το πρόβλημα της περβολικής εξομάλυνσης, εφαρμόζουμε το σταθμισμένο σχήμα του ML-GCN που ορίζεται ως:

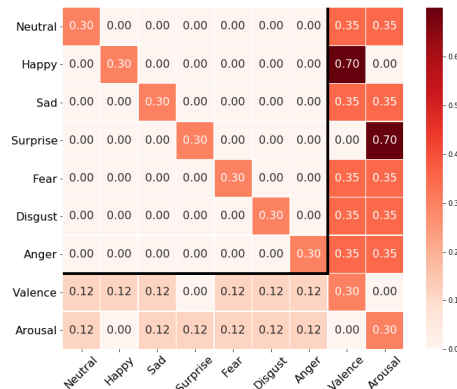
$$A''_{ij} = \begin{cases} (p / \sum_{i \neq j}^9 A'_{ij}) \times A'_{ij}, & \text{if } i \neq j \\ 1 - p, & \text{if } i = j \end{cases}$$

όπου η μεταβλητή  $p$  καθορίζει τα βάρη που εκχωρούνται σε έναν κόμβο και τους γειτονικούς του κόμβους.

## Συμπεράσματα

Στον παρακάτω πίνακα βλέπουμε την απόδοση των καλύτερων μοντέλων που προτείνονται στη διπλωματική στο κατηγορικό μοντέλο του συνόλου δεδομένων AffectNet. Παρατηρούμε ότι οι μέθοδοι μετρικής μάθησης και μάθησης πολλαπλών εργασιών που προτείνουμε βελτιώνουν την συνολική επίδοση του συστήματος με το Emotion-GCN να πετυχαίνει την μεγαλύτερη ακρίβεια.

Μοντέλο	Ακρίβεια
Baseline	64.37
Island loss	65.11
Local subclass loss with margin	65.29
Multi-task CCC	65.63
Multi-task Eight	65.97
<b>Emotion-GCN</b>	<b>66.46</b>



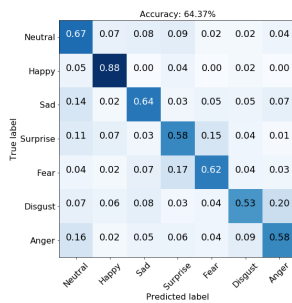
Ο πίνακας γειτνίασης του Emotion-GCN.



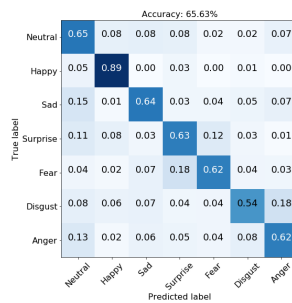
Στον επόμενο πίνακα βλέπουμε μια σύγκριση του καλύτερου μας μοντέλου (Emotion-GCN) με τα state-of-the-art μοντέλα αναγνώρισης στο σύνολο δεδομένων AffectNet. Παρατηρούμε ότι το Emotion-GCN ξεπερνάει σε απόδοση όλα τα άλλα μοντέλα πετυχαίνοντας ακρίβεια **66.46%** στο validation set του συνόλου δεδομένων AffectNet.

Μέθοδος	Ακρίβεια
Facial Motion Prior Network [Che+19a]	61.52
CAKE [Ker+18]	61.7
OADN [DZC20]	61.89
CNNs and BOVW + global SVM [GIP19]	63.31
Siamese [HNM19]	64
<b>Emotion-GCN</b>	<b>66.46</b>

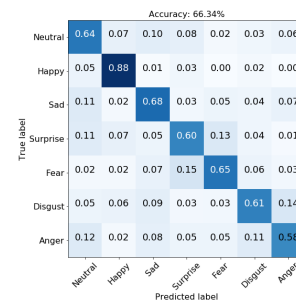
Οι πίνακες σύγχυσης για τα μοντέλα μας παρουσιάζονται στο παρακάτω σχήμα. Παρατηρούμε ότι η προτεινόμενη μέθοδος αυξάνει την ακρίβεια για τις περισσότερες κλάσεις, ενώ το βασικό μας δίκτυο αποδίδει καλύτερα μόνο στην ουδέτερη κλάση.



Single-task



Multi-task + CCC



Emotion-GCN

Πίνακας σύγχυσης διάφορων μοντέλων στο σύνολο δεδομένων AffectNet.



# Table of contents

<b>Table of Contents</b>	<b>xxvii</b>
<b>List of acronyms</b>	<b>xxix</b>
<b>List of figures</b>	<b>xxxi</b>
<b>List of tables</b>	<b>xxxiv</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Affective Computing	1
1.1.1 Modalities of Emotion	1
1.1.2 Emotion models	2
1.2 Facial Expressions	3
1.2.1 Universality	3
1.2.2 Other related tasks	4
1.3 Applications	5
1.3.1 Human-Robot Interaction	5
1.3.2 Digital Entertainment	5
1.3.3 Health Care	5
1.3.4 Advertisement	6
1.3.5 Education	6
1.4 Challenges	6
1.5 Deep Learning	6
1.5.1 Machine Learning	7
1.5.2 From Machine Learning to Deep Learning	7
1.5.3 Deep learning architectures	8
1.6 Other related areas	9
1.6.1 Multi-task Learning	9
1.6.2 Metric Learning	9
1.6.3 Support Vector Machines	9
1.7 Contributions	10
1.8 Structure of the thesis	10
<b>2 Literature Review</b>	<b>11</b>
2.1 Preprocessing	11
2.1.1 Face Detection	11
2.1.2 Face Alignment	12
2.1.3 Face Normalization	12
2.1.4 Data Augmentation	12
2.2 FER on the categorical model	13
2.2.1 Handcrafted features	13
2.2.2 Deep features	14
2.3 FER on the dimensional model	17

2.3.1	First attempts	17
2.3.2	Deep features	18
2.4	Combining emotion representations	19
2.4.1	Databases	20
2.5	Facial analysis toolkits	23
<b>3</b>	<b>Baseline Models</b>	<b>25</b>
3.1	Preprocessing	25
3.2	Architecture	27
3.3	Loss Function	27
3.4	Soft Loss	28
3.5	Results	29
<b>4</b>	<b>Metric learning models</b>	<b>31</b>
4.1	Variations	31
4.2	Center Loss	32
4.3	Extensions of center loss	33
4.3.1	Island loss	34
4.3.2	Local Subclass loss	34
4.4	Contrastive loss	36
4.5	Using VA in metric learning	38
4.5.1	Motivation	38
4.5.2	Example	38
4.5.3	Proposed Loss function	39
4.6	Results	40
<b>5</b>	<b>Multitask learning models</b>	<b>41</b>
5.1	Baseline architectures	41
5.1.1	Dimensional model	41
5.1.2	Facial Landmarks	42
5.2	Advanced architectures	43
5.2.1	Divide VA space	43
5.2.2	Concordance Correlation Coefficient	44
5.2.3	Different fusion configurations	44
5.3	Sampler	45
5.3.1	Age Sampler	45
5.3.2	Expression Sampler	46
5.4	Focal Loss	46
5.5	Results	47
<b>6</b>	<b>Emotion-GCN</b>	<b>49</b>
6.1	Overview of GCN	49
6.1.1	A simple example	49
6.1.2	Embedding the karate club network	50
6.2	Inspiration	51
6.3	Emotion-GCN	52
6.3.1	Feature extraction	53
6.3.2	Dependent classifiers and regressors	53
6.4	Results	55
6.4.1	Ablation Study	55
6.4.2	Visualization	56
6.4.3	Dependence between classifiers	57
6.4.4	Comparison with the State of the Art	58
<b>7</b>	<b>Conclusion and Future Work</b>	<b>59</b>
7.1	Conclusion	59

7.2 Future Work .....	59
<b>A Bibliography</b>	<b>61</b>



# List of Acronyms

<b>AUs</b>	Action Units
<b>CCC</b>	Concordance Correlation Coefficient
<b>CNN</b>	Convolutional Neural Network
<b>FACS</b>	Facial Action Coding System
<b>FER</b>	Facial Expression Recognition
<b>GAN</b>	Generative Adversarial Network
<b>GCN</b>	Graph Convolutional Network
<b>HCI</b>	Human Computer Interaction
<b>LBP</b>	Local Binary Pattern
<b>MSE</b>	Mean Squared Error
<b>MTL</b>	Multi-Task Learning
<b>SVM</b>	Support Vector Machine





# List of figures

1.1.1 Ways of modelling emotion in affective computing. . . . .	3
1.2.1 Upper face AUs. . . . .	4
1.4.1 Illustration of FER challenges using samples of AffectNet. . . . .	7
1.5.1 Deep architectures that are used in the thesis. . . . .	8
2.2.1 Handcrafted features on the categorical model of affect (Source: [Zha+98]). . . . .	13
2.2.2 Supervised Scoring Ensemble (Source: [Hu+17]). . . . .	14
2.2.3 Island loss layer (Source: [Cai+18]). . . . .	15
2.2.4 Feature-level ensemble (Source: [Bar+16]). . . . .	15
2.2.5 MSCNN model (Source: [Zha+17a]). . . . .	16
2.2.6 AUDN model (Source: [Liu+13]). . . . .	17
2.3.1 AffWildNet (Source: [Kol+19]). . . . .	18
2.4.1 A/V-MT-VGG-RNN: A Multi-Modal and Multi-Task model (Source: [KZ19]). . . . .	19
2.4.2 The holistic (multi-task, multi-domain, multi-label) FaceBehaviorNet (Source: [KZ21], [KSZ21]). . . . .	20
2.5.1 Example of using the OpenFace toolkit (Source: [Bal+18]). . . . .	24
3.1.1 A common numbering of the 68 facial landmark coordinates. . . . .	25
3.1.2 An example of the preprocessing in a sample from Affectnet. . . . .	26
3.1.3 Data augmentation techniques that are used in our models. . . . .	26
3.2.1 Building blocks of the deep architectures that are used. . . . .	27
3.4.1 Sample from AffectNet that depicts anger and the output of our best model. . . . .	28
4.1.1 Example of large intra-class and small inter-class variation. . . . .	31
4.2.1 The distribution of learned features under the joint supervision of softmax and center loss (Source: [Wen+16]). . . . .	32
4.2.2 Deep learned features (a) without extra loss, with (b) center loss and (c) center local loss (The squares denote the learned centers). . . . .	33
4.3.1 Deep features learned by (a) softmax loss, (b) softmax loss + center loss, and (c) softmax loss + island loss in the feature space (Source: [Cai+18]). . . . .	34
4.3.2 Deep learned features with (a) island loss and (b) euclidean island loss (The squares denote the learned centers). . . . .	35
4.3.3 Deep learned features with (a) local subclass loss and (b) local subclass loss with margin (The squares denote the learned centers). . . . .	36
4.4.1 Positive contrastive loss. . . . .	37
4.4.2 Negative contrastive loss. . . . .	37
4.5.1 Distribution of the training samples of AffectNet in valence-arousal space. . . . .	38
4.5.2 Feature space of an imaginary model before applying the center loss. . . . .	39
4.5.3 (a) Ideal and (b) real form of the feature space after applying the center loss. . . . .	39
5.1.1 Proposed multi-task architecture using VA. . . . .	42
5.1.2 Inputs of the mask-based multi-task framework. . . . .	42
5.1.3 Proposed multi-task architecture using facial masks. . . . .	43
5.2.1 Angular division of valence-arousal space in eight parts. . . . .	43
5.2.2 Proposed multi-task architecture using division of the VA space. . . . .	44

5.2.3 Different configurations for multi-task learning. . . . .	44
5.3.1 (a) Label distribution and (b) mean accuracy in different age groups. . . . .	45
5.4.1 Focal loss for several values of $\gamma$ . . . . .	46
6.1.1 Applying a GCN model in Zachary’s Karate club network (Source: [Bra+07]) . . . . .	50
6.2.1 A graph over the labels to model label dependencies in multi-label image recognition. . . . .	51
6.2.2 Overall framework of ML-GCN model for multi-label image recognition.. . . . .	52
6.3.1 Illustration of the dependencies between the categorical and the dimensional model. . . . .	52
6.3.2 Overall architecture of our Emotion-GCN model for FER in the wild. . . . .	53
6.3.3 Adjacency matrix of our Emotion-GCN. . . . .	54
6.4.1 Confusion matrices of our Emotion-GCN on the validation set of AffectNet. . . . .	55
6.4.2 Predictions of our models on samples of AffectNet. . . . .	57
6.4.3 Visualization of the cosine similarity between the learned classifiers and regressors by Emotion-GCN on AffectNet. . . . .	57

# List of tables

2.1	Different types of face alignment (Source: [LD20]). . . . .	12
2.2	Emotion databases annotated for the categorical model of affect. . . . .	21
2.3	Emotion databases annotated for the dimensional model of affect. . . . .	22
2.4	Emotion databases annotated for both the categorical and the dimensional model of affect. . . . .	23
2.5	Summary of freely available toolkits for facial analysis. . . . .	24
3.1	Performance of our baseline models on the categorical model of AffectNet. . . . .	29
3.2	Performance of models using soft and logit loss. . . . .	29
4.1	Performance of metric learning models on the categorical model of AffectNet. . . . .	40
5.1	Performance of multi-task learning models on the categorical model of AffectNet. . . . .	47
5.2	Performance of models using proposed sampling techniques and focal loss. . . . .	47
5.3	Performance of models combining metric and multi-task learning on the categorical model of AffectNet. . . . .	48
5.4	Performance of models using SVM. . . . .	48
6.1	Performance of Emotion-GCN on the categorical model of AffectNet and Aff-Wild2. . . . .	55
6.2	Classification accuracy of Emotion-GCN on the categorical model of AffectNet using different values for $\tau$ , $p$ and $L$ (number of GCN layers). In the first table $p = 0.7$ and in the second table $\tau = 0.1$ . . . . .	56
6.3	Performance of Emotion-GCN on the dimensional model of AffectNet and Aff-Wild2. . . . .	56
6.4	Comparison of Emotion-GCN with state-of-the-art methods on AffectNet (7-way classification). . . . .	58
6.5	Performance of Emotion-GCN using BReG-NeXt as the backbone network on AffectNet. . . . .	58
7.1	Comparison of our best proposed models with state-of-the-art methods on AffectNet. . . . .	60



# Chapter 1

## Introduction

In this chapter we introduce the reader to the scientific areas that are connected to FER. First, we present an overview of the general domain of Affective Computing focusing on the facial expressions and discuss the applications and the challenges of emotion recognition. Then, we make a brief introduction to deep learning, metric learning and multi-task learning. Finally, we highlight the contributions, the structure and the notation of the thesis.

### 1.1 Affective Computing

According to Oxford Dictionaries, emotion is a strong feeling deriving from one's circumstances, mood or relationships with others. It is impossible to imagine life without emotions since they color people's life experiences and give those experiences meaning and flavor. Hence, emotion analysis has been a topic of scientific research in psychology for over a century [Dar15] [Can27].

Emotions can be expressed either verbally through emotional vocabulary or by expressing non-verbal cues such as facial expressions, the intonation of voice and gestures. Most of the latest HCI systems are deficient in interpreting these emotional cues and suffer from the lack of emotional intelligence. In other words, they are unable to identify human emotional states and use this information in deciding upon proper actions to execute [Koe+11]. The goal of Affective Computing is to fill this gap by developing systems and devices that can recognize, interpret, process and simulate human affects. It is an interdisciplinary field spanning computer science, psychology and cognitive science.

#### 1.1.1 Modalities of Emotion

The first step towards developing affect-aware systems is to study the ways that people convey their emotional state because these forms of emotional expression will be the input signals of an affect-aware intelligent system.

##### Facial Expressions

Facial expressions are one of the most powerful, natural and universal signals for human beings to convey their emotional states and intentions [Dar15] [LD20]. The face can express emotion sooner than people verbalize or even realize their feelings. While the cultural and ethnic background of a person can affect his expressive style, Ekman indicated that humans perceive certain basic emotions in the same way regardless of their culture [EF71], [Ekm94]. These six universal facial expressions (happiness, sadness, surprise, fear, disgust and anger) constitute the categorical model. Since the thesis concerns emotion recognition through facial expression, further analysis of facial expressions is presented in the next section.

##### Body Gestures

Gestures are also an important form of emotion expression. They include movements of hands, head and other parts of the body allowing individuals to communicate a variety of feelings, thoughts and emotions

[Nor+21]. First, the hands are probably the richest source of body language information. For example, based on the position of hands one is able to determine whether a person is honest (one will turn the hands inside towards the interlocutor) or insincere (hiding hands behind the back). Exercising open-handed gestures during conversation can give the impression of a more reliable person. It is a trick often used in debates and political discussions.

Head positioning also reveals a lot of information about the emotional state. Many research studies indicate that people are prone to talk more if the listener encourages them by nodding. The pace of the nodding can signal patience or lack of it. In a neutral position, the head remains still in front of the interlocutor. If the chin is lifted it may mean that the person is displaying superiority or even arrogance. Exposing the neck might be a signal of submission. In [Dar15] Charles Darwin noted that like animals, people tilt their heads when they are interested in something. That is why women perform this gesture when they are interested in men, an additional display of submission results in greater interest from the opposite sex, e.g. a lowered chin signals a negative or aggressive attitude.

The torso is probably the least communicative part of the body. However, its angle with the body is an indicative attitude. For example, placing the torso frontally to the interlocutor can be considered as a display of aggression. By turning it at a slight angle one may be considered self-confident and devoid of aggression. Leaning forward, especially when combined with nodding and smiling, is the most distinct way to show curiosity. We refer readers to [Nor+21] for a more detailed survey on Emotional Body Gesture Recognition.

## Speech

The signal of speech is one of the fastest and most natural methods of communication between humans. It conveys not only linguistic messages, but also emotional information. For example, the sympathetic arousal associated with an anger state often produces changes in respiration and an increase in muscle tension, which influence the vibration of the vocal folds and vocal tract shape, affecting the acoustic characteristics of the speech, which in turn can be used by the listener to infer the respective state [Sch86]. We refer readers to [EKK11] for a more detailed survey on Speech Emotion Recognition.

## Text

Nowadays communication and human interaction are carried out using social sites and applications in textual form. Emotion is expressed through these texts and analyzed by natural language processing techniques. We refer readers to [AWN20] for more information on text-based emotion recognition.

### 1.1.2 Emotion models

Numerous computational models describing human emotional states have been proposed by the psychology community. However, we have no clear evidence as to which representation is more appropriate for analyzing facial expressions. Therefore, it is important to discuss the benefits and drawbacks of each emotion representation.

#### Categorical model

Classifying emotions into a set of distinct classes that can be recognized and described easily in daily language has been common since at least the time of Darwin. More recently, influenced by the research of Paul Ekman [Ekm94] [EF71] a dominant view upon affect is based on the underlying assumption that humans universally express and recognize a set of discrete primary emotions which include happiness, sadness, fear, anger, disgust, and surprise. Mainly because of its simplicity and its universality claim, the universal primary emotions hypothesis has been by far the first choice for affective computing research and has been extensively exploited. However, restricting human emotion to a predefined set of discrete categories has its drawbacks. Keeping in mind that human emotion is ambiguous and subjective, the categorical model creates large intra-class variations and small inter-class differences.

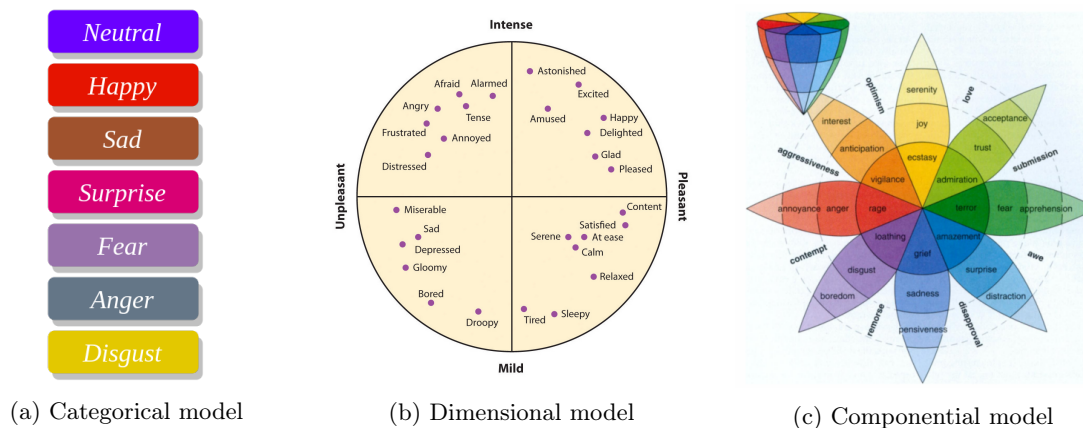


Figure 1.1.1: Ways of modelling emotion in affective computing.

## Dimensional

Another popular approach is to model emotions along a set of latent dimensions [Rus80]. These dimensions include valence (how pleasant or unpleasant a feeling is) activation (how likely is the person to take action under the emotional state) and control (the sense of control over the emotion). Due to their continuous nature, such models can theoretically describe more complex and subtle emotions. Unfortunately, the richness of the space is more difficult to use for automatic recognition systems because it can be challenging to link such described emotion to an expression of affect. This is why, many automatic systems based on dimensional representation of emotion simplified the problem by dividing the space in a limited set of categories like positive vs negative or quadrants of the 2D space.

## Componential models

Somehow in-between categorical and dimensional models in terms of descriptive capacity, componential models of affect, arrange emotions in a hierarchical fashion where each superior layer contains more complex emotions which can be composed of emotions of previous layers. The best example of componential models was proposed by Plutchik [Plu01]. According to his theory, more complex emotions are combinations of pairs of more basic emotions, called dyads. For example, love is considered to be a combination of joy and trust. Primary dyads, e.g. optimism=anticipation+joy, are often felt, secondary dyads, e.g. guilt=joy+fear, are sometimes felt and tertiary dyads, e.g. delight=joy+surprise, are seldom felt. These types of models are rarely used in affective computing literature compared to the previous two but should be taken into consideration due to their effective compromise between ease of interpretation and expressive capacity.

## 1.2 Facial Expressions

Since the purpose of the thesis is to recognize the human emotional state through facial expressions, we start with an introduction to the facial expressions of humans. A facial expression is one or more motions or positions of the muscles beneath the skin of the face. Humans can adopt a facial expression voluntarily or involuntarily, and the neural mechanisms responsible for controlling the expression differ in each case. Voluntary facial expressions are often socially conditioned and follow a cortical route in the brain. Conversely, involuntary facial expressions are believed to be innate and follow a subcortical route in the brain.

### 1.2.1 Universality

A vital issue in emotion analysis is the universality of facial expressions. Ekman et al. [Ekman+87] discovered strong evidence of the universality of some facial expressions of emotion as well as why expressions may appear differently across cultures. Through continued cross-cultural studies, they noticed that many of the apparent differences in facial expressions across cultures were due to context. To describe this phenomenon,













Upper Face Action Units					
AU 1	AU 2	AU 4	AU 5	AU 6	AU 7
					
<b>Inner Brow Raiser</b>	<b>Outer Brow Raiser</b>	<b>Brow Lowerer</b>	<b>Upper Lid Raiser</b>	<b>Cheek Raiser</b>	<b>Lid Tightener</b>
<b>*AU 41</b>	<b>*AU 42</b>	<b>*AU 43</b>	<b>AU 44</b>	<b>AU 45</b>	<b>AU 46</b>
					
<b>Lid Droop</b>	<b>Slit</b>	<b>Eyes Closed</b>	<b>Squint</b>	<b>Blink</b>	<b>Wink</b>

Figure 1.2.1: Upper face AUs.

they coined the term display rules: rules we learn in the course of growing up about when, how, and to whom it is appropriate to show our emotional expressions.

## 1.2.2 Other related tasks

Besides FER, there are other facial tasks that are similar and usually share similar learning techniques. Here, we present some of them in order to introduce the reader to the variety of facial tasks that can be combined with our task in many cases.

### Action Units Detection

Instead of recognizing the meaning that underlies a displayed behavior, another representation of facial behavior is the FACS. FACS is a comprehensive and anatomical system that could encode various facial movements by the combination of basic AUs and makes the emotion categories much wider. AUs define certain facial configurations caused by contraction of one or more facial muscles, and they are independent of the interpretation of emotions [ZLZ20]. Over the years many researchers have focused on facial behavior analysis based on automatic AUs recognition. FACS consists of 44 action units. Thirty are anatomically related to the contraction of a specific set of facial muscles. The anatomic basis of the remaining 14 is unspecified. These 14 are referred to in FACS as miscellaneous actions. Many action units may be coded as symmetrical or asymmetrical. For action units that vary in intensity, a 5-point ordinal scale is used to measure the degree of muscle contraction. Figure 1.2.1 shows some examples of AUs.

### Facial expression synthesis

An important research area of affect analysis is facial expression synthesis where the goal is to generate face images with specific expressions for a specified human subject. It has drawn much attention in the field of computer graphics, computer vision and pattern recognition. Synthesizing photo-realistic facial expression images has been of great value for both academic and industrial communities, and has been widely applied in facial animations, face editing, face data augmentation and face recognition [Son+18]. During the last two decades, many facial expression synthesis methods have been proposed, which can be roughly divided into two categories. The first category mainly resorts to computer graphics technique to directly warp input faces to target expressions [Yan+12] [Yeh+16] [Zha+05] or re-use sample patches of existing images [MPK09], while the other aims to build generative models to synthesize images with predefined attributes [DSC18] [Sus+08].

For the first category, a lot of research efforts have been devoted to finding correspondence between existing facial textures and target images. Earlier approaches usually generate new expressions by creating fully textured 3D facial models [Bla+03] [MPK09], warping face images via feature correspondence [The+09] and optical flow [Yan+12], or compositing face patches from an existing expression dataset [Li+13]. Although this kind of methods can usually produce realistic images with high resolution, their elaborated yet complex processes often result in expensive computation. The representative methods in the second category are deep



generative models that have recently obtained impressive results for image synthesis applications [Iso+17] [Hua+17b] [Zhu+17] [KZ20a] [Son+18]. However, images generated by such methods sometimes lack fine details and tend to be blurry or of low resolution. Besides, target expression attributes are usually encoded in a latent feature space, where certain directions are aligned with semantic properties.

### Facial micro-expression recognition

A less explored research area is facial micro-expression recognition. Micro-expressions are brief and involuntary rapid facial emotions that are elicited to hide a certain true emotion [EF69]. They usually last between 1/5 to 1/25 of a second and occur in only specific parts of the face [EF71]. Aside from the short duration of micro-expressions, they also possess low intensity. There is a vast range of applications that can benefit from the study of micro-expressions [Tak+18]. A primary reason for the strong interest in micro-expressions is that it proves to be an important clue for lie detection. For example, in situations when the suspects are being questioned, a micro-expression fleeting across the face can tell the police that the criminal is pretending to be innocent. It can also benefit the border security officers for identifying suspicious behavior of the individuals during usual interviews of checking for potential dangers. In the study of psychotherapy, micro-expressions have been proved very helpful in understanding the genuine emotions of the patients. Micro-expression recognition systems are sometimes also used as an additional module for user authentication. In other fields, such as marketing, distance learning, and many more, micro-expressions can be used as recognition to reflect human reactions and feedback to advertisements, products, services and learning materials.

### Face analysis

Face analysis is a challenging and actively researched problem with applications to face recognition, emotion analysis, biometrics, security, etc. It includes tasks like face alignment, head-pose estimation, gender and smile recognition and its techniques can many times be used in our task. All these tasks have been approached either as separate problems [Che+16b] [WBF19] or using multi-task architectures [Ran+17] [RPC17].

## 1.3 Applications

An essential part of every research area is its application to real-world problems. A system that can effectively recognize facial expressions has useful applications in many domains. Some of them are briefly mentioned here.

### 1.3.1 Human-Robot Interaction

An important challenge in the field of human-robot interaction is the possibility to endow robots with emotional intelligence in order to make the interaction more intuitive, genuine, and natural [SPR20]. However, the presence itself of a robot represents a bias in the recognition task since the robot presence, embodiment, and behavior could affect empathy [Kwa+13], elicit emotions [Guo+19] [SN19] and impact experience [Cam+15]. Also, most FER datasets are not suited for emotion recognition in real settings since the visual field of view of the robot may not be aligned to the images stored in the dataset and robot movements may even occlude its field of view.

### 1.3.2 Digital Entertainment

Emotion recognition enables the creation of personalized and more interactive forms of digital entertainment. In the field of game development, FER can be used to generate game spaces (i.e. levels) such that the spaces optimize player challenge for the individual player without explicitly asking for player feedback during gameplay [Blo+14].

### 1.3.3 Health Care

Since human facial expressions change with different states of health a FER system can be beneficial to a healthcare framework. There are several works in the literature related to expression recognition systems for healthcare in smart cities [Muh+17] [Alh16]. Dantcheva et al. [Dan+16] were motivated by the growing

number of elderly people worldwide and a need to provide an improved healthcare service for them. They developed an approach for detecting facial expressions of Alzheimer’s disease patients. Mano et al. [Man+16] explored the deployment of the embedded computing in Health Smart Homes to improve in-home healthcare using the Internet of Things.

### 1.3.4 Advertisement

Face has been shown to display discriminative valence information. Greater zygomatic major muscle (occurring in smiles) activity is observed during ads with a positive emotional tone and greater corrugator muscle (brow furrow) activity is observed during ads with negative emotional tone [BLP01]. Therefore, FER would be beneficial in understanding the relationship between emotional responses to content and measures of advertising effectiveness. Over the previous years many researchers have analysed facial responses to video content and their relationship to marketing effectiveness [McD+14] [MS17] [OMT18].

### 1.3.5 Education

In e-learning environments, a FER system can help teachers to identify whether students understand the teaching content based on students’ different expressions and can adjust their teaching programs. Also, using emotions in software systems for e-learning would considerably increase performance if the software could adapt to the emotional state of the learner. Hence, many models have been proposed for FER in virtual learning environments by using webcams [Yan+18] [BNW16].

## 1.4 Challenges

In order to develop robust systems that can recognize facial expressions under real-world conditions, we should first consider the challenges that our task presents (Figure 1.4.1).

1. **Variations:** A central challenge in FER is to disentangle the various factors of variation that are present in an image.
  - *Identity.* A facial image contains a lot of information that is unrelated to the depicted expression and has to do with the personal characteristics of the person. For example, accessories, hairs or a beard are regarded as noise from the system and can hurt its total performance.
  - *Head pose.* The position of the head can vary drastically in real-world environments while most FER datasets are composed mostly by frontal images.
  - *Illumination.* Depending on the location of the source of light with respect to the camera and the captured face, facial images are seen with different illumination patterns overlaid on top of the image of the face.
  - *Intra-class.* A single emotion can be expressed through various ways. This inherent characteristic of the task leads to a large intra-class and small inter-class variation of our learned features. As shown in Figure 1.4.1d, surprise can be expressed through ways that are much different from each other (open and close mouth).
2. **Occlusion:** The face region may be occluded by hair, eyeglasses and clothes such as scarf or handkerchief.
3. **Subjectivity:** Annotating a facial expression dataset is imposed to subjectivity since there are expressions that do not correspond to only a single type of emotion. For example, an image of a man that widely opens his mouth and his eyes could be classified either as an expression of surprise or fear since both emotions can be expressed through similar facial characteristics (Fig. 1.4.1f).

## 1.5 Deep Learning

To implement accurate affect-aware systems in the wild deep learning methods are employed. Therefore, we first introduce the basic concepts of deep learning and the deep architectures that will be used in the thesis.

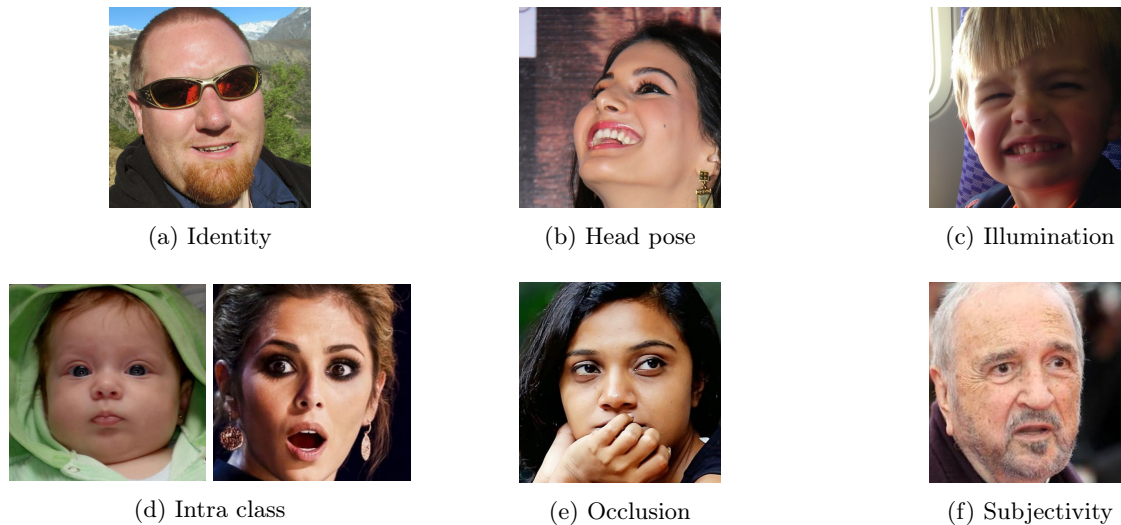


Figure 1.4.1: Illustration of FER challenges using samples of AffectNet.

### 1.5.1 Machine Learning

A machine learning algorithm is an algorithm that is able to learn from data. According to Mitchell et al. [Mit+97] “A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ ”. More formally, the goal of a machine learning model is to learn a set of parameters using a set of input data points. While there are many kinds of tasks that can be solved with machine learning, we present the ones that will be used in the thesis:

- **Classification:** In this type of task, the model is asked to specify which of  $k$  categories some input belongs to. To solve this task, the learning algorithm is usually asked to produce a function  $f : \mathbb{R}^n \rightarrow \{1, \dots, k\}$ . When  $y = f(x)$ , the model assigns an input described by vector  $x$  to a category identified by numeric code  $y$ . There are other variants of the classification task, for example, where  $f$  outputs a probability distribution over classes. An example of a classification task is object recognition, where the input is an image (usually described as a set of pixel brightness values), and the output is a numeric code identifying the object in the image.
- **Regression:** In this type of task, the computer program is asked to predict a numerical value given some input. To solve this task, the learning algorithm is asked to output a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . This type of task is similar to classification, except that the format of the output is different. An example of a regression task is the prediction of the expected claim amount that an insured person will make (used to set insurance premiums) or the prediction of future prices of securities.

### 1.5.2 From Machine Learning to Deep Learning

Conventional machine learning techniques are limited in their ability to process natural data in its raw form. For decades, constructing a pattern recognition or machine learning system required careful engineering and considerable domain expertise to design a feature extractor that transformed the raw data (such as the pixel values of an image) into a suitable internal representation or feature vector from which the learning subsystem, often a classifier, could detect or classify patterns in the input.

Representation learning is a set of methods that allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep learning methods are representation learning methods with multiple levels of representation, obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. With the composition of enough such transformations, very complex functions can be learned. For classification tasks, higher layers of representation amplify aspects of the input that are important for discrimination and suppress irrelevant variations. An image, for example, comes in

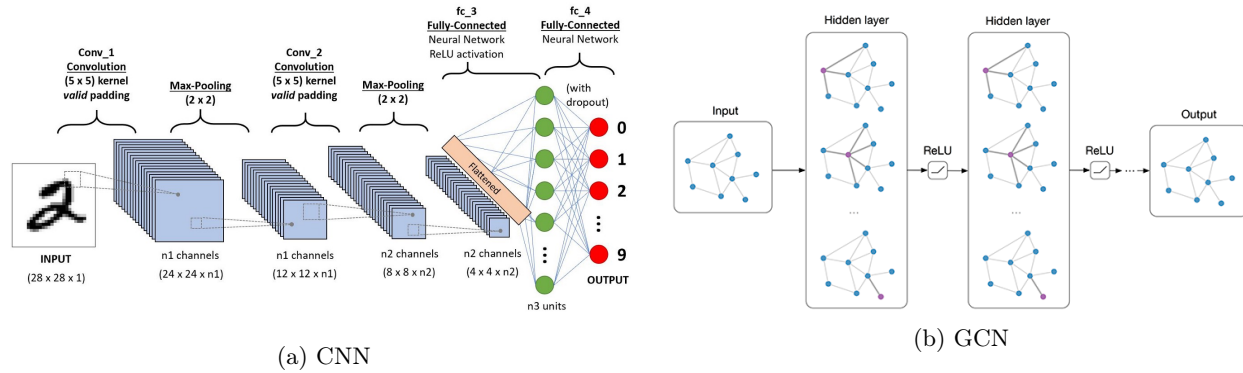


Figure 1.5.1: Deep architectures that are used in the thesis.

the form of an array of pixel values, and the learned features in the first layer of representation typically represent the presence or absence of edges at particular orientations and locations in the image. The second layer typically detects motifs by spotting particular arrangements of edges, regardless of small variations in the edge positions. The third layer may assemble motifs into larger combinations that correspond to parts of familiar objects, and subsequent layers would detect objects as combinations of these parts. The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure. For more information about deep learning, we refer readers to [LBH15] and [Goo+16].

### 1.5.3 Deep learning architectures

Over the years many deep learning architectures have been proposed by the research community. Here we briefly present the deep learning architectures that are used in the thesis for feature extraction (Figure 1.5.1).

#### Convolutional Neural Networks

CNN architectures are probably the most well-known deep learning models used to solve computer vision tasks, in particular image classification. A typical architecture consists of repetitions of a stack of several convolution layers and a pooling layer, followed by one or more fully connected layers.

- **Convolutional Layer:** A convolution layer is a fundamental component of the CNN architecture that performs feature extraction, which typically consists of a combination of linear and nonlinear operations, i.e., convolution operation and activation function.
- **Pooling Layer:** A pooling layer provides a typical downsampling operation that reduces the in-plane dimensionality of the feature maps in order to introduce a translation invariance to small shifts and distortions, and decrease the number of subsequent learnable parameters. It is of note that there is no learnable parameter in any of the pooling layers, whereas filter size, stride, and padding are hyperparameters in pooling operations, similar to convolution operations.
- **Fully Connected Layer:** The output feature maps of the final convolution or pooling layer is typically flattened, i.e., transformed into a one-dimensional (1D) array of numbers (or vector), and connected to one or more fully connected layers, also known as dense layers, in which every input is connected to every output by a learnable weight. Once the features extracted by the convolution layers and downsampled by the pooling layers are created, they are mapped by a subset of fully connected layers to the final outputs of the network, such as the probabilities for each class in classification tasks. The final fully connected layer typically has the same number of output nodes as the number of classes. Each fully connected layer is followed by a nonlinear function, such as ReLU.

## Graph Convolutional Networks

A lot of research has been done lately towards generalizing neural networks to work on arbitrarily structured graphs. Generally, the goal of a GCN is to learn a function  $f$  on a graph  $G = (V, E)$  that takes as input a feature description for each node of the graph  $H^l \in \mathbb{R}^{n \times d}$  ( $n = |V|$ ) and a correlation matrix  $A \in \mathbb{R}^{n \times n}$  and produces new node-level features  $H^{l+1} \in \mathbb{R}^{n \times d'}$ . The update rule is formulated as follows:

$$H^{l+1} = h(\hat{A}H^lW^l) \quad (1.5.1)$$

where  $W^l \in \mathbb{R}^{d \times d'}$  is the weight matrix for the  $l$ -th GCN layer,  $\hat{A}$  is the normalized version of matrix  $A$  such that all rows sum to one and  $h(\cdot)$  is LeakyRELU [MHN13]. For more information on GCN we refer readers to [KW17].

## 1.6 Other related areas

In our proposed models we combine ideas from various research areas. Here we make a brief introduction to some of these research areas.

### 1.6.1 Multi-task Learning

Human can learn multiple tasks simultaneously and during this learning process, human can use the knowledge learned in a task to help the learning of another task. For example, according to our experience in learning to play tennis and squash together, we find that the skill of play tennis can help learn to play squash and vice versa. Inspired by such human learning ability, MTL [Car97], a learning paradigm in machine learning, aims to jointly learn multiple related tasks so that the knowledge contained in a task can be leveraged by other tasks with the hope of improving the generalization performance of all the tasks at hand [ZY21]. One reason that MTL is effective is that it utilizes more data from different learning tasks when compared with single-task learning. With more data, MTL can learn more robust and universal representations for multiple tasks and more powerful models, leading to better knowledge sharing among tasks, better performance of each task, and a low risk of overfitting in each task. For more information on MTL we refer readers to [ZY21].

### 1.6.2 Metric Learning

Metric learning is an approach based directly on a distance metric that aims to establish similarity or dissimilarity between objects. While metric learning aims to reduce the distance between similar objects, it also aims to increase the distance between dissimilar objects [KB19]. In supervised metric learning, we seek an appropriate metric by formulating an optimization objective function to exploit supervised information of the training samples, where the objective functions are designed for different specific tasks [LHZ17]. However, most conventional metric learning methods usually learn a linear mapping to project samples into a new feature space, which suffer from the nonlinear relationship of data points in metric learning. While the kernel trick can be adopted to address this nonlinearity problem, this type of method suffers from the scalability problem because the kernel trick has two major issues: 1) choosing a kernel is typically difficult and quite empirical and 2) the expression power of kernel functions is often not flexible enough to capture the nonlinearity in the data.

Motivated by the fact that deep learning is an effective solution to model the nonlinearity of samples, several deep metric learning methods have been proposed in recent years. The key idea of deep metric learning is to explicitly learn a set of hierarchical nonlinear transformations to map data points into other feature space for comparing or matching by exploiting the architecture of neural networks in deep learning, which unifies feature learning and metric learning into a joint learning framework. For more information on deep metric learning for computer vision tasks we refer readers to [LHZ17], [KB19].

### 1.6.3 Support Vector Machines

A SVM is a statistical learning method based on the structural risk minimization principle [BGV92]. It uses the concept of decision planes that utilize decision boundaries to optimally separate data into different

categories. More formally, a SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier.

In deep learning, SVM is a widely used alternative to softmax for classification. Using SVMs in combination with convolutional nets have been proposed in the past as part of a multistage process. In particular, a deep convolutional net is first trained using supervised/unsupervised objectives to learn good invariant hidden latent representations. The corresponding hidden variables of data samples are then treated as input and fed into linear (or kernel) SVMs.

## 1.7 Contributions

The main contributions of the thesis are summarized below:

1. We present an overview of the latest methods on FER in-the-wild discussing their novelties and limitations.
2. We implement a baseline deep learning model focusing mainly on the data preprocessing and the training pipeline.
3. We investigate the types of variations that are connected to the task and use metric learning techniques to reduce the impact of these variations.
4. We explore the relationship between the categorical and the dimensional emotion representation and train multi-task learning networks to exploit their emotional dependencies.
5. Inspired by recent work in multi-label image recognition we propose Emotion-GCN, a novel multi-task learning framework that uses a GCN to capture the emotional dependencies.
6. We train and evaluate all our methods under real-world settings using AffectNet dataset, the largest in-the-wild database of facial expressions.

A part of our work has been accepted to the IEEE International Conference on Automatic Face and Gesture Recognition 2021 (FG 2021) with the authors being Panagiotis Antoniadis, Panagiotis Paraskevas Filntisis and Petros Maragos [AFM21]. Also, the same authors along with Ioannis Pikoulis participated in ICCV ABAW2 competition where we leveraged facial, bodily and context information to recognize emotion in real-world videos [Ant+21].

## 1.8 Structure of the thesis

The thesis consists of the following chapters:

- In chapter 1 we introduce the reader to the research areas that are related to the thesis making a brief overview of each area without getting into deeper analysis.
- In chapter 2 we present an overview of the related work on FER in-the-wild from the preprocessing steps and the datasets to the deep learning methods on the categorical and the dimensional model of affect.
- In chapter 3 we present a baseline deep learning model for FER in-the-wild focusing mainly on the data preprocessing and the training pipeline.
- In chapter 4 we explore metric learning techniques.
- In chapter 5 we explore multi-task learning techniques.
- In chapter 6 we propose our novel Emotion-GCN model.
- In chapter 7 we conclude the thesis.

# Chapter 2

## Literature Review

In this chapter we present an overview of the previous work on FER. We start by discussing the standard preprocessing steps and then we proceed to the previously proposed techniques for recognizing facial expressions in the categorical and the dimensional model of affect. Finally, we present the facial databases and the facial analysis toolkits that are available.

### 2.1 Preprocessing

A necessary stage in every FER system is preprocessing where the face region of the person is detected, cropped and processed in order to reduce the impact of factors that may hurt the overall performance of the system. Here we present an overview of the methods that are used in the literature regarding the preprocessing stage.

#### 2.1.1 Face Detection

The first step is to detect the face and remove the background and non-face areas of the image. It can be regarded as a specific case of object detection, where the task is to find the location and sizes of all objects in an image that belongs to a given class. However, different from generic object detection, face detection features smaller variations in the aspect ratio, but much larger variations in scale (from several pixels to thousand pixels). Early face detection efforts were mainly based on the classical approach, in which hand-crafted features were extracted from the image and were fed into a classifier to detect likely face regions. Two landmark classical works for face detection are the Haar Cascades classifier [VJ01] and the Histogram of Oriented Gradients followed by SVM [DT05]. With the great success of deep learning in computer vision, researchers proposed several promising model architectures over the past years. Based on their main technical contributions to face detection, these works can be organized in the following five categories according to Minaee et al. [Min+21].

- Cascade-CNN Based Models: Li et al. [Li+15] proposed one of the early deep models for face detection based on a CNN cascade. The proposed CNN cascade operates at multiple resolutions, quickly rejects the background regions in the fast low resolution stages, and carefully evaluates a small number of candidates in the last high resolution stage. To improve localization effectiveness and reduce the number of candidates at later stages, they introduce a CNN-based calibration stage after each of the detection stages in the cascade. Many works that also rely on cascaded CNN architectures followed [Zha+16] [Zha+17b] [Qin+16].
- R-CNN and Faster-RCNN Based Models: Region proposal-based CNN models have been very successful for object detection, and have also been applied to face detection by several works. Chen et al. [Che+16a] proposed Supervised Transformer Network where the first stage is a multi-task Region Proposal Network, which simultaneously predicts candidate face regions along with associated facial landmarks. The candidate regions are then warped by mapping the detected facial landmarks to their

Table 2.1: Different types of face alignment (Source: [LD20]).

	Type	Points	Real-time
Holistic	AAM [CET01]	68	×
	MoT [ZR12]	39/68	×
Part-based	DRMF [Ast+13]	66	×
	SDM [XD13]	49	✓
Cascaded regression	3000 fps [Ren+14]	68	✓
	Incremental [Ast+14]	49	✓
Deep learning	cascaded CNN [SWT13]	5	✓
	MTCNN [Zha+16]	5	✓

canonical positions to better normalize the face patterns. The second stage, which is a R-CNN, then verifies if the warped candidate regions are valid faces or not. In many studies [JL17] [Wan+17] [SWH18], faster R-CNN model [Ren+15] has been effectively used for face detection.

- **Single Shot Detector Models:** Single stage detection (SSD) is another popular and major direction in deep learning based face detection. Unlike two-stage proposal-classification detectors, such as R-CNN models, SSD detects faces in a single stage directly from the early convolutional layers in a classification network. Some examples are SSH [Naj+17], S3FD [Zha+17d], Faceboxes [Zha+17c], RefineFace [Zha+20b] and YOLO-face [Che+20].
- **Feature Pyramid Network Based Models:** A feature pyramid is a neural network structure which combines semantically weak features with semantically strong features using skip-connections [Lin+17a]. Inspired by the Feature Pyramid Network many networks for face detection have been proposed like FANet [Zha+20a], PyramidBox [Tan+18] and Selective Refinement Network [Chi+19].
- **Other models:** There are many proposed networks that do not fall into any of the above categories like HyperFace [RPC17] and Faceness-Net [Yan+17].

## 2.1.2 Face Alignment

The next step in our preprocessing stage is to perform face alignment in order to obtain a canonical alignment of the face based on translation, scale and rotation. Some methods try to impose a (pre-defined) 3D model and then apply a transform to the input image such that the landmarks on the input face match the landmarks on the 3D model. Other, more simplistic methods rely only on the facial landmarks themselves to obtain a normalized rotation, translation, and scale representation of the face. In Table 2.1 a summary of different types of face alignment is presented.

## 2.1.3 Face Normalization

Changes in illumination and head pose can introduce large variations in images and hurt the performance. Therefore, illumination and pose normalization methods are employed to ameliorate the influence of these variations. Illumination and contrast can vary in different images even from the same person with the same expression, especially in unconstrained environments, which can result in large intra-class variances. Choi et al. [CKR16] used a DCNN model to eliminate the illumination effect and maximize the discriminative power for feature representation. Thakare and Thakare [TT11] used a fuzzy-neural network to deal with depth information of face images for feature matching. Some studies have employed pose normalization techniques to yield frontal facial views for FER. Face Identity-Preserving [Zhu+13], Multi-View Perceptron [Zhu+14] and Controlled Pose Feature [Yim+15] are three methods that can be used to handle both pose variations and illumination changes

## 2.1.4 Data Augmentation

Deep neural networks require sufficient training data to ensure generalizability to a given recognition task. In order to train a deep learning model in a more diverse dataset, data augmentation techniques are used



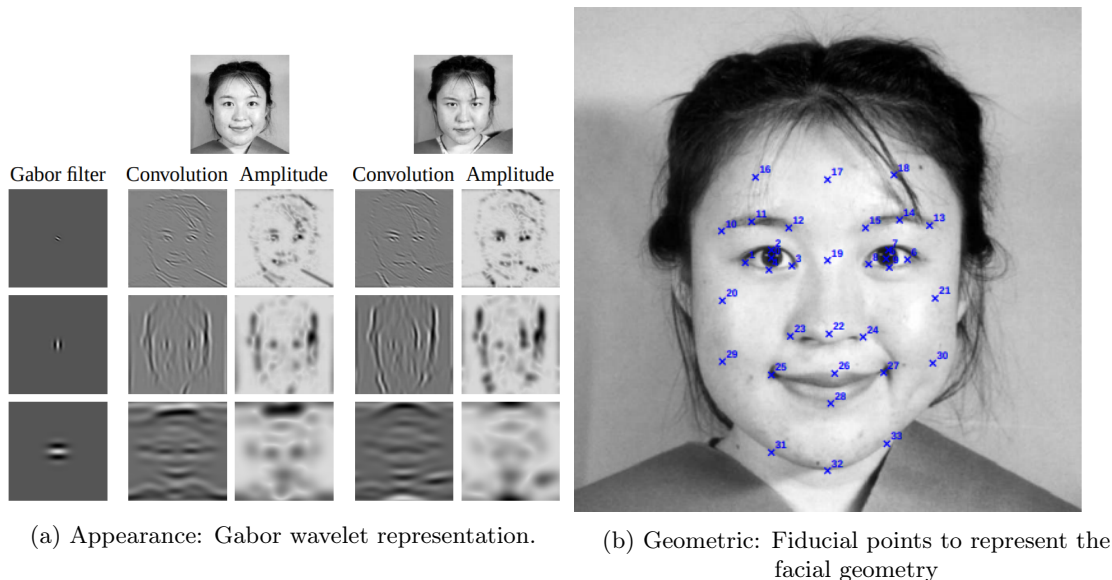


Figure 2.2.1: Handcrafted features on the categorical model of affect (Source: [Zha+98]).

that are divided in two categories:

- On-the-fly: These methods are applied in the input sample during the training step like random cropping, horizontal flip, random rotation, etc.
- Offline: The most frequently used offline operations include random perturbations and transforms, e.g. rotation, shifting, skew, scaling, noise, contrast and color jittering. For example, common noise models, salt & pepper and speckle noise [Pit+17] and Gaussian noise [Lop+17] are employed to enlarge the data size. Also, GANs [Goo+14] can also be applied to augment data by generating diverse appearances varying in poses and expressions.

## 2.2 FER on the categorical model

Every FER system consists of three parts: face preprocessing, feature learning and classification. The first step was discussed in the previous section. Here we present an overview of the methods that are used for feature learning and classification on the categorical model of affect.

### 2.2.1 Handcrafted features

The early works on FER are mostly based on handcrafted features that are either geometric or appearance-based. Geometric features present the shape and locations of facial components (including mouth, eyes, brows, nose). The facial components or facial feature points are extracted to form a feature vector that represents the face geometry [EP97] [PR00]. The appearance features present the appearance (skin texture) changes of the face, such as wrinkles and furrows. Gabor wavelets are widely used to extract the facial appearance changes as a set of multiscale and multiorientation coefficients. The Gabor filter may be applied to specific locations on a face [Lyo+98] or to the whole face image [Bar+01]. Also, LBP has been used for FER by analyzing the contrast within sub-regions of an image [ZP07]. In the standard configuration, a pixel is compared with the eight neighboring pixels. This yields a binary pattern of 8bit. The LBP descriptor can be stored as a histogram. Each bin of the histogram corresponds to one binary pattern configuration that represents a facial feature. In this way, a 256-dimensional descriptor is obtained. A generalization of appearance features across different persons is not trivial. This is one of the major drawbacks of appearance-based approaches. Zhang et al. [Zha+98] was the first to compare two types of features to recognize expressions, the geometric positions of 34 fiducial points on a face and 612 Gabor wavelet coefficients extracted from the face image at these 34 fiducial points (Figure 2.2.1). The recognition rates for six emotion-specified expressions were significantly

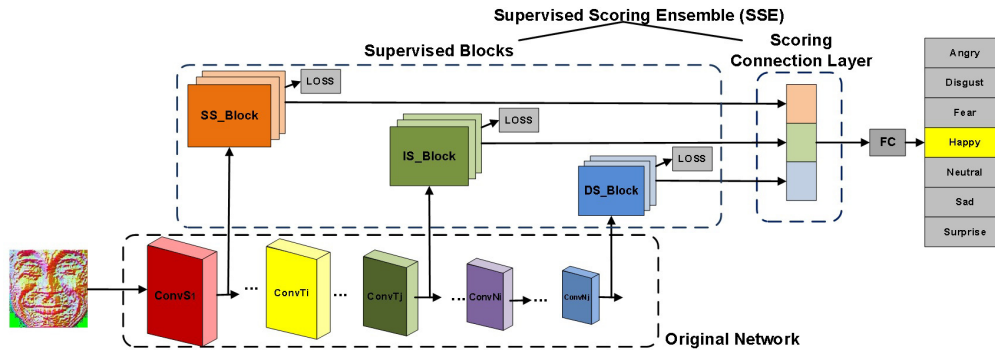


Figure 2.2.2: Supervised Scoring Ensemble (Source: [Hu+17]).

higher for Gabor wavelet coefficients. Donato et al. [Don+99] compared several techniques for recognizing six single upper face AUs and six lower face AUs. These techniques included optical flow, principal component analysis, independent component analysis, local feature analysis and Gabor wavelet representation. The best performances were obtained using a Gabor wavelet representation and independent component analysis. All in all, while there is a lot of intuition behind handcrafted features and their performance on several lab-controlled databases is impressive, they lack generalizability and sufficient learning capacity.

### 2.2.2 Deep features

Due to their success in the field of computer vision, deep learning methods have also been applied in FER and achieve impressive performance in very challenging scenarios. Deep learning attempts to capture high-level abstractions through hierarchical architectures of multiple nonlinear transformations and representations. Since a large volume of studies have been conducted for expression recognition tasks based on static images we divide these works in the following categories based on [LD20].

#### Auxiliary blocks & layers

Based on the foundation architecture of CNN, several studies have proposed the addition of well-designed auxiliary blocks or layers to enhance the expression-related representation capability of learned features. Yao et al. [Yao+16] designed Holonet where CReLU [Sha+16] was combined with the powerful residual structure [He+16] to increase the network depth without efficiency reduction and an inception-residual block [Sze+16] was uniquely designed for FER to learn multi-scale features to capture variations in expressions. Hu et al. [Hu+17] proposed Supervised Scoring Ensemble (Figure 2.2.2) where three types of supervised blocks were embedded in the early hidden layers of the mainstream CNN for shallow, intermediate and deep supervision, respectively. Zhao et al. [Zha+18c] proposed Feature Selection Network that automatically extracts and filters facial features by embedding a feature selection mechanism inside the AlexNet. The designed feature selection mechanism effectively filters irrelevant features and emphasizes correlated features according to learned feature maps. Interestingly, Zeng et al. [ZSC18] pointed out that the inconsistent annotations among different FER databases are inevitable which would damage the performance when the training set is enlarged by merging multiple datasets. To address this problem, they proposed IPA2LT framework where an end-to-end trainable LTNet is designed to discover the latent truths from the human annotations and the machine annotations trained from different datasets by maximizing the log-likelihood of these inconsistent annotations.

#### Loss layers

The traditional softmax loss layer in CNNs simply forces features of different classes to remain apart, but FER in real-world scenarios suffers from not only high inter-class similarity but also high intra-class variation. Therefore, several works have proposed novel loss layers for FER. Based on the center loss by Wen et al. [Wen+16] that minimizes the euclidean distance between deep features and their corresponding class centers many task-specific losses have been proposed for FER. Cai et al. [Cai+18] designed island loss to reduce the intra-class variations while enlarging the inter-class differences simultaneously (Figure 2.2.3). Specifically, the

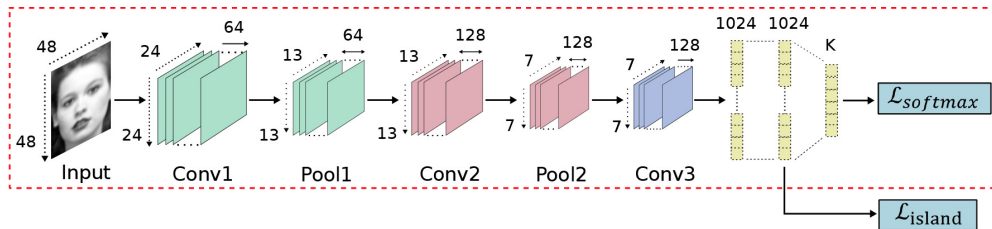


Figure 2.2.3: Island loss layer (Source: [Cai+18]).

loss pulls the samples towards their corresponding class centers to achieve intra-class compactness and at the same time, pushes the centers away from each other to make the clusters as isolated “islands”. Similarly, Li et al. [LDD17] proposed Locality-preserving loss that pulls the locally neighboring features of the same class together so that the intra-class local clusters of each class are compact. Also, some losses employ positive and negative samples to learn discriminative features inspired by the triplet loss [SKP15]. Guo et al. [Guo+16] designed exponential triplet-based loss that gives difficult samples more weight when updating the network and Liu et al. [Liu+17] proposed (N+M)-tuples cluster loss that alleviates the difficulty of anchor selection and threshold validation in the triplet loss.

### Network ensemble

A common technique that improves the generalization performance of a single network is to combine multiple networks in a network ensemble. Two key factors should be considered when implementing network ensembles: (1) sufficient diversity of the networks to ensure complementarity and (2) an appropriate ensemble method that can effectively aggregate the committee networks. To achieve the first goal, different kinds of training data and various network parameters or architectures are considered to generate diverse committees. By changing the size of filters, the number of neurons and the number of layers of the networks, and applying multiple random seeds for weight initialization, the diversity of the networks can be enhanced [Kim+15]. Also, different architectures of networks can be used to enhance the diversity like in [HBW15] where Hamester et al. trained a CNN in a supervised way and a convolutional autoencoder in an unsupervised way. Regarding the second factor, the networks of the ensemble can be combined either at a feature level or a decision level. For feature-level ensembles, the most commonly adopted strategy is to concatenate features learned from different networks. Bargal et al. [Bar+16] learned features from different deep networks and concatenated them to obtain a single feature vector that described the input image (Figure 2.2.4). For decision-level ensembles, three widely-used rules are applied:

1. Majority voting that determines the class with the most votes using the predicted label yielded from each individual.
2. Simple average that determines the class with the highest mean score using the posterior class probabilities yielded from each individual with the same weight.
3. Weighted average that determines the class with the highest weighted mean score using the posterior

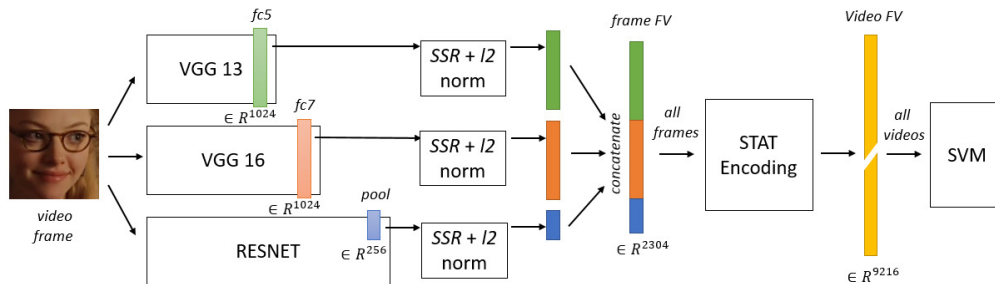


Figure 2.2.4: Feature-level ensemble (Source: [Bar+16]).

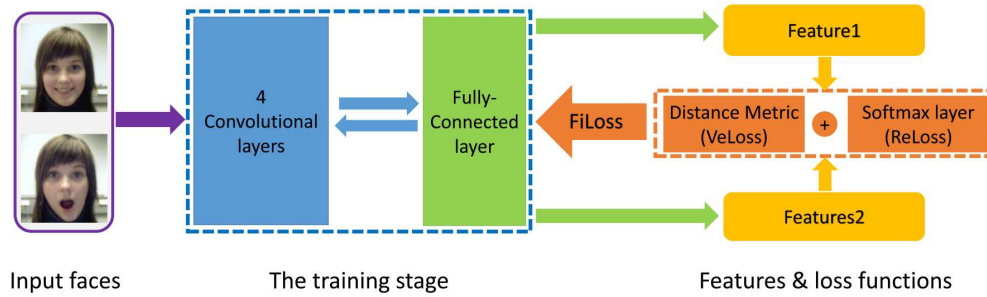


Figure 2.2.5: MSCNN model (Source: [Zha+17a]).

class probabilities yielded from each individual with different weights.

The most widely used technique is weighted average where many methods have been proposed to find an optimal set of weights. Kahou et al. [Kah+13] proposed a random search method to weight the model predictions for each emotion type. Yu et al. [YZ15] used the log-likelihood loss and hinge loss to adaptively assign different weights to each network. Kim et al. [Kim+15] proposed an exponentially weighted average based on the validation accuracy to emphasize qualified individuals. Pons et al. [PM17] used a CNN to learn weights for each individual model.

### Multi-task networks

Many existing networks for FER focus on a single task and learn features that are sensitive to expressions without considering interactions among other latent factors. However, in the real world, FER is intertwined with various factors, such as head pose, illumination, and subject identity (facial morphology). To solve this problem, multi-task learning is introduced to transfer knowledge from other relevant tasks and to disentangle nuisance factors.

Many works suggested that simultaneously conducted FER with other tasks, such as facial landmark localization and facial AUs detection, can jointly improve FER performance [PM18] [DBT14]. Also, many works employed multi-task learning for identity-invariant FER. Meng et al. [Men+17] proposed IACNN that contains two identical sub-CNNs. One stream uses expression-sensitive contrastive loss to learn expression-discriminative features, and the other stream uses identity-sensitive contrastive loss to learn identity-related features for identity-invariant FER. Zhang et al. [Zha+17a] designed MSCNN that was trained under the supervision of both FER and face verification tasks and forces the model to focus on expression information (Figure 2.2.5). Furthermore, Ranjan et al. [Ran+17] proposed an all-in-one CNN model to simultaneously solve a diverse set of face analysis tasks including smile detection. The network was first initialized using the weights pre-trained on face recognition, then task-specific sub-networks were branched out from different layers with domain-based regularization by training on multiple datasets. Specifically, as smile detection is a subject-independent task that relies more on local information available from the lower layers, the authors proposed to fuse the lower convolutional layers to form a generic representation for smile detection. Conventional supervised multi-task learning requires training samples labeled for all tasks. To relax this, Zhang et al. [Zha+18b] proposed a novel attribute propagation method which can leverage the inherent correspondences between facial expression and other heterogeneous attributes despite the disparate distributions of different datasets.

### Cascaded networks

In a cascaded network, various modules for different tasks are combined sequentially to construct a deeper network, where the outputs of the former modules are utilized by the latter modules. Related studies have proposed combinations of different structures to learn a hierarchy of features through which factors of variation that are unrelated to expressions can be gradually filtered out.

Most commonly, different networks or learning methods are combined sequentially and individually, and each of them contributes differently and hierarchically. Rifai et al. [Rif+12] proposed a multiscale contractive convolutional network to obtain local-translation-invariant representations. Then, the contractive autoencoder

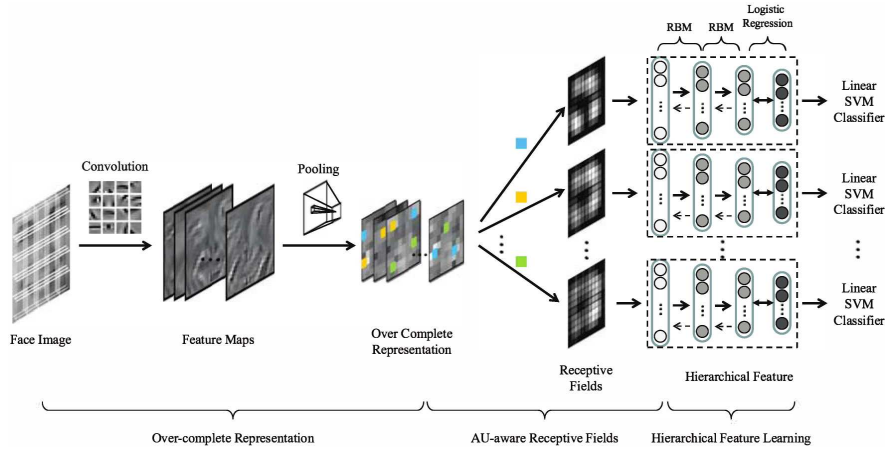


Figure 2.2.6: AUDN model (Source: [Liu+13]).

was designed to hierarchically separate out the emotion-related factors from subject identity and pose. Liu et al. [Liu+13] first learned over-complete representations using a CNN architecture and then exploited a multilayer RBM to learn higher-level features for FER (Figure 2.2.6).

### Generative adversarial network

Recently, GAN-based methods have been successfully used in image synthesis to generate impressively realistic faces, numbers, and a variety of other image types, which are beneficial to training data augmentation and the corresponding recognition tasks. Several works have proposed novel GAN-based models for pose-invariant FER and identity-invariant FER. For pose-invariant FER, Lai et al. [LL18] proposed a GAN-based face frontalization framework, where the generator frontalizes input face images while preserving the identity and expression characteristics and the discriminator distinguishes the real images from the generated frontal face images. Zhang et al. [Zha+18a] proposed a GAN-based model that generates images with different expressions under arbitrary poses for multi-view FER. For identity-invariant FER, Yang et al. [YZY18] proposed an Identity-Adaptive Generation model with two parts. The upper part generates images of the same subject with different expressions using cGANs. Then, the lower part conducts FER for each single identity sub-space without involving other individuals, thus identity variations can be well alleviated. Chen et al. [CKI18] proposed a Privacy-Preserving Representation-Learning Variational GAN that combines VAE and GAN to learn an identity-invariant representation that is explicitly disentangled from the identity information and generative for expression-preserving face image synthesis. Yang et al. [YCY18] proposed a De-expression Residue Learning procedure to explore the expressive information, which is filtered out during the de-expression process but still embedded in the generator. Then, the model extracts this information from the generator directly to mitigate the influence of subject variations and improve the FER performance.

## 2.3 FER on the dimensional model

A number of researchers have shown that in everyday interactions people exhibit non-basic, subtle and rather complex affective states like thinking, embarrassment or depression. Such subtle and complex affective states can be expressed via tens (or possibly hundreds) of anatomically possible facial expressions. Accordingly, a single label (or any small number of discrete classes) may not reflect the complexity of the affective state conveyed by such rich sources of information. Hence, a number of researchers advocate the use of dimensional description of human affect, where an affective state is characterized in terms of a small number of latent dimensions. Here we briefly present an overview of the FER methods on the dimensional model of affect.

### 2.3.1 First attempts

At first, the common strategy in automatic dimensional affect classification was to simplify the problem of classifying the six basic emotions to a three-class valence-related classification problem: positive, neutral

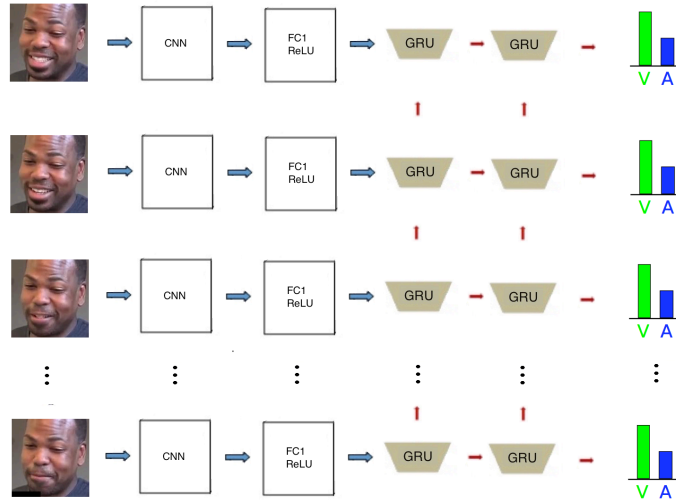


Figure 2.3.1: AffWildNet (Source: [Kol+19]).

and negative emotion classification [GP10]. A similar simplification is to reduce the dimensional emotion classification problem to a two-class problem (positive vs. negative and active vs. passive) or a four-class problem (quadrants of 2D A-V space) [Car+06] [FT05]. Glowinski et al. [Glo+08] analysed four emotions, each belonging to one quadrant of the A-V emotion space: high arousal positive valence (joy), high arousal negative valence (anger), low arousal positive valence (relief), and low arousal negative valence (sadness). Kleinsmith and Bianchi-Berthouze [KB07] used a back-propagation algorithm to build a separate model for each of the affective dimensions for discriminating between levels of affective dimensions from posture (high-low, high-neutral, and low-neutral). Wollmer et al. [Wöl+08] used Conditional Random Fields for discrete emotion recognition by quantising the continuous labels for valence and arousal to four and/or seven arbitrary levels. Kulic and Croft [KC07] perform quantization into 3 categories (low/medium/high), and Chanel et al. [CAP07] consider 3 classes, namely, excited-negative, excited-positive, and calm-neutral. Karpouzis et al. [Kar+07] focused on positive vs. negative or active vs. passive classes.

### 2.3.2 Deep features

One of the first deep learning architectures for valence and arousal estimation was proposed by Khorrami et al. [Kho+16] where both frame-based CNN and CNN plus RNN architectures were proposed. The CNN consisted of 3 convolutional layers; the first two layers were followed by max pooling layers and the third by a quadrant pooling layer. A fully connected layer was then used, followed by the output layer. The CNN plus RNN architectures consisted of the previously described CNN network (keeping its weights fixed) without the top regression layer, followed by a single RNN layer that gave the final estimates. This methodology achieved very high valence and arousal correlations in a part of the RECOLA database [Rin+13]. Later Chen et al. [Che+17] fused handcrafted and deep learning features using different modalities (acoustic, visual, and textual). They also considered the interlocutor influence (a person's influence on the interacting partner's behaviors) for the acoustic features. Recently, Kollias et al. [Kol+19] proposed AffWildNet that performs prediction of continuous emotion dimensions based on visual cues. It includes convolutional and recurrent neural network layers, exploiting the invariant properties of convolutional features, while also modeling temporal dynamics that arise in human behavior via the recurrent layers (Figure 2.3.1). The same authors [KZ20b] presented a CNN-RNN based approach which exploits multiple CNN features for dimensional emotion recognition in-the-wild, utilizing the One-Minute Gradual-Emotion dataset [Bar+18]. Low, mid and high level features are extracted from the trained CNN component and are exploited by RNN subnets in a multi-task framework. Their outputs constitute an intermediate level prediction; final estimates are obtained as the mean or median values of these predictions.

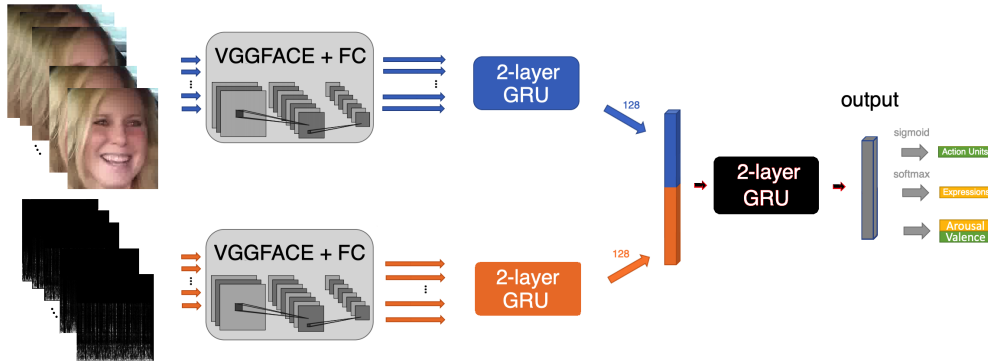


Figure 2.4.1: A/V-MT-VGG-RNN: A Multi-Modal and Multi-Task model (Source: [KZ19]).

## 2.4 Combining emotion representations

Due to the variety of emotion representations, there are models that try to leverage more than one representation to recognize human emotion. Multi-task learning has been mostly used to simultaneously recognize emotions in different representations. Chang et al. [CHC17] proposed FATAUVA-Net method that performs sequential facial attribute recognition, AU detection, and VA estimation on videos. Specifically, the deep network consists of a core layer, an attribute layer, an AU layer and a valence-arousal layer that are all trained sequentially. The core layer is a series of convolutional layers, followed by the attribute layer which extracts facial area’s features (face, eye, eyebrow, mouth). These layers are used in supervised learning of AUs. Finally, AUs are employed as mid-level representations to estimate the intensity of valence and arousal. Xiaohua et al. [Xia+19] trained a 2Att-2Mt model in AffectNet for facial emotion estimation on static images based on a two-level attention with a two-stage framework. Firstly, the features of the corresponding region (position level features) are extracted and enhanced automatically by first-level attention mechanism. Then, they utilize Bi-directional Recurrent Neural Network with self-attention (second-level attention) to make full use of the relationship features of different layers (layer-level features) adaptively. And then, they propose a two-stage multi-task learning structure, which exploits categorical representations to ameliorate the dimensional representations and estimate valence and arousal simultaneously in view of the inherent complexity of dimensional representations and correlation of the two targets. Lately, there has been a lot of research by Kollias et al. [KZ19] [KZ21] [KSZ21] to effectively apply multi-task learning for facial expression recognition. In [KZ19] they conducted multi-task experiments on Aff-Wild2 database, the first large scale in-the-wild database containing annotations for all 3 main behavior tasks. The binary cross-entropy loss was used for AU detection. The MSE and CCC losses were used for VA estimation. The standard loss for expression classification was the categorical cross entropy. They developed multi-task CNNs, multi-task CNN-RNNs and multi-modal, multi-task CNN-RNNs, which were trained on Aff-Wild2 and then evaluated to 10 publicly available databases beating the state-of-the-art on emotion recognition in some of these databases. The final Multi-Modal and Multi-Task model consists of two identical streams that extract features directly from raw input images and spectrograms, respectively. The features from the two streams are concatenated, forming a 256-dimensional feature vector that is passed through a 2-layer GRU layer with 128 units in each layer, in order to fuse the information of the audio and visual streams. The output layer follows on top of it (Figure 2.4.1). Also, an important issue when using MTL on facial databases is the fact that most times not all samples are annotated with both labels. Kollias et. al [KZ21] [KSZ21] explored task-relatedness as a means for co-training, in a weakly-supervised way, tasks that contain little, or even non-overlapping annotations. Task-relatedness is introduced in MTL, either explicitly through prior expert knowledge, or through data-driven studies. They proposed a novel distribution matching approach, in which knowledge exchange is enabled between tasks, via matching of their predictions’ distributions. Based on this approach, they built FaceBehaviorNet, the first framework for large-scale face analysis, by jointly learning all facial behavior tasks. FaceBehaviorNet (Figure 2.4.2) was trained for joint basic expression recognition, action unit detection and valence-arousal estimation achieving impressive results. Finally, the Affective Behavior Analysis in-the-wild (ABAW) 2020 Competition was the first Competition aiming at automatic analysis of the three main behavior tasks of valence-arousal estimation, basic expression recognition and action unit

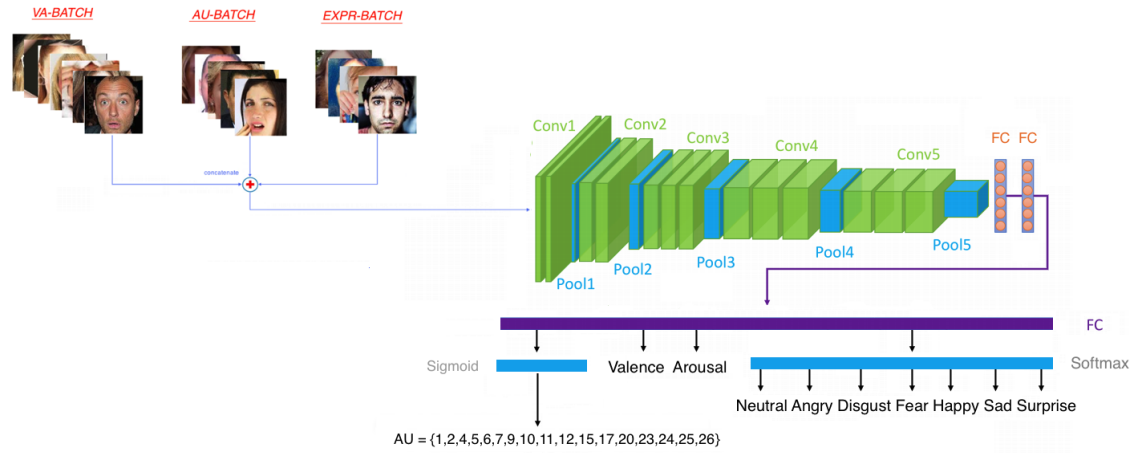


Figure 2.4.2: The holistic (multi-task, multi-domain, multi-label) FaceBehaviorNet (Source: [KZ21], [KSZ21]).

detection. The winners of the competition proposed novel MTL methods on FER [Kol+20].

Apart from MTL, research on the dependencies between the categorical and the dimensional emotion representations is limited. Recently, Kervadec et al. [Ker+18] studied the link between the two representations and proposed a 3-dimensional representation of emotion learned in a multi-domain fashion.

### 2.4.1 Databases

The majority of the FER models are based on supervised machine learning methodologies. These systems require annotated image samples for training. Therefore, researchers have created databases of human actors portraying various emotions. At first, most of these databases contained posed expressions acquired in a controlled lab environment. However, studies show that posed expressions can be different from unposed expressions in configuration, intensity, and timing. Then, researchers captured unposed facial behavior while the subject is watching a short video, engaged in laboratory-based emotion inducing tasks, or interacted with a computer-mediated tutoring system. Although a large number of frames can be obtained by these approaches, the diversity of these databases is limited due to the number of subjects, head position, and environmental conditions. Therefore, there was a demand to develop systems that are based on natural, unposed facial expressions. Recently, to address this demand databases of facial expression and affect in the wild received much attention. These databases are either captured from movies or the Internet, and annotated with the categorical model, the dimensional model or the FACS model. Most of the in-the-wild databases covert only one model of affect and techniques that exploit the dependencies between these models cannon be applied. Therefore, there are some efforts to annotate large in-the-wild databases in more than one models. A brief overview of the FER databases follows focusing on the categorical and the dimensional model.

#### Categorical model

Cohn-Kanade (CK) database was released in 2000 for the purpose of promoting research into automatically detecting individual facial expressions. Later, the **Extended Cohn-Kanade (CK+)** database [Luc+10] was released to address some limitations of CK. The number of sequences was increased by 22% and the number of subjects by 27%. The target expression for each sequence was fully FACS coded and emotion labels were revised and validated. In addition to this, non-posed sequences for several types of smiles and their associated metadata were added.

CMU **MultiPie** face database [Gro+10] contains around 750,000 images of 337 people recorded in up to four sessions over the span of five months. Subjects are imaged under 15 view points and 19 illumination conditions while displaying a range of facial expressions.



Table 2.2: Emotion databases annotated for the categorical model of affect.

Database	Information	Subjects	Condition
CK+ [Luc+10]	327 frontal and 30-degree images	123	Controlled
MultiPie [Gro+10]	Around 750,000 images	337	Controlled
MMI [Pan+05]	2,900 videos and high-resolution images	75	Controlled
SFEW [Dha+11b]	700 images taken from videos	330	Wild
FER-2013 [Goo+13]	Images queried from web	35,887	Wild
EmotionNet [FSM16]	Images queried from web	~100,000	Wild
FER-Wild [Mol+16]	Images queried from web	~24,000	Wild
RAF-DB [LDD17]	Images queried from web	29,672	Wild

MMI database [Pan+05] addresses a number of key omissions in other databases of facial expressions. In particular, it contains recordings of the full temporal pattern of a facial expression, from neutral, through a series of onset, apex, and offset phases and back again to a neutral face. The database consists of over 2,900 videos and high-resolution still images of 75 subjects. It is fully annotated for the presence of AUs in videos (event coding), and partially coded on frame-level, indicating for each frame whether an AU is in either the neutral, onset, apex or offset phase. A small part is annotated for audio-visual laughter. The database is freely available to the scientific community.

**SFEW** [Dha+11b] has been developed by selecting frames from AFEW [Dha+11a], a dynamic temporal facial expressions data corpus consisting of close to real world environment extracted from movies. The database covers unconstrained facial expressions, varied head poses, large age range, occlusions, varied focus, different resolution of face and close to real world illumination. Frames were extracted from AFEW sequences and labelled based on the label of the sequence. In total, SFEW contains 700 images that have been labelled for six basic expressions by two independent annotators.

**Facial Expression Recognition 2013 (FER-2013)** dataset [Goo+13] was introduced in ICML 2013 Workshop on Challenges in Representation Learning. The dataset was created using the Google image search API to search for images of faces that match a set of 184 emotion-related keywords like “blissful”, “enraged”, etc. These keywords were combined with words related to gender, age or ethnicity, to obtain nearly 600 strings which were used as facial image search queries. The first 1,000 images returned for each query were kept for the next stage of processing. OpenCV face recognition was used to obtain bounding boxes around each face in the collected images. Human labelers then rejected incorrectly labeled images, corrected the cropping if necessary, and filtered out some duplicate images. Approved, cropped images were then resized to  $48 \times 48$  pixels and converted to grayscale. The resulting dataset contains 35,887 images.

**EmotionNet** was presented in [FSM16] based on a novel computer vision algorithm to annotate a large database of one million images of facial expressions of emotion in the wild. Then, the authors used WordNet to download 1,000,000 images of facial expressions with associated emotion keywords from the Internet. These images were then automatically annotated with AUs, AU intensities and emotion categories by their algorithm. The result is EmotionNet a highly useful database that can be readily queried using semantic descriptions for applications in computer vision, affective computing, social and cognitive psychology and neuroscience.

**FER-Wild** [Mol+16] was collected using three search engines that were queried using 1,250 emotion related keywords in six different languages. The retrieved images were mapped by two annotators to six basic expressions and neutral. Deep neural networks and noise modeling were used in three different training scenarios to find how accurately facial expressions can be recognized when trained on noisy images collected from the web using query terms. The dataset contains around 24,000 annotated images in a wild setting.

**RAF-DB** [LDD17] contains about 30,000 facial images from thousands of individual. Specifically, well-trained annotators were asked to label face images with one of the seven basic categories and each face was annotated enough times independently, i.e. about 40 times in the experiment. Then, the noisy labels were filtered by an EM based reliability evaluation algorithm, through which each image can be represented reliably by a 7-dimensional emotion probability vector. By analyzing 1.2 million labels of 29,672 great-diverse facial images downloaded from the Internet, these Real-world Affective Faces (RAF) were naturally

Table 2.3: Emotion databases annotated for the dimensional model of affect.

Database	Information	Subjects	Condition
DEAP [Koe+11]	1-minute videos shown to subjects	32	Controlled
MAHNOB-HCI [Sol+11]	Gaze, video, audio, physiological signals	27	Controlled
SEMAINE [McK+11]	Conversations between person and agent	150	Controlled
RECOLA [Rin+13]	Audio, video, ECG and EDA	46	Controlled
AFEW-VA [Kos+17a]	600 videos from films	600	Wild
OMG-Emotion [Bar+18]	420 Youtube videos of around a minute	420	Wild
Aff-Wild [Kol+19]	298 videos from YouTube	200	Wild

categorized into two types: basic expression with single-modal distribution and compound emotions with bimodal distribution.

### Dimensional model

**DEAP** [Koe+11] is a multimodal dataset for the analysis of human affective states. The electroencephalogram and peripheral physiological signals of 32 participants were recorded as they watched 40 one-minute long excerpts of music videos. Participants rated each video in terms of the levels of arousal, valence, like/dislike, dominance and familiarity. For 22 of the 32 participants, frontal face video was also recorded. A novel method for stimuli selection was used, utilising retrieval by affective tags from the last.fm website, video highlight detection and an online assessment tool.

In **MAHNOB-HCI** dataset [Sol+11] 30 participants were shown fragments of movies and pictures, while monitoring them with 6 video cameras, a head-worn microphone, an eye gaze tracker, as well as physiological sensors measuring ECG, EEG (32 channels), respiration amplitude, and skin temperature. Each experiment consisted of two parts. In the first part, fragments of movies were shown, and a participant was asked to annotate its own emotive state after each fragment on a scale of valence and arousal. In the second part of the experiment, images or video fragments were shown together with a tag at the bottom of the screen. In some cases, the tag correctly described something about the situation. However, in other cases the tag did not actually apply to the media item. After each item, the participants were asked to press a green button if they agreed with the tag being applicable to the media item, or press a red button if not. During the whole experiment, audio, video, gaze data and physiological data were recorded simultaneously with accurate synchronisation between sensors. The database is freely available to the research community.

**SEMAINE** [McK+11] is a large audiovisual database as part of an iterative approach to building agents that can engage a person in a sustained, emotionally coloured conversation, using the Sensitive Artificial Listener (SAL) paradigm. Data used to build the system came from interactions between users and an 'operator' simulating a SAL agent, in different configurations: Solid SAL (designed so that operators displayed appropriate non-verbal behaviour) and Semiautomatic SAL (designed so that users' experience approximated interacting with a machine). Having built the system, the authors recorded user interactions with the most communicatively competent version and baseline versions with reduced nonverbal skills. High quality recording was provided by five high-resolution, high frame rate cameras, and four microphones, recorded synchronously. It contains recordings of 150 participants, for a total of 959 conversations with individual SAL characters, lasting approximately 5 minutes each. Solid SAL recordings are transcribed and extensively annotated: 6-8 raters per clip traced five affective dimensions and 27 associated categories. Other scenarios are labelled on the same pattern, but less fully.

**RECOLA** [Rin+13] is a multimodal corpus of spontaneous collaborative and affective interactions in French. Participants were recorded in dyads during a video conference while completing a task requiring collaboration. Different multimodal data, i.e. audio, video, ECG and EDA, were recorded continuously and synchronously. In total, 46 participants took part in the test, for which the first 5 minutes of interaction were kept to ease annotation. In addition to these recordings, 6 annotators measured emotion continuously on two dimensions: arousal and valence, as well as social behavior labels on five dimensions.

**AFEW-VA** [Kos+17a] is a dataset of highly accurate per-frame annotations of valence and arousal for 600 challenging video clips extracted from feature films (also used in part for the AFEW dataset [Dha+11a]).

Table 2.4: Emotion databases annotated for both the categorical and the dimensional model of affect.

Database	Information	Subjects	Condition
AffectNet	Images queried from web	~450,000	Wild
Aff-Wild 2	548 videos	458	Wild

For each video clip, further per-frame annotations of 68 facial landmarks are provided.

**OMG-Emotion** [Bar+18] is composed of Youtube videos which are around a minute in length and are annotated taking into consideration a continuous emotional behavior. The videos were selected using a crawler technique that uses specific keywords based on long-term emotional behaviors such as "monologues", "auditions", "dialogues" and "emotional scenes". After the videos were selected, an algorithm was created to identify if the video has at least two different modalities which contribute for the emotional categorization: facial expressions, language context, and a reasonably noiseless environment. A total of 420 videos were selected, totaling around 10 hours of data.

**Aff-Wild** [Kol+19] is a large scale in-the-wild database annotated in terms of valence and arousal. The authors capitalized on the abundance of data available in video-sharing websites such as YouTube and selected videos that display the affective behavior of people, for example videos that display the behavior of people when watching a trailer, a movie, a disturbing clip, or reactions to pranks. They collected 298 videos displaying reactions of 200 subjects, with a total video duration of more than 30 hours. This database has been annotated by 8 lay experts with regards to two continuous emotion dimensions, i.e. valence and arousal.

### Categorical and Dimensional model

**AffectNet** [MHM17] is by far the largest database of facial expressions that provides both categorical and VA annotations. The facial images are collected from the Internet by querying different search engines (Google, Bing, and Yahoo) using 1250 emotion related tags in six different languages (English, Spanish, Portuguese, German, Arabic, and Farsi). AffectNet contains more than one million images with faces and extracted facial landmark points. Twelve human experts manually annotated 450,000 of these images in both categorical and dimensional (valence and arousal) models and tagged the images that have any occlusion on the face. It is a very challenging database as it contains images of people from different races and ethnic groups as well as high variety in the background, lighting, pose, point of view, etc.

**Aff-Wild2** [KZ19] is the first ever database annotated for valence-arousal estimation, action unit detection and basic expression classification. It consists of 548 videos collected from YouTube and shows both subtle and extreme human behaviours in real-world settings. It is an extension of Aff-Wild [Kol+19] by collecting a new dataset consisting of 260 YouTube videos, with 1,413,000 frames and a total length of 13 hours and 5 minutes. The new videos have wide range in subjects: age (from babies to elderly people), ethnicity (caucasian/hispanic/latino/asian/black/african american), profession (e.g. actors, athletes, politicians, journalists), head pose, illumination conditions and occlusions.

## 2.5 Facial analysis toolkits

Past years have seen huge progress in FER and other related facial analysis tasks. However, very few tools are available to the research community that can recognize all of them. There is a large gap between state-of-the-art algorithms and freely available toolkits. This is especially true when real-time performance is needed that is a necessity for interactive systems. In Table 2.5, we briefly present the freely available toolkits for facial analysis. TAUD is the implementation of a LPQ-TOP-based AU detector [JVP11] and is the first toolkit that recognizes facial expressions. It is developed as a WIN32 executable and the input can either be a video sequence (.avi) or a set of images that form a video (.png). However, TAUD is trained on a small database and the system assumes that the image/video contains a frontal view of the face. Later, OpenFace [BRM16] was developed being the first toolkit capable of facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation with available source code for both running and training the models. Also, it is capable of real-time performance and is able to run from a simple webcam without any specialist hardware. OpenFace 2.0 [Bal+18] is an extension of the OpenFace toolkit. While OpenFace is able to perform

Table 2.5: Summary of freely available toolkits for facial analysis.

Tool	Approach	Real-time	Free
TAUD	[JVP11]		✓
OpenFace [BRM16]	[BBR13] [BMR15]	✓	✓
OpenFace 2.0 [Bal+18]	[Woo+15] [Zad+17] [Zha+16]	✓	✓

the above mentioned tasks, it struggles when the faces are non-frontal or occluded and in low illumination conditions. OpenFace 2.0 is able to cope with such conditions through the use of a new CNN based face detector and a new and optimized facial landmark detection algorithm. This leads to improved accuracy for facial landmark detection, head pose tracking, AU recognition and eye gaze estimation.

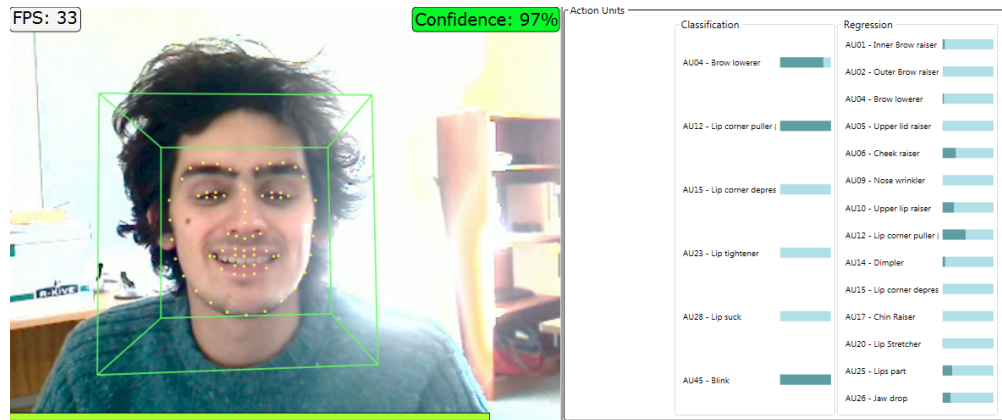


Figure 2.5.1: Example of using the OpenFace toolkit (Source: [Bal+18]).

In Figure 2.5.1 we can see an example of using OpenFace for three facial analysis tasks. We observe that the toolkit manages to effectively predict the three tasks even though the illumination condition of the video is bad. However, more research should be conducted to continuously update the facial analysis toolkits using the novel techniques that are proposed in the literature.

# Chapter 3

## Baseline Models

### 3.1 Preprocessing

The first stage in training any machine learning model is to preprocess the training data. In our case, we deal with images that depict the face of a human. Since we care only about the facial expression, we should first crop the facial region of the image and remove the non-facial parts. Nowadays, face detectors have achieved very high accuracy and the problem of detecting the face is almost solved (of course there are cases of bad illumination conditions or extreme occlusions). Therefore, this information is either provided by the creators of the database or is acquired using a simple face detection software.

The second stage is face alignment that usually boosts the performance because it reduces the high variation of the dataset. We perform landmark-based face alignment to obtain a normalized rotation, translation, and scale representation of each face. Our alignment method relies only on the facial landmarks and is fast so it can be executed in real-time without adding any extra computational burden to the system. First, we detect the eye landmarks that correspond to the points between 37 and 48 (Figure 3.1.1). Then, we calculate the average mean of the six detected landmarks of each eye and compute the angle between the two centroids:

$$angle = \tan^{-1} \left( \frac{r_y - l_y}{r_x - l_x} \right) \tag{3.1.1}$$

where  $(r_x, r_y)$  and  $(l_x, l_y)$  are the coordinates of the right and the left centroid respectively. After rotation, the eyes should also be equidistant from the edges of the image. Therefore, scaling is also applied:

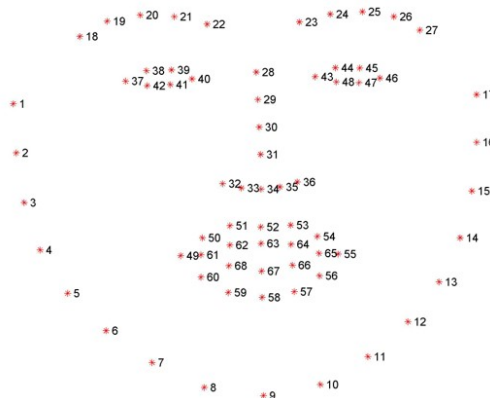


Figure 3.1.1: A common numbering of the 68 facial landmark coordinates.

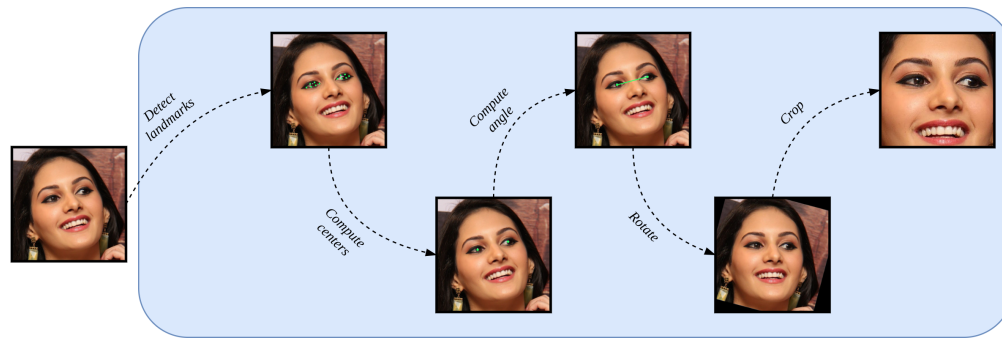


Figure 3.1.2: An example of the preprocessing in a sample from Affectnet.

$$scale = \frac{desired\_distance}{\sqrt{(r_x - l_x)^2 + (r_y - l_y)^2}} \quad (3.1.2)$$

Then, we apply an affine transformation with the computed angle and scale to the image to get the aligned image. Finally, the face region of the aligned image is cropped again to keep only the emotion-related parts of the face. In Figure 3.1.2 we can see the whole procedure of crop and alignment in a sample of AffectNet.

The usual final stage in preprocessing is data augmentation that are techniques to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data. It acts as a regularizer and helps reduce overfitting when training a machine learning model. It is closely related to oversampling in data analysis. In our models, six types of data augmentation are applied:

1. Flip: Horizontally flip the given image randomly with a given probability.
2. Brightness: Randomly change the brightness of an image.
3. Contrast: Randomly change the contrast of an image. Contrast can be thought of as the degree to which light and dark colours in the image differ.
4. Rotation: Rotate the image by a defined angle.
5. Hue: Randomly change the hue of an image. Hue can be thought of as the ‘shade’ of the colors in an image.
6. Saturation: Randomly change the saturation of an image. Saturation can be thought of as the ‘amount’ of color in an image.

In Figure 3.1.3 we can see the effect of the above data augmentation techniques in a sample image from AffectNet.

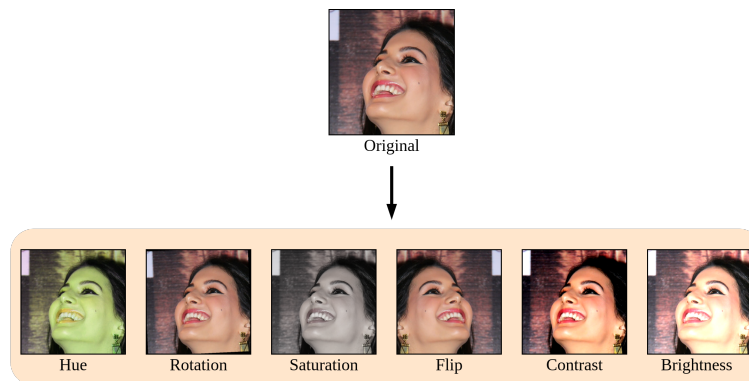


Figure 3.1.3: Data augmentation techniques that are used in our models.

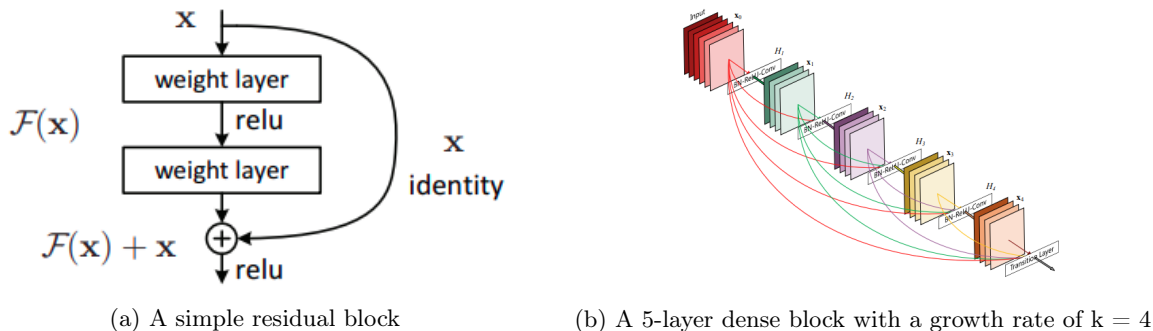


Figure 3.2.1: Building blocks of the deep architectures that are used.

## 3.2 Architecture

Recognizing facial expressions in-the-wild presents a lot of variations and learning discriminative features constitutes a challenging task. Hence, deep CNN architectures are employed to capture the diversity of the dataset and extract discriminative learned features. As a baseline model we employ a Densely Connected Convolutional Network (DenseNet) [Hua+17a] that is an extension of Residual Networks (ResNets) [He+16]. The core idea of a ResNet proposed by He et al. is the introduction of a so-called identity shortcut connection that skips one or more layers (Figure 3.2.1a). They argue that stacking layers should not degrade the network performance, because we could simply stack identity mappings upon the current network, and the resulting architecture would perform the same. This indicates that the deeper model should not produce a training error higher than its shallower counterparts. They hypothesize that letting the stacked layers fit a residual mapping is easier than letting them directly fit the desired underlying mapping. Therefore, ResNet architecture enjoys the accuracy gains from greatly increased depths while at the same time it eases the convergence of the training algorithm by using the residual layers. Although these architectures have worked well for many computer vision problems, the most recently proposed DenseNet architecture has obtained significant improvements and yields many benefits. In this residual architecture, the input of each layer consists of the feature maps of all earlier layers and its output is passed to each subsequent layer. The feature maps are aggregated with depth-concatenation. Other than tackling the vanishing gradients problem, the authors argue that this architecture also encourages feature reuse, making the network highly parameter-efficient. One simple interpretation is that in ResNet the output of the identity mapping is added to the next block, which might impede information flow if the feature maps of two layers have very different distributions. Therefore, concatenating feature maps can preserve them all and increase the variance of the outputs, encouraging feature reuse (Figure 3.2.1b).

## 3.3 Loss Function

After discussing the preprocessing steps and the network architecture of our baseline model, we should then define the objective functions that will guide the training of our models. In the categorical model we have to recognize the depicted emotion out of the seven basic emotion categories. The output of our network is seven scores  $s_1, \dots, s_7$  one for each emotion. To convert these scores into probabilities, the softmax function is applied:

$$\hat{y}_i = \frac{e^{s_i}}{\sum_{k=1}^7 e^{s_k}} \quad (3.3.1)$$

where  $\hat{y}_i$  is the probability of emotion  $i$ . The most frequent loss function for a classification task is categorical cross-entropy loss that for a training sample  $x_i$  is defined as:

$$L_{CE} = - \sum_{i=1}^7 y_i \log(\hat{y}_i) \quad (3.3.2)$$

where  $y_i = 1$  if class  $i$  is the ground truth expression. In most emotion databases the class distribution is highly imbalanced because the majority of the images from search engines depict happy or neutral faces. We believe that this happens because people tend to post online images with positive expressions rather than negative ones. To deal with the imbalance problem, our networks are trained using a weighted version of the traditional categorical cross-entropy loss. In other words, the network is penalized more for misclassifying samples from under-represented classes than from well-represented classes:

$$L_{CE_w} = - \sum_{i=1}^7 w_i y_i \log(\hat{y}_i) \quad (3.3.3)$$

$$w_i = \frac{f_i}{f_{min}} \quad (3.3.4)$$

where  $f_i$  is the number of samples of the  $i$ -th class and  $f_{min}$  is the number of samples in the most under-represented class i.e. Disgust.

### 3.4 Soft Loss

In Figure 3.4.1 we see a sample from AffectNet that depicts the anger expression and the output of our best model. We observe that our model wrongly predicts the sad expression. However, due to the inherent intra-class variation of our task the expression of sadness and anger are easily confused in this example. Inspired by ideas in domain transfer by Tzeng et al. [Tze+15] we design a soft loss that uses soft labels in CE Loss as follows:

$$L_{soft} = - \sum_{c=1}^M l_i^c \log(p_{i,c}) \quad (3.4.1)$$

and logit loss that pushes logits to match the label distribution using a MSE:

$$L_{logit} = \frac{1}{2} \sum_{c=1}^M \|p_{i,c} - l_i^c\|_2^2 \quad (3.4.2)$$

where  $l_i^c = P(c | v_i, a_i)$  and  $p_{i,c}$  denotes the probability that sample  $i$  belongs to class  $c$ . The next step is to estimate  $P(c | v_i, a_i)$  that defines the class distribution given the valence  $v_i$  and arousal  $a_i$  in the bin  $i$ . First, we divide the VA space in a predefined number of square bins. For each bin  $i$  we calculate the frequency of each class based on the training set. These histograms can be considered an estimate of the class distribution in each region of the VA space. Finally, we use these values as the variable  $l_i^c$  of the soft loss and the logit loss.

Class	$P(class image)$
Neutral	0.0957
Happy	0.0009
<b>Sad</b>	<b>0.4913</b>
Surprise	0.0019
Fear	0.0018
Disgust	0.0051
<b>Anger</b>	<b>0.4034</b>

Figure 3.4.1: Sample from AffectNet that depicts anger and the output of our best model.



Table 3.1: Performance of our baseline models on the categorical model of AffectNet.

Architecture	# of classes	Accuracy
ResNet50	8	47.8%
ResNet50 + Weighted Loss	8	53.2%
Pretrained ResNet50 + Weighted Loss	8	55.4%
Pretrained DenseNet121 + Weighted Loss	8	57.3%
* + Alignment	8	59.5%
* + Alignment + Augmentation	8	<b>61.4%</b>
* + Alignment + Augmentation	7	<b>64.37%</b>

\*Pretrained DenseNet121 + Weighted Loss

## 3.5 Results

We train and evaluate our models on AffectNet [MHM17], the largest facial database that contains more than one million facial images collected from the Internet. It is a very challenging database as it contains images of people from different races and ethnic groups as well as high variety in the background, lighting, pose, point of view, etc. Since the test set of AffectNet is not publicly available, we evaluate our approaches on the validation set that contains 500 images for each emotion. The mean class accuracy is used as the evaluation metric for the classification task because the validation set is balanced. Also, all networks are trained for 10 epochs using a batch size of 35 and a learning rate of 0.001. Stochastic Gradient Descent is adopted as the optimization algorithm with a momentum of 0.9 and PyTorch is used as our deep learning framework.

In Table 3.1 we present the performance of our baseline models on the categorical model of AffectNet. We should note that this is the only chapter where we present our results on the 8-way classification task. We observed that the majority of past studies ([ZSC18], [Ker+18], [HNM19], etc.) excludes the emotion of contempt from the database and train the models on the images with neutral and the 6 basic emotions (approximately 280,000 training samples). Therefore, we decided to deal with the 7-way classification problem (without contempt) in the rest of the thesis.

We observe that the accuracy increases when we are using the weighted version of the CE loss. Also, the pretraining and data preprocessing steps we discussed earlier benefit the overall performance of the network. Our best baseline model achieves **61.4%** in the 8-way classification task and **64.37%** in the 7-way classification task. In the following chapters, we will use these models as our baselines for comparison with the proposed models.

In Table 3.2 we see the performance of the baseline model using the soft and the logit loss. We observe that the proposed losses increase the accuracy of the baseline model indicating that the integration of the VA distribution in a loss function is beneficial.

Table 3.2: Performance of models using soft and logit loss.

Loss	Parameters	Accuracy
WCE	-	64.37
WCE + Soft	$\lambda = 0.1$ , bins=5	64.68
WCE + Logit	$\lambda = 1$ , bins=10	<b>65.05</b>



# Chapter 4

## Metric learning models

### 4.1 Variations

A facial image contains a lot of information that is unrelated to the depicted expression. Parts that are related to the identity of the person (identity bias), occlusions due to accessories and pose variations act as noise for the FER system and reduce its total performance. These variations are usually called appearance variations. Also, in our task there are two other types of variations that originate from the nature of the human expression. Specifically, human expressions change in a continuous way while the categorical model restricts the expressions in discrete emotion categories. We know that a single emotion can be expressed through various ways. For example, anger can be expressed through ways that are much different to each other and closer to the expression of other emotions like disgust (in respect to the facial characteristics). This inherent feature of the categorical model leads to a large intra-class and small inter-class variation of the learned features (Figure 4.1.1).

1. **Intra-class variation:** Samples from the same class do not have close representations to the feature space. In our case, each class corresponds to a certain emotion. We know that a single emotion is expressed through various ways. For example, surprise can be expressed through raised and arched eyebrows, dilated eyes or a dropped jaw. Even a very good recognition model will not project all these images to the same region of the feature space to balance between discriminativeness and generalizability. Therefore, these features will be far apart in the feature space although they express the same emotion.
2. **Inter-class variation:** Samples from different classes have close representations to the feature space. Respectively, there are characteristics of the face that convey different emotions. For example, when the eyebrows are lowered and knit together the person may feel angry, sad or even fear. Therefore, there will be representations of these emotions very close in the feature space resulting in small inter-class variation.

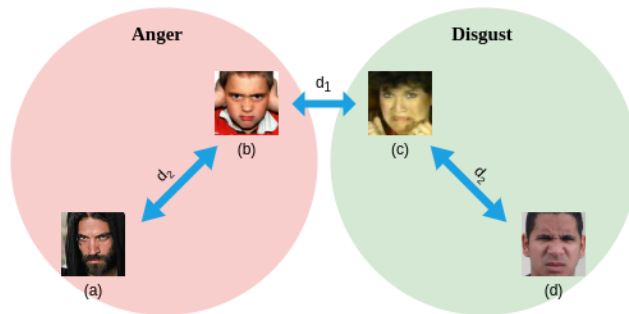


Figure 4.1.1: Example of large intra-class and small inter-class variation.

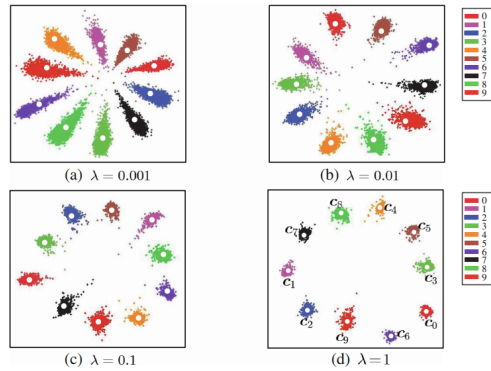


Figure 4.2.1: The distribution of learned features under the joint supervision of softmax and center loss (Source: [Wen+16]).

## 4.2 Center Loss

As mentioned before, a reason that FER in the wild is still a challenging problem is the variations caused by diversity in head pose, illumination, occlusions, and personal attributes. Traditional CNNs are optimized using a softmax loss, which penalizes the misclassified samples and thus forces the features of different classes to stay apart. However, due to high intra-class variations, the features in each cluster are often scattered. Furthermore, the clusters overlap because of high inter-class similarities. Therefore, additional losses are employed to reduce these variations and improve the recognition accuracy. Wen et al. [Wen+16] introduced center loss into CNNs to reduce the intra-class variations of the learned features for face recognition. The scope of a center loss is to pull the learned features of the same class towards a class center. To achieve this, the center loss penalizes the distances between the extracted deep features and their corresponding class centers. It is defined as:

$$L_c = \frac{1}{2} \sum_{i=1}^m \|x_i - c_{y_i}\|_2^2 \quad (4.2.1)$$

where  $x_i$  is the feature vector of sample  $i$ ,  $y_i$  is the class of sample  $i$  and  $c_{y_i}$  is the center of class  $y_i$ . The formulation effectively characterizes the intra-class variations. Ideally, the  $c_{y_i}$  should be updated as the deep features changed. In other words, we need to take the entire training set into account and average the features of every class in each iteration, which is inefficient even impractical. Therefore, the center loss can not be used directly. The authors perform the update of the centers based on mini-batch instead of updating the centers with respect to the entire training set. In each iteration, the centers are computed by averaging the features of the corresponding classes (In this case, some of the centers may not update). Also, to avoid large perturbations caused by few mislabelled samples, they use a scalar  $\alpha$  to control the learning rate of the centers. The gradients of  $L_c$  with respect to  $x_i$  and update equation of  $c_{y_i}$  are computed as:

$$\frac{\partial L_c}{\partial x_i} = x_i - c_{y_i} \quad \Delta c_j = \frac{\sum_{i=1}^m \delta(y_i = j)(c_j - x_i)}{1 + \sum_{i=1}^m \delta(y_i = j)} \quad (4.2.2)$$

where  $\delta(\text{condition}) = 1$  if the *condition* is satisfied and  $\delta(\text{condition}) = 0$  if not. The model is optimized using both the center and the softmax loss function:

$$L = L_{CE} + \lambda L_c \quad (4.2.3)$$

A scalar  $\lambda$  is used for balancing the two loss functions. In Figure 4.2.1 we can see the effect of parameter  $\lambda$  to a face recognition database. With proper  $\lambda$ , the discriminative power of deep features can be significantly enhanced. Moreover, features are discriminative within a wide range of  $\lambda$ . Therefore, the joint supervision benefits the discriminative power of deeply learned features, which is crucial for face recognition.

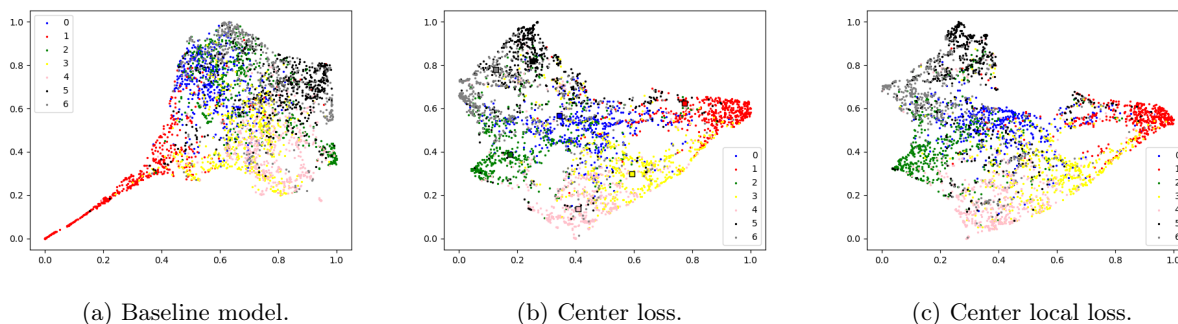


Figure 4.2.2: Deep learned features (a) without extra loss, with (b) center loss and (c) center local loss (The squares denote the learned centers).

In our case we applied the center loss on the FER task to reduce the high intra-class variation. In order to qualitatively explore the effect of center loss in our system, we apply UMAP [MHM18] dimensionality reduction technique to project the learned features from  $R^{1024}$  to  $R^2$ . The UMAP algorithm is based on three assumptions about the data:

- The data is uniformly distributed on Riemannian manifold.
- The Riemannian metric is locally constant (or can be approximated as such).
- The manifold is locally connected.

From these assumptions, it is possible to model the manifold with a fuzzy topological structure. The embedding is found by searching for a low dimensional projection of the data that has the closest possible equivalent fuzzy topological structure. For more information, we refer readers to the paper [MHM18]. After applying UMAP to our features, we plot the 2-D projected learned features to explore the compactness of each class. As we can see in Figure 4.2.2 the center loss successfully pulls the learned features of each class to each other and creates more compact clusters. A modification of center loss is to define the coordinates of the centers explicitly instead of learning them as extra parameters. Hayale et al. [HNM19] used a center local loss that computes the center of a class for each mini-batch as the mean value of the features that belong to this class:

$$c_j = \frac{\sum_{i=1}^m \delta(y_i = j) x_i}{\sum_{i=1}^m \delta(y_i = j)} \quad (4.2.4)$$

A drawback of center local loss is that it highly depends on the data in contrast to the center loss where the centers are learned. So, the misclassified representations pull the computed mean in a wrong direction.

### 4.3 Extensions of center loss

While the formulation of center loss is intuitive, it comes with some problems that we should carefully consider:

1. It only deals with the problem of large inter-class variations and ignores the small inter-class differences. The samples of a certain class come closer to the center of the class. However, the center loss does not guarantee that centers of different classes will be apart.
2. It does not respect the inherent intra-class variation of our task. The center loss reduces the intra-class variation of every class bringing each sample closer to its respective center. However, our task (and most tasks) requires the existence of some variation according to the emotion. For example, the emotion of happiness is expressed by more ways than the emotion of fear. Therefore, happy class should be able to learn features with more variations than the fear class.

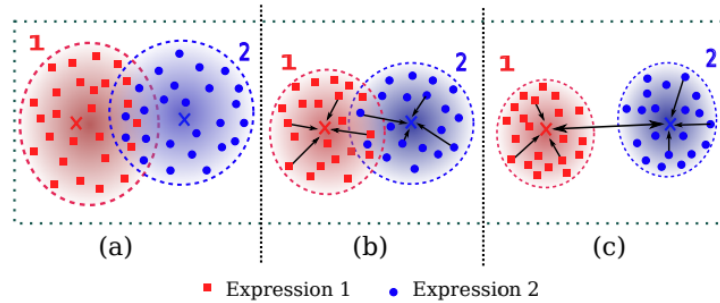


Figure 4.3.1: Deep features learned by (a) softmax loss, (b) softmax loss + center loss, and (c) softmax loss + island loss in the feature space (Source: [Cai+18]).

### 4.3.1 Island loss

To deal with the first problem, Cai et al. [Cai+18] proposed Island loss. The island loss is defined as the summation of the center loss and the pairwise distances between class centers in the feature space:

$$L_{island} = L_c + \lambda_1 \sum_{c_j \in N} \sum_{\substack{c_k \in N \\ c_k \neq c_j}} \left( \frac{c_k \cdot c_j}{\|c_k\|_2 \|c_j\|_2} + 1 \right) \quad (4.3.1)$$

where  $N$  is the set of expression labels;  $c_k$  and  $c_j$  denote the  $k$ -th and  $j$ -th center with  $L_2$  norm  $\|c_k\|_2$  and  $\|c_j\|_2$ , respectively;  $(\cdot)$  represents the dot product. Specifically, the first term penalizes the distance between the sample and its corresponding center and the second term penalizes the similarity between expression centers. The hyperparameter  $\lambda_1$  is used for balancing the two terms. By minimizing the island loss, the samples of the same expression will get closer to each other and those of different expressions will be pushed apart (Figure 4.3.1). Based on SGD, the update of the  $j$ -th class center can be calculated as:

$$\Delta c_j = \frac{\sum_{i=1}^m \delta(y_i = j)(c_j - x_i)}{1 + \sum_{i=1}^m \delta(y_i = j)} + \frac{\lambda_1}{|N| - 1} \sum_{c_k \in N, c_k \neq c_j} \frac{c_k}{\|c_k\|_2 \|c_j\|_2} - \left( \frac{c_k \cdot c_j}{\|c_k\|_2 \|c_j\|_2^3} \right) c_j \quad (4.3.2)$$

where  $|N|$  is the total number of expressions. In this manner, the class centers can be updated iteratively in each mini-batch with a learning rate  $\alpha$ :

$$c_j^{t+1} = c_j^t - \alpha \Delta c_j^t \quad (4.3.3)$$

In Equation 4.3.1 we can see that island loss uses the cosine similarity of the centers in order to push them apart. During training we combine island loss with the aforementioned center loss that is based on the euclidean distance of samples. We are not sure if the simultaneous use of a cosine-based and euclidean-based loss on the feature space is optimal. To deal with this problem, we propose a slight modification of the island loss to an euclidean version as follows:

$$L_{island_{eucl}} = L_c + \lambda_1 \sum_{c_j \in N} \sum_{\substack{c_k \in N \\ c_k \neq c_j}} \max(m - \|c_k - c_j\|, 0)^2 \quad (4.3.4)$$

In Figure 4.3.2 the deep features learned by the island and the euclidean island loss are presented. We observe that the island loss effectively manages not only to reduce the intra-class variation but also to increase the inter-class differences.

### 4.3.2 Local Subclass loss

To deal with the second problem of center loss (not respect the inherent intra-class variation), Luo et al. [LHD18] proposed local subclass loss. Specifically, they assume that each emotion category consists of some

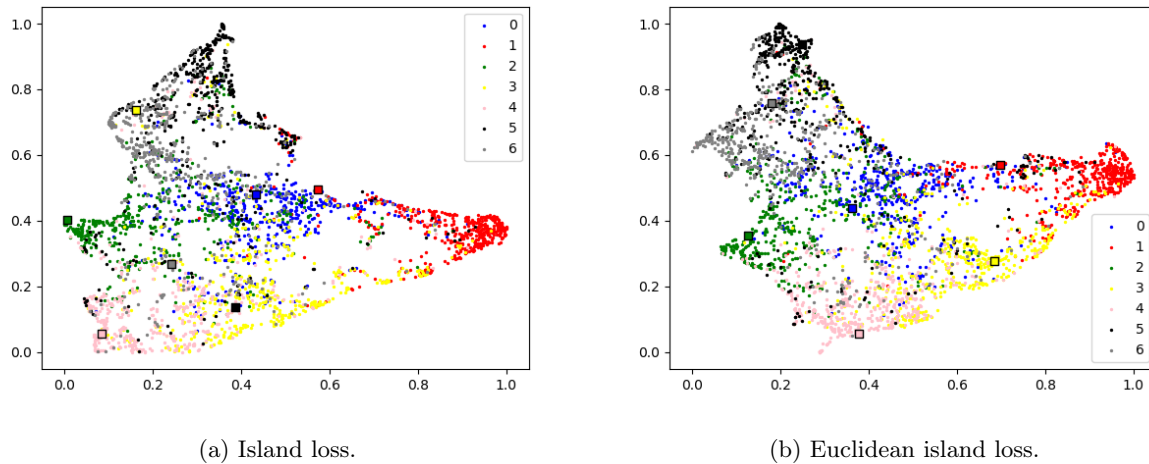


Figure 4.3.2: Deep learned features with (a) island loss and (b) euclidean island loss (The squares denote the learned centers).

subclasses. Different subclasses may represent different attributes or modalities. Local subclass loss aims at alleviating the discrepancy of samples which belong to the same subclass rather than constraining all intra-class samples to get close like the center loss. However, there are many problems with the above hypothesis. Firstly, how to represent each subclass and how to define the number of subclass of each category. Actually, we do not know the distribution of intra-class samples so that we can not know the exact number of subclass of each category. Secondly, without real tagging information, we can not decide which subclass the samples belong to directly. Thirdly, if we know which subclass the samples belong to, how to constrain their variance. To address these problems, the authors made the following definitions and assumptions:

1. They use cluster centers to represent subclasses. The cluster centers are vectors that have the same dimensions as the feature vector in the feature space. The value of cluster centers is initialized randomly at the beginning of training and the parameters are updated according to the proposed loss function. They directly define that there are  $K$  subclasses in a basic expression category. And for convenience of calculations, they assume that different expression categories have the same number of subclass.
2. In Euclidian space, the distance between a sample and all its intra-class cluster centers is calculated. Then, the nearest cluster center of this sample is selected.
3. The variance of samples belong to the same subclass is reduced by minimizing the Euclidian distance between feature vectors and the subclass centers in feature space.

Based on the above, the local subclass loss is defined as follows:

$$L_{subclass} = \frac{1}{2} \sum_{i=1}^m \|x_i - c_{y_i}^{min}\|_2^2 \quad (4.3.5)$$

where  $x_i$  represents the feature vector of  $i$ -th sample in a mini batch,  $y_i$  is its label and  $c_{y_i}^{min}$  denotes the nearest intra-class center of  $x_i$ . The distance between the feature vector and subclass center is calculated in Euclidean space.  $m$  is the sample number of each mini batch. This method aims at learning locally compact feature space by minimizing the distance between feature vectors and its nearest subclass center. Obviously, the exact value of subclass centers is hard to define by handcrafting and it should be updated as the deep features change. In practice, the subclass centers are initialized randomly at the beginning of training. The loss function is derivable and parameter can be updated by gradient descent strategy. The gradient of  $L_{subclass}$  in respect to  $x_i$  and each center vector are:

$$\frac{\partial L_{subclass}}{\partial x_i} = x_i - c_{min}^{y_i} \quad (4.3.6)$$

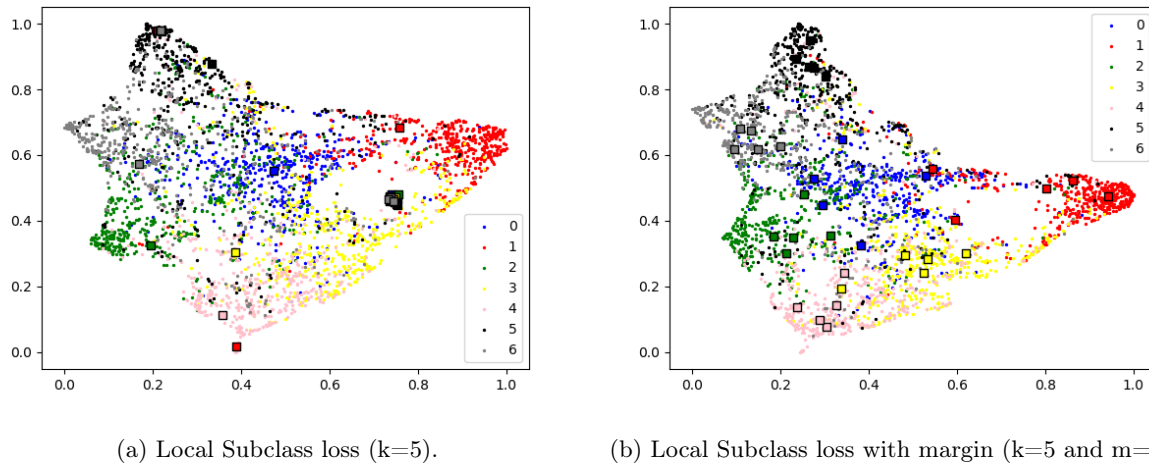


Figure 4.3.3: Deep learned features with (a) local subclass loss and (b) local subclass loss with margin (The squares denote the learned centers).

$$\frac{\partial L_{subclass}}{\partial c_k^j} = \frac{\sum_{i=1}^m 1\{x_i, c_k^j\}(c_k^j - x_i)}{1 + \sum_{i=1}^m 1\{x_i, c_k^j\}} \quad (4.3.7)$$

Here,  $c_k^j$  means the  $k$ -th subclass center of class  $j$ . Function  $1\{x_i, c_k^j\}$  equals to 1 if  $c_k^j$  is the nearest insta-class center of  $x_i$ , otherwise equals to 0. Obviously, many subclass centers may not be selected to constrain  $x_i$  in a mini batch, and their gradients are zero, thus these centers will not be updated in this iteration. The constant 1 is added to the denominator to avoid the overflow from division by 0. In the training process of CNNs, the number of subclass centers is fixed assuming that each class has the same number of subclass. When each class has only one subclass center, the method is identical to the original center loss.

A possible drawback of local subclass loss is that it does not constrain the distance between the subclass centers. Due to the random initialization of centers, some subclass centers may end up not be selected on any iteration. We can deal with this by adding a margin in the distance of the centers that belong to the same class. Intuitively, centers from the same class should not be far in the feature space. To achieve this, we propose an additional term in the loss function that keeps the centers of the same class close to each other:

$$L_{subclass_m} = L_{subclass} + \lambda_1 \sum_{k \in K} \sum_{c_j \in N_k} \sum_{\substack{c_i \in N_k \\ c_i \neq c_j}} \max(\|c_k^j - c_k^i\| - m, 0)^2 \quad (4.3.8)$$

In Figure 4.3.3 the deep features learned by the local subclass loss with and without the margin are presented. We observe that the intra-class variation is not suppressed so much as in the island loss indicating that the local subclass loss enables more variation inside each class. Also, when we are using the additional margin term, the learned centers are more uniformly distributed in the feature space.

## 4.4 Contrastive loss

Instead of using center-based losses to reduce variation, we can use Siamese networks that aim to increase the distance of samples from different classes. Siamese networks are incredibly powerful networks, responsible for significant improvements in face recognition, signature verification, etc. They are a special class of neural networks that have the above properties:

1. They contain two (or more) identical subnetworks.
2. These subnetworks have the same architecture, parameters and weights.



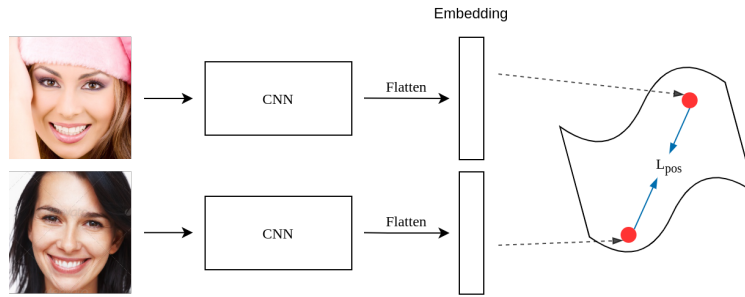


Figure 4.4.1: Positive contrastive loss.

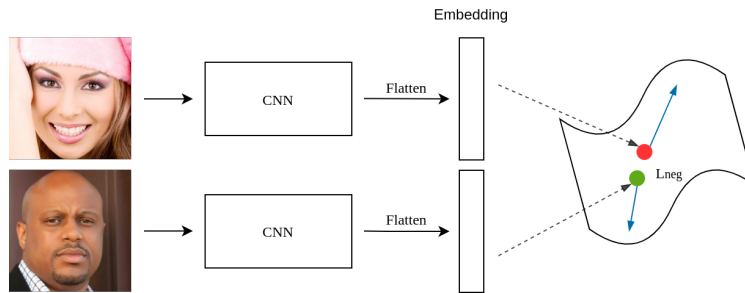


Figure 4.4.2: Negative contrastive loss.

3. Any parameter updates are mirrored across both subnetworks, meaning if you update the weights on one, then the weights in the other are updated as well.

When training Siamese networks we need to have a positive and a negative set:

1. Positive set that contains pairs of images that belong to the same class.
2. Negative set that contains pairs of images that belong to different classes.

In each mini-batch, we construct the positive and negative sets. Then, to reduce the distance of the positive pairs in the feature space we apply the positive contrastive loss that penalizes the distance of each positive pair so as to bring their feature vectors closer.

$$L_{pos} = \frac{1}{2|C_{pos}|} \sum_{i=1}^m \sum_{j=i+1}^m \delta(y_i = y_j) \|x_i - x_j\|^2 \quad (4.4.1)$$

To increase the distance of the negative pairs in the feature space we apply the negative contrastive loss that penalizes the opposite of the distance of two negative pairs so as to move their feature vectors away. Also, a margin  $m$  is used in order to remove the influence of the negative loss when the distance is large.

$$L_{neg} = \frac{1}{2|C_{neg}|} \sum_{i=1}^m \sum_{j=i+1}^m \delta(y_i \neq y_j) \max(0, m - \|x_i - x_j\|_2)^2 \quad (4.4.2)$$

In Figures 4.4.1 and 4.4.2 we can see the general pipeline of a siamese network using the positive and negative contrastive loss respectively. When the two samples depict the same expression (happy) their feature representations are pulled closer. Respectively, when the two samples depict different expressions (happy and neutral) their feature representations are pushed away.

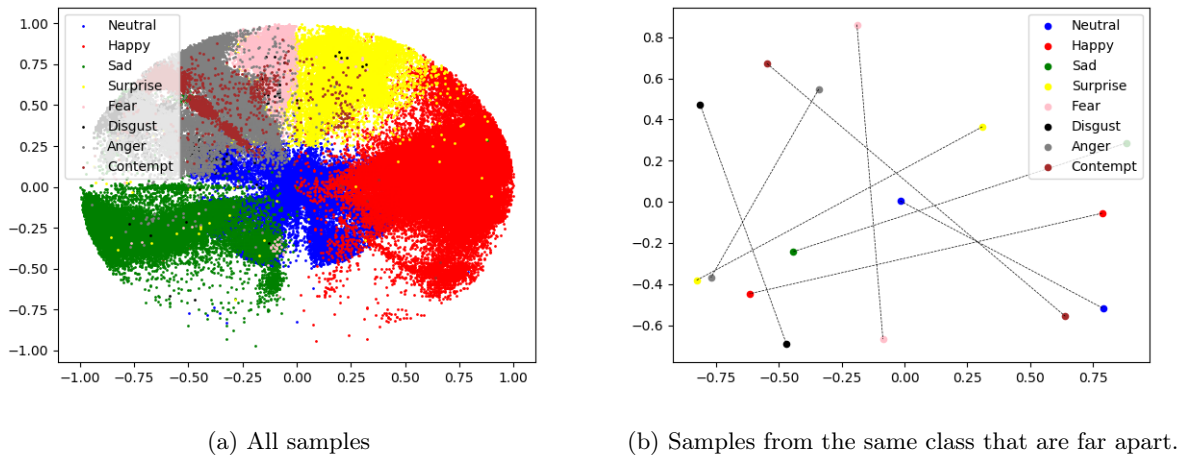


Figure 4.5.1: Distribution of the training samples of AffectNet in valence-arousal space.

## 4.5 Using VA in metric learning

### 4.5.1 Motivation

Although center loss aims to reduce the intra-class variation of the task and learn discriminative features, it fails many times to improve the total performance of the system. A possible limitation of the center loss in our case is that it deals with all samples of a class equally. Specifically, it does not take into account that the same emotion can be expressed in different intensities and using different parts of the face. For example, no matter if a person is very happy or a bit happy, the center loss pulls both samples in the center of the happy class. It is obvious that the FER task presents an inherent intra-class variation that should be respected and not be reduced by the loss function. In Figure 4.5.1 we show the distribution of the training samples of AffectNet in the valence-arousal space. As we can see, the samples of a certain class are not always gathered in a central region of the VA space. On the contrary, there are samples from the same class that are far apart in the VA space. This is directly explained by our previous observation; there are different ways and intensities to express a certain emotion. Therefore, a type of inherent variation is present that is suppressed by the center loss hurting the overall performance.

### 4.5.2 Example

To better understand the concept of inherent variation, we can think of the following theoretical example. In Figure 4.5.2 we see the position of some samples in the feature space of an imaginary model. The pink circle denotes the region that the samples of the anger class are gathered and the green circle defines the region that the samples of the disgust class are gathered in the feature space. Using only the cross-entropy loss during training, the intra-class variation will not be reduced and the pairs of images  $a, b$  and  $c, d$  will be far in the feature space. Also, images  $b$  and  $c$  are very close although they depict different emotions. As mentioned before, this situation motivated researchers to propose the center loss so as to bring samples from the same class closer in the feature space. In Figure 4.5.3a the ideal feature space learned by the center loss is shown. As expected, the distance between the pairs of the same class is smaller by reducing the distance of each sample from its corresponding center. However, in reality the learned feature space would not have this form due to the effect of inherent variation. In particular, we observe that images  $b$  and  $c$  present similar facial expression characteristics although they depict different emotions because anger and disgust are sometimes expressed in similar ways like tight mouth and half closed eyes. As a result, their distance in the feature space cannot be reduced so much by the center loss resulting in the feature space of Figure 4.5.3b.

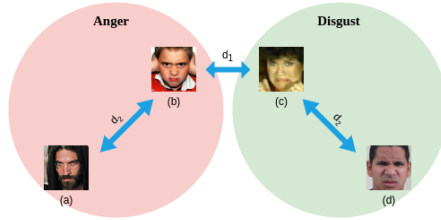


Figure 4.5.2: Feature space of an imaginary model before applying the center loss.

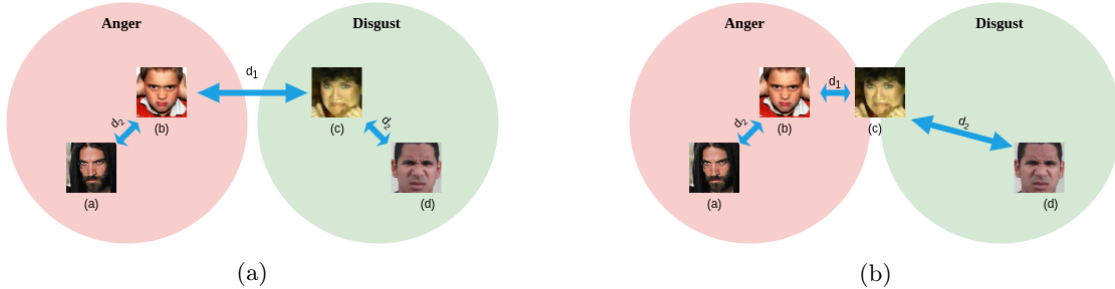


Figure 4.5.3: (a) Ideal and (b) real form of the feature space after applying the center loss.

### 4.5.3 Proposed Loss function

To deal with this problem, we should design a loss function that acts like center loss but also takes into account these situations. Specifically we want:

- ✓ To reduce variations from pose, illumination, identity, etc.
- ✗ Not to reduce the inherent variation of the task.

To take into account the inherent variation of the task we exploit the values of valence and arousal. In particular, samples that are close in the VA space share similar expression characteristics although they may belong to different classes. Respectively, samples that are far in the VA space share different expression characteristics although they may belong to the same class. Therefore, our proposed loss should be high when the samples are close in the VA space and low when they are far. We denote it as VA-based center loss and is computed as follows:

$$L_c^{va} = \frac{1}{2} \sum_{i=1}^m w_i \|x_i - c_{y_i}\|_2^2 \quad (4.5.1)$$

where  $w_i$  should follow the above rules; high when the VA values of  $x_i$  and  $c_{y_i}$  are close and low when their VA values are far based on a defined metric function. The problem here is that  $c_{y_i}$  is not a real sample but the feature vector of the center of class  $y_i$ . Therefore, it does not have a value for valence and arousal. Therefore, we should make an assumption regarding the VA values of the class centers. To overcome this, we assume that the valence and arousal of the center equals the mean valence and arousal of the samples that belong to each class:

$$v_j^m = \frac{\sum_{i=1}^m \delta(y_i = j) v_i}{\sum_{i=1}^m \delta(y_i = j)}, \quad a_j^m = \frac{\sum_{i=1}^m \delta(y_i = j) a_i}{\sum_{i=1}^m \delta(y_i = j)} \quad (4.5.2)$$

where  $v_i$  and  $a_i$  are the valence and arousal of data  $i$  respectively. Then, we add a weight factor for each point according to how close its valence and arousal is to that of the class center. If they are close, a large weight is applied to come even closer. If the values of valence and arousal are far then the impact of the loss is smaller. The weight  $w_i$  is defined as:

$$w_i = \frac{(d_{max} - d_i)^2}{4} \quad d_i = \sqrt{(v_i - v_{y_i}^m)^2 + (a_i - a_{y_i}^m)^2} \quad d_{max} = 2\sqrt{2} \quad (4.5.3)$$

where  $d_{max}$  denotes the maximum possible euclidean distance of two points in the VA space and  $d_i$  corresponds to the euclidean distance of data  $i$  from its class center  $y_i$  in the VA space.

## 4.6 Results

In Table 4.1 we can see the performance of all the metric learning models on the categorical model of AffectNet. We observe that almost all our proposed models achieve better recognition accuracy than the baseline model. The best performance is achieved by the local subclass loss with margin that not only reduces the variations of the task but also allows some intra-class variation. Also, the fact that the euclidean version of the island loss performs worse than the pure island loss is unexpected based on the intuition behind this approach. Although the learned feature space may seem better organized in the euclidean version, the final learned model does not perform better. Finally, the VA-based center loss achieves better performance than the pure center loss indicating that the integration of the VA values in the center loss benefits the system.

Table 4.1: Performance of metric learning models on the categorical model of AffectNet.

Architecture	Accuracy
Best baseline*	64.37
Center loss ( $\lambda = 0.1$ )	64.37
Center local loss ( $\lambda = 0.1$ )	64.77
Island loss ( $\lambda = 0.1$ )	<b>65.11</b>
Euclidean Island loss ( $\lambda = 0.1, m=2$ )	65.05
Euclidean Island loss ( $\lambda = 0.1, m=5$ )	65.06
Local subclass loss with margin ( $\lambda = 0.1, k=2, m=0.5$ )	64.43
Local subclass loss with margin ( $\lambda = 0.1, k=3, m=0.5$ )	64.43
Local subclass loss with margin ( $\lambda = 0.1, k=5, m=0.5$ )	65.26
Local subclass loss with margin ( $\lambda = 0.1, k=10, m=0.5$ )	<b>65.29</b>
VA-based Center local loss (only V)	64.11
VA-based Center local loss (only A)	64.6
VA-based Center local loss	<b>64.91</b>

\*Pretrained Densenet121 + Weighted Loss + Alignment + Augmentation

# Chapter 5

## Multitask learning models

### 5.1 Baseline architectures

The introduction of multi-task learning is particularly relevant in computer vision, as it has proved to successfully boost the performance of an individual task with the inclusion of other correlated tasks in the training process. Thus, the main task can benefit from other tasks by sharing a common feature representation and transferring knowledge between different domains. In facial expression recognition, a model can benefit by being trained simultaneously with other face-related tasks, such as AU detection, face recognition or landmark localization. Therefore, it is essential to explore the FER task using multi-task learning techniques.

#### 5.1.1 Dimensional model

Numerous models describing the human emotional states have been proposed by the psychology community. However, we have no clear evidence as to which representation is more appropriate and the majority of FER systems use either the categorical or the dimensional model of affect. It is shown that both emotion models have their respective benefits and drawbacks. Therefore, more and more studies try to leverage both representations through multi-task learning. By using a shared common feature representation we can capture the strong dependence between the categorical and the dimensional model improving the overall performance of the system. Since AffectNet dataset provides the values of valence and arousal for each sample, we can simultaneously train our model both in recognizing the basic facial expressions and regressing these continuous values. Since the two emotion models are correlated, multi-task learning can benefit the recognition task. Our initial multi-task architecture consists of a DenseNet backbone network that extracts the feature vector for each image. This shared vector is then used both for regressing the values of valence and arousal and recognizing the basic emotion. We denote as  $x$  the feature vector extracted from the backbone network. For the classification task, the predicted scores are computed as:

$$\hat{y}_i = \frac{e^{w_i^c \cdot x}}{\sum_{k=1}^7 e^{w_k^c \cdot x}} \quad (5.1.1)$$

where  $\hat{y}_i$  is the probability of emotion  $i$  and  $w_i^c$  is the classifier of emotion  $i$ . The network is trained using a weighted version of the traditional categorical cross entropy loss  $L^c$  since the dataset is highly imbalanced. In other words, the network is penalized more for misclassifying samples from under-represented classes than from well-represented classes as follows:

$$L^c = - \sum_{i=1}^7 w_i y_i \log(\hat{y}_i) \quad (5.1.2)$$

$$w_i = \frac{f_i}{f_{min}} \quad (5.1.3)$$

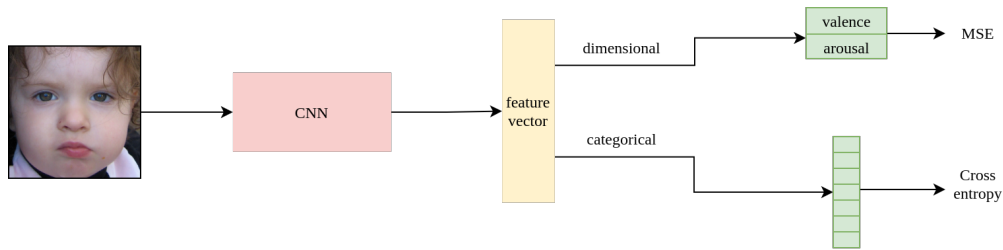


Figure 5.1.1: Proposed multi-task architecture using VA.

where  $y_i = 1$  if class  $i$  is the ground truth expression,  $f_i$  is the number of samples of the  $i$ -th class and  $f_{min}$  is the number of samples in the most under-represented class i.e. Disgust. For the regression task, the predicted values are computed as:

$$\hat{p}_v = w_v^r \cdot x, \quad \hat{p}_a = w_a^r \cdot x \quad (5.1.4)$$

where  $\hat{p}_v, \hat{p}_a$  are the predicted values of valence and arousal respectively and  $w_v^r, w_a^r$  correspond to the VA regressors. The classical MSE is used as a loss function that is defined as:

$$L^r = \frac{1}{2} \|\hat{p}_v - p_v\|_2^2 + \frac{1}{2} \|\hat{p}_a - p_a\|_2^2 \quad (5.1.5)$$

where  $p_v$  and  $p_a$  are the ground truth values of valence and arousal respectively. The overall loss function of our network training is defined as  $L = L^c + L^r$  and the overall architecture is shown in Figure 5.1.1.

## 5.1.2 Facial Landmarks

Facial landmark localization is a fundamental component in many face analysis tasks. The position of facial landmarks in the region of the eyes and the mouth are considered to be crucial in expression recognition since these parts convey rich emotional information. Therefore, facial landmarks localization can be integrated into a multi-task learning framework with our task. By learning to localize the position of the eyes and the mouth along with the FER task, more attention in these emotion-related regions is given with possible benefits in the final performance. However, directly regressing 24 facial landmarks (6 for each eye and 12 for the mouth region) would overkill the training and underestimate our main task. Generally in multi-task learning we want to combine tasks that are equally challenging so as to avoid situations where one task overshadows the other. Therefore, we should transform the landmark localization task in an easier task. To achieve this, we construct for each image a binary mask that depicts the regions of eyes and mouth in a more abstract level than the landmarks. In the mask, each pixel is equal to 1 when it is located in the eye or in the mouth region, otherwise it is equal to 0. An example of a mask is shown in Figure 5.1.2.

By converting the general landmark localization task to a mask localization task, we decrease the difficulty of the problem. At the same time, we train the FER task with a problem that is highly correlated and can benefit the overall performance. During training, the per-pixel binary cross-entropy loss is applied between the predicted masks and the ground truth for the landmark localization task along with the traditional weighted cross-entropy for the FER task. In Figure 5.1.3 the proposed architecture is presented.

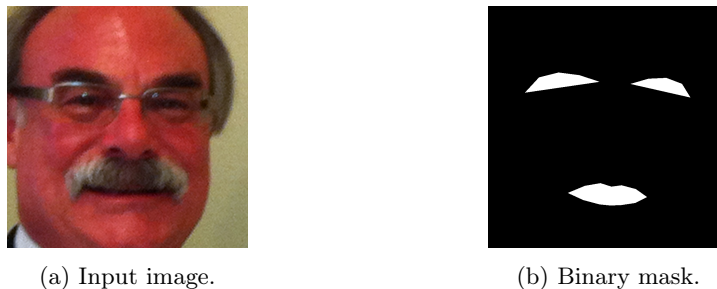


Figure 5.1.2: Inputs of the mask-based multi-task framework.

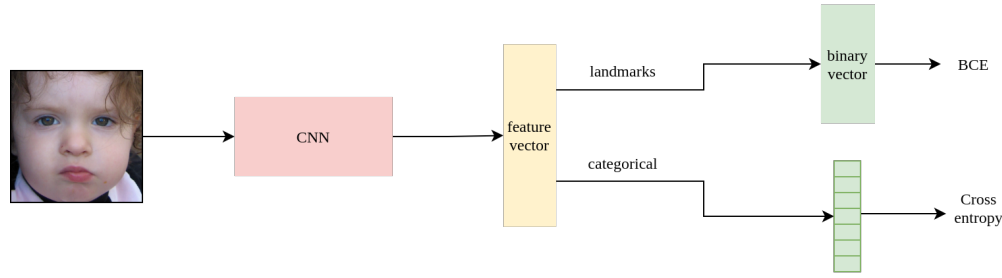


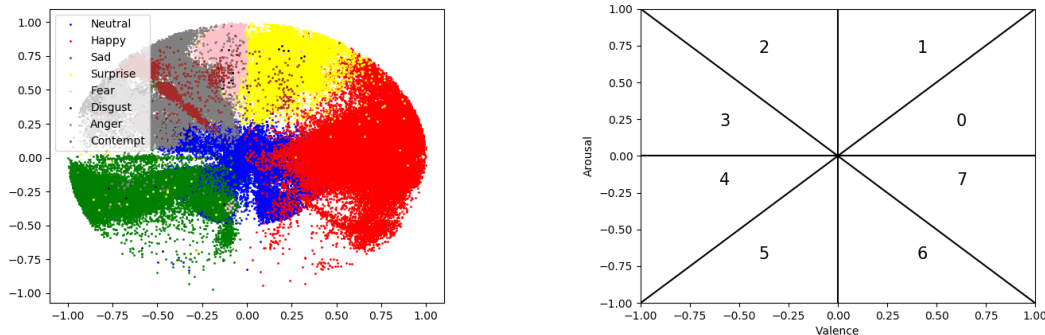
Figure 5.1.3: Proposed multi-task architecture using facial masks.

## 5.2 Advanced architectures

In the previous section, we investigated some baseline multi-task learning models for the FER task. Our next goal is to improve these models and propose architectures that capture the dependencies of the tasks more effectively.

### 5.2.1 Divide VA space

As mentioned before, in multi-task learning the tasks should be balanced i.e. their difficulty should be equal. However, simultaneously regressing the VA values and classifying the basic expression is not a balanced multi-task setting and hurts the performance. Therefore, we should transform the regression task as we did with the landmark localization task. We propose to divide the VA space in certain regions and convert the regression task into a classification one. The problem that occurs here is how to effectively divide the VA space to take full advantage of the benefits that VA values can provide. The most natural way is to divide the VA space in 4 quarters according to the sign of the values of valence and arousal and assign a class to each quarter. However, the categorical model contains 7 classes. That means that the ideal number of parts that we should divide the VA space is 7. For practical and computational reasons, we divide the space in 8 parts. The next parameter is the division strategy. If we observe the distribution of the samples in the VA space (Figure 5.2.1), the basic emotions are located around the neutral emotion that appears when valence and arousal are close to zero. Therefore, an angular division fits better. By training the network to predict the region of each sample in the valence-arousal space along with our main recognition task the network benefits from the former and the accuracy increases. The proposed architecture is shown in Figure 5.2.2 where CE loss is used for both tasks.



(a) Distribution of training sample in the VA space.

(b) Division strategy of VA space.

Figure 5.2.1: Angular division of valence-arousal space in eight parts.

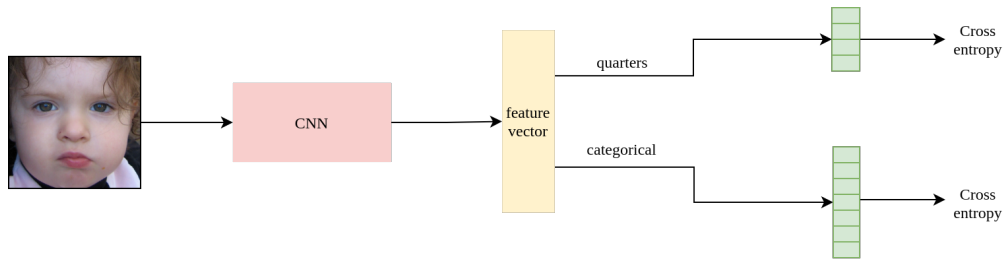


Figure 5.2.2: Proposed multi-task architecture using division of the VA space.

## 5.2.2 Concordance Correlation Coefficient

Another way to improve our baseline multi-task learning models is to use a better loss function for the regression task. While MSE is the most frequently used loss function for continuous tasks, affective computing researchers argue that using a correlation-based metric to evaluate the performance of dimensional emotion recognition is more appropriate than calculating its errors. The CCC is often used to measure the performance of dimensional emotion recognition since it takes the bias into Pearson's correlation coefficient. CCC measures the agreement between the true emotion dimension with predicted emotion degree. If the predictions shifted in value, the score is penalized in proportion to deviation. Hence, CCC is more reliable than Pearson correlation, MAE and MSE to evaluate the performance of dimensional speech emotion recognition. CCC is defined as:

$$\rho_c = \frac{2s_{xy}}{s_x^2 + s_y^2 + (\bar{x} - \bar{y})^2} \quad (5.2.1)$$

where  $s_x$  and  $s_y$  denote the variance of the predicted and ground truth values respectively,  $\bar{x}$  and  $\bar{y}$  are the corresponding mean values and  $s_{xy}$  is the respective covariance value. The range of CCC is from -1 (perfect disagreement) to 1 (perfect agreement). Hence, in our case we define  $L^r$  as:

$$L^r = 1 - \frac{\rho_v + \rho_a}{2} \quad (5.2.2)$$

where  $\rho_v$  and  $\rho_a$  are the respective CCC of valence and arousal.

## 5.2.3 Different fusion configurations

A crucial parameter in MTL is how to combine the different tasks at the feature level. When the tasks share all layers (excluding the last prediction layer), they have a lot of common knowledge to share. In our case, both tasks deal with a FER problem but using a different representation of the emotion. Therefore, there is a lot of information to be shared and the concatenation is done in the last FC layer (Figure 5.2.3a). In our case that we care more about the performance on the categorical model, we can change the configuration so as to give more attention there. We propose an architecture where the last classification layer takes as input both the output of the last FC layer and the VA predictions (Figure 5.2.3b). By doing this, we give an extra boost to the classification task. Finally, another configuration is to have a different last FC layer

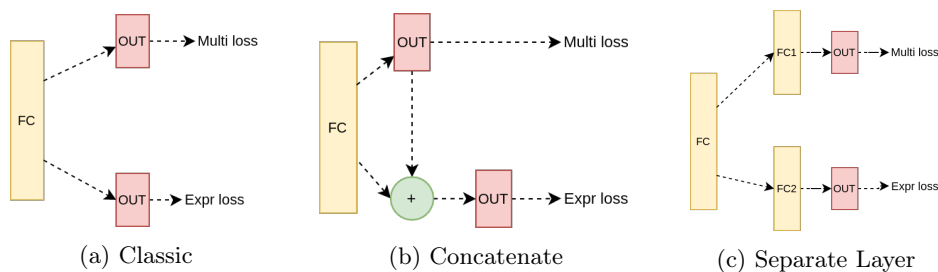


Figure 5.2.3: Different configurations for multi-task learning.



for each task (Figure 5.2.3c). In this configuration we enable the two tasks to learn more specific knowledge and share less information.

## 5.3 Sampler

An important part of training deep learning networks is data loading. The way that the loader handles the training data and passes them to the model influences a lot the optimization algorithm and the final performance of the system. In our previous models, we follow the traditional way of data loading where:

1. The training set is randomly subdivided into batches.
2. The model evaluates a batch by calculating its loss function (forward propagation).
3. The model estimates the gradients and actualizes its weights (back propagation).
4. The model repeats steps 2 and 3 until it reaches the end of the training set.
5. The model repeats the whole process as many times the user requires it. Optionally, the examples can be shuffled before training another epoch, thus making new batches reducing the risks of overfitting the data.

We observe that the traditional way of data loading deals with all samples equally. However, there are samples that are considered difficult (high loss) or easy (low loss). To take this into account we can define our own way of sampling (step 1) instead of passing all samples randomly.

### 5.3.1 Age Sampler

Generally, we can define the sampling strategy based on features that are correlated with the accuracy of the model or the distribution of the labels. In a classification task if we know beforehand that the difficult samples present a certain feature we can sample these examples more frequently. In our case, one intuitive approach is to take into account the age of a person before recognizing the emotion. We can assume that young people tend to depict more positive emotions than older people. Since there has been a lot of research in the age estimation task, we can easily validate our assumption in the AffectNet dataset. In Figure 5.3.1 we present the distribution of the expressions and the mean accuracy in different age groups. We observe that our assumption is invalid and the age is not correlated to the depicted expression. Also, our system performs almost the same among different age groups. Therefore, the information of age cannot benefit the system through sampling.

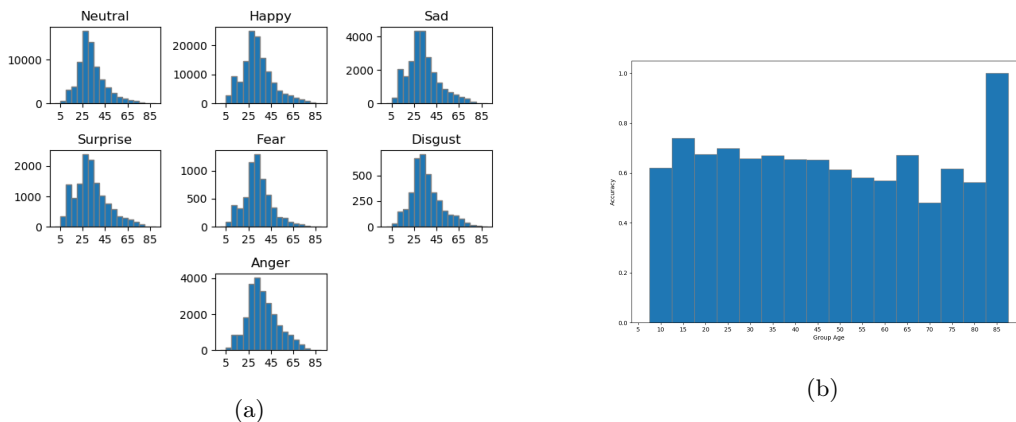


Figure 5.3.1: (a) Label distribution and (b) mean accuracy in different age groups.

### 5.3.2 Expression Sampler

An alternative idea is to make use of sampling to deal with the imbalance of the emotion datasets. As mentioned before, emotion databases usually contain more positive than negative emotions since people tend to share photos of themselves in a positive mood. In the previous models, we handle the imbalance problem using a weighted version of the CE loss. Alternatively, we can use the classical CE loss and sample each example  $i$  based on its weight  $w_i$  that (as in the weighted CE loss) is defined as:

$$w_i = \frac{f_{y_i}}{f_{min}} \quad (5.3.1)$$

where  $y_i$  is the class of sample  $i$ ,  $f_i$  is the number of samples of the  $i$ -th class and  $f_{min}$  is the number of samples in the most under-represented class. As a result, examples from underrepresented classes that are considered more difficult will be picked with more frequently by the sampler.

## 5.4 Focal Loss

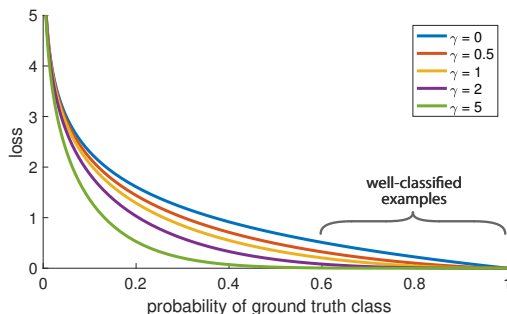


Figure 5.4.1: Focal loss for several values of  $\gamma$ .

In the previous deep models, we dealt with the problem of class imbalance using a weighted version of the CE loss. However, the extreme class imbalance encountered during training of our networks overwhelms the CE loss. Easily classified happy and neutral samples comprise the majority of the loss and dominate the gradient. Lin et al. [Lin+17b] proposed focal loss to deal with class imbalance in 1-stage object detection where there is an extreme imbalance between foreground and background classes during training. We can use the proposed focal loss in our task since most emotion datasets are imbalanced and our models achieve the lowest accuracy in the less represented class. Focal loss in a binary classification problem is defined as:

$$L_{focal} = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (5.4.1)$$

$$p_t = \begin{cases} p & , \text{ if } y = 1 \\ 1 - p & , \text{ otherwise} \end{cases} \quad (5.4.2)$$

where it adds a factor  $(1 - p_t)^\gamma$  to the standard cross entropy criterion. Setting  $\gamma > 0$  reduces the relative loss for well-classified examples ( $p_t > 0.5$ ), putting more focus on hard, misclassified examples. The focal loss is visualized for several values of  $\gamma \in [0, 5]$  in Figure 5.4.1. We should note two properties of the focal loss:

1. When an example is misclassified and  $p_t$  is small, the modulating factor is near 1 and the loss is unaffected. As  $p_t \rightarrow 1$ , the factor goes to 0 and the loss for well-classified examples is down-weighted.
2. The focusing parameter  $\gamma$  smoothly adjusts the rate at which easy examples are downweighted. When  $\gamma = 0$ , FL is equivalent to CE, and as  $\gamma$  is increased the effect of the modulating factor is likewise increased. Intuitively, the modulating factor reduces the loss contribution from easy examples and extends the range in which an example receives low loss. For instance, with  $\gamma = 2$ , an example classified with  $p_t = 0.9$  would have 100 times lower loss compared with CE and with  $p_t \approx 0.968$  it

Table 5.1: Performance of multi-task learning models on the categorical model of AffectNet.

Architecture	Accuracy		
	Classic	Concatenate	Separate
Best baseline*		64.37	
Best baseline* + MSE	64.8	64.74	-
Best baseline* + CCC	65.63	65.43	65.2
Best baseline* + Half (MSE)	-	64.89	-
Best baseline* + Half (BCE)	-	65.26	-
Best baseline* + Quarter	65.14	65.63	-
Best baseline* + Eight	<b>65.86</b>	<b>65.97</b>	65.43
Best baseline* + Twelve	-	65.11	-

\*Pretrained Densenet121 + Weighted Loss + Alignment + Augmentation

would have 1000 times lower loss. This in turn increases the importance of correcting misclassified examples (whose loss is scaled down by at most 4 times for  $p_t \leq .5$  and  $\gamma = 2$ ).

## 5.5 Results

In Table 5.1 we can see the performance of all the multitask learning models on the categorical model of AffectNet. First, we observe that all multi-task learning models perform better than the baseline single-task model verifying that the classification task benefits from the shared feature representation and the integration of the VA values as an additional task. Also, in agreement with recent studies, using the CCC loss for the regression task increases the classification accuracy of the model verifying that in MTL a correlation-based regression loss performs better. Finally, dividing the VA space in 4 or 8 parts is more effective than directly regressing the VA values illustrating the benefits of our idea to transform the regression task into a classification task.

In Table 5.2 we can see the effect of different samplers in a multi-task learning model. We can see the neither the age or the expression sampler manage to improve the performance of the model. Also, in the same table we can see the accuracy of the multi-task learning models before and after using the focal loss. We observe that the benefits of focal loss are not present in our task and do not improve the overall performance.

Since our proposed multi-task and metric learning models improved the accuracy of our baseline model, we can combine them in a single framework in order to exploit their respective benefits. Specifically, we can apply the proposed metric losses to the shared feature representation of the multi-task learning networks. In Table 5.3 we see the performance of models combining metric and multi-task learning on the categorical model of AffectNet. Despite the intuition behind this approach, the performance of the models is worse than just using a separate metric or multi-task learning model. We believe that the decrease in the performance is

Table 5.2: Performance of models using proposed sampling techniques and focal loss.

Expr Loss	Multi Loss	Sampler	Accuracy
WCE	WCE	-	65.97
WCE	WCE	Expression	60.74
WCE	WCE	Age	62.86
CE	WCE	Expression	62.77
CE	CE	Expression	62.43
Focal	WCE	-	64
Focal	Focal	-	64.74

Model: Multi-task learning model in 8-division

Table 5.3: Performance of models combining metric and multi-task learning on the categorical model of AffectNet.

Multi-task	Metric	Accuracy		
		Classic	Concatenate	Separate
CCC	Island loss	65.6	65.57	-
CCC	Local subclass loss with margin (k=3)	-	64.89	-
CCC	Local subclass loss with margin (k=5)	65.77	65.43	65.71
Eight	Island loss	65.25	65.2	-
Eight	Local subclass loss with margin (k=3)	-	64.86	-
Eight	Local subclass loss with margin (k=5)	64.37	65.69	65.69

due to the fact that the shared feature space of the multi-task models is not proper for metric learning since in our metric losses we assumed that the feature space is learned only for the categorical model of affect.

Previously our deep learning models were trained end-to-end for classification. Specifically, the input image  $I$  passed through a DenseNet to obtain  $1024 \times 7 \times 7$  feature maps. Then, a global max-pooling function was applied to get the feature vector  $x$  of the image. Then, each classifier acted as a weight vector that was multiplied with the feature  $x$  and then passes through a softmax activation function assigning a probability score to its emotion category. Therefore, the classification and the feature extraction part were simultaneously optimized end-to-end. A different approach is to learn the feature vectors for each sample as before but then train another classifier using the learned features as input. In particular, we train an SVM classifier on the extracted features of the deep model. To train the SVM classifier on a GPU we used ThunderSVM<sup>1</sup> that exploits GPUs and multi-core CPUs to achieve high efficiency. By separating the classification from the feature extraction part, we intend to take advantage of the benefits of SVM that searches for a hyperplane in an N-dimensional space to distinctly classify the data points. In the proposed metric learning models especially, by following this technique we exploit more effectively the discriminative feature space that the metric losses create. In Table 5.4 we see the accuracy of different proposed deep models before and after using SVM. We observe that there is a consistent increase in the performance of almost every model. However, the increase is negligible concerning the extra training time and computing power that the training of the SVM classifier requires.

Table 5.4: Performance of models using SVM.

Architecture	Accuracy	
	Softmax	SVM
Eight + Island loss	65.25	65.71
Eight + Local subclass loss with margin (k=5)	65.69	65.63
CCC + Island loss	65.6	65.69
CCC + Local subclass loss with margin (k=5)	65.77	65.8

<sup>1</sup><https://github.com/Xtra-Computing/thundersvm>

# Chapter 6

## Emotion-GCN

In this chapter we present Emotion-GCN, a novel MTL framework that exploits the dependencies between the categorical and dimensional model of affect using a GCN to recognize facial expressions in-the-wild. First, we make an introduction to the architecture of GCNs. Then, we present ML-GCN model for multi-label image recognition that was the inspiration of our proposed model. Finally, we discuss in detail the architecture of Emotion-GCN, its experimental results, providing a clear view of the accuracy improvements introduced by our method.

### 6.1 Overview of GCN

Many important real-world datasets come in the form of graphs or networks: social networks, knowledge graphs, protein-interaction networks, the World Wide Web, etc. Before GCNs, very little attention had been devoted to the generalization of neural network models to such structured datasets. Generalizing well-established neural models like RNNs or CNNs to work on arbitrarily structured graphs is a challenging problem. Kipf et al. [KW17] started from the framework of spectral graph convolutions and introduced simplifications that in many cases allow both for significantly faster training times and higher predictive accuracy, reaching state-of-the-art classification results on a number of benchmark graph datasets. Currently, most GCNs have a somewhat universal architecture in common. For these models, the goal is to learn a function of signals/features on a graph  $G = (V, E)$  which takes as input

- a feature description  $x_i$  for every node  $i$ ; summarized in a  $N \times D$  feature matrix  $X$  ( $N$ : number of nodes,  $D$ : number of input features).
- a representative description of the graph structure in matrix form; typically in the form of an adjacency matrix  $A$  (or some function thereof).

and produces a node-level output  $Z$  (an  $N \times F$  feature matrix, where  $F$  is the number of output features per node). Every neural network layer can then be written as a non-linear function:

$$H^{l+1} = f(H^l, A) \tag{6.1.1}$$

with  $H^{(0)} = X$  and  $H^{(L)} = Z$  (or  $z$  for graph-level outputs),  $L$  being the number of layers. The specific models then differ only in how  $f$  is chosen and parameterized.

#### 6.1.1 A simple example

As an example, we consider the following very simple form of a layer-wise propagation rule:

$$f(H^{(l)}, A) = \sigma(A H^{(l)} W^{(l)}) \tag{6.1.2}$$

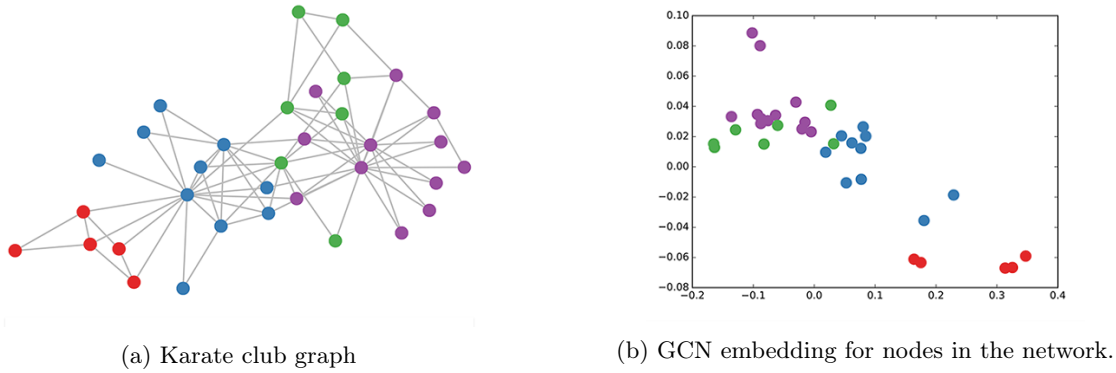


Figure 6.1.1: Applying a GCN model in Zachary’s Karate club network (Source: [Bra+07])

where  $W^{(l)}$  is a weight matrix for the  $l$ -th neural network layer and  $\sigma(\cdot)$  is a non-linear activation function like the ReLU.

Despite its simplicity, this model is already quite powerful. But first, let us address two limitations of this simple model: multiplication with  $A$  means that, for every node, we sum up all the feature vectors of all neighboring nodes but not the node itself (unless there are self-loops in the graph). We can fix this by enforcing self-loops in the graph: we simply add the identity matrix to  $A$ . The second major limitation is that  $A$  is typically not normalized and therefore the multiplication with  $A$  will completely change the scale of the feature vectors (we can understand that by looking at the eigenvalues of  $A$ ). Normalizing  $A$  such that all rows sum to one, i.e.  $D^{-1}A$ , where  $D$  is the diagonal node degree matrix, gets rid of this problem. Multiplying with  $D^{-1}A$  now corresponds to taking the average of neighboring node features. In practice, dynamics get more interesting when we use a symmetric normalization, i.e.  $D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$  (as this no longer amounts to mere averaging of neighboring nodes). Combining these two tricks, we essentially arrive at the propagation rule introduced in [KW17]:

$$f(H^{(l)}, A) = \sigma(\hat{D}^{\frac{1}{2}} \hat{A} \hat{D}^{\frac{1}{2}} H^{(l)} W^{(l)}) \quad (6.1.3)$$

with  $\hat{A} = A + I$ , where  $I$  is the identity matrix and  $\hat{D}$  is the diagonal node degree matrix of  $\hat{A}$ .

### 6.1.2 Embedding the karate club network

Now, we will take a look at how the above simple GCN model works on a well-known graph dataset: Zachary’s karate club network (Problem taken from [here](#)). We take a 3-layer GCN with randomly initialized weights. Now, even before training the weights, we simply insert the adjacency matrix of the graph and  $X = I$  (i.e. the identity matrix, as we don’t have any node features) into the model. The 3-layer GCN now performs three propagation steps during the forward pass and effectively convolves the 3rd-order neighborhood of every node (all nodes up to 3 hops away). Remarkably, the model produces an embedding of these nodes that closely resembles the community-structure of the graph as shown in Figure 6.1.1b. It is remarkable that we have initialized the weights completely at random and have not yet performed any training updates so far.

This might seem somewhat surprising. We can shed some light on this by interpreting the GCN model as a generalized, differentiable version of the well-known Weisfeiler-Lehman algorithm on graphs. The (1-dimensional) Weisfeiler-Lehman algorithm works as follows:

For all nodes  $v_i \in G$ :

- Get features  $h_{v_j}$  of neighboring nodes  $v_j$
- Update node features  $h_{v_i} \leftarrow \text{hash}\left(\sum_j h_{v_j}\right)$ , where  $\text{hash}(\cdot)$  is (ideally) an injective hash function

Repeat for  $k$  steps or until convergence.

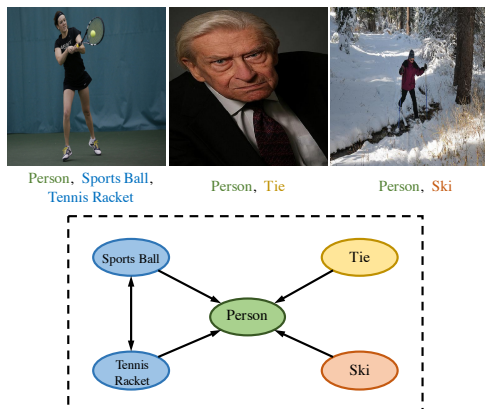


Figure 6.2.1: A graph over the labels to model label dependencies in multi-label image recognition.

In practice, the Weisfeiler-Lehman algorithm assigns a unique set of features for most graphs. This means that every node is assigned a feature that uniquely describes its role in the graph. Exceptions are highly regular graphs like grids, chains, etc. For most irregular graphs, this feature assignment can be used as a check for graph isomorphism (i.e. whether two graphs are identical, up to a permutation of the nodes). Going back to our Graph Convolutional layer-wise propagation rule (now in vector form):

$$h_{v_i}^{(l+1)} = \sigma \left( \sum_j \frac{1}{c_{ij}} h_{v_j}^{(l)} W^{(l)} \right) \quad (6.1.4)$$

where  $j$  indexes the neighboring nodes of  $v_i$ .  $c_{ij}$  is a normalization constant for the edge  $(v_i, v_j)$  which originates from using the symmetrically normalized adjacency matrix  $D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$  in our GCN model. We now see that this propagation rule can be interpreted as a differentiable and parameterized (with  $W^{(l)}$ ) variant of the hash function used in the original Weisfeiler-Lehman algorithm. If we now choose an appropriate non-linearity and initialize the random weight matrix such that it is orthogonal, this update rule becomes stable in practice (also thanks to the normalization with  $c_{ij}$ ). And we make the remarkable observation that we get meaningful smooth embeddings where we can interpret distance as (dis-)similarity of local graph structures.

## 6.2 Inspiration

Here we will present the model that inspired us to implement our proposed Emotion-GCN model. In multi-label classification, there has been a lot of research on how to properly capture and explore the correlation between labels. For example, a key step in multi-label image recognition is to model the label dependencies since the objects normally co-occur in the physical world. As shown in Figure 6.2.1, the labels *person* co-occurs many times with the labels *racket*, *tie* or *ski*. However, the labels *tie* and *ski* do not co-occur in the physical world. By capturing these dependencies, a model can improve its performance.

A naive way to address the multi-label recognition problem is to treat the objects in isolation and convert the multi-label problem into a set of binary classification problems to predict whether each object of interest presents or not. However, these methods are essentially limited by ignoring the complex topology structure between objects. This stimulates research for approaches to capture and explore the label correlations in various ways. Chen et al. [Che+19b] proposed ML-GCN to capture the label correlations for multi-label image recognition, which properties with scalability and flexibility impossible for competing approaches. Instead of treating object classifiers as a set of independent parameter vectors to be learned, they propose to learn inter-dependent object classifiers from prior label representations, e.g., word embeddings, via a GCN based mapping function. In the following, the generated classifiers are applied to image representations generated by another sub-net to enable end-to-end training. As the embedding-to-classifier mapping parameters are shared across all classes (i.e., image labels), the gradients from all classifiers impact the GCN based classifier generation function. This implicitly models the label correlations. Furthermore, to explicitly model the label

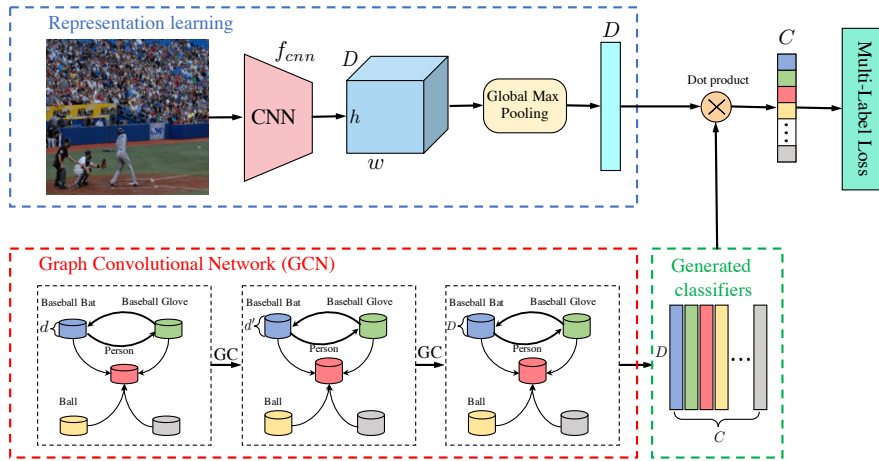


Figure 6.2.2: Overall framework of ML-GCN model for multi-label image recognition..

dependencies for classifier learning, they design an effective label correlation matrix to guide the information propagation among nodes in GCN. Specifically, they propose a re-weighted scheme to balance the weights between a node and its neighborhood for node feature update, which effectively alleviates overfitting and over-smoothing. The experiments on two multi-label image recognition datasets show that their approach obviously outperforms existing state-of-the-art methods. The overall framework of ML-GCN is shown in Figure 6.2.2, which is composed of two main modules, i.e., the image representation learning and GCN based classifier learning modules. Inspired by this architecture, we propose Emotion-GCN a novel GCN based MTL framework for FER in the wild.

### 6.3 Emotion-GCN

We believe that the strong dependence between the categorical and the dimensional model is not fully exploited when they only share a feature representation in a multi-task learning setting. Figure 6.3.1 illustrates this relation using the validation set of AffectNet. Inspired by the proposed architecture of ML-GCN for multi-label classification, we propose Emotion-GCN a novel GCN based MTL framework for FER in the wild. The main idea in Emotion-GCN is to generate dependent expression classifiers and valence-arousal (VA) regressors through a GCN based mapping function instead of learning them as separate parameter vectors. The generated vectors are then applied to an extracted image representation to enable end-to-end training. Hence, the dependence between the categorical and the dimensional emotion models is explicitly

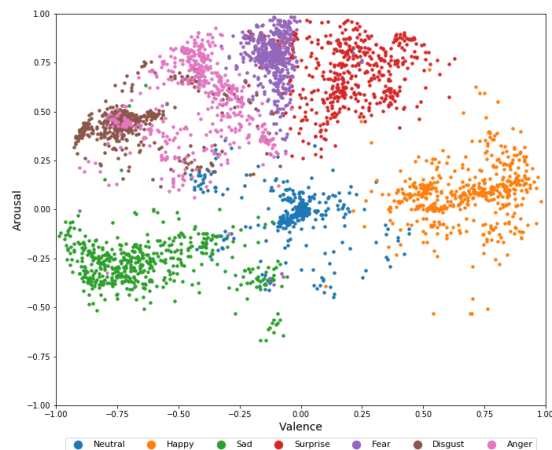


Figure 6.3.1: Illustration of the dependencies between the categorical and the dimensional model.



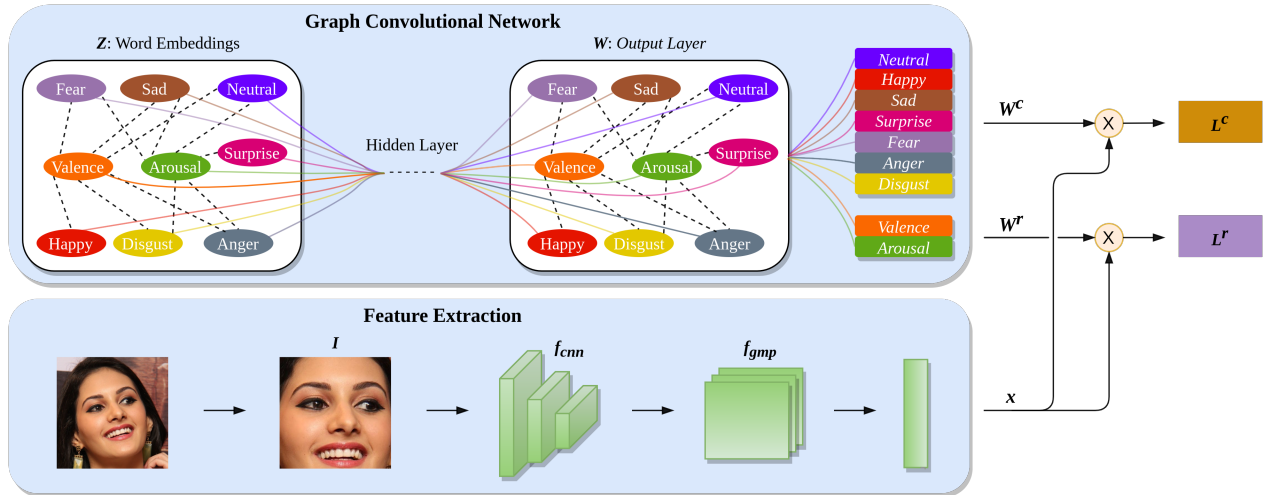


Figure 6.3.2: Overall architecture of our Emotion-GCN model for FER in the wild.

captured through both a shared feature representation and the dependent classifiers and regressors. The overall architecture of the proposed Emotion-GCN is shown in Figure 6.3.2.

### 6.3.1 Feature extraction

Given an input image  $I$  of size  $227 \times 227$  pixels, we obtain  $1024 \times 7 \times 7$  feature maps using a DenseNet. Each layer in DenseNet obtains additional inputs from all preceding layers and passes on its own feature maps to all subsequent layers to ensure maximum information flow between layers. Then, a global max-pooling function is applied to get the feature vector  $x$  of the image:

$$x = f_{gmp}(f_{cnn}(I)) \in R^D \quad (6.3.1)$$

where  $f_{cnn}$  corresponds to the convolution layers of the DenseNet,  $f_{gmp}$  to the the global max-pooling function and  $D = 1024$ .

### 6.3.2 Dependent classifiers and regressors

Given the feature vector  $x$  of the image, we want to learn one classifier for each facial expression (classification task) and one regressor for each dimension in the VA space (regression task). Each classifier acts as a weight vector that is multiplied with the feature  $x$  and then passes through a softmax activation function assigning a probability score to its emotion category. In parallel, each regressor follows the same procedure without passing through an activation function since its output is a predicted continuous value for either valence or arousal. We denote the expression classifiers by  $W^c \in R^{7 \times D}$  and the VA regressors by  $W^r \in R^{2 \times D}$  where each row of the matrices corresponds to a classifier and a regressor respectively. Traditionally, these vectors are optimized as individual parameters of the deep learning network. To capture the correlation between the categorical and the dimensional model, we propose to learn these vectors using a GCN that maps their word embeddings to dependent classifiers and regressors retaining the shared information between the two tasks.

#### GCN formulation

For an overview in GCN refer to section 6.1. In our case, the nodes of the graph correspond to the seven expression labels and the two VA dimensions i.e.  $V = \{\text{Neutral, Happy, Sad, Surprise, Fear, Disgust, Anger, Valence, Arousal}\}$ . Hence, the input is  $Z \in R^{9 \times d}$  that contains the word embedding of each node ( $d$  is the dimensionality of the embeddings). Each GCN layer  $l$  takes the node representations from the previous layer  $H^l$  as inputs and outputs new node representations  $H^{l+1}$ . Finally, the output representation of the last layer  $W \in R^{9 \times D}$  contains the dependent classifiers and regressors. The first seven rows of the output matrix  $W$  constitute the classification part  $W^c$  and the rest two the regression part  $W^r$ .

### Adjacency Matrix Design

According to the update rule of a GCN the feature of a node in the graph is the weighted sum of its own feature and the adjacent nodes' features. Since the purpose behind the use of a GCN is to exploit the dependencies between the categorical and the dimensional model, we should design the adjacency matrix  $A$  to this direction. As before, we assume that the first seven rows of  $A$  correspond to the basic expressions and the last two are the continuous dimensions i.e. valence and arousal. Since we deal with a multi-task and not a multi-label recognition problem, we are interested only in the correlation between the categorical and the dimensional model. Hence, we set the other pairs of  $A$  to zero except for the diagonal to allow self-loops. Also, we take the absolute value of the correlation to ignore its type (positive or negative) and focus on its amplitude. The correlation matrix  $A \in R^{9 \times 9}$  can be written as:

$$A_{ij} = \begin{cases} 1, & \text{if } i = j \\ |c_{ij}|, & \text{if } i \in \text{Cat} \wedge j \in \text{Dim} \\ |c_{ij}|, & \text{if } j \in \text{Cat} \wedge i \in \text{Dim} \\ 0, & \text{else} \end{cases} \quad (6.3.2)$$

where  $\text{Cat}$  and  $\text{Dim}$  are the set of indices of the categorical and the dimensional labels respectively. As a correlation metric, we use the Spearman's rank correlation coefficient [Spe61] that for two variables  $X = \{x_1, \dots, x_N\}$  and  $Y = \{y_1, \dots, y_N\}$  is defined as:

$$c_{xy} = \frac{\sum_{k=1}^N x_{k,r} y_{k,r}}{\sqrt{\sum_{k=1}^N x_{k,r}^2 \sum_{k=1}^N y_{k,r}^2}} \quad (6.3.3)$$

where  $x_{k,r}$  and  $y_{k,r}$  are the rank transformation of the initial values  $x_k$  and  $y_k$ . Following the ideas in ML-GCN [Che+19b], we use a threshold  $\tau$  to filter the noisy edges as follows:

$$A'_{ij} = \begin{cases} 1, & \text{if } A_{ij} \geq \tau \\ 0, & \text{if } A_{ij} < \tau \end{cases} \quad (6.3.4)$$

where  $\tau = 0.1$  to enable the propagation of information between weakly correlated nodes. As shown in [LHW18], the learned features of each node may be over-smoothed and become indistinguishable when a binary correlation matrix is used. To alleviate the over-smoothing problem, we apply the re-weighted scheme of ML-GCN that is defined as:

$$A''_{ij} = \begin{cases} (p / \sum_{j=1}^9 A'_{ij}) \times A'_{ij}, & \text{if } i \neq j \\ 1 - p, & \text{if } i = j \end{cases} \quad (6.3.5)$$

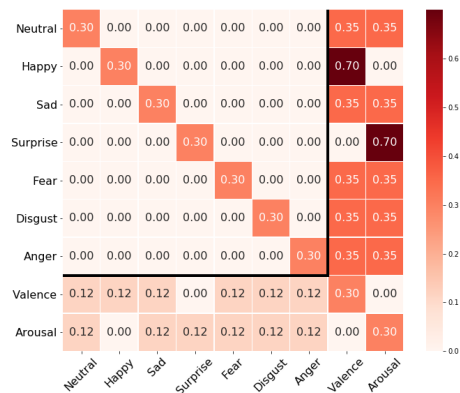


Figure 6.3.3: Adjacency matrix of our Emotion-GCN.

Table 6.1: Performance of Emotion-GCN on the categorical model of AffectNet and Aff-Wild2.

Method	AffectNet	Aff-Wild2
Single-task	64.37	45.06
Multi-task + MSE	64.8	43.1
Multi-task + CCC	65.69	43.33
Emotion-GCN	<b>66.46</b>	<b>48.92</b>

where the variable  $p$  determines the weights assigned to a node itself and its adjacent nodes. We set  $p = 0.7$  to increase the influence of the neighborhood nodes. Figure 6.3.3 shows the output adjacency matrix  $A''$  of Emotion-GCN using the training set of AffectNet. As we expected, pairs like happy-valence and surprise-arousal are strongly connected since the presence of the one usually denotes the presence of the other (Figure 6.3.1).

## 6.4 Results

### 6.4.1 Ablation Study

First, we investigate the performance of four different networks on the categorical and the dimensional model of affect to present the improvements that Emotion-GCN introduces.

#### Categorical Model

We trained (i) a single-task network for discrete FER using the weighted CE loss. Then, two multi-task networks were trained for discrete and continuous FER using a weighted CE loss for the classification task and a (ii) MSE or (iii) CCC loss for the regression task. Finally, we trained (iv) the proposed Emotion-GCN model that generates dependent classifiers and regressors using a 2-layer GCN as presented in Figure 6.3.2. Table 6.1 shows the results of our experiments. In AffectNet we can see that learning to predict the VA values as an additional task in a MTL framework increases the accuracy by 1.32% since the shared representation improves the generalization of the network. Also, in agreement with recent studies, using the CCC loss for the regression task increases the classification accuracy of the model verifying that in MTL a correlation-based regression loss performs better. Finally, Emotion-GCN achieves a total accuracy of **66.46%**, since the dependent classifiers manage to effectively capture the dependencies between the two emotion representations. The confusion matrices for these models are shown in Figure 6.4.1. It can be seen that our proposed method increases the accuracy for most classes while the single-task network performs better only in the neutral class. Similarly, in Aff-Wild2 the total evaluation metric increases by 3.86% indicating that the benefits of Emotion-GCN generalize in more datasets. To investigate how different choices for the parameters of the GCN affect the performance, we perform additional experiments in Table 6.2. We observe that the number of GCN layers ( $L$ ) is the most crucial parameter since using only one layer decreases the accuracy a lot.

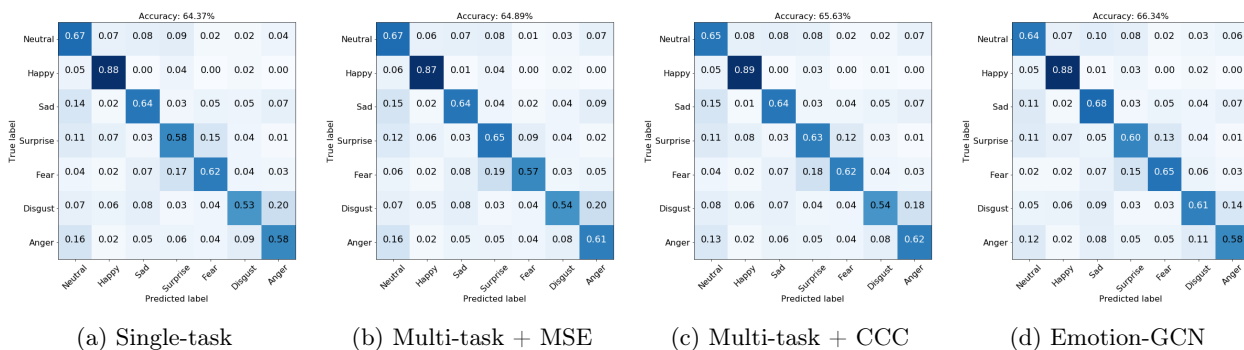


Figure 6.4.1: Confusion matrices of our Emotion-GCN on the validation set of AffectNet.

Table 6.2: Classification accuracy of Emotion-GCN on the categorical model of AffectNet using different values for  $\tau$ ,  $p$  and  $L$  (number of GCN layers). In the first table  $p = 0.7$  and in the second table  $\tau = 0.1$ .

		$\tau$			
		0	0.05	0.1	0.15
$L$	1	53.09	43.85	53.71	54.94
	2	65.29	65.31	<b>66.46</b>	65.69
	3	65.77	64.94	64.74	65.74

		$p$						
		0.2	0.3	0.4	0.5	0.6	0.7	0.8
$L$	1	49.6	50.94	52.57	52.66	52.86	53.71	59.31
	2	65.71	66.14	65.29	66.29	66.09	<b>66.46</b>	65.06
	3	65.29	65.54	65.89	66.09	65.54	64.74	65.83

Table 6.3: Performance of Emotion-GCN on the dimensional model of AffectNet and Aff-Wild2.

Method	AffectNet		Aff-Wild2	
	CCC-V	CCC-A	CCC-V	CCC-A
Single-task	0.761	0.628	0.416	0.501
Multi-task + MSE	0.752	0.572	0.435	0.378
Multi-task + CCC	<b>0.768</b>	<b>0.651</b>	0.408	0.481
Emotion-GCN	0.767	0.649	<b>0.457</b>	<b>0.514</b>

## Dimensional Model

For the evaluation of our models on the VA space, the networks (ii), (iii) and (iv) are the same since they are trained for both discrete and continuous FER. Additionally, a single-task network for VA regression was trained using the CCC loss. Table 6.3 shows the performance evaluation of our experiments on the dimensional model. In AffectNet we can see that the multi-task network with the CCC loss improves the regression performance along with the classification one verifying that both tasks benefit from the shared feature representation. Actually, we observe that the increase in the performance of arousal prediction is higher than that of valence. This is due to the fact that most emotions have positive value of arousal (Figure 6.3.1) and a simultaneous emotion classification provides more useful information in the task of arousal regression. Finally, our GCN based approach achieves similar performance with that of the multi-task network on the dimensional model. In Aff-Wild2, our proposed model surpasses the performance of both the single-task and multi-task networks. Overall, our method presents significant improvements in both the categorical and the dimensional model. In Figure 6.4.2 we present some positive and negative results of our Emotion-GCN model on AffectNet. The first column indicates the ground truth values. The second and the third column present the predictions of the multi-task network trained with CCC loss and our Emotion-GCN respectively. To examine which network better exploits the emotional dependencies, we selected samples where the predictions of the networks on the dimensional model are close. In the first four samples our Emotion-GCN model successfully recognizes the depicted emotion while the multi-task network fails indicating that our proposed model effectively captured the dependencies presented in the VA space (Figure 6.3.1). However, there are cases where our network fails (last two samples).

## 6.4.2 Visualization

To further analyze the effectiveness of our approach, we investigate the similarity between the dependent classifiers and regressors. In Figure 6.4.3 the cosine similarity of the learned vectors on AffectNet by our single-task networks, our best multi-task network and our Emotion-GCN are presented. As we can see, learning a shared representation for both tasks through MTL slightly increases the similarity between the expression classifiers and the VA regressors. Our proposed method increases their similarity even more in consistence with the dependencies presented in Figure 6.3.1. Specifically, the regressor of valence comes closer to the classifier of happy and the regressor of arousal closer to the classifiers of anger, disgust, fear and surprise. These emotions appear in regions of VA space where the values of valence or arousal are high.






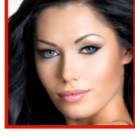
	Ground Truth	Multi-Task	Emotion-GCN		Ground Truth	Multi-Task	Emotion-GCN
	Neutral V: 0.0 A: 0.2	Surprise V: 0.12 A: 0.15	Neutral V: 0.15 A: 0.15		Fear V: -0.05 A: 0.93	Surprise V: -0.07 A: 0.95	Fear V: -0.07 A: 1.0
	Happy V: 0.64 A: 0.02	Neutral V: 0.47 A: -0.12	Happy V: 0.42 A: -0.11		Surprise V: 0.13 A: 0.66	Surprise V: 0.11 A: 0.98	Fear V: 0.09 A: 0.92
	Sad V: -0.68 A: -0.37	Neutral V: -0.32 A: -0.12	Sad V: -0.34 A: -0.13		Happy V: 0.77 A: -0.09	Happy V: 0.46 A: -0.13	Neutral V: 0.42 A: -0.14

Figure 6.4.2: Predictions of our models on samples of AffectNet.

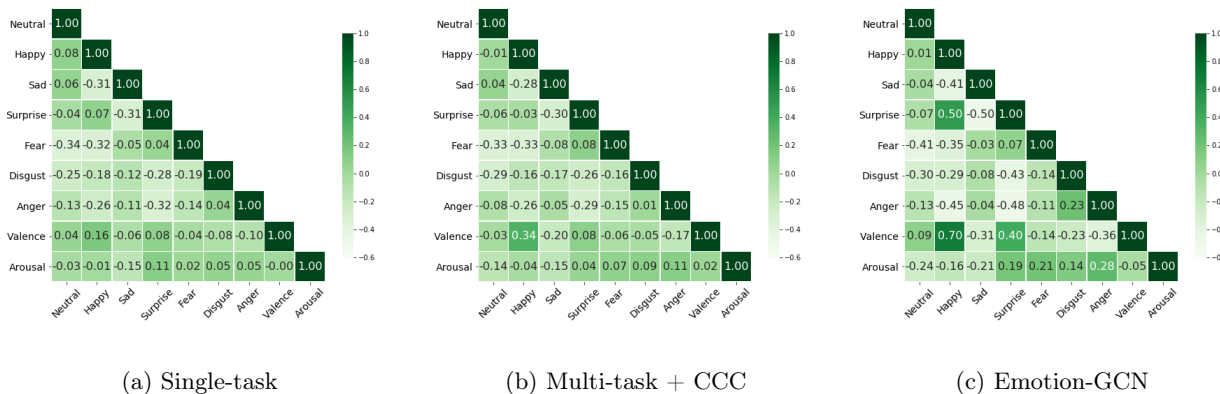


Figure 6.4.3: Visualization of the cosine similarity between the learned classifiers and regressors by Emotion-GCN on AffectNet.

Therefore, the proposed network has successfully captured the dependence between the categorical and the dimensional emotion representation.

We also observe an increase in the similarity between the pairs valence-surprise and happy-surprise while their respective nodes are not adjacent in the graph (Figure 6.3.3). These dependencies are successfully captured by our network since in a 2-layer GCN a node incorporates information from a 2-hop neighborhood [DMT18]. To further examine whether the dependence between the emotions of happiness and surprise is reasonable, we compute their co-occurrence in a multi-label emotion dataset. The EMOTIC dataset [Kos+17b] is a collection of images of people in unconstrained environments annotated according to their apparent emotional states. Each person is annotated for 26 discrete categories, with multiple labels assigned to each image. About 25% of the samples labeled as happy are labeled as surprise too that indicates that these two emotions are strongly related indeed.

### 6.4.3 Dependence between classifiers

Inspired by the fact that there are dependencies between the basic emotions, we trained a similar GCN based network but we enabled this time the direct propagation of information between the seven emotion classifiers. Instead of setting their correlation to zero as in Emotion-GCN, we compute the conditional probability matrix of the basic emotions like in ML-GCN. The model achieves a total accuracy of 66.23% on the categorical model of AffectNet. Despite the intuition behind this approach, the recognition performance is slightly lower

Table 6.4: Comparison of Emotion-GCN with state-of-the-art methods on AffectNet (7-way classification).

Method	Accuracy
IPA2LT [ZSC18]	57.31
gACNN [Li+18]	58.78
Facial Motion Prior Network [Che+19a]	61.52
CAKE [Ker+18]	61.7
OADN [DZC20]	61.89
CNNs and BOVW + global SVM [GIP19]	63.31
Siamese [HNM19]	64
<b>Emotion-GCN (ours)</b>	<b>66.46</b>

than that of our proposed method. We believe that by enabling the propagation of information between the basic emotions the dependence between the categorical and dimensional model is ignored. Also, we deal with a single-label recognition task and the possible benefits of this approach are more suitable in a multi-label recognition task.

#### 6.4.4 Comparison with the State of the Art

In Table 6.4, we compare the performance of our Emotion-GCN model with several state-of-the-art methods for FER on the categorical model of AffectNet. Regarding Aff-Wild2, we only used the subset which contains both categorical and continuous annotations, while methods in the bibliography typically report results on the whole dataset, making direct comparisons not possible. In IPA2LT [ZSC18], the authors designed LTNet to discover the latent truths from the human annotations and the machine annotations trained from different FER datasets. In gACNN [Li+18] and OADN [DZC20] attention networks were proposed for occlusion aware FER. Facial Motion Prior Network in [Che+19a] generates a facial mask so as to focus on facial muscle moving regions. In [GIP19] deep (CNNs) and handcrafted features (BOVW) were combined and in [HNM19] a deep Siamese network along with a supervised loss function was used to reduce the intra-class variation of the task. Closer to our work, CAKE [Ker+18] proposed a 3-dimensional representation of emotion learned in a multi-domain fashion. Our Emotion-GCN model considerably outperforms these recent state-of-the-art methods, achieving an accuracy of **66.46%**.

In [HNM20] BReG-NeXt achieved state-of-the-art accuracy in 8-way classification on AffectNet (including contempt) by introducing a residual-based network architecture. To investigate the ability of Emotion-GCN to generalize over other model architectures as well, we replaced our DenseNet backbone network with BReG-NeXt using the publicly available code. Since the provided code does not include the data preprocessing strategy, we followed our preprocessing and training pipeline described before (same with DenseNet), which explains the different results compared to [HNM20]. As we can see in Table 6.5, our proposed model outperforms the single-task and multi-task models when using BReG-NeXt as the backbone network indicating that Emotion-GCN generalizes across different model architectures as well.

Table 6.5: Performance of Emotion-GCN using BReG-NeXt as the backbone network on AffectNet.

Method	Network	Accuracy
Single-task	BReG-NeXt	60.49
Multi-task+CCC	BReG-NeXt	61.14
Emotion-GCN	BReG-NeXt	<b>61.94</b>

# Chapter 7

## Conclusion and Future Work

### 7.1 Conclusion

In the current thesis we deal with the task of recognizing human emotion using the facial expressions as the input signal. First, we present early methods that use traditional feature extractors and then we move on to the latest proposed deep learning techniques. Then, we analyze the emotion representations that we should use in our models with the categorical and the dimensional model being the most widespread. After we realized that both emotion representations are very powerful and their correlation can benefit the overall performance, we use both representation in most of our methods.

Initially, we approach the task from a metric learning perspective after illustrating the variation characteristics that the FER task presents. We implement already proposed metric losses and evaluate them on the largest facial expression database, AffectNet. Then, we extend the ideas of island and local subclass loss by proposing some modifications of them improving the overall performance. Also, we integrate the VA information in metric learning by designing a VA-based center loss.

Then we move on to multi-task learning architectures using mainly the categorical and the dimensional emotion representation. After presenting various multi-task architectures, we end up in our best proposed network, Emotion-GCN. This novel GCN based MTL model learns a shared feature representation for both discrete and continuous expression recognition to exploit the dependencies between the categorical and the dimensional model of affect. To further capture these dependencies, the expression classifiers and the VA regressors are learned through a GCN that maps their word representation to dependent vectors inspired by recent work in multi-label image recognition. Experimental results on AffectNet have demonstrated that our Emotion-GCN outperforms the performance of the recent state-of-the-art methods for discrete FER. Therefore, we submitted part of our work regarding Emotion-GCN to the IEEE International Conference on Automatic Face and Gesture Recognition 2021 (FG 2021) with the authors being Panagiotis Antoniadis, Panagiotis Paraskevas Filntis and Petros Maragos [AFM21]. In Table 7.1 we can see the accuracy improvements of our proposed models on the categorical model of AffectNet.

### 7.2 Future Work

The results of our experiments show that our proposed methods are capable of improving the performance across different datasets and backbone architectures. Also, we surpass the previous state-of-the-art methods on the categorical model of AffectNet. However, more research on FER is required in order to develop systems with performance closer to human performance. To achieve this goal, we mention the following ideas as future work:

1. **Extend proposed techniques for videos:** In the thesis, we deal with the problem of recognizing facial expression in static images. However, facial expression recognition can benefit from the temporal correlation of consecutive frames in a sequence. Therefore, more and more studies propose architectures for FER in dynamic image sequences (videos). It would be interesting to train and evaluate our proposed

Table 7.1: Comparison of our best proposed models with state-of-the-art methods on AffectNet.

<b>Model</b>	<b>Accuracy</b>
Facial Motion Prior Network [Che+19a]	61.52
CAKE [Ker+18]	61.7
OADN [DZC20]	61.89
CNNs and BOVW + global SVM [GIP19]	63.31
Siamese [HNM19]	64
Local subclass loss (ours)	65.29
Multi-task CCC (ours)	65.63
Multi-task Eight (ours)	65.97
<b>Emotion-GCN (ours)</b>	<b>66.46</b>

methods on videos. Regarding the multi-task networks, the backbone network can just be substituted by a deep network for videos. In metric learning, the proposed losses may need to be adjusted to the dynamic sequence since their design is mainly focused on single frames. Finally, an extension of Emotion-GCN in videos would be interesting. Although we evaluated the proposed Emotion-GCN in Aff-Wild2 that contains videos, we handle each frame as a single example without exploiting the temporal correlation of the frames.

2. **Suppress wrong annotations:** One of the challenges of FER that we did not deal with in the thesis is the uncertain or wrong annotations. Annotating a qualitative large-scale facial expression dataset is extremely difficult due to the uncertainties caused by ambiguous facial expressions, low-quality facial images, and the subjectiveness of annotators. These uncertainties degrade the performance of FER systems and their suppression has not been explored a lot [Wan+20]. Therefore, reducing these uncertainties and re-training the proposed models in the new datasets is a useful future extension of the thesis.



# Appendix A

## Bibliography

- [AWN20] Acheampong, F. A., Wenyu, C., and Nunoo-Mensah, H. “Text-based emotion detection: Advances, challenges, and opportunities”. In: *Engineering Reports 2.7* (2020), e12189.
- [Alh16] Alhussein, M. “Automatic facial emotion recognition using weber local descriptor for e-Healthcare system”. In: *Cluster Computing 19.1* (2016), pp. 99–108.
- [AFM21] Antoniadis, P., Filntisis, P. P., and Maragos, P. “Exploiting Emotional Dependencies with Graph Convolutional Networks for Facial Expression Recognition”. In: *arXiv preprint arXiv:2106.03487* (2021).
- [Ant+21] Antoniadis, P. et al. “An Audiovisual and Contextual Approach for Categorical and Continuous Emotion Recognition In-the-Wild”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. 2021, pp. 3645–3651.
- [Ast+13] Asthana, A. et al. “Robust discriminative response map fitting with constrained local models”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 3444–3451.
- [Ast+14] Asthana, A. et al. “Incremental face alignment in the wild”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1859–1866.
- [BNW16] Bahreini, K., Nadolski, R., and Westera, W. “Towards multimodal emotion recognition in e-learning environments”. In: *Interactive Learning Environments 24.3* (2016), pp. 590–605.
- [Bal+18] Baltrušaitis, T. et al. “Openface 2.0: Facial behavior analysis toolkit”. In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE. 2018, pp. 59–66.
- [BBR13] Baltrušaitis, T., Banda, N., and Robinson, P. “Dimensional affect recognition using continuous conditional random fields”. In: *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE. 2013, pp. 1–8.
- [BMR15] Baltrušaitis, T., Mahmoud, M., and Robinson, P. “Cross-dataset learning and person-specific normalisation for automatic action unit detection”. In: *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. Vol. 6. IEEE. 2015, pp. 1–6.
- [BRM16] Baltrušaitis, T., Robinson, P., and Morency, L.-P. “Openface: an open source facial behavior analysis toolkit”. In: *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2016, pp. 1–10.
- [Bar+16] Bargal, S. A. et al. “Emotion recognition in the wild from videos using images”. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 2016, pp. 433–436.
- [Bar+18] Barros, P. et al. “The OMG-emotion behavior dataset”. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2018, pp. 1–7.
- [Bar+01] Bartlett, M. S. et al. *Automatic analysis of spontaneous facial behavior: A final project report*. Tech. rep. Technical Report UCSD MPLab TR 2001.08, University of California, San Diego, 2001.
- [Bla+03] Blanz, V. et al. “Reanimating faces in images and video”. In: *Computer graphics forum*. Vol. 22. 3. Wiley Online Library. 2003, pp. 641–650.

- [Blo+14] Blom, P. M. et al. “Towards personalised gaming via facial expression recognition”. In: *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*. Vol. 10. 1. 2014.
- [BLP01] Bolls, P. D., Lang, A., and Potter, R. F. “The effects of message valence and listener arousal on attention, memory, and facial muscular responses to radio advertisements”. In: *Communication research* 28.5 (2001), pp. 627–651.
- [BGV92] Boser, B. E., Guyon, I. M., and Vapnik, V. N. “A training algorithm for optimal margin classifiers”. In: *Proceedings of the fifth annual workshop on Computational learning theory*. 1992, pp. 144–152.
- [Bra+07] Brandes, U. et al. “On modularity clustering”. In: *IEEE transactions on knowledge and data engineering* 20.2 (2007), pp. 172–188.
- [Cai+18] Cai, J. et al. “Island loss for learning discriminative features in facial expression recognition”. In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE. 2018, pp. 302–309.
- [Cam+15] Cameron, D. et al. “Presence of life-like robot expressions influences children’s enjoyment of human-robot interactions in the field”. In: *Proceedings of the AISB Convention 2015*. The Society for the Study of Artificial Intelligence and Simulation of Behaviour. 2015.
- [Can27] Cannon, W. B. “The James-Lange theory of emotions: A critical examination and an alternative theory”. In: *The American journal of psychology* 39.1/4 (1927), pp. 106–124.
- [Car+06] Caridakis, G. et al. “Modeling naturalistic affective states via facial and vocal expressions recognition”. In: *Proceedings of the 8th international conference on Multimodal interfaces*. 2006, pp. 146–154.
- [Car97] Caruana, R. “Multitask learning”. In: *Machine learning* 28.1 (1997), pp. 41–75.
- [CAP07] Chanel, G., Ansari-Asl, K., and Pun, T. “Valence-arousal evaluation using physiological signals in an emotion recall paradigm”. In: *2007 IEEE International Conference on Systems, Man and Cybernetics*. IEEE. 2007, pp. 2662–2667.
- [CHC17] Chang, W.-Y., Hsu, S.-H., and Chien, J.-H. “FATAUVA-Net: An integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation”. In: *Proc. CVPRW*. 2017, pp. 17–25.
- [Che+16a] Chen, D. et al. “Supervised transformer network for efficient face detection”. In: *European Conference on Computer Vision*. Springer. 2016, pp. 122–138.
- [CKI18] Chen, J., Konrad, J., and Ishwar, P. “Vgan-based image representation learning for privacy-preserving facial expression recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 1570–1579.
- [Che+16b] Chen, J.-C. et al. “A cascaded convolutional neural network for age estimation of unconstrained faces”. In: *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE. 2016, pp. 1–8.
- [Che+17] Chen, S. et al. “Multimodal multi-task learning for dimensional and continuous emotion recognition”. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 2017, pp. 19–26.
- [Che+20] Chen, W. et al. “YOLO-face: a real-time face detector”. In: *The Visual Computer* (2020), pp. 1–9.
- [Che+19a] Chen, Y. et al. “Facial motion prior networks for facial expression recognition”. In: *Proc. VCIP*. 2019, pp. 1–4.
- [Che+19b] Chen, Z.-M. et al. “Multi-label image recognition with graph convolutional networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5177–5186.
- [Chi+19] Chi, C. et al. “Selective refinement network for high performance face detection”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 8231–8238.
- [CKR16] Choi, Y., Kim, H.-I., and Ro, Y. M. “Two-step learning of deep convolutional neural network for discriminative face recognition under varying illumination”. In: *Electronic Imaging* 2016.11 (2016), pp. 1–5.
- [CET01] Cootes, T. F., Edwards, G. J., and Taylor, C. J. “Active appearance models”. In: *IEEE Transactions on pattern analysis and machine intelligence* 23.6 (2001), pp. 681–685.

- 
- [DT05] Dalal, N. and Triggs, B. "Histograms of oriented gradients for human detection". In: *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*. Vol. 1. Ieee. 2005, pp. 886–893.
- [Dan+16] Dantcheva, A. et al. "Emotion facial recognition by the means of automatic video analysis". In: *Gerontechnology* (2016).
- [Dar15] Darwin, C. *The expression of the emotions in man and animals*. University of Chicago press, 2015.
- [DMT18] Derr, T., Ma, Y., and Tang, J. "Signed graph convolutional networks". In: *Proc. ICDM*. 2018, pp. 929–934.
- [DBT14] Devries, T., Biswaranjan, K., and Taylor, G. W. "Multi-task learning of facial landmarks and expression". In: *2014 Canadian conference on computer and robot vision*. IEEE. 2014, pp. 98–103.
- [Dha+11a] Dhall, A. et al. "Acted facial expressions in the wild database". In: *Australian National University, Canberra, Australia, Technical Report TR-CS-11 2* (2011), p. 1.
- [Dha+11b] Dhall, A. et al. "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark". In: *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE. 2011, pp. 2106–2112.
- [DSC18] Ding, H., Sricharan, K., and Chellappa, R. "Exprgan: Facial expression editing with controllable expression intensity". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.
- [DZC20] Ding, H., Zhou, P., and Chellappa, R. "Occlusion-adaptive deep network for robust facial expression recognition". In: *Proc. IJCB*. 2020, pp. 1–9.
- [Don+99] Donato, G. et al. "Classifying facial actions". In: *IEEE Transactions on pattern analysis and machine intelligence* 21.10 (1999), pp. 974–989.
- [Ekm94] Ekman, P. "Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique." In: *Psychological bulletin* 115.2 (1994), pp. 268–287.
- [EF69] Ekman, P. and Friesen, W. V. "Nonverbal leakage and clues to deception". In: *Psychiatry* 32.1 (1969), pp. 88–106.
- [EF71] Ekman, P. and Friesen, W. V. "Constants across cultures in the face and emotion." In: *Journal of personality and social psychology* 17.2 (1971), p. 124.
- [Ekm+87] Ekman, P. et al. "Universals and cultural differences in the judgments of facial expressions of emotion." In: *Journal of personality and social psychology* 53.4 (1987), p. 712.
- [EKK11] El Ayadi, M., Kamel, M. S., and Karray, F. "Survey on speech emotion recognition: Features, classification schemes, and databases". In: *Pattern recognition* 44.3 (2011), pp. 572–587.
- [EP97] Essa, I. A. and Pentland, A. P. "Coding, analysis, interpretation, and recognition of facial expressions". In: *IEEE transactions on pattern analysis and machine intelligence* 19.7 (1997), pp. 757–763.
- [FSM16] Fabian Benitez-Quiroz, C., Srinivasan, R., and Martinez, A. M. "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5562–5570.
- [FT05] Fragopanagos, N. and Taylor, J. G. "Emotion recognition in human–computer interaction". In: *Neural Networks* 18.4 (2005), pp. 389–405.
- [GIP19] Georgescu, M.-I., Ionescu, R. T., and Popescu, M. "Local learning with deep and handcrafted features for facial expression recognition". In: *IEEE Access* 7 (2019), pp. 64827–64836.
- [Glo+08] Glowinski, D. et al. "Technique for automatic emotion recognition by body gesture analysis". In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE. 2008, pp. 1–6.
- [Goo+16] Goodfellow, I. et al. *Deep learning*. Vol. 1. 2. MIT press Cambridge, 2016.
- [Goo+13] Goodfellow, I. J. et al. "Challenges in representation learning: A report on three machine learning contests". In: *International conference on neural information processing*. Springer. 2013, pp. 117–124.
- [Goo+14] Goodfellow, I. J. et al. "Generative adversarial networks". In: *arXiv preprint arXiv:1406.2661* (2014).
- [Gro+10] Gross, R. et al. "Multi-pie". In: *Image and vision computing* 28.5 (2010), pp. 807–813.
-

- [GP10] Gunes, H. and Pantic, M. “Automatic, dimensional and continuous emotion recognition”. In: *International Journal of Synthetic Emotions (IJSE)* 1.1 (2010), pp. 68–99.
- [Guo+19] Guo, F. et al. “The effect of a humanoid robot’s emotional behaviors on users’ emotional responses: Evidence from pupillometry and electroencephalography measures”. In: *International Journal of Human–Computer Interaction* 35.20 (2019), pp. 1947–1959.
- [Guo+16] Guo, Y. et al. “Deep neural networks with relativity learning for facial expression recognition”. In: *2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2016, pp. 1–6.
- [HBW15] Hamster, D., Barros, P., and Wermter, S. “Face expression recognition with a 2-channel convolutional neural network”. In: *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–8.
- [HNM20] Hasani, B., Negi, P. S., and Mahoor, M. “BReG-NeXt: Facial affect computing using adaptive residual networks with bounded gradient”. In: *IEEE Trans. on Affective Computing* (2020).
- [HNM19] Hayale, W., Negi, P., and Mahoor, M. “Facial Expression Recognition Using Deep Siamese Neural Networks with a Supervised Loss function”. In: *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–7.
- [He+16] He, K. et al. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.
- [Hu+17] Hu, P. et al. “Learning supervised scoring ensemble for emotion recognition in the wild”. In: *Proceedings of the 19th ACM international conference on multimodal interaction*. 2017, pp. 553–560.
- [Hua+17a] Huang, G. et al. “Densely connected convolutional networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.
- [Hua+17b] Huang, R. et al. “Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2439–2448.
- [Iso+17] Isola, P. et al. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.
- [JVP11] Jiang, B., Valstar, M. F., and Pantic, M. “Action unit detection using sparse appearance descriptors in space-time video volumes”. In: *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 314–321.
- [JL17] Jiang, H. and Learned-Miller, E. “Face detection with the faster R-CNN”. In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 650–657.
- [Kah+13] Kahou, S. E. et al. “Combining modality specific deep neural networks for emotion recognition in video”. In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. 2013, pp. 543–550.
- [Kar+07] Karpouzis, K. et al. “Modeling naturalistic affective states via facial, vocal, and bodily expressions recognition”. In: *Artificial intelligence for human computing*. Springer, 2007, pp. 91–112.
- [KB19] Kaya, M. and Bilge, H. Ş. “Deep metric learning: A survey”. In: *Symmetry* 11.9 (2019), p. 1066.
- [Ker+18] Kervadec, C. et al. “Cake: Compact and Accurate K-dimensional representation of Emotion”. In: *CoRR* (2018).
- [Kho+16] Khorrani, P. et al. “How deep neural networks can improve emotion recognition on video data”. In: *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 619–623.
- [Kim+15] Kim, B.-K. et al. “Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition”. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 2015, pp. 427–434.
- [KW17] Kipf, T. N. and Welling, M. “Semi-Supervised Classification with Graph Convolutional Networks”. In: *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, 2017.
- [KB07] Kleinsmith, A. and Bianchi-Berthouze, N. “Recognizing affective dimensions from body posture”. In: *International conference on affective computing and intelligent interaction*. Springer, 2007, pp. 48–58.
- [Koe+11] Koelstra, S. et al. “Deap: A database for emotion analysis; using physiological signals”. In: *IEEE transactions on affective computing* 3.1 (2011), pp. 18–31.

- 
- [KSZ21] Kollias, D., Sharmanska, V., and Zafeiriou, S. “Distribution Matching for Heterogeneous Multi-Task Learning: a Large-scale Face Study”. In: *arXiv preprint arXiv:2105.03790* (2021).
- [KZ19] Kollias, D. and Zafeiriou, S. “Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface”. In: *arXiv preprint arXiv:1910.04855* (2019).
- [KZ20a] Kollias, D. and Zafeiriou, S. “VA-StarGAN: Continuous Affect Generation”. In: *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2020, pp. 227–238.
- [KZ21] Kollias, D. and Zafeiriou, S. “Affect Analysis in-the-wild: Valence-Arousal, Expressions, Action Units and a Unified Framework”. In: *arXiv preprint arXiv:2103.15792* (2021).
- [KZ20b] Kollias, D. and Zafeiriou, S. P. “Exploiting multi-cnn features in cnn-rnn based dimensional emotion recognition on the omg in-the-wild dataset”. In: *IEEE Transactions on Affective Computing* (2020).
- [Kol+19] Kollias, D. et al. “Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond”. In: *International Journal of Computer Vision* 127.6 (2019), pp. 907–929.
- [Kol+20] Kollias, D. et al. “Analysing affective behavior in the first abaw 2020 competition”. In: *arXiv preprint arXiv:2001.11409* (2020).
- [Kos+17a] Kossaifi, J. et al. “AFEW-VA database for valence and arousal estimation in-the-wild”. In: *Image and Vision Computing* 65 (2017), pp. 23–36.
- [Kos+17b] Kosti, R. et al. “EMOTIC: Emotions in Context dataset”. In: *Proc. CVPRW*. 2017, pp. 61–69.
- [KC07] Kulic, D. and Croft, E. A. “Affective state estimation for human-robot interaction”. In: *IEEE Transactions on Robotics* 23.5 (2007), pp. 991–1000.
- [Kwa+13] Kwak, S. S. et al. “What makes people empathize with an emotional robot?: The impact of agency and physical embodiment on human empathy for a robot”. In: *2013 IEEE RO-MAN*. IEEE, 2013, pp. 180–185.
- [LL18] Lai, Y.-H. and Lai, S.-H. “Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition”. In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 263–270.
- [LBH15] LeCun, Y., Bengio, Y., and Hinton, G. “Deep learning”. In: *nature* 521.7553 (2015), pp. 436–444.
- [Li+15] Li, H. et al. “A convolutional neural network cascade for face detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 5325–5334.
- [Li+13] Li, K. et al. “A data-driven approach for facial expression retargeting in video”. In: *IEEE Transactions on Multimedia* 16.2 (2013), pp. 299–310.
- [LHW18] Li, Q., Han, Z., and Wu, X.-M. “Deeper insights into graph convolutional networks for semi-supervised learning”. In: *Proc. AAAI Conference on Artificial Intelligence*. 2018.
- [LD20] Li, S. and Deng, W. “Deep facial expression recognition: A survey”. In: *IEEE Transactions on Affective Computing* (2020), pp. 1–1.
- [LDD17] Li, S., Deng, W., and Du, J. “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2852–2861.
- [Li+18] Li, Y. et al. “Occlusion aware facial expression recognition using cnn with attention mechanism”. In: *IEEE Transactions on Image Processing* 28.5 (2018), pp. 2439–2450.
- [Lin+17a] Lin, T.-Y. et al. “Feature pyramid networks for object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2117–2125.
- [Lin+17b] Lin, T.-Y. et al. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.
- [Liu+13] Liu, M. et al. “Au-aware deep networks for facial expression recognition”. In: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, 2013, pp. 1–6.
- [Liu+17] Liu, X. et al. “Adaptive deep metric learning for identity-aware facial expression recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 20–29.
- [Lop+17] Lopes, A. T. et al. “Facial expression recognition with convolutional neural networks: coping with few data and the training sample order”. In: *Pattern Recognition* 61 (2017), pp. 610–628.
- [LHZ17] Lu, J., Hu, J., and Zhou, J. “Deep metric learning for visual understanding: An overview of recent advances”. In: *IEEE Signal Processing Magazine* 34.6 (2017), pp. 76–84.
-

- [Luc+10] Lucey, P. et al. “The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression”. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE. 2010, pp. 94–101.
- [LHD18] Luo, Z., Hu, J., and Deng, W. “Local subclass constraint for facial expression recognition in the wild”. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE. 2018, pp. 3132–3137.
- [Lyo+98] Lyons, M. et al. “Coding facial expressions with gabor wavelets”. In: *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE. 1998, pp. 200–205.
- [MHN13] Maas, A. L., Hannun, A. Y., and Ng, A. Y. “Rectifier nonlinearities improve neural network acoustic models”. In: *Proc. ICML*. 2013, p. 3.
- [Man+16] Mano, L. Y. et al. “Exploiting IoT technologies for enhancing Health Smart Homes through patient identification and emotion recognition”. In: *Computer Communications* 89 (2016), pp. 178–190.
- [MS17] McDuff, D. and Soleymani, M. “Large-scale affective content analysis: Combining media content features and facial reactions”. In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE. 2017, pp. 339–345.
- [McD+14] McDuff, D. et al. “Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads”. In: *IEEE Transactions on Affective Computing* 6.3 (2014), pp. 223–235.
- [MHM18] McInnes, L., Healy, J., and Melville, J. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).
- [McK+11] McKeown, G. et al. “The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent”. In: *IEEE transactions on affective computing* 3.1 (2011), pp. 5–17.
- [Men+17] Meng, Z. et al. “Identity-aware convolutional neural network for facial expression recognition”. In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE. 2017, pp. 558–565.
- [Min+21] Minaee, S. et al. “Going Deeper Into Face Detection: A Survey”. In: *arXiv preprint arXiv:2103.14983* (2021).
- [Mit+97] Mitchell, T. M. et al. “Machine learning”. In: (1997).
- [MPK09] Mohammed, U., Prince, S. J., and Kautz, J. “Visio-lization: generating novel facial images”. In: *ACM Transactions on Graphics (ToG)* 28.3 (2009), pp. 1–8.
- [MHM17] Mollahosseini, A., Hasani, B., and Mahoor, M. H. “Affectnet: A database for facial expression, valence, and arousal computing in the wild”. In: *IEEE Transactions on Affective Computing* 10.1 (2017), pp. 18–31.
- [Mol+16] Mollahosseini, A. et al. “Facial expression recognition from world wild web”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2016, pp. 58–65.
- [Muh+17] Muhammad, G. et al. “A facial-expression monitoring system for improved healthcare in smart cities”. In: *IEEE Access* 5 (2017), pp. 10871–10881.
- [Naj+17] Najibi, M. et al. “Ssh: Single stage headless face detector”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4875–4884.
- [Nor+21] Noroozi, F. et al. “Survey on Emotional Body Gesture Recognition”. In: *IEEE Transactions on Affective Computing* 12 (2021), pp. 505–523.
- [OMT18] Okada, G., Masui, K., and Tsumura, N. “Advertisement effectiveness estimation based on crowd-sourced multimodal affective responses”. In: *Proceedings of the IEEE Conference on computer vision and pattern recognition workshops*. 2018, pp. 1263–1271.
- [PR00] Pantic, M. and Rothkrantz, L. J. “Expert system for automatic analysis of facial expressions”. In: *Image and Vision Computing* 18.11 (2000), pp. 881–905.
- [Pan+05] Pantic, M. et al. “Web-based database for facial expression analysis”. In: *2005 IEEE international conference on multimedia and Expo*. IEEE. 2005, 5–pp.
- [Pit+17] Pitaloka, D. A. et al. “Enhancing CNN with preprocessing stage in automatic emotion recognition”. In: *Procedia computer science* 116 (2017), pp. 523–529.
- [Plu01] Plutchik, R. “The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice”. In: *American scientist* 89.4 (2001), pp. 344–350.

- 
- [PM17] Pons, G. and Masip, D. “Supervised committee of convolutional neural networks in automated facial expression analysis”. In: *IEEE Transactions on Affective Computing* 9.3 (2017), pp. 343–350.
- [PM18] Pons, G. and Masip, D. “Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition”. In: *arXiv preprint arXiv:1802.06664* (2018).
- [Qin+16] Qin, H. et al. “Joint training of cascaded CNN for face detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3456–3465.
- [RPC17] Ranjan, R., Patel, V. M., and Chellappa, R. “Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.1 (2017), pp. 121–135.
- [Ran+17] Ranjan, R. et al. “An all-in-one convolutional neural network for face analysis”. In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE. 2017, pp. 17–24.
- [Ren+14] Ren, S. et al. “Face alignment at 3000 fps via regressing local binary features”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2014, pp. 1685–1692.
- [Ren+15] Ren, S. et al. “Faster r-cnn: Towards real-time object detection with region proposal networks”. In: *arXiv preprint arXiv:1506.01497* (2015).
- [Rif+12] Rifai, S. et al. “Disentangling factors of variation for facial expression recognition”. In: *European Conference on Computer Vision*. Springer. 2012, pp. 808–822.
- [Rin+13] Ringeval, F. et al. “Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions”. In: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE. 2013, pp. 1–8.
- [Rus80] Russell, J. A. “A circumplex model of affect.” In: *Journal of personality and social psychology* 39.6 (1980), p. 1161.
- [SN19] Saunderson, S. and Nejat, G. “How robots influence humans: A survey of nonverbal communication in social human–robot interaction”. In: *International Journal of Social Robotics* 11.4 (2019), pp. 575–608.
- [Sch86] Scherer, K. R. “Vocal affect expression: A review and a model for future research.” In: *Psychological bulletin* 99.2 (1986), p. 143.
- [SKP15] Schroff, F., Kalenichenko, D., and Philbin, J. “Facenet: A unified embedding for face recognition and clustering”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 815–823.
- [Sha+16] Shang, W. et al. “Understanding and improving convolutional neural networks via concatenated rectified linear units”. In: *international conference on machine learning*. PMLR. 2016, pp. 2217–2225.
- [Sol+11] Soleymani, M. et al. “A multimodal database for affect recognition and implicit tagging”. In: *IEEE transactions on affective computing* 3.1 (2011), pp. 42–55.
- [Son+18] Song, L. et al. “Geometry guided adversarial facial expression synthesis”. In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 627–635.
- [Spe61] Spearman, C. “The proof and measurement of association between two things.” In: (1961).
- [SPR20] Spezialetti, M., Placidi, G., and Rossi, S. “Emotion recognition for human-robot interaction: recent advances and future perspectives”. In: *Frontiers in Robotics and AI* 7 (2020).
- [SWH18] Sun, X., Wu, P., and Hoi, S. C. “Face detection using deep learning: An improved faster RCNN approach”. In: *Neurocomputing* 299 (2018), pp. 42–50.
- [SWT13] Sun, Y., Wang, X., and Tang, X. “Deep convolutional network cascade for facial point detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 3476–3483.
- [Sus+08] Susskind, J. M. et al. “Generating facial expressions with deep belief nets”. In: *Affective Computing, Emotion Modelling, Synthesis and Recognition* (2008), pp. 421–440.
- [Sze+16] Szegedy, C. et al. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.
- [Tak+18] Takalkar, M. et al. “A survey: facial micro-expression recognition”. In: *Multimedia Tools and Applications* 77.15 (2018), pp. 19301–19325.
- [Tan+18] Tang, X. et al. “Pyramidbox: A context-assisted single shot face detector”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 797–813.
-

- [TT11] Thakare, N. M. and Thakare, V. M. “An innovative hybrid approach to construct fuzzy-neural network for 3D face recognition system”. In: *2011 11th International Conference on Hybrid Intelligent Systems (HIS)*. IEEE. 2011, pp. 463–467.
- [The+09] Theobald, B.-J. et al. “Mapping and manipulating facial expression”. In: *Language and speech* 52.2-3 (2009), pp. 369–386.
- [Tze+15] Tzeng, E. et al. “Simultaneous deep transfer across domains and tasks”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4068–4076.
- [VJ01] Viola, P. and Jones, M. “Rapid object detection using a boosted cascade of simple features”. In: *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. IEEE. 2001, pp. I–I.
- [Wan+17] Wang, H. et al. “Face r-cnn”. In: *arXiv preprint arXiv:1706.01061* (2017).
- [Wan+20] Wang, K. et al. “Suppressing uncertainties for large-scale facial expression recognition”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 6897–6906.
- [WBF19] Wang, X., Bo, L., and Fuxin, L. “Adaptive wing loss for robust face alignment via heatmap regression”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 6971–6981.
- [Wen+16] Wen, Y. et al. “A discriminative feature learning approach for deep face recognition”. In: *European conference on computer vision*. Springer. 2016, pp. 499–515.
- [Wöl+08] Wöllmer, M. et al. “Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies”. In: *Proc. 9th Interspeech 2008 incorp. 12th Australasian Int. Conf. on Speech Science and Technology SST 2008, Brisbane, Australia*. 2008, pp. 597–600.
- [Woo+15] Wood, E. et al. “Rendering of eyes for eye-shape registration and gaze estimation”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2015, pp. 3756–3764.
- [Xia+19] Xiaohua, W. et al. “Two-level attention with two-stage multi-task learning for facial emotion recognition”. In: *Journal of Visual Communication and Image Representation* 62 (2019), pp. 217–225.
- [XD13] Xiong, X. and De la Torre, F. “Supervised descent method and its applications to face alignment”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 532–539.
- [Yan+18] Yang, D. et al. “An emotion recognition model based on facial recognition in virtual learning environment”. In: *Procedia Computer Science* 125 (2018), pp. 2–10.
- [Yan+12] Yang, F. et al. “Facial expression editing in video using a temporally-smooth factorization”. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2012, pp. 861–868.
- [YCY18] Yang, H., Ciftci, U., and Yin, L. “Facial expression recognition by de-expression residue learning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2168–2177.
- [YZY18] Yang, H., Zhang, Z., and Yin, L. “Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks”. In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE. 2018, pp. 294–301.
- [Yan+17] Yang, S. et al. “Faceness-net: Face detection through deep facial part responses”. In: *IEEE transactions on pattern analysis and machine intelligence* 40.8 (2017), pp. 1845–1859.
- [Yao+16] Yao, A. et al. “HoloNet: towards robust emotion recognition in the wild”. In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 2016, pp. 472–478.
- [Yeh+16] Yeh, R. et al. “Semantic facial expression editing using autoencoded flow”. In: *arXiv preprint arXiv:1611.09961* (2016).
- [Yim+15] Yim, J. et al. “Rotating your face using multi-task deep neural network”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 676–684.
- [YZ15] Yu, Z. and Zhang, C. “Image based static facial expression recognition with multiple deep network learning”. In: *Proceedings of the 2015 ACM on international conference on multimodal interaction*. 2015, pp. 435–442.
- [Zad+17] Zadeh, A. et al. “Convolutional experts constrained local model for 3d facial landmark detection”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 2519–2528.



- 
- [ZSC18] Zeng, J., Shan, S., and Chen, X. “Facial expression recognition with inconsistently annotated datasets”. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 222–237.
- [Zha+18a] Zhang, F. et al. “Joint pose and expression modeling for facial expression recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3359–3368.
- [Zha+20a] Zhang, J. et al. “Feature agglomeration networks for single stage face detection”. In: *Neurocomputing* 380 (2020), pp. 180–189.
- [Zha+17a] Zhang, K. et al. “Facial expression recognition based on deep evolutionary spatial-temporal networks”. In: *IEEE Transactions on Image Processing* 26.9 (2017), pp. 4193–4203.
- [Zha+16] Zhang, K. et al. “Joint face detection and alignment using multitask cascaded convolutional networks”. In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503.
- [Zha+17b] Zhang, K. et al. “Detecting faces using inside cascaded contextual cm”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 3171–3179.
- [Zha+05] Zhang, Q. et al. “Geometry-driven photorealistic facial expression synthesis”. In: *IEEE Transactions on visualization and computer graphics* 12.1 (2005), pp. 48–60.
- [Zha+17c] Zhang, S. et al. “Faceboxes: A CPU real-time face detector with high accuracy”. In: *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE. 2017, pp. 1–9.
- [Zha+17d] Zhang, S. et al. “S3fd: Single shot scale-invariant face detector”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 192–201.
- [Zha+20b] Zhang, S. et al. “Refineface: Refinement neural network for high performance face detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).
- [ZY21] Zhang, Y. and Yang, Q. “A survey on multi-task learning”. In: *IEEE Transactions on Knowledge and Data Engineering* (2021).
- [Zha+18b] Zhang, Z. et al. “From facial expression recognition to interpersonal relation prediction”. In: *International Journal of Computer Vision* 126.5 (2018), pp. 550–569.
- [Zha+98] Zhang, Z. et al. “Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron”. In: *Proceedings Third IEEE International Conference on Automatic face and gesture recognition*. IEEE. 1998, pp. 454–459.
- [ZP07] Zhao, G. and Pietikainen, M. “Dynamic texture recognition using local binary patterns with an application to facial expressions”. In: *IEEE transactions on pattern analysis and machine intelligence* 29.6 (2007), pp. 915–928.
- [Zha+18c] Zhao, S. et al. “Feature Selection Mechanism in CNNs for Facial Expression Recognition.” In: *BMVC*. 2018, p. 317.
- [ZLZ20] Zhi, R., Liu, M., and Zhang, D. “A comprehensive survey on automatic facial action unit analysis”. In: *The Visual Computer* 36.5 (2020), pp. 1067–1093.
- [Zhu+17] Zhu, J.-Y. et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232.
- [ZR12] Zhu, X. and Ramanan, D. “Face detection, pose estimation, and landmark localization in the wild”. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 2879–2886.
- [Zhu+13] Zhu, Z. et al. “Deep learning identity-preserving face space”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2013, pp. 113–120.
- [Zhu+14] Zhu, Z. et al. “Multi-view perceptron: a deep model for learning face identity and view representations”. In: *Advances in Neural Information Processing Systems*. 2014, pp. 217–225.
-