

University of Nevada, Reno

Attributes in Face Processing: Novel Methods for Explanation, Training, and Representation

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Computer Science and Engineering

by

Nathan Thom

Dr. Emily Hand, Advisor

May, 2024

© by Nathan Thom
All Rights Reserved



THE GRADUATE SCHOOL

We recommend that the dissertation
prepared under our supervision by

Nathan Thom

entitled

**Attributes in Face Processing: Novel Methods for Explanation,
Training, and Representation**

be accepted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

Emily Hand, Ph.D.
Advisor

Alireza Tavakkoli, Ph.D.
Committee Member

George Bebis, Ph.D.
Committee Member

Frederick C. Harris Jr., Ph.D.
Committee Member

Mark Lescroart, Ph.D.
Graduate School Representative

Markus Kemmelmeier, Ph.D., Dean
Graduate School

May, 2024

Abstract

Facial attribute recognition, the automatic detection of human-describable visual features from face images, has important applications across numerous domains including biometrics, visual search, and accessibility. While deep learning has revolutionized the field of facial recognition, the learned representations often lack interpretability. This dissertation argues for an approach that explicitly models face images as semantically meaningful facial attributes. Representing faces using attribute vectors instead of embeddings can yield more interpretable models that facilitate the identification and mitigation of biases, while also reducing the need for frequent retraining.

The key contributions of this dissertation span four areas: (1) A novel technique for interpreting the visual features learned by deep face attribute models, based on concepts from human cognition research. (2) An improved facial attribute recognition method that constrains deep networks to only utilize information from spatially relevant regions for each attribute. (3) An unsupervised approach to discover the most visually discriminative groupings of images, to address issues of attribute choice in existing datasets. (4) DoppelVer, a challenging new face recognition benchmark comprised of look-alike individuals, which reveals the difficulty of modeling fine-grained similarity between highly similar classes.

Through extensive experiments, this dissertation demonstrates the effectiveness of the proposed techniques for improving the performance, generalization, and interpretability of facial attribute recognition.

The overarching conclusion is that facial attribute recognition benefits from a research paradigm that combines deep learning with attribute modeling. Such an approach yields face recognition systems that are interpretable, fair, and efficient. This dissertation motivates further research into novel model architectures, training schemes, and benchmarks to extend these results and realize the full potential of facial attribute recognition for both scientific progress and real-world impact in domains like biometrics, human-computer interaction, and accessibility technology.

Acknowledgments

I began to study on the campus of UNR nearly 15 years ago as a 7th grader embarking in homeschooling while my Dad, Mom, and brothers pursued their degrees. Much of my independent life and development has occurred on this campus. My experiences have been overwhelmingly positive because of the extraordinary people that I sojourned with!

To begin, I would like to thank Phi Bya for teaching me a bit about nearly everything from selecting conferences to installing operating systems. I also want to express gratitude to Cooper Flourens for his honest review of my work and encouragement during difficult times. I think of these two like brothers and I am lucky to know them. I have fond memories with too many others to name, but a few are Jayam Sutariya, Cayler Miley, Bryson Lingenfelter, Nathan Michelotti, Shaine Hirsh, Sara Davis, Gaetano Evangelista, Batyr Charyyev, and Jeremy Speth.

I am grateful to my dissertation committee Dr. Emily Hand, Dr. Alireza Tavakkoli, Dr. George Bebis, Dr. Frederick Harris Jr., and Dr. Mark Lescroart. Each of you contributed time and energy into improving this dissertation and encouraging me along the way. Special thanks to my advisor, Dr. Emily Hand, for her trust in my ability and patient guidance.

Finally, I send my sincere gratitude to my family. My wife Kathleen, as always, provided unwavering support every step of the way. Whether she was listening attentively, lovingly motivating, or offering valuable feedback she was always there for me. She never shied away from taking weight off of my shoulders. Sincere appreciation to my brothers Ben, Max, and Nick as well as their spouses. Each of them has been a profound positive influence and helped me find my way. Thank you to the Scafidi family for all of the love and advice. Thanks to my Mom, Shendry Thom, for being the daily cheerleader who could always make time to talk. In particular I would like to thank my Dad, Jay Thom. I would not be the man that I am without his dedicated mentorship, teaching, and friendship. I will never forget these great times. A bright future awaits!

This material is based upon work supported by the National Science Foundation under Grant Number 1909707.

Contents

Abstract	i
Acknowledgments	ii
List of Tables	vii
List of Figures	viii
1 Introduction	1
2 Background	5
2.1 Facial Attributes	5
2.1.1 Related Research	8
2.1.2 Theory and Application	10
2.1.3 Attribute Recognition with Traditional Methods	10
2.1.4 Attribute Recognition with Deep Learning	13
2.2 Model Interpretability	20
2.3 Semantic Segmentation	22
2.4 Image Clustering	23
2.5 Face Recognition and Existing Benchmarks	24
2.5.1 Face Recognition	24
2.5.2 Existing Benchmarks	25
3 Deep Vision Model Perception of Gender From Faces	27
3.1 Introduction	27
3.2 Proposed Method	28
3.3 Experiments	31
3.3.1 Model	31
3.3.2 Data	32
3.4 Results	35
3.4.1 Low Resolution Image Data	35
3.4.2 High Resolution Image Data	36
3.4.3 Generalizable Behaviors	36
3.4.4 Incorporating Heat Maps	38

3.4.5	Discussion: A Comparison with Human Perception	40
4	Parsing Faces with Semantic Segmentation for Improved Facial Attribute Recognition	45
4.1	Introduction	45
4.2	Proposed Methods	47
4.2.1	Segmentation Label Generation	48
4.2.2	Attribute Segmentation and Recognition	50
4.3	Experiments and Results	53
4.3.1	Datasets	53
4.3.2	AttParseNet Training	56
4.3.3	Baseline Model Training	57
4.3.4	Experimental Setup	58
4.3.5	Results on CelebA	58
4.3.6	Generalization to LFWA and UMD-AED	61
5	Attribute Data and Consensus Subspace Clustering	65
5.1	Introduction	65
5.1.1	Attribute Data Problems	65
5.2	Unsupervised Image Labeling	67
5.3	Methodology	68
5.3.1	Feature Extraction	69
5.3.2	Denoising Module	70
5.3.3	Variational Autoencoder	71
5.3.4	Basic Subspace clustering	72
5.3.5	Consensus Clustering	73
5.4	Experiments and Results	73
5.4.1	Datasets	74
5.4.2	Methods for Comparison	74
5.4.3	Metrics	74
5.4.4	Results	75
6	DoppelVer: A Benchmark for Face Verification	78
6.1	Introduction	78
6.2	Proposed Method	81
6.2.1	Dataset Collection	81
6.2.2	Data Preparation	81
6.2.3	Protocol Generation	83
6.2.4	Intended Use	86
6.3	Experiments	87
6.3.1	Evaluation Model	88
6.3.2	Training and Evaluation Process	88
6.3.3	Discussion of Results	89

7 Conclusion	91
7.1 Future Research	92
References	93
Appendices	112
A List of Publications	112

List of Tables

2.1	SOTA CelebA Attribute Accuracy	20
3.1	ResNet-50 Model hyperparameters	32
3.2	Samples from each of the occlusion datasets used in analysis (Part 1)	33
3.3	Samples from each of the occlusion datasets used in analysis (Part 2)	34
4.1	Table attributes predicted more accurately by AttParseNet	60
5.1	Label assignment accuracy of CSC compared with existing methods .	75
6.1	Face verification accuracy of SOTA models on benchmark datasets . .	89
6.2	Face verification AUC of SOTA models on benchmark datasets	89

List of Figures

2.1	Sample images from the FaceTracer dataset	11
2.2	Sample images from the PubFig dataset	12
2.3	Sample images from the CelebA dataset	14
2.4	Sample images from the LFW dataset	15
2.5	Sample images from the UMD-AED dataset	19
3.1	Processing pipeline for interpreting vision model perceptions	31
3.2	Gender recognition heat map generated with CAM	39
3.3	CAM heat maps with different image regions occluded	40
3.4	A non-informative CAM heat map	41
4.1	Locations of facial landmarks utilized	47
4.2	Examples of the 10 base semantic segmentation regions	49
4.3	AttParseNet, our multi-task learning architecture	51
4.4	Sample images from the CelebA dataset	54
4.5	Sample images from the LFWA dataset	55
4.6	Sample images from the UMD-AED dataset	56
4.7	Accuracy of AttParseNet and baseline for attributes in CelebA	59
4.8	Accuracy of AttParseNet and baseline for attributes in LFWA	62
4.9	Accuracy of AttParseNet and baseline for attributes in UMD-AED	63
5.1	Consensus Subspace Clustering pipeline	69
5.2	Feature extraction using convolutional autoencoder	70
5.3	Flattened feature denoising module	71
5.4	Feature extraction using variational autoencoder	72
5.5	UMAP visualization of raw USPS images vs. CSC extracted features	76
6.1	Image samples from DoppelVer’s evaluation protocols	85
6.2	Image samples from existing benchmark datasets	85

Chapter 1

Introduction

Facial attributes are human-describable visual features of the face such as age, gender, ethnicity, emotion, hair color, nose size, and presence of a beard. The automatic detection of these attributes has important applications across many domains including law enforcement, security, human-computer interaction, visual search, and marketing. From a research perspective, facial attributes are utilized for topics like facial recognition, biometrics, image retrieval, image generation or augmentation, human computer interaction, and few shot learning.

Research into automatically recognizing high-level facial attributes like gender and race dates back to at least 1990 [41]. Facial attribute recognition was first popularized in 2008 with the FaceTracer search engine [73]. Despite the relative youth of the field, research interest in facial attribute recognition has declined over the past decade. This is largely due to the transformative impact of deep learning, which has enabled end-to-end models that operate holistically, without explicit facial attribute detection [24, 149, 160].

The allure of deep learning is understandable. Deep neural networks can automatically learn to map from raw pixels to target outputs, obviating the need for intermediate steps like face detection, facial feature extraction, and attribute classification. What's more, deep learning models frequently outperform traditional machine learning pipelines. However, this power comes at a price. The feature representations learned by deep networks are inscrutable to humans, making it difficult to understand how the model is making its decisions. Spurious correlations and biases in the

training data can be picked up and amplified. Generalization to new identities and environments is unreliable. Perhaps most importantly, improvements in deep model performance do not necessarily translate into insights that grow human knowledge.

Machine learning, at its core, began with the promise of modeling human perception. Deep learning has evolved this goal into something else entirely, shifting the focus to performance on benchmark tasks. In this dissertation, I am interested in illuminating the mechanisms behind deep networks to encourage a better understanding of human recognition and perception.

The use of semantically meaningful attributes as an interim task provides four distinct advantages over holistic deep learning:

- Transparency and interpretability
- Bias detection
- Data efficiency
- Scientific understanding

To be clear, we are not arguing against deep learning as a tool. Deep neural networks have revolutionized many aspects of artificial intelligence, from computer vision to natural language processing. However, we believe that for facial analysis tasks, it is important to consider deep learning's limitations and explore alternative paradigms that prioritize interpretability, fairness, efficiency, and human understanding.

In addition to the aforementioned benefits of attributes in deep learning recognition frameworks, facial attributes provide significant value for accessible technology. With an aging population in the United States and countries across the world, more and more individuals are losing their sight [63, 81]. Work in facial attribute recognition can benefit these individuals. Wearable devices can be developed that can provide feedback to users, giving detailed descriptions of faces. Imagine a grandmother meeting her new grandchild for the first time with such technology. Facial

attributes, as well as human describable features of visual scenes more broadly, can give individuals back some form of their sight through language.

This dissertation presents research that advances the field of facial attribute recognition. The overarching goal is to enhance facial attribute recognition capabilities through a deeper comprehension of the descriptive visual characteristics of the human face. The key contributions are as follows:

1. A novel approach to interpret the visual features utilized by deep facial attribute models, drawing insights from human cognition research.
2. An improved facial attribute recognition method that constrains deep networks to utilize information only from spatially relevant regions for each attribute.
3. An unsupervised technique to discover the most visually discriminative groupings of images, addressing the issue of attribute selection in existing datasets.
4. DoppelVer, a challenging new face recognition benchmark comprising look-alike individuals, which highlights the difficulty of modeling fine-grained similarity between highly similar classes.

The organization of this dissertation is as follows: In Chapter 2 we provide necessary background to the topics presented and research discussed. Specifically, it reviews the field of attribute recognition within computer vision, as well as feature extraction for face verification as a whole. Chapter 3 explores the perceptions of deep vision models when making facial attribute predictions. By utilizing concepts from human cognition research we develop a technique for interpreting the visual features utilized for attribute prediction. Based on our findings, Chapter 4 details an improved method of facial attribute recognition. The augmentation to the learning process enforces that deep vision models only utilize information from spatial locations on the input image which contain information that is relevant for detecting each attribute. Chapter 5 addresses issues in publicly available attribute recognition datasets. An issue shared by nearly all publicly available attribute data is that of attribute choice.

We present a method for automatically grouping images into the most relevant categories based on visual information. Chapter 6 introduces a novel benchmark for facial recognition tasks. The dataset is comprised of doppelgangers, which are individuals with high visual similarity. DoppelVer features a novel axis of difficulty when compared to other benchmarks. We show that state-of-the-art facial recognition methods fail to effectively model similarity between similar classes. Chapter 7 concludes the dissertation with a summary of our findings and discusses the meaningful trends, which might motivate important future research directions.

Chapter 2

Background

Parts of this chapter have been previously published. The relevant works are detailed in Appendix A: 1, 2, and 3.

This chapter provides background research related to the key concepts discussed in this dissertation. We begin by thoroughly reviewing the literature on visual attributes, with a particular emphasis on facial attributes. This information is relevant throughout the dissertation. Next, we offer insights into gender recognition in humans and model interpretability, which are primarily pertinent to the content covered in Chapter 3. We then discuss works from the field of semantic segmentation, as this background is relevant to Chapter 4. Furthermore, we provide an overview of image clustering, which is related to the work presented in Chapter 5. Finally, we review facial verification and existing benchmark datasets, establishing the necessary background for Chapter 6.

2.1 Facial Attributes

Facial attributes – human-describable features of faces – were introduced to the computer vision community in 2008, with their first application being image search [73]. Kumar et al. identified a problem with the image search engines of the time, realizing that simple descriptive search terms would not produce expected face image results. Attributes were then used for face recognition and verification as well as, again, for image search and retrieval [74, 75, 141] before attribute recognition itself became the

focus of research.

Well before the introduction of facial attributes, recognition of gender and age from faces were well-established problems in the computer vision community [37, 115, 126]. One of the earliest works on gender recognition from faces utilized a neural network that predicted sex directly from image pixels [41]. This method, like many others, required the face images to be scaled, aligned and cropped in order to perform well. In [21], the authors also took a holistic view of the face, creating so-called holons – reduced feature vectors learned via an auto-encoder – to perform identity, emotion and gender recognition. Research has shown that both age and ethnicity play a big role in gender recognition. For example, gender recognition performance has been shown to degrade when models are trained on a mixture of ethnicities, rather than focused on a target ethnicity [38]. In addition, gender recognition performance significantly depends on age, with young males and older females posing challenges for the models [8, 43].

One of the earliest works on age recognition focused on cranio-facial development theory, developing models to describe the changing shape of the face as it aged [76]. Focusing on texture as well as shape, active appearance models – statistical models – were developed for age recognition from face images [78]. Success was also found in age estimation by considering a collection of images from an individual in order to determine the aging pattern for that person [39]. Age estimation from face images remains a very challenging problem in the computer vision community because each person ages differently, and so it is a profoundly individual problem [125]. Adding to the challenge of age recognition problems, they can be considered a categorical classification problem (e.g. to what age group does this face belong?) or a regression problem (e.g. what is the age of this face?), depending on the context and data available.

Attributes exist in domains other than faces, including pedestrians, objects, and actions. An attribute is simply a describable feature, and so it lends itself nicely to many problems in computer vision. Pedestrian attributes can include clothing,

gender, hair color and length, as well as part visibility and pose [183]. Attributes of objects can include multiple categories: shape, color, texture, part, and material as well as global or local presence of the attributes [9, 31]. Attributes of actions include pedestrian attributes, object attributes as well as action-specific attributes such as environment and motion [89].

The problem of attribute recognition has gained a lot of attention in the research community over the past decade, mostly due to the wide applicability of attribute prediction for real-world applications. Pedestrian attributes have been used in surveillance for re-identifying individuals and searching for suspects based on a description of their visual attributes [98]. The application of attribute detection to surveillance is ultimately an image search problem where a query of attributes is provided and the most relevant results are returned. Thus, most image search techniques can directly correlate with surveillance, and many have utilized facial attributes [137, 141]. Applying attribute recognition to surveillance can lead to quicker identification and save a significant amount of human hours.

Another application of attribute recognition is in human computer interaction (HCI). Many applications that involve HCI benefit from knowledge of the user. For example, proper greetings rely on gender information (e.g. Mr., Ms.). A user's expression can determine whether or not they are enjoying an application (e.g. smiling or frowning). More specifically, attributes have been used by companies such as Facebook in order to improve accessibility of their platform, providing image descriptions to those with visual impairments [113]. Other applications of facial attributes in HCI include active authentication, the process of continuously authenticating a user on a device. Attributes have already been successfully deployed for this problem [134, 135, 136]. Being describable features by definition makes facial attributes widely applicable to many real-world problems.

Focus in attribute research has been dedicated to the general discovery of attributes for building datasets and vocabularies for preexisting data. Note that these approaches operate in a broader scope than just facial attribute recognition. [9] made

strides to automatically aggregate and label data from noisy internet sources. They cite work that uses gender, race, and other attributes to improve face verification and search. Expanding on this concept, the authors mine websites that have diverse images and automatically label the images based on captions. This allows for diverse vocabulary discovery and improved predictions across multiple object types. A similar approach, with the goal of learning attributes in large datasets, draws connections between semantically unrelated objects by looking at their visually describable attributes [131]. For example, zebras, beetles, and street crossings all share the stripes attribute. With the development of several large-scale labeled datasets for the problem of facial attribute recognition, the field has grown significantly.

In the following sections we detail the research in facial attribute recognition from images and videos. We will also present work in the general field of attribute recognition, when applicable. All the while we will be introducing datasets and discussing methods based on traditional machine learning and computer vision as well as those based on deep learning.

2.1.1 Related Research

Attributes are not solely applicable to faces. They have been successfully applied to objects, pedestrians and actions. Here we provide a brief history of each field, from early works to state-of-the-art.

Attributes of objects include textures, colors, patterns, shapes and many other describable features. Early methods for attribute recognition were focused on aiding object recognition. These initial works recognized basic patterns, textures, and colors [33, 59, 79, 93]. As researchers became more active in the field, the focus shifted to describing objects, rather than simply naming them [31, 120]. Many researchers utilized object attributes for few and zero-shot learning as they provide a compact description of objects that a system may not have seen previously. In [165], Wang et al. focused on dependencies between objects and attributes, improving both attribute and object recognition. [147] identifies attributes (e.g. shape, color, material) of 3D

objects with the goal of helping autonomous robots understand and interact with the world around them.

In [157], the authors present a dataset (AirplanOID) and a method for understanding objects in fine-grain detail, using attributes. The dataset contains images of airplanes and attributes such as *facing direction*, *is-airline*, *location*, etc. More recently, Wang et al. further explore attributes for object recognition [166]. Their approach utilizes attributes as additional information during model training, requiring no attribute labels at test time.

Attributes of actions include descriptive features such as environment, pose, objects involved, etc. that can be used to break an action down into its component parts. One of the first works in action attribute recognition modeled the human visual cortex. This was accomplished by applying motion-direction sensitive units to video inputs, thereby recognizing human body, head, hand and general animal actions [58]. [89, 175] and [140] all focused on identifying action parts in still images. In [89], the authors used attributes of a scene to understand actions. Yao et al. used a combination of given action verbs (e.g. bending, squatting, riding, etc.) along with poselets and objects to predict actions from still images [175]. In [140], the authors presented a method which learns a template for a variety of actions in order to localize actions in a frame. Zhang et al. presented a multi-task learning method in which attributes and actions are learned simultaneously [184]. [148] and [13] focused on attributes for action recognition in 3D. State-of-the-art methods rely on supervised deep learning in order to recognize attributes of actions [17, 66].

Attributes of pedestrians include whole-body attributes such as clothing, pose, etc. as well as facial attributes. Identifying attributes in this context can be challenging due to viewpoint and extreme pose changes. With a focus on gait analysis, [180] used K Nearest Neighbors and spectral clustering to identify attributes such as gender and age from gait information including speed, acceleration, rhythm, etc. In work done by Deng et al., support vector machines were trained on a large-scale dataset to recognize attributes of pedestrians [25, 26]. The authors collected the PETA dataset

for these works, which is still a benchmark in the field [26]. In recent years, deep learning has become the standard for pedestrian attribute recognition, with the focus on convolutional and recurrent neural networks [18, 98, 178, 186, 190]. Convolutional neural networks are useful for localization of pedestrian attributes, while recurrent neural networks are successful in identifying attribute relationships. Automatic recognition of pedestrian attributes has applications in soft-biometrics, surveillance and autonomous vehicle guidance.

2.1.2 Theory and Application

We review work on facial attribute recognition from images and video and separate work into two categories: traditional methods and deep learning.

2.1.3 Attribute Recognition with Traditional Methods

Prior to the advent of deep learning in all aspects of computer vision, other traditional methods, such as support vector machines were used for attribute recognition. In 2008, Kumar et al. built a face search engine that they called FaceTracer [73]. This search engine operated on user queries involving one or more of the available attributes. For example, “smiling Asian men with glasses.” The search engine would then return face images that exhibited the desired traits. The search engine was built on a set of attribute classifiers, capable of identifying binary facial attributes in an image. The attribute classifiers were built on four feature sets: face region, pixel data color space (e.g. RGB, HSV), normalization method, and data aggregation method. Support vector machines (SVMs) were then trained for every region, feature type, and parameter combination. Adaboost is then run on this set of “local SVMs” to generate a set of strong classifiers. Finally, a global SVM is trained by finding the union of the strong classifiers. Along with the FaceTracer search engine, Kumar et al. introduced a dataset by the same name. At the time of publication, the dataset consisted of over 3.1 million face images, 17,000 of which were manually labeled with 10 attributes: age, gender, race, hair color, eye wear, mustache, expression, blurry,



Figure 2.1: Sample images from the FaceTracer dataset [73].

lighting, and environment. Sample images from the FaceTracer dataset are shown in Figure 2.1.

A year later, the same group shifted their research focus toward face verification using attributes [74]. Face verification aims to address the following question: given two images, do they belong to the same person? The authors developed two different methods to generate descriptions of faces, using attribute and simile classifiers. SVMs are used as attribute classifiers, trained on a collection of low-level features, similar to the previous work [73]. They introduced additional low-level features such as edge magnitudes and gradient directions. As a part of this work, the authors introduced a new dataset, PubFig, for face verification. Sample images from the PubFig dataset are shown in Figure 2.2. Additional data was collected in order to train facial attribute classifiers. 1000 images were labeled for each of 65 binary attributes using Amazon Mechanical Turk. Simile classifiers were used as well to identify the similarity of a face to a set of reference faces. For each reference individual, a classifier is trained on each region to distinguish that region from the same region on other faces. These simile classifiers allowed for comparisons between faces without requiring additional labels. The final face verification system utilized a hybrid of attribute and simile classifiers and achieved state-of-the-art accuracy. After the release of PubFig, the authors tested their attribute classifiers on all images in the dataset, providing 65 attribute scores for each image along with the image data. Some methods utilized these scores as labels in order to train attribute classifiers [136]. In 2011, the same



Figure 2.2: Sample images from the PubFig dataset [74].

group again used facial attributes for improved face verification and image search [75], extending the set of attributes to 73.

Several groups realized the potential of facial attributes to improve image search and retrieval with natural queries [73, 137, 141, 156]. With a focus on surveillance, Vaquero et al. utilized pedestrian attributes for search and retrieval in low-quality video [156]. Others have explored different ways to perform multi-attribute search queries [137, 141]. [141] improved over previous ranking methods that required individual models for each search term. Instead they used correlations amongst attributes to provide additional information to the search query. Their Multi-Attribute Retrieval and Ranking (MARR) method benefited from the strong relationship amongst attributes. The authors labeled a subset of the Labeled Faces in the Wild (LFW) dataset [56] (9992 images) with 27 binary attributes (a subset from the work of [74]). A year later, [137] focused on developing a meaningful way to combine different attribute scores. They construct a normalized score space based on Extreme Value Theory. The authors aimed to convert raw SVM output scores to a normalized score that would be more consistent with human labeling as well as with the scores of other attributes. After converting the scores, they were able to fuse them to allow for multi-attribute queries in a shared score space. The authors use the method of [75] to extract attribute scores from face images.

All of the publicly available datasets up to this point considered facial attributes to have binary values, that is, the attribute is either present or it is not. This can

be a very challenging way to view the problem when many attributes are subjective or exist on a gradient (e.g. some hair is more blond than others). Parikh et al. aimed to address this issue in [121], focusing on so-called “relative attributes.” The authors utilize the concept of relative attributes to generate a ranking function for each attribute allowing for a new type of zero-shot learning in which they can describe unseen objects relative to previously seen ones. They propose a learning-to-rank formulation that learns a desired ordering of the training images. This learning framework resulted in a model that could better capture the strength of a particular attribute compared to a binary classification model. The authors utilized a subset of the PubFig dataset in order to learn their ranking functions.

Labeling attributes is a time-consuming process, with each image needing multiple (over 70 in the case of [75]) labels. This became a limiting factor in facial attribute research very quickly. In [10], the authors introduce a likeness measure as a way to utilize describable features without requiring an extensive labeling process. The goal of this work is to improve face verification performance. For each pair of subjects, the authors create a classifier that is capable of distinguishing between the two subjects. This process results in many likeness classifiers, or “Tom v. Pete” classifiers, as they call them. Face images are classified using this collection of “Tom v. Pete” classifiers giving a set of scores that indicate the person’s likeness to a particular subject in a pair classifier. This set of scores is then used as a subject’s feature vector which is in turn used for face verification. This work resulted in human-describable features of faces, in the form of their likeness to other individuals. That is, “this person looks more like subject 1 than subject 2, and more like subject 3 than subject 2” etc. [10] built on the concept of automatically generated attributes as seen in the simile classifiers of [74].

2.1.4 Attribute Recognition with Deep Learning

In 2015, Liu et al. introduced two large-scale benchmark datasets for the problem of facial attribute recognition in unconstrained images – CelebFaces Attributes (CelebA)

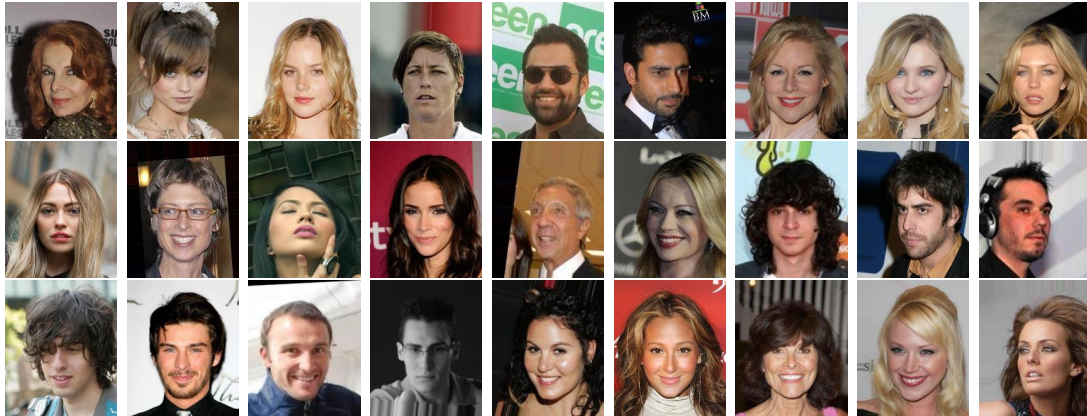


Figure 2.3: Sample images from the CelebA dataset [96].

and Labeled Faces in the Wild Attributes (LFWA) [96]. CelebA contains over 200,000 images each labeled with forty binary attributes, which are a subset from those used in [75]. The CelebA dataset contains a wide range of images including full body and close cropped faces. The dataset includes these original images as well as cropped and aligned face images. Sample cropped and aligned images from CelebA are shown in Figure 2.3. Along with CelebA, attribute labels were added to the popular face verification benchmark LFW, creating LFWA. LFWA contains roughly 13,000 images, labeled with the same forty binary attributes from CelebA. Some sample images from LFW can be seen in Figure 2.4. CelebA and LFWA were the first (and only to date) large-scale datasets introduced for the problem of facial attribute recognition from images. Prior to CelebA and LFWA, no dataset labeled with attributes was large enough to effectively train deep neural networks. With the introduction of this dataset, many deep learning methods were used for facial attribute recognition [51, 96, 130, 134, 161].

Along with CelebA and LFWA, Liu et al. introduced a method for attribute recognition that involves two networks: LNet and ANet. LNet is a localization network that localizes the face with weak attribute supervision, and ANet uses the localized face to predict facial attributes. LNet was built on the widely popular AlexNet [72] architecture trained on the ImageNet object recognition dataset [23]. After being pre-trained on the large-scale Imagenet dataset, LNet was fine-tuned on



Figure 2.4: Sample images from the LFW dataset [56, 96].

the original (full body, unaligned) CelebA images using weak attribute supervision. With the weak supervision from CelebA’s facial attributes, LNet was able to accurately localize the face in a given image. Once the face was localized by LNet, ANet was trained from the cropped face image. ANet was also built on the AlexNet architecture, pre-trained on ImageNet, and fine-tuned on CelebA. This two-network scheme produced impressive results on CelebA and LFWA.

In 2016, Wang et al. introduced a method dubbed “Walk and Learn” in which they utilized face tracks as additional supervision for facial attribute recognition [161]. The authors collected additional data by attaching wearable cameras to their bodies and walking around different areas of New York City. They used face tracking to identify individuals in every frame and used face verification to pre-train their network. Their deep network was then fine-tuned on the CelebA dataset, producing improved results on some challenging attributes over [96].

Just a year after the release of CelebA and LFWA, many researchers began to notice some very serious label imbalance issues. In particular, [130] focused on adjusting the label imbalance during network training. The authors introduced a Mixed Objective Optimization Network (MOON) capable of learning all attributes at once while at the same time adjusting for label imbalance. We note that prior to this work, individual models were learned for each attribute, including all of the methods previously discussed [73, 74, 75, 96, 161]. This was incredibly inefficient and did not take advantage of a shared representation for all attributes. MOON addressed

both of these issues by utilizing the popular VGG-16 network architecture [142] and training from random initialization on CelebA. MOON was the first to combine attribute learning into one network, address dataset imbalance, and train on CelebA from scratch rather than fine-tune. MOON addressed label imbalance by calculating source and target distributions for each attribute and applying a weight to the backpropagation within a euclidean loss in order to adjust for the distribution discrepancies. The source distribution for an attribute was the distribution of positive and negative instances of the attribute in CelebA, and the target distribution could be set to any desired distribution, though the authors experimented with an even target distribution. MOON produced impressive results on CelebA and highlighted the severe imbalance issues associated with it.

[30] also tackled the problem of multi-task learning for facial attribute recognition, utilizing a Restricted Boltzmann Machine (RBM) rather than a CNN. Their model is trained with both the aligned face images from CelebA and facial landmark points as inputs. The authors extend RBMs to handle multiple tasks and multiple inputs naming it the Multi-Task Multimodal RBM (MTM-RBM). The MTM-RBM compares favorably with [96]. To date this is the only method for facial attribute recognition that utilizes an RBM model.

With the introduction of deep learning in the facial attribute domain, many began to wonder how robust these models truly are. In [129] they aim to address this question by introducing an adversary. They develop a Fast Flipping Attribute (FFA) method that generates adversarial examples that cause classification errors. The FFA method identifies directions which can generate adversarial examples by inverting the classification score and calculating the gradient with respect to the inverted score. Searching along those gradient directions results in images that produce classification errors. The authors found that some attributes (e.g. *wavy hair* and *wearing necklace*) were more robust to adversarial attacks than others (e.g. *big nose* and *young*).

Several groups began to address the problem of facial alignment in attribute recognition. In [48], the authors introduce the Alignment-Free Facial Attribute Clas-

sification Technique (AFFACT), which performs data augmentation allowing a deep convolutional neural network to recognize attributes without first aligning the face images. The AFFACT method performs augmentation of the dataset through scaling, rotation, shifting, and blurring. Training ResNet [52] architectures, the authors applied AFFACT data augmentation to CelebA and were able to achieve state-of-the-art performance. [28] also aimed to address the problem of attribute recognition from unaligned face images by utilizing a cascade network capable of identifying different regions of the face and recognizing attributes without alignment. Their face region localization network is capable of detecting face regions based on weakly supervised attribute data. Rather than performing data augmentation like [48], this work focused on part-based approach to attribute prediction.

Only two years after the introduction of CelebA and LFWA, performance on the benchmark datasets began to plateau. In [51], the authors aimed to improve attribute recognition accuracy by taking advantage of relationships amongst attributes both implicitly and explicitly. The authors introduce a new deep CNN architecture for attribute recognition: Multitask CNN (MCNN). MCNN had fewer than 16 million parameters compared to the 138 million parameters in the VGG-16 model used for MOON. MCNN took advantage of attribute relationships by learning a shared representation at the lower levels of the network and branching off into spatial attribute groupings at the higher levels of the network. Finally, attribute relationships were learned at the score level with an auxiliary network (AUX) that was attached to the trained MCNN. The combined network, MCNN-AUX utilizes attribute relationships in three different ways and produced state-of-the-art results on CelebA and LFWA.

Aiming to utilize localization cues to improve facial attribute prediction, [61] combined the problem of facial attribute recognition with that of semantic segmentation. Semantic segmentation requires predicting a label for every pixel in an image, producing a class map over the entire image. The authors aggregated face segments, provided as a part of the Helen Dataset [80], to create seven segments: background, hair, face skin, eyes, eyebrows, mouth and nose. They utilize a gating mechanism

to focus the attribute recognition network on regions of interest for a particular attribute. For example, they focus mouth-related attributes (e.g. smiling, mouth open) on the mouth segment provided by the segmentation method.

Along a similar vein, [53] uses generative adversarial networks (GANs) to generate abstraction images that are then used to improve facial attribute recognition through a multi-stream network acting on the abstraction and original images. The abstraction images produce a kind of facial segmentation with textual information, localizing parts and providing additional supervision to the facial attribute recognition task. The multi-stream abstraction image formulation for attribute recognition outperformed the recent work of segmentation for improved facial attribute recognition [61].

As a follow-up to MOON, [49] introduced a method called “Selective Learning” to perform balancing of multi-label datasets during training of a deep neural network in order to address the label imbalance in CelebA. The authors introduce a new CNN – Attribute CNN (AttCNN), which has roughly 6 million parameters, compared to the 16 from MCNN [51]. In MOON, the labels were balanced by considering an overall dataset source distribution for each attribute. In [49], the authors note that this does not fully address the problem as each batch that is used to train the CNN may be more or less balanced than the overall training set. The author’s solution was to perform label balancing at the batch level. Every attribute in each batch was balanced according to a desired target distribution by sampling from the over-represented class and weighting the underrepresented class. The Selective Learning method produced comparable results to MOON on CelebA and LFWA. The authors also introduced a new evaluation dataset: the University of Maryland Attribute Evaluation Dataset (UMD-AED). UMD-AED consists of roughly 3,000 images sparsely labeled with facial attributes. Some sample images from UMD-AED are shown in Figure 2.5. Each of the forty attributes in CelebA has fifty positive and fifty negative instances in UMD-AED, allowing for balanced testing of facial attribute recognition methods. Selective Learning and AttCNN significantly outperformed MOON on UMD-AED.

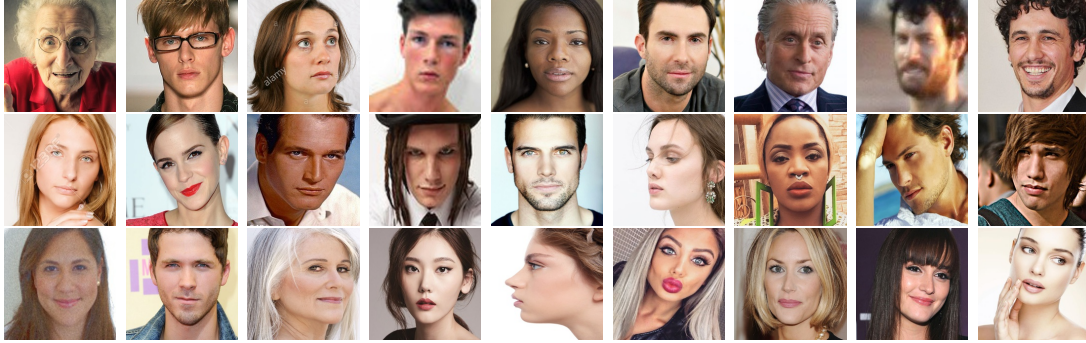


Figure 2.5: Sample images from the UMD-AED dataset [49].

Most research in facial attribute recognition focused on unconstrained images. Hand et al. shifted the focus to video in [50] using weakly labeled video to train attribute prediction models. The authors labeled four frames in every video of YouTube Faces [172] – a video dataset collected for face verification – with the forty binary facial attributes from CelebA. They introduced several methods for utilizing weakly labeled frames to improve attribute prediction in video: Motion Attention and Temporal Coherence. Their motion attention mechanism focused attribute models on areas of motion in the video, reducing overfitting, and the temporal coherence constraint encouraged nearby frames to have similar network responses, relying on the fact that nearby frames in a video will likely have similar – but perhaps not the same – attributes. Combining motion attention and temporal coherence, the authors were able to train a deep CNN on unlabeled video frames from YouTube Faces, outperforming traditional fine-tuning methods. [50] was the first, and only to date, attempt to utilize video for facial attribute recognition.

We present the average accuracy over all attributes in CelebA for all state-of-the-art methods in Table 2.1. We can see that since the introduction of the dataset in [96], only a four percent gain in accuracy has been achieved on average. This emphasizes that there are many challenges that have yet to be addressed in the field of facial attribute recognition.

The field of facial attribute recognition is still a very young one, having been introduced just over a decade ago. Since its introduction, huge strides have been

made, with current systems capable of recognizing facial attributes in unconstrained images and video. There are many open research directions that will lead to significant improvements in the state-of-the-art in facial attribute prediction. With many applications relying heavily on the recognition of human-describable features, the field of facial attribute recognition will be of great interest for many years to come.

Table 2.1: Average attribute classification accuracy across all forty attributes in CelebA for current state-of-the-art methods.

Method	Accuracy
Liu et al. (LNet+ANet) [96]	87.30%
Ehrlich et al. (MTM-RBM) [30]	87.00%
Wang et al. (Walk and Learn) [161]	88.00%
Rudd et al. (MOON) [130]	90.94%
Gunther et al. (AFFACT) [48]	91.97%
Hand et al. (MCNN-AUX) [51]	91.30%
Kalayeh et al. [61]	91.80%
Ding et al. [28]	91.23%
Hand et al. (AttCNN) [49]	91.05%
He et al. [53]	91.81%

2.2 Model Interpretability

The rapid growth of complex and opaque decision systems, particularly CNNs, has led to their widespread application in critical areas such as medicine, security, and finance. The increasing accessibility of these models through software libraries like Tensorflow [1] and PyTorch [122] has further contributed to their ubiquity in both research and industry. However, the lack of interpretability in these “black box” systems, where the internal logic is hidden from the user, poses significant practical and ethical challenges [2].

Adadi et al. identify four key motivations driving the field of explainable AI (XAI): justification, control, improvement, and discovery [2]. While various definitions of interpretability exist within the machine learning community, a formal technical meaning remains elusive. Lipton suggests that an interpretable model can be

characterized by its transparency (how does the model work?) and post-hoc explanation (what more can the model tell us?) [88]. In this work, we consider a model explainable if its decisions are consistent with human intuition and can be compared to human decision processes.

Existing approaches to interpretability can be broadly categorized into reverse engineering and design of explanations. Reverse engineering techniques reconstruct an explanation for a decision based on a dataset of training decision records, while design methods develop an interpretable predictor model alongside the decision set [42]. A third approach, simplifying models to fit within a small class of explainable systems, often sacrifices accuracy and performance for the sake of simplicity and does not contribute to the advancement of XAI.

Reverse engineering methods, particularly those tailored to CNN architectures, include visualizations such as heat maps and bounding boxes [88]. Heat mapping techniques, such as DeConvNet [181], Guided Backprop [102, 145], CAM, and Grad-CAM [16, 118, 138, 189], highlight important regions within an input image for discrimination. However, these constructions are heuristic notions of image saliency [35] and are limited to isolated examples, unable to provide overarching generalizations of a model’s attention.

Design techniques, such as LIME and SHAP, explain classifier predictions by learning comprehensible sparse local predictors around the decision [99, 128]. However, Slack et al. demonstrate that post-hoc explanation techniques relying on input perturbations can be unreliable and easily manipulated to provide innocuous explanations that do not reflect underlying biases [143].

Kim et al. demonstrate that saliency maps should be utilized for explainability with caution, as some methods produce very similar explanations for trained models as they do for models with randomized network weights [3]. When using visual analysis of explanations alone, inductive bias can play a role in our interpretation of visual features which are most important to the model. Another issue with gradient based approaches is that they are only applicable to models based on artificial

neural networks. Finally, many methods (reverse engineering or design techniques) which produce an interpretation of the input features that are important to a model’s prediction only provide an explanation for a single sample at a time.

There has been limited work in the explanation of classifiers used in facial processing. Most effort has been applied to facial recognition systems. In the context of facial processing a common predictive task is that of face verification (i.e. given a pair of images, predict if they depict the same identity). In this case, explanations should be provided to show which visual features were utilized to make a prediction of same or different. An issue in explaining model decisions is that humans utilize contrastive explanations [2]. For example, ”identity A’s eyes are *darker* than identity B’s eyes.”

Rajpal et al. provide the performance of LIME on various facial recognition datasets [124]. The authors tasked human labelers with scoring the utility of the explanations. Another approach to providing meaningful interpretations of vision model perceptions is found in the work of Williford et al. [171]. An explainer is provided with three images: a probe, a mate, and a non-mate. The probe and the mate are real images of the same identities. The non-mate is an image of the identity which has been altered such that a prominent visual region has been swapped with that of another person. Their algorithms are trained to produce a saliency map which reflects the pixels which make the mate more similar to the probe than the non-mate.

A significant missing contribution is that of global, model agnostic explainability for facial processing models. Global explainability is important as it provides insight into the features used by the model across many visual environments. Model agnosticity is important as visual processing systems may have varying computational constraints.

2.3 Semantic Segmentation

Semantic segmentation is a valuable task in the field of computer vision. The task is to assign one or more labels to each pixel in an input. Unlike a classification task, which

predicts which semantic concepts are depicted in an image, semantic segmentation also finely localizes where such concepts occur. Example applications of semantic segmentation are autonomous driving, pedestrian recognition, and computer aided diagnostics [12, 32, 34, 155, 191, 192].

Face parsing, also known as facial semantic segmentation, is a specialized sub-field of semantic segmentation that focuses on segmenting facial regions into semantically meaningful parts, such as skin, hair, eyes, nose, and mouth. Traditionally, Conditional Random Fields were used by all state-of-the-art methods for face parsing [60, 144, 168]. As in many other fields, deep learning became the new state-of-the-art [85, 91, 94, 100, 144].

2.4 Image Clustering

In this section, we detail two broad classes of methods for image clustering. The first class of methods that we describe jointly learn to compress images into dense representations and cluster the dense representations into classes. Fard et al. [174, 108] propose methods that tune an autoencoder to generate k-means friendly representations. In [173], Xie et al. pass samples through an encoder to generate representations, cluster with k-means and correct the cluster assignments with a clustering loss based on a KL divergence between soft assignments and their target distribution. Borrowing from Xie et al., [44] and [45] use the same learning framework with an undercomplete autoencoder to preserve the local structure of input data. Wang et al. [164] pass the input image through an orthogonal autoencoder prior to applying spectral clustering. Affeldt et al. [5] use multiple autoencoder architectures to generate multiple representations from the input data. The representations are then clustered with spectral clustering. The authors of [15] propose an architecture in which a neural network reduces the dimension of input images. The learned representations are clustered and the corresponding pseudo labels are used as supervision for training the network.

The second group of works that we highlight are miscellaneous techniques for improving cluster performance. Li et al. [83] use a boosting method to train on easier

samples, then gradually expose the model to more challenging data. [47] utilizes an ensemble of classifiers to generate cluster assignments and compute a similarity graph. Finally the similarity graph is pruned to extract high confidence cluster assignments. [27] uses a modified VAE in which the latent space is sampled from a mixture of Gaussian distributions. Clustering is achieved by calculating how far the mixture distribution is from the normal distribution. Lastly, Li et al.[84] implement multi-view autoencoders for multi-view data with shared weights. Their network structure has a deep embedding clustering layer which recalculates cluster centers each iteration.

2.5 Face Recognition and Existing Benchmarks

2.5.1 Face Recognition

Face recognition is separated into three well-defined steps: (1) face detection and localization, (2) extraction of features from the detected face, and (3) classification (verification or identification) [70]. The first task is to decide whether or not there are faces in an image. If there are one or more faces, then the system identifies bounding boxes for each face. The feature extraction step generates a feature vector from the localized face. This feature vector should be discriminative enough to separate images of one identity from images of other identities. Lastly, there is the classification step. This is separated into two classes of techniques: identification and verification. In the identification scenario the system is aware of a finite number of identities and it should learn to match each image sample to one identity class. For the verification task the model is only provided with supervision in the form of a binary label which represents either same or different, and so pairs of images are compared at each step. Any face recognition system that is meant to be deployed in “the wild” will need to perform all three of these steps. That being said, each step is commonly considered an active research topic.

2.5.2 Existing Benchmarks

There are a large number of datasets collected and presented for the purpose of facial feature extraction and classification. Many of these datasets are designed either for training or evaluation. Here we describe the major datasets that already exist for the purpose of model evaluation and benchmarking and compare them with the proposed DoppelVer dataset.

Labeled Faces in the Wild (LFW) [56]: The LFW dataset was created by Huang et al. in 2007. At the time of publishing, many face recognition datasets were collected by small teams of researchers with the intent of collecting facial images in constrained settings. LFW however was meant for studying the problem of recognizing faces in unconstrained settings. The dataset contains 13,233 images and 7,549 identities. The researchers behind LFW contributed significantly to the field by presenting a dataset organization that focused on the honest reporting of results for the task of open-set face recognition. Their dataset contains a development view and an evaluation view as well as splits for 10 fold cross-validation. The current SOTA accuracy on LFW is 99.8% (± 0.2001) [6].

AgeDB [110]: This dataset was introduced in 2017, with a focus on accurate hand-labeling of age. This is a useful database when performing tasks such as age-invariant face verification, age estimation, and face age progression. The database contains 16,488 images of 568 identities with accurate-to-the-year age labels. The average number of images per individual is 29, with an age range of 1 to 101 years old, the average age for an individual being 50.3 years. AgeDB provides four face verification protocols, each split into 10 folds following LFW’s process. These four protocols restrict the age variance across sample pairs. The provided protocols cap age range to 5, 10, 20 and 30 years respectively. The current SOTA accuracy on AgeDB 30 is 98.7% [7]

Cross-Age LFW (CA-LFW) [188]: The authors of this database posit that methods reporting accuracy on LFW’s benchmark are optimistic. To show this, CA-

LFW has both positive and negative pairs which depict a large age gap, while also providing negative pairs which are of the same race and gender. These visually similar negative pairs emphasize the effect of age difference on classifier performance. This dataset contains the same identifies as LFW with 6,000 image pairs. The current SOTA accuracy on CA-LFW is 95.87% [24]

Cross-Pose LFW (CP-LFW) [187]: CP-LFW was proposed by the same authors as CA-LFW and was released one year later. This publication shifts focus to the important task of face verification in the presence of extreme pose. They note that nearly all images in LFW are near-frontal, suggesting that results on LFW provide a poor representation of a face recognition method’s performance when deployed into a real setting. The current SOTA accuracy on CA-LFW is 92.08% [24]

Chapter 3

Deep Vision Model Perception of Gender From Faces

3.1 Introduction

In recent years, convolutional neural networks (CNNs) have achieved remarkable performance on a wide range of computer vision tasks. However, this advancement has come hand-in-hand with a substantial increase in model complexity [19], making it difficult to understand how these models perceive and process visual information. This lack of interpretability is a major challenge, as it hinders our ability to trust and rely on these systems, especially in critical applications [2]. To be clear, interpretability is an important problem for many models. CNNs and deep networks have simply exemplified the issue of models which are capable of modeling highly complex patterns.

In this chapter, we present a novel architecture-agnostic approach for interpreting the perceptions of vision models when making predictions on facial data. We draw inspiration from cognitive science research of human face perception to develop techniques for analyzing and explaining the perceptions of vision models. As a basis for exploring this topic we utilize a CNN trained to predict gender with a large face dataset.

Our experimental design, motivated by human perception research [150], systematically occludes key facial regions and measures the impact on model performance. This technique allows us to identify the most influential visual features and under-

stand how the model’s attention shifts based on available information. Comparison to well-established characteristics of human face perception demonstrates that our method produces interpretations that are consistent with the human decision-making processes [107]. As such, our explanations are accessible and understandable to a wide audience.

By bridging the gap between vision models and human cognition, this chapter contributes to the development of more interpretable and trustworthy artificial intelligence systems, paving the way for their wider adoption in critical applications. Additionally, we cultivate an understanding of the importance of specific visual data to the prediction of particular facial attributes. We apply this knowledge in Chapter 4.

The key contributions of this chapter are as follows:

1. We introduce a novel occlusion-based technique for interpreting the perceptions of vision models, which is architecture-agnostic and applicable to a wide range of models.
2. We demonstrate the effectiveness of our approach by analyzing ResNet-50 [52] on the task of gender classification using the large-scale CelebA dataset [96].
3. We identify key facial regions and information encoding patterns that are most influential for the model’s predictions, and show how the model’s attention shifts based on the available information.
4. We draw parallels between the model’s perceptions and well-established characteristics of human face perception, demonstrating that our interpretations are consistent with human decision-making processes.

3.2 Proposed Method

In this section, we introduce a novel occlusion-based technique for interpreting the perceptions of vision models. Our approach is architecture-agnostic and applicable to a wide range of models that accept images as input data. While we focus on

gender recognition in this chapter, our method can be easily extended to other facial attributes.

We consider a black box classification model that maps the feature space (RGB images) to the decision space (output predictions) through a non-transparent learning process. Our goal is to identify image regions that significantly impact the model’s output by occluding targeted areas of the input images and measuring the corresponding changes in model performance. The selection of occlusion regions is based on research of human perception. To this end we define five primary facial regions consistent with cue-driven human perception [111, 150]: eyebrows, eyes, nose, mouth, and chin.

Occlusions are modeled as the removal of visual information from a target region by replacing all pixel values in the region with a pixel value of 0. This operation restricts the model’s ability to extract meaningful visual features from the occluded area.

We aim to search for informative occlusion regions that cause a significant decrease in the target score for predicting a given class. Instead of relying solely on class probability, we utilize accuracy, precision, and recall scores as indicators of meaningful regions. Accuracy provides information on the general importance of a region to the model’s predictions, while precision, and recall score indicate potential model bias and class-skewing when data is obscured.

We adopt the language of *configural* and *featural* importance to describe the role of different facial regions in the model’s decision-making process. A region with high configural importance will produce lower metric scores when it is isolated and occluded, indicating that it significantly contributes to the model’s overall performance in combination with other image regions. Conversely, a region with high featural importance will itself be predicted with higher accuracy than a non-featurally important region, demonstrating that it contains sufficient information for prediction on its own. Notably, featurally important regions exhibit high intra-class similarity, showing strong distinctions between samples of different classes while resembling each

other within the same group.

To assess the configural and featural importance of facial regions, we train two distinct predictive models. The first model is trained on images with all visual features available, without any occlusions. The second model is trained on images with all visual data removed, except for the target region. Evaluation of both models is performed with both non-occluded images and occluded images. For the first model, the absolute difference between performance metrics is used to select regions of high configural importance. For the second model, high metric scores for non-occluded images indicate the featural importance of the target region. Performance on the occluded images indicates the distinctness of the target region between labeled classes. Figure 3.1 provides a visual aid for understanding the system.

The two models serve to eliminate spurious correlations and local explanations that are not robust to artifacts. The first model assesses how well the base classifier can classify an image with a region absent, analogous to the neuro-scientific concept of configural importance. The second model tests the classifier’s ability to accurately determine the class of the image based on an isolated region, referred to as featural importance in human studies.

As our method primarily seeks to express decision justifications using transparent human vocabulary, we employ extensive analysis to generalize understandable patterns in network behavior. A behavior that has been observed and validated through experimentation is referred to as a trend. These trends comprise a set of rules describing the logic behind the black box model, thereby providing interpretability at a global level. To ensure the generalizability of our explanations, we test our trends using 5-fold cross-validation on both high (224x224) and low (32x32) resolution datasets.

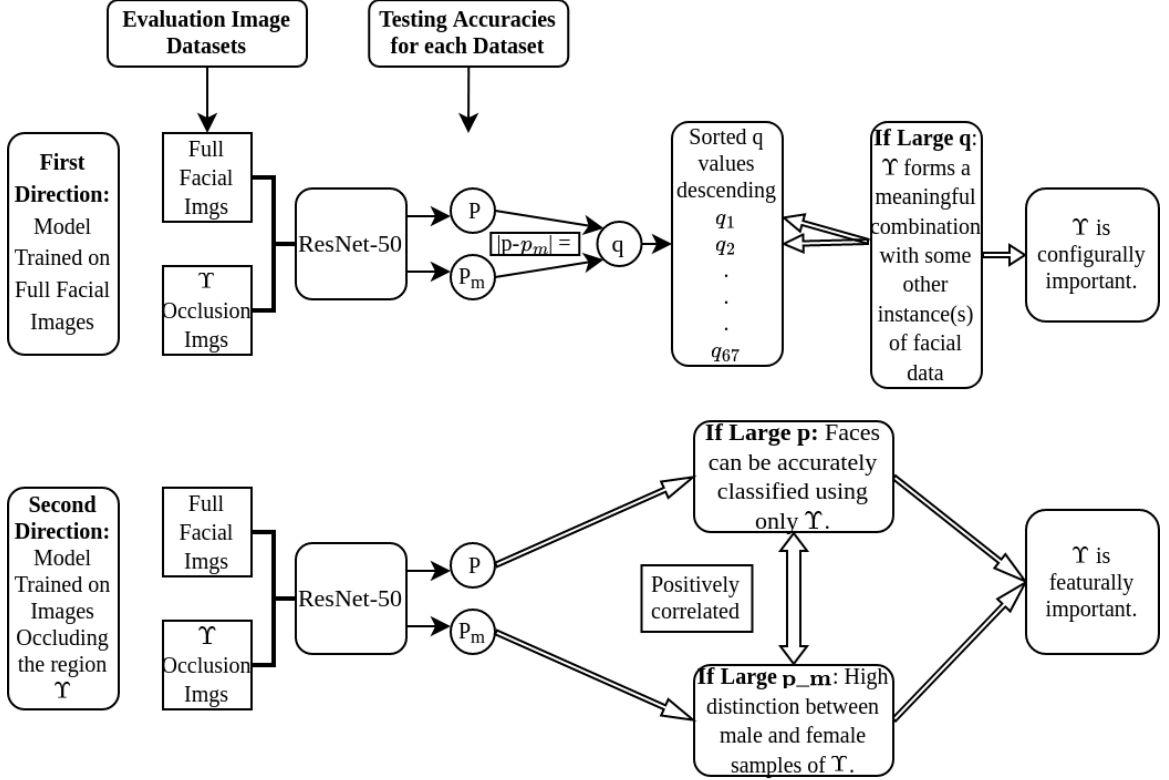


Figure 3.1: **The Two Directions of Evaluation.** The figure shows the testing phase for two ResNet-50 instances: one trained on full facial images (first row) and another trained on occlusion images of a given region Υ (second row). The test accuracy for the full facial dataset is p , and the accuracy for occlusion images is p_m . In the first scenario, the absolute difference between these values, q , measures the configural importance of Υ . In the second scenario, p indicates the featural importance of Υ , while p_m determines the distinctness of Υ samples between classes.

3.3 Experiments

3.3.1 Model

To demonstrate the capability of our technique we train ResNet-50 [52], from scratch, on the task of gender recognition. This is accomplished by replacing the final fully connected layer with a linear layer which produces a single output. The sigmoid function is applied to the output of the network. During training, the output is passed into the binary cross entropy loss function and the model weights are updated using the Adam optimizer [67]. During inference we apply a threshold of 0.5 to the

output to produce an attribute prediction. We implement this network with PyTorch [122]. Hyperparameter selections are provided in Table 3.1.

Table 3.1: ResNet-50 Model hyperparameters, selected based on a grid search.

Field	Low Res. Model	High Res. Model
Image Size	32x32	224x224
Batch Size	128	128
Learning Rate	0.0005	0.001
Epochs	50	50
Momentum	0.4	0.5
Dropout	0.5	0.3

3.3.2 Data

Our experiments utilize the challenging, publicly available dataset, CelebA [96]. Originally collected for attribute classification, CelebA contains roughly 200,000 images. The images are split into training, validation, and test splits with approximately 80%, 10%, and 10% of samples respectively. The samples in CelebA vary widely with regard to subject pose, illumination and image quality. The class balance of CelebA is 42% male and 58% female. We want to mimic the settings for cognitive experiments in order to be able compare machine and human perceptions. To this end we define five primary facial regions consistent with cue-driven human perception [111, 150], and index them accordingly: 1. Eyebrows, 2. Eyes, 3. Nose, 4. Mouth, 5. Chin.

To do this we extract 68 landmarks using the DLib landmarking tool provided in OpenCV [11]. To increase the robustness of local explanations, we consider an important region within the context of its surroundings. This includes analyses of a feature (e.g. only the nose), the horizontal extension of the feature (e.g., the nose and cheeks, from ear to ear) and distinct combinations of the prior. Furthermore, symmetrical and axis-distributed information encoding is tested for by systematically removing horizontal and vertical image data. 67 total variations of the CelebA dataset are generated. Descriptions of all datasets can be found in Tables 3.2 and 3.3.

Table 3.2: Samples from each of the occlusion datasets used in analysis (Part 1). We generate 67 total variations of the original image set.


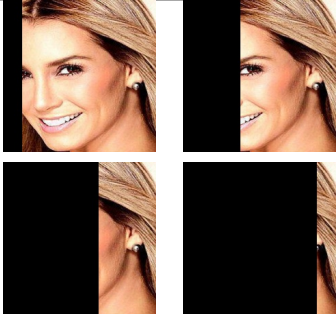






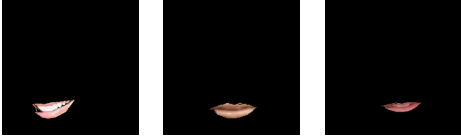

Dataset	Occlusion	Examples
Right_to_left <i>Segments 1 - 7</i>	The segments successively occlude increasing vertical image percentages in increments of 1/7 the image size, moving from right to left.	
Left_to_Right <i>Segments 1 - 7</i>	The segments successively occlude increasing vertical image percentages in increments of 1/7 the image size, moving from left to right.	
Top_to_bottom <i>Segments 1 - 5</i>	The segments consecutively reveal horizontal strips surrounding facial regions, moving from top to bottom.	
Bottom_to_top <i>Segments 1 - 5</i>	The segments consecutively reveal horizontal strips surrounding facial regions, moving from bottom to top.	
Permutation_blackout_pairs <i>Segments 1 - 10</i>	The segments remove every possible unique combination of two distinct facial regions, along with their immediate horizontal surroundings.	

Table 3.3: Samples from each of the occlusion datasets used in analysis (Part 2). We generate 67 total variations of the original image set.

Dataset	Occlusion	Examples
Permutation_blackout_triples <i>Segments 1 - 10</i>	The segments remove every possible unique combination of three distinct facial regions, along with their immediate horizontal surroundings.	
Just_region <i>Segments 1 - 5</i>	Each of the five segments retains only the horizontal strip containing a single facial region.	
Region_blackout <i>Segments 1 - 5</i>	Each of the five segments removes only one horizontally extended facial region.	
Just_region_contoured <i>Segments 1 - 5</i>	Each of the five segments display only the contoured facial region without any surrounding facial information (e.g. the eyebrows absent the forehead.)	
Region_blackout_contoured <i>Segments 1 - 5</i>	Each of the five segments removes one contoured facial region (e.g. occluding the nose while preserving the cheeks.)	

3.4 Results

We conducted experiments on both low (32x32) and high (224x224) resolution versions of the CelebA dataset. The significant results from each are reviewed in the following sections.

3.4.1 Low Resolution Image Data

ResNet-50 performs gender prediction on low-resolution, aligned, facial images with 99.99% accuracy. We generate a q -list, a sorted descending list of changes in testing accuracies, corresponding to when the full facial image is visible compared to when the region given by the map m is occluded (the first direction in our method). The q -list quantitatively characterizes the relationship between the input perturbation defined by deleting an isolated region and the model’s performance.

The list indicates that the ranking of regions, in order of greatest to least effect on classification accuracy, is: Nose, Mouth, Eyebrows, Eyes, Chin. Testing the model’s ability to predict the gender of only an individual region (the second direction in our method) yields an interesting polarity. The ranking, from most to least significant, is: Eyes, Eyebrows, Mouth, Chin, Nose. These two lists illustrate the difference between a region’s configural importance (its predictive effect when missing) and its featural importance (its intraclass similarity). For example, the nose is the most configurally important and least featurally important region, implying that it enhances the effect of other facial elements on performance and is more powerful when combined with other regions than when used alone for gender prediction.

The low-resolution model relies heavily on the mouth for discrimination. When this region is occluded along with any one or two others, it produces lower classification accuracies than any other combinations of two or three regions. Images with two regions occluded, where one is the mouth, are classified with an average accuracy of only 49%, while the average accuracy of images with any two regions except for

the mouth occluded is 73.5%. As long as the mouth is preserved, we can remove up to three facial regions and still outperform 80% of the permutation blackouts that remove only two. Metrics demonstrate over a 26% increase in accuracy when providing lower (nose and below) as opposed to upper (strictly above the nose) facial data. These behaviors show that the model prioritizes lower facial data when evaluated on low-resolution images.

3.4.2 High Resolution Image Data

With high-resolution data, ResNet-50 once again reaches near-perfect (99.97%) accuracy for gender prediction. However, the introduction of high-resolution images causes the model to shift focus from the lower to the upper face. The q -list generated from individual region occlusions reflects the expected change in high-priority regions, with the order from most to least being: Eyes, Nose, Eyebrows, Mouth, Chin. Interestingly, the ranking of region intraclass similarities remains entirely unchanged from the 32x32 case: Eyes, Eyebrows, Mouth, Chin, Nose. This suggests that while the featural contribution of each region is consistent between high and low resolution, the model’s attention shifts towards the upper face when processing high-resolution images.

3.4.3 Generalizable Behaviors

Several trends uniformly hold across both high and low-resolution data:

1. **Vertical vs. Horizontal:** Classification accuracy decreases consistently when information is removed vertically, but more aggressively when deleted horizontally by landmark. Vertical cross-sections of the face are generally only informative when at least 57% of the face is shown, with accuracy steadily decreasing in average increments of 5.5% with each occlusion before this threshold. In contrast, removing horizontal cross-sections consecutively causes stronger decreases in accuracy (an average of 15%) and does not result in the same uniform decrease seen with vertical removals.

2. **Spatial Distribution of Information:** Performance changes symmetrically when occluding vertical regions of the face from right to left or left to right. This property is not observed when deletion regions shift horizontally from bottom to top or from top to bottom, with each horizontal removal corresponding to a unique landmarked region and causing a distinct change in accuracy. The potency of a vertical cross-section is given by the density of facial features within it, while the importance of a horizontal component is determined by the q -list ranking of the contained region.
3. **Classification Bias:** Classifying images with three or more upper facial regions occluded results in low precision (around 0.4) and high recall (around 0.9), indicating a strong model bias toward the male class when information is removed. When classifying the lower vs. upper face, ResNet-50 switches from a very high to a very low recall, with the appearance of information contained in the nose/cheeks shifting the balance of class predictions and resulting in significantly more positively-labeled (male) samples. The upper face is predicted female with extreme bias, while the lower face is frequently categorized as male.
4. **Highly Influential Regions:** In both low and high-resolution situations, one region is heavily prioritized for discrimination (the mouth in the former case, eyes in the latter). By identifying this feature, model decisions can be reconstructed with good accuracy (an average minimum threshold of 39% accuracy with the worst-performing combination of this region and one other, and an average maximum of 74% accuracy with its best-performing combination).

These trends describe high-impact factors on the decision processes of ResNet-50, with potential applications in data compression and feature selection. For example, by eliminating the mouth and chin regions, which are of low priority to the high-resolution model, we can discard nearly 40% of the image data and still maintain 98% classification accuracy. Similarly, training with only the extended horizontal eye region visible, which on average covers only 18% of the face, we can reach 77%

accuracy, significantly outperforming any previous facial isolations.

Compared to existing local interpretability techniques that often require impractically large numbers of input perturbations, the proposed method simultaneously finds highly influential decision regions, explicitly measures the model’s sensitivity to each region, and preserves comparative speed and simplicity. Furthermore, by averaging our metrics over the entirety of the CelebA dataset, we avoid the pitfalls of noisy input and chance correlation found in methods that analyze by example, demonstrating the effectiveness of the proposed featural scheme.

3.4.4 Incorporating Heat Maps

For comparison, we use the Class Activation Mapping (CAM) method [123] to generate saliency maps of sample images. An example is displayed in Figure 3.2. Features that are highly weighted by the model in the final convolution layer, and thus considered influential for discrimination, appear red in the corresponding visualizations.

The simplicity of our method allows for seamless combination with many other techniques. The proposed method integrates with heat-mapping to yield a visual representation of how the model shifts attention based on the available information.

The saliency masks shown in Figure 3.3 summarize where the CNN looks within an image to make predictions. The results support many of our evaluations. The third and fourth images in each row show alternate portions of the face occluded, but with the eyes still visible. In these cases, the model discriminates using the active eye region. When the eyes are occluded, as in the second column (displaying only lower facial data), the model uses information found around the mouth to determine classifications.

While some generalizations can be made from these saliency maps, the first column of Figure 3.3 depicts a prominent issue with them: inconsistency. Each sample produces a unique map, but due to lack of standardization, we can only draw imprecise conclusions about saliency from individual samples. There is also a risk of variation among sample constructions.

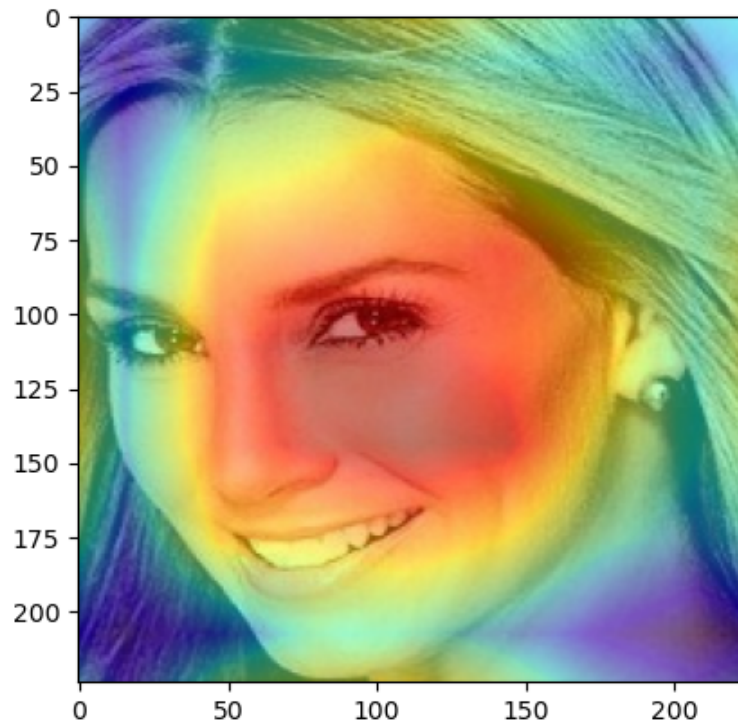


Figure 3.2: *Example heat map generated with CAM on ResNet-50 trained with high resolution images. The red regions are highly weighted by the model (in the final convolution layer) during classification. Best viewed in color.*

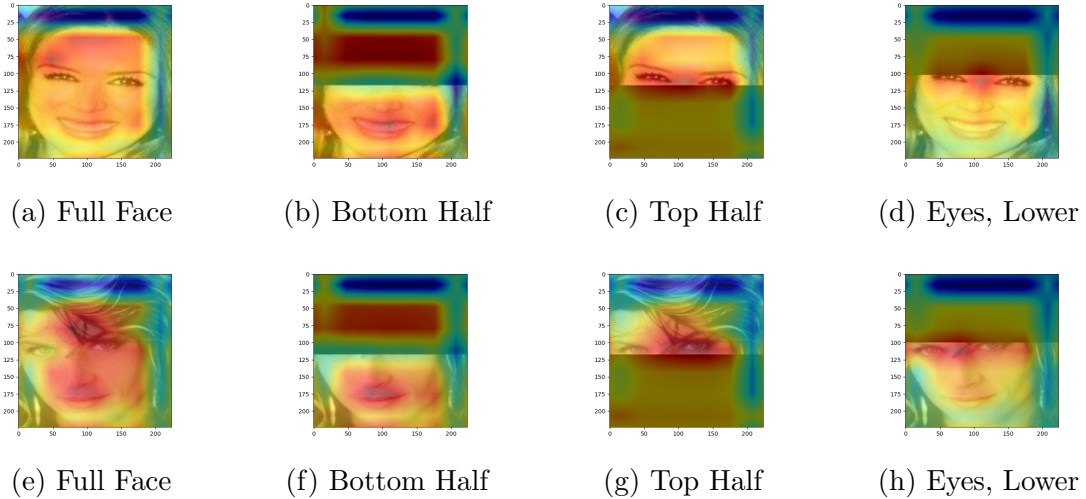


Figure 3.3: A selection of CAM heat maps generated with various occlusions. The maps show the redirection of model attention to new regions when previously prioritized information is no longer available.

To exemplify the issue, see the saliency map depicted in Figure 3.4. This map provides little specificity regarding featural importance or information distribution. The figure is not representative of all potential saliency maps, but this itself indicates a larger problem: Heat maps can only provide local interpretability, and if this one were to be chosen in a random selection and used for explanation, almost no statements could be made about model interpretability.

3.4.5 Discussion: A Comparison with Human Perception

The development of deep neural networks has led to near-human performance in automated gender recognition systems [115]. However, the specific visual cues and their contributions to gender categorization have been a topic of interest in psychology and neuroscience. These cues, which can be classified as either shape or surface cues [112], have been studied to understand their role in human perception of gender. Research in neuroscience supports three key observations regarding gender recognition in humans:

1. Shape cues, both in-plane and three-dimensional, provide significant information about gender that human observers can exploit and are prioritized over

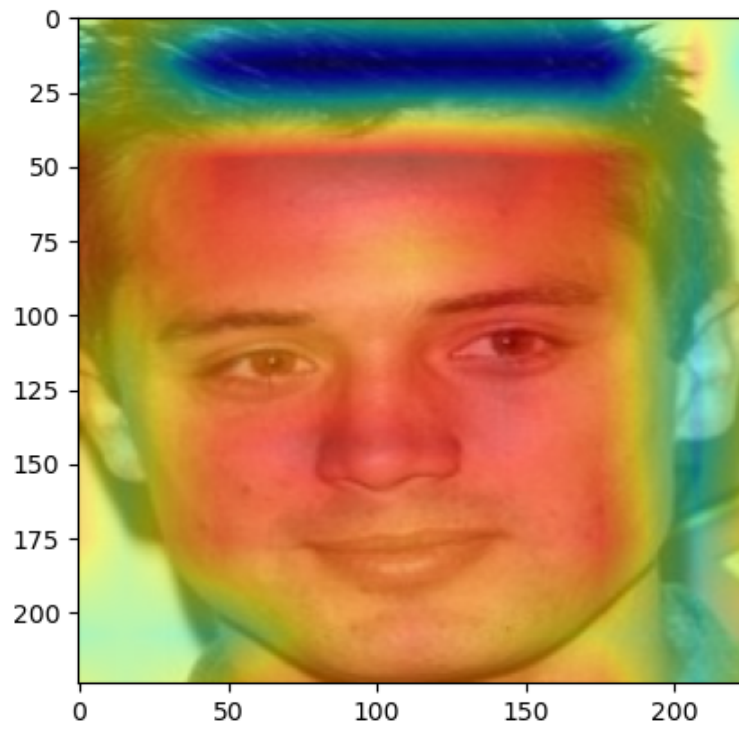


Figure 3.4: A non-informative CAM heat map.

surface cues, forming the basis for distinguishing facial parts [112].

2. Face recognition is typically guided by the diagnosticity of distinct local features such as the eyes and mouth [111].
3. Face perception normally emphasizes holistic or configural aspects of the face over specific features [36, 103, 162], but missing information can disrupt this process and lead to a more feature-based approach.

The majority of research has focused on organizing the contribution of various visual cues, with occlusion being the most popular method of evaluation. Human studies on cognition produce results which, by construction, can be directly compared to our own.

1. **Eyes as Predictors:** Nestor [112] and Russell [132] propose that the luminance difference between the main features - the eyes and mouth - and the rest of the face generates a pattern more typical of female than male faces. In other words, the greater the contrast between the luminance of the eyes and the other regions, the more likely a face will be considered female. The authors claim that the use of cosmetics is highly persuasive in this regard. As almost every candidate in CelebA darkens the eye region, this implies that human predictions of the eye region will be significantly female over male. This hypothesis is congruent with ResNet-50's observed class-skewing of eyes as female, which almost exactly follows the class distribution of genders in CelebA. Dupuis-Roy [29] applied the Bubbles technique to show that the eyes/eyebrow are the most important facial cue for accurate gender discrimination. In summary, both the human and machine are more likely to identify eyes as male or female, and correctly classify gender on that basis.
2. **Determinative Featural Importance:** Nestor et al.[111] attempt to use feature segmentation to diagnose the relative use of distinct local features, such

as the eyes and mouth. Since we conclude that ResNet-50 prioritizes the learning of a specific facial region in horizontal passes, both machine and human processes use a few featurally important sections for gender recognition.

3. **Switching Attention:** Tanaka [150] experiments with facial inversions to suggest that the previously observed bias towards the eye region was attentional, and could be overridden by redirecting participant attention to the mouth. This almost identically parallels the heat maps in the previous section: when eye information is available, it is used deterministically for classification. However, when the eye/brows region is obscured, both the model and the human switch to reliance on the mouth for predictions.
4. **Common Occlusions:** Freud et al. [36, 117] test the effects of masks and sunglasses on human face processing, and remark on the difference between lower and upper facial data availability. They determine that due to lack of test subject exposure, humans recognize faces wearing masks much less accurately than those wearing sunglasses. Our analysis using permutation blackouts is versatile enough to form similar conclusions: ResNet-50 will recognize masked faces *more* accurately than those wearing sunglasses. Our method provides explanations whose complexity can be scaled up or down depending on the need of the target audience, making them widely accessible.

In this chapter we present a novel framework for explaining the decisions of deep learning models targeting gender recognition. We use a methodological occlusion technique to construct machine explanations that closely resemble cataloged human decision justifications. By converging on highly influential facial regions and extracting spatial information encoding, we show both the simplicity of our method and the informativeness of our results in comparison to existing works on interpretability.

The findings in this work suggest the importance of specific visual features for the recognition of particular facial attributes. Results suggest that information from the entire face is not necessary for effectively classifying the presence of an attribute. In

the next chapter we will explore improving facial attribute recognition by intentionally confining the model's attention to regions of the face which contain visual information related to the direct attribute being predicted.

Chapter 4

Parsing Faces with Semantic Segmentation for Improved Facial Attribute Recognition

4.1 Introduction

In 2015, deep learning methods became popular for facial attribute recognition, with the introduction of two large-scale datasets for the problem: CelebA and LFWA [96]. With the introduction of these large scale datasets, the number of deep learning methods for the problem of attribute recognition increased dramatically. Although CelebA allowed for significant progress to be made in the field, it has been shown to have significant label imbalance and noise [86, 87]. We detail issues with available data in more detail in Chapter 5. The label imbalance has resulted in methods which report optimistically high metric scores due to significant skew towards negative sample count. Methods which are trained to perform well on CelebA and LFWA are unlikely to generalize well to unseen images.

To address this issue, we propose a joint learning architecture in which attribute recognition is combined with semantic segmentation. We call this architecture AttParseNet. Semantic segmentation is the problem of classifying every pixel in an image as belonging to one or multiple classes. State of the art methods for semantic segmentation utilize convolutional neural networks (CNNs) and seek to identify a class for each pixel in an image [40, 97, 116]. Semantic segmentation requires that

the trained model learns to localize high-level visual features in diverse input images.

To enable our method we have generated a novel, weak labeling of attribute segments for the CelebA dataset. These label segments are derived from automatically detected facial landmarks. These segments are weak because there is no manual revision of the extracted facial landmark data. Regardless, these attribute segments provide meaningful additional supervision for our classifier.

We utilize semantic segmentation as a means to enforce that AttParseNet learn to localize the facial regions which are associated with each attribute. This task guides model attention to the correct region of the input image. In addition, the predicted segmentation masks are used as the only input to the attribute classification heads at the end of our network architecture. This technique restricts the visual data which are used for making attribute predictions.

As we demonstrated in the previous chapter, the availability of various visual data can significantly alter the model’s predictions for a given facial attribute. The restriction of visual data to regions which are valuable for a given task has a significant positive effect on the model’s performance and generalizability. We also showed that our gender predicting CNN attends to features of the entire face despite singular facial regions being responsible for much of the model’s ability to discern gender. We apply these findings by reducing the model’s attention to finite areas. Our results show that this technique achieves increased generalization to unseen data.

Very few works address the problem of semantic segmentation of faces. Kalayeh et al. propose segmenting the face into parts for improved attribute recognition [61]. This however still differs from our approach, which makes attribute predictions based on segmentation labels. This effectively enforces that the model make predictions about attribute presence based restricted visual data.

To summarize, this work’s contributions include:

- AttParseNet: a multi-task CNN for simultaneous attribute localization and recognition using a weakly labeled training approach.

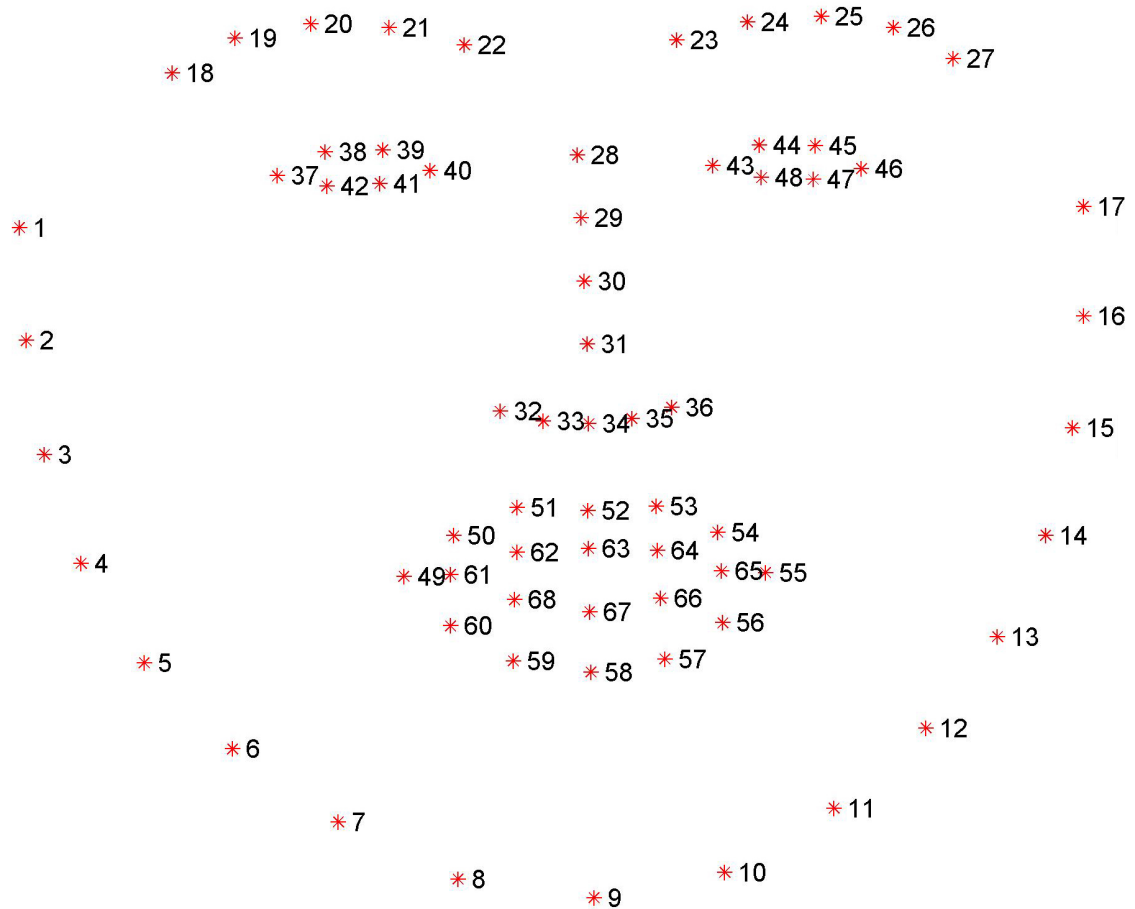


Figure 4.1: Layout of facial landmarks extracted from OpenCV and OpenFace.

- A framework for generating semantic segmentation labels in the context of facial attributes.
- Weak attribute segments for the full CelebA dataset.

4.2 Proposed Methods

The proposed method consists of two main parts: the generation of weak segmentation labels, and the multi-task learning framework. In the following sections we detail each.

4.2.1 Segmentation Label Generation

Teaching the model to localize facial attributes is facilitated by semantic segmentation labels. This form of labeling assigns classes to each pixel in an input image. For the scope of this work, the pixels of each image in CelebA are labeled with the presence or absence of 40 attribute classes. The binary labels for the classes are provided along with the CelebA dataset. Example attribute classes are *smiling*, *wavy hair*, *young*, etc. Our semantic segmentation labels are represented as masks of the same height and width as the input images and a depth of 40 channels (one for each attribute class). Segment masks have a value of 255 in regions where the attribute is present and 0 everywhere else. Hand-labeling this data is expensive and slow, so we opt to automate the process by introducing a weak labeling strategy, which requires no human supervision.

Generation of segment labels begins by extracting a set of facial landmarks from each image in CelebA. Figure 4.1 shows the layout of the 68 facial landmarks that are used. We utilize the OpenCV and OpenFace landmark detectors to extract these points [11, 179].

OpenCV’s facial landmark detector is a pre-trained model that localizes 68 key facial points in an image. It extracts Histogram of Oriented Gradients (HOG) features and applies a cascade of regression trees to iteratively refine the landmark positions, starting from an initial estimate. OpenFace’s landmark detector is a pre-trained set of CNNs which produce response maps without knowledge of other landmark positions. These response maps are produced from expert models at a variety of scales and angles. These response maps are considered jointly to produce point estimates for each landmark.

In both cases, the final output is a set of 68 facial landmark coordinates that accurately identify the jawline, eyebrows, nose, eyes, and mouth. This approach is fast, efficient, and robust to variations in facial pose, expression, and lighting conditions. We first pass all images through the OpenCV detector. Images which



Figure 4.2: Examples of the 10 base regions used to generate weak semantic segmentation labels. These regions are overlaid with the original images for visualization purposes. The segment regions are show in blue and landmark points are red.

do not receive predictions are passed through the OpenFace detector. This technique yields fiducial points for over 99% of CelebA. The remaining images are hand-labeled with landmarks.

The set of collected 68 facial landmarks are used to define a set of base facial regions. The base regions are *below chin*, *chin*, *cheeks*, *mouth*, *above mouth*, *nose*, *eyes*, *eyebrows*, *ears*, and *top of head*. The *chin*, *mouth*, *nose*, *eyes* and *eyebrows* regions are precise because they are defined directly from the 68 landmark points. The remaining 5 regions are established by combining these precise regions with information about facial geometry. For example, the *top of head* region is created by using landmarks from the eyebrows and information about facial geometry, since no landmarks for the forehead are given. We refer to these regions as *rough segments*. Figure 4.2 shows the different regions used in the generation of attribute segments. For example, the *mouth* region is defined as the polygon which has vertices at landmark points {49-60}.

Each of the 40 attribute labels in CelebA are mapped to a set of base regions which contain the visual features necessary for detecting a given attribute. We assume

that the attribute *Smiling* occurs in the base region of *mouth*. Some attributes, such as *No_Beard*, are located in multiple facial regions: *below chin*, *chin*, *cheeks*, and *above mouth*.

Combining this information with the attribute labels in CelebA enables a nearly automatic system for producing segmentation labels for the entire dataset. This method is significant because it provides a framework for producing additional layers of supervision on arbitrary attribute recognition tasks.

To generate the segmentation masks we begin by referencing the 40 attribute labels provided with the CelebA dataset. Each segmentation mask begins as a black image, all pixels set to 0. If the attribute is labeled as present, we fetch the base regions which the attribute is mapped to. The correct polygons are formed and filled with pixel values of 255. Each image in CelebA receives 40 segmentation masks, resulting in over 8 million segmentation masks total.

We consider the segmentation labels to be weak for two reasons: 1) our rule-based method for generating segments relies on automated facial landmark extraction, which may result in imprecise landmarks and regions. For example, misaligned mouth landmarks can lead to incorrect mouth segments. Moreover, the absence of hair landmarks means that all hair-related attributes (e.g., *brown hair*, *wavy hair*) have rough segments derived from the *top of head* region. 2) the physical manifestation of several attributes is unclear, leading to proposed segments that may not provide adequate coverage. For instance, there is ongoing debate in the field of expression and micro-expression recognition regarding the indicators of a smile: whether it is solely the mouth or if other facial deformations around the eyes also play a role. In this work, we assume that the mouth is responsible for mouth-related attributes, potentially missing out on other facial cues.

4.2.2 Attribute Segmentation and Recognition

Once the weakly labeled attribute segments have been generated, the next step is to build a model that learns to recognize attributes. Attribute recognition and segmen-

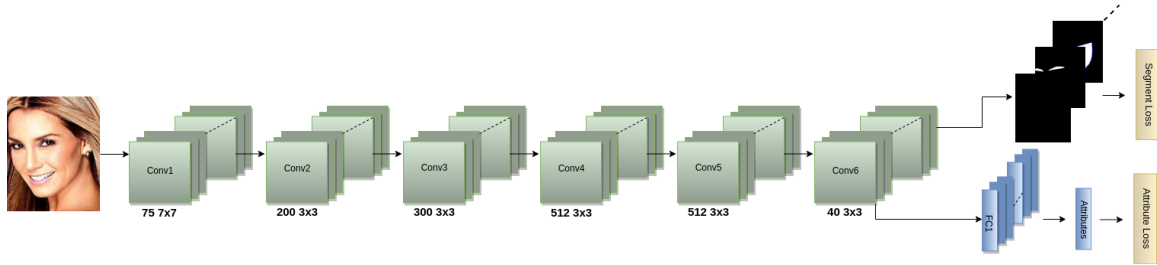


Figure 4.3: AttParseNet, our multi-task learning architecture. Input of an image is provided and is passed through 6 shared convolutional layers. The network outputs segmentation masks and attribute predictions.

tation are learned jointly with a CNN architecture that we call AttParseNet. The task of semantic segmentation is used to improve our model’s attribute recognition accuracy and generalizability.

The proposed multitask attribute segmentation and recognition model is an eight-layer CNN. The architecture for the CNN is shown in Figure 4.3. The model consists of six convolution layers, the first using filters of size 7x7, and the remaining layers using filters of size 3x3. The number of filters in each layer is as follows: 75, 200, 300, 512, 512, and 40. Max pooling is performed after the first convolution layer. After the final convolution layer, the model produces 40 feature maps, each of size 96x76. Each feature map represents a facial attribute and the location in which it occurs.

The generated segmentation labels are the same size as the input images. To compare the masks with the feature maps from the final convolution layer, we perform downsampling to a size of 96x76. The downsampling operation utilizes nearest neighbor interpolation, which assigns each pixel in the downsampled image the value of the nearest pixel in the original image without any averaging or blending. This method is chosen to retain the binary nature of the segmentation masks, as it preserves sharp edges and avoids introducing intermediate values. These downsampled feature maps are then passed into two loss computation modules: the semantic segmentation loss and the facial attribute recognition loss.

The semantic segmentation loss is formulated as mean squared error (MSE) between the output feature maps and segment labels. In this context, MSE loss is

used to measure the reconstruction error between the predicted segmentation masks and the ground truth masks. Given an input image, the model generates C feature maps of size $h \times w$, where C represents the number of attributes or classes in the dataset. These feature maps are then compared with the downsampled ground truth segmentation masks of the same size. The MSE loss for a single image is calculated as follows:

$$\text{MSE} = \frac{1}{C \times h \times w} \sum_{c=1}^C \sum_{i=1}^h \sum_{j=1}^w (y_{c,i,j} - \hat{y}_{c,i,j})^2$$

where $y_{c,i,j}$ represents the value of the downsampled ground truth mask at position (i, j) for attribute c , and $\hat{y}_{c,i,j}$ represents the corresponding predicted value from the model's output feature map. By minimizing the MSE loss during training, the model learns to generate segmentation masks that closely match the ground truth masks, thereby improving its ability to identify the spatial location of visual features necessary for attribute recognition.

For the recognition task, the feature maps are flattened but not concatenated. Each flattened feature map is passed into a separate fully connected layer with a shape of 7296×1 . This results in a final 40-dimensional output prediction. The facial attribute recognition loss is calculated using binary cross-entropy (BCE) loss. BCE loss is commonly used for multi-label classification tasks, where each sample can belong to multiple classes simultaneously. In this case, each facial attribute is treated as an independent binary classification problem. The model predicts the presence or absence of each attribute based on the flattened feature maps.

Given a batch of N images, the BCE loss for the facial attribute recognition task is computed as follows:

$$\text{BCE} = -\frac{1}{N} \sum_{n=1}^N \sum_{c=1}^C [y_{n,c} \log(\hat{y}_{n,c}) + (1 - y_{n,c}) \log(1 - \hat{y}_{n,c})]$$

where $y_{n,c}$ is the ground truth label for attribute c in image n , and $\hat{y}_{n,c}$ is the predicted probability of attribute c being present in image n . The BCE loss penalizes

the model for incorrect predictions and encourages it to learn the correct attribute labels.

For AttParseNet to learn from both the segmentation and recognition tasks in one framework, each task has its own loss function. During training, the total loss for the model is computed as an equally weighted sum of the semantic segmentation loss and the facial attribute recognition loss: $\text{Total Loss} = \text{MSE} + \text{BCE}$

The proposed multitask CNN architecture offers several advantages. By sharing features across both tasks, the model can learn more robust and generalized representations. The weak semantic segmentation task in AttParseNet provides an added level of supervision to the problem of attribute recognition for free. By “free,” we mean that there is a very small amount of human labeling required, and the segments are generated using facial landmark points and weakly labeled using the image-level attribute labels provided with CelebA.

Adding weakly labeled semantic segmentation to AttParseNet forces the model to activate on regions of interest when learning attribute representations, which leads to a more robust and generalizable attribute model. We showcase this in our experiments. **It is important to note that the weak segments are used only at training time and are not needed during testing.**

4.3 Experiments and Results

4.3.1 Datasets

We begin our experimentation on the CelebA dataset. Introduced by Liu et al. [96], CelebA consists of 202,599 celebrity face images, each annotated with 40 binary attributes such as gender, age, hair color, and facial features like smiling or wearing glasses. The dataset was carefully curated to provide a diverse set of images with varying poses, backgrounds, and lighting conditions, making it a challenging and representative dataset for evaluating the performance of attribute recognition models. Example images can be seen in Figure 4.4.



Figure 4.4: Sample images from the CelebA dataset [96].

The images in CelebA were sourced from the Internet and cover a wide range of real-world scenarios. The dataset is divided into three subsets: a training set containing approximately 162,000 images, a validation set with 20,000 images, and a test set with the remaining 20,000 images. This split allows for proper model development, hyperparameter tuning, and unbiased evaluation of the final trained models.

One notable aspect of the CelebA dataset is that it features both cropped and aligned images, as well as full body, unaligned images. In our experiments, we crop the full body, unaligned images with our extracted landmark points. These images are used for training AttParseNet, as this allows the model to learn from more natural and unconstrained facial representations. For training the baseline network, we use the cropped and aligned images, which provide a more focused view of the facial region. Both AttParseNet and the baseline network require input images with dimensions of 218x178 pixels. Therefore, we resize the cropped images and segmentation labels to 218x176 and 96x76 pixels, respectively.

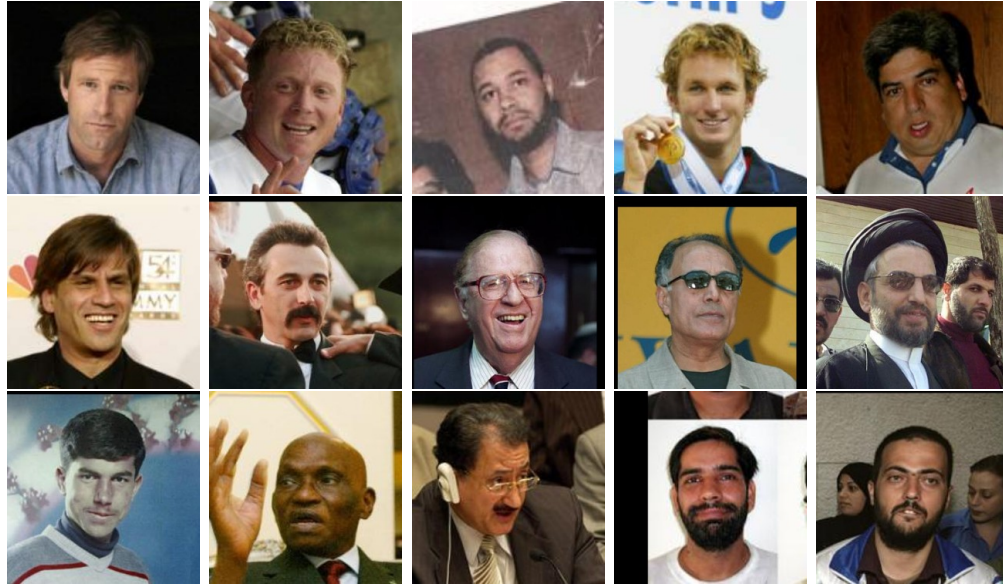


Figure 4.5: Sample images from the LFWA dataset [96].

To further validate the generalization capability of our trained models, we also evaluate their performance on two additional datasets: LFWA [96] and UMD-AED [49]. The LFWA dataset contains 13,232 face images of 5,749 identities, annotated with the same 40 attributes as in CelebA. We report results on the entire LFWA dataset to assess how well our models perform on a different distribution of images. The UMD-AED dataset, despite its modest size of 2,800 facial images, proves to be a powerful tool for exposing vulnerabilities in attribute models. Each image is annotated with a subset of the 40 attributes found in CelebA and LFWA datasets. A unique characteristic of UMD-AED is its balanced distribution of positive and negative samples for each attribute, with 50 instances of each. This equilibrium enables the dataset to effectively uncover the limitations of attribute models. Images of LFWA and UMD-AED can be found in Figures 4.5 and 4.6 respectively.

By evaluating our models on multiple datasets with varying characteristics, we aim to provide a comprehensive analysis of their robustness and ability to generalize to different domains.



Figure 4.6: Sample images from the UMD-AED dataset [49].

4.3.2 AttParseNet Training

AttParseNet is trained exclusively on the unaligned, cropped images from the CelebA training split, without any additional external data. This approach ensures that the model learns to extract relevant features and perform attribute recognition solely based on the information available within the CelebA dataset.

The training process for AttParseNet consists of two stages. In both stages, we use the Adam optimizer [67] with a learning rate of $1E-3$ to update the network weights. The Adam optimizer is chosen for its adaptive learning rate capabilities and efficient convergence properties.

In the first stage, the model is trained for 10 epochs, during which the network weight updates are based solely on the MSE loss computed from the semantic segmentation task. The semantic segmentation task involves predicting a coarse segmentation mask for each facial attribute, providing a high-level understanding of the spatial distribution of attributes. By training the model initially on this task, we allow the network weights to warm up and converge to a reasonable starting point. This stage helps in reducing the MSE loss to a level where the values of the BCE loss

become comparable, facilitating effective joint learning in the subsequent stage.

The second stage of training involves a multi-task learning approach, where AttParseNet is trained simultaneously on both the segmentation and recognition tasks for 22 epochs. During this stage, the MSE loss from the segmentation task and the BCE loss from the attribute recognition task are summed to form the total loss. The BCE loss measures the discrepancy between the predicted attribute probabilities and the ground-truth attribute labels, while the MSE loss ensures that the model maintains its ability to generate accurate segmentation masks. By optimizing both losses jointly, AttParseNet learns to capture the intricate relationships between facial attributes and their spatial localizations. We hypothesize that joint learning reduces the risk of learning spurious correlations between the occurrence of facial attributes and other visual features of images which might co-occur with attributes in the training data.

We emphasize that during the validation and testing phases, we do not use the segment labels. The model’s performance is evaluated solely based on its ability to predict the presence or absence of facial attributes given an input image. This approach aligns with real-world scenarios where ground-truth segmentation masks are not available during inference.

4.3.3 Baseline Model Training

The baseline model is trained using the aligned images from the CelebA training split, ensuring a fair comparison with AttParseNet. By utilizing the aligned dataset, the baseline model benefits from the implicit alignment provided by the image-level attribute labels, which serves as weak segment supervision.

The training process for the baseline model consists of a single stage, where the model is trained solely on the attribute recognition task for 22 epochs. We employ the Adam optimizer [67] with a learning rate of 1E-3 to update the network weights, leveraging its adaptive learning rate capabilities and efficient convergence properties.

During training, the BCE loss is computed based on the discrepancy between the

predicted attribute probabilities and the ground-truth attribute labels. The model learns to capture the relationships between facial attributes and their corresponding visual features present in the aligned images.

It is important to note that the baseline model shares an identical architecture with AttParseNet, including the same hyperparameters. The key difference lies in the absence of the segmentation learning task, which allows us to isolate the effects of learning localization alongside attribute recognition. By focusing exclusively on the attribute recognition task, the baseline model serves as a reference to evaluate the impact of joint learning and localization on AttParseNet’s performance.

In the following sections, we present and analyze the experimental results, comparing the performance of AttParseNet with the baseline model across different evaluation metrics and datasets.

4.3.4 Experimental Setup

We implemented both the proposed AttParseNet architecture and baseline attribute classifier using PyTorch [122]. The CelebA dataset was split into training, validation, and test sets according to the provided partitions. Training was accelerated using two NVIDIA GTX-1080 TI GPUs. To prevent overfitting, we employed early stopping by monitoring the loss on the training and validation sets, stopping training when the losses became comparable.

4.3.5 Results on CelebA

The baseline model, trained on aligned CelebA images without segmentation, achieved an average attribute accuracy of 86% on the aligned test set. In comparison, AttParseNet achieved an average accuracy of 87% on the unaligned test set. While the absolute improvement is small, it is substantial considering the accuracy is averaged over 40 attributes. Figure 4.7 shows the accuracy achieved by both networks for each attribute. Table 4.1 details which attributes specifically benefited from joint training with segmentation.

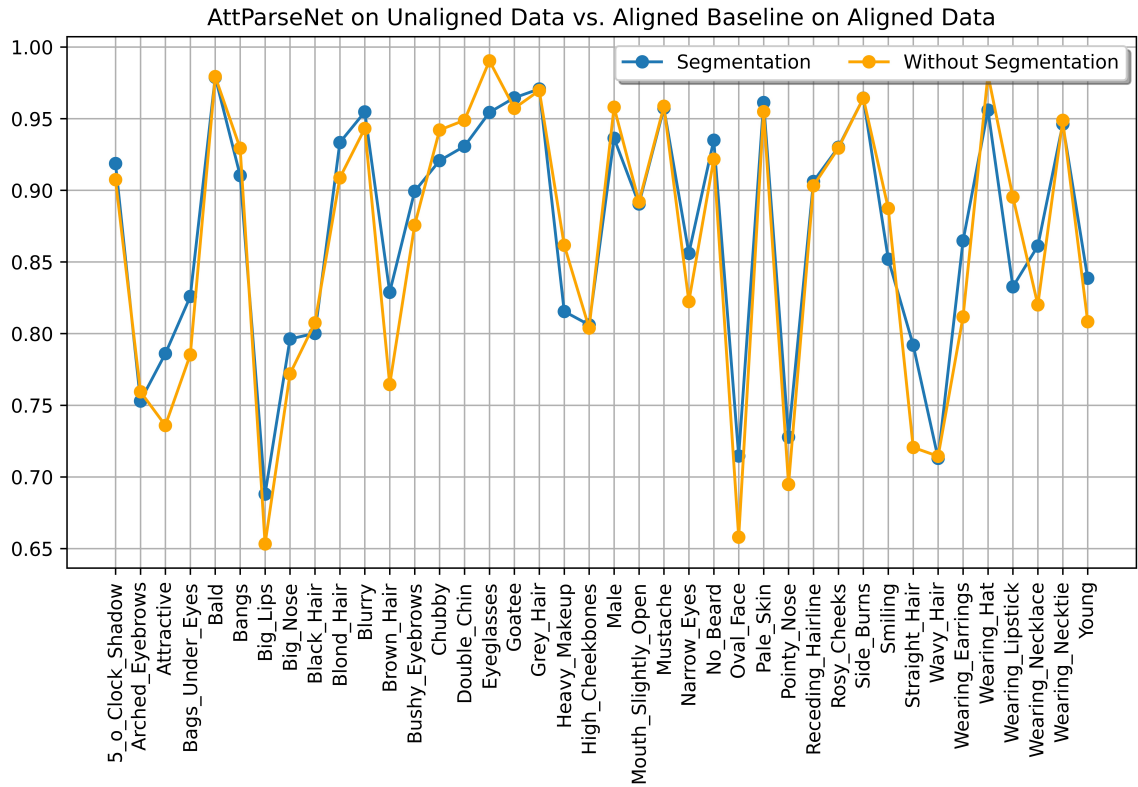


Figure 4.7: Average accuracy achieved on each facial attribute for the proposed architecture and a baseline model. The models are evaluated on the unaligned and aligned data sets respectively. AttParseNet is trained with the weak semantic segmentation task (best viewed in color).

Table 4.1: Attribute improvement comparison table. A ✓ represents an improvement of average accuracy score for AttParseNet over the aligned baseline.

Attribute Name	CelebA	LFWA	UMD-AED
5.o_Clock_Shadow	✓	✓	✓
Arched_Eyebrows	×	✓	✓
Attractive	✓	✓	✓
Bags_Under_Eyes	✓	✓	✓
Bald	×	✓	✓
Bangs	×	✓	✓
Big_Lips	✓	✓	✓
Big_Nose	✓	✓	✓
Black_Hair	×	✓	✓
Blond_Hair	✓	✓	✓
Blurry	✓	✓	✓
Brown_Hair	✓	✓	✓
Bushy_Eyebrows	✓	✓	✓
Chubby	×	✓	×
Double_Chin	×	✓	×
Eyeglasses	×	×	✓
Goatee	✓	✓	×
Gray_Hair	✓	✓	✓
Heavy_Makeup	×	✓	✓
High_Cheekbones	✓	✓	✓
Male	×	✓	✓
Mouth_Slightly_Open	×	✓	✓
Mustache	×	✓	✓
Narrow_Eyes	✓	×	✓
No_Beard	✓	✓	✓
Oval_Face	✓	×	✓
Pale_Skin	✓	×	✓
Pointy_Nose	✓	×	✓
Receding_Hairline	✓	✓	✓
Rosy_Cheeks	✓	✓	✓
Sideburns	✓	✓	×
Smiling	×	✓	✓
Straight_Hair	✓	✓	✓
Wavy_Hair	×	✓	✓
Wearing_Earrings	✓	✓	✓
Wearing_Hat	×	✓	✓
Wearing_Lipstick	×	✓	✓
Wearing_Necklace	✓	✓	×
Wearing_Necktie	×	✓	×
Young	✓	✓	✓
Total Improved	24	35	34

Interestingly, about half of the attributes that did not improve occur in rough face segments (segments constructed from predicted landmark augmentations, see Figure 4.1). These are always on the face periphery. The lack of improvement may be due to our tighter face cropping compared to the aligned CelebA crop, reducing available contextual information for attributes like Mouth Slightly Open, Smiling, and Eyeglasses.

4.3.6 Generalization to LFWA and UMD-AED

To evaluate the generalization of AttParseNet, we tested it on the LFWA and UMD-AED datasets, both relevant for facial attribute recognition. LFWA is widely used, while UMD-AED has nearly perfect attribute label balance. Tests are completed by collecting predictions from AttParseNet and the baseline model for all data in each dataset, then accuracy is calculated based on the ground truth labels.

LFWA is examined first. See Figure 4.8. We note that all attribute classes show increased performance besides *Eyeglasses*, *Narrow Eyes*, *Oval Face*, *Pale Skin*, and *Pointy Nose*. The accuracy differences for the latter three were minor (less than 1%), while some attributes are recognized by AttParseNet as much as 30% more accurately. *Eyeglasses* is an attribute which would benefit from an expanded segmentation mask for additional visual features.

Next, results on the UMD-AED dataset are analyzed. Accuracy for this trial is shown in Figure 4.9. Here we see improvement on all attributes besides *chubby*, *double chin*, *goatee*, *Sideburns*, *wearing necklace* and *wearing necktie*. Each of the attributes that are not improved upon show less than 1% difference of accuracy, on average. This being said, AttParseNet and the aligned baseline are separated by nearly 40% accuracy for some attributes. It is of note that many of the accuracy scores for the aligned baseline classifier are within 5% of 50% accuracy score, suggesting it learned a degenerate majority-class output function for 23 attributes.

These experiments suggest that the joint learning of semantic segmentation

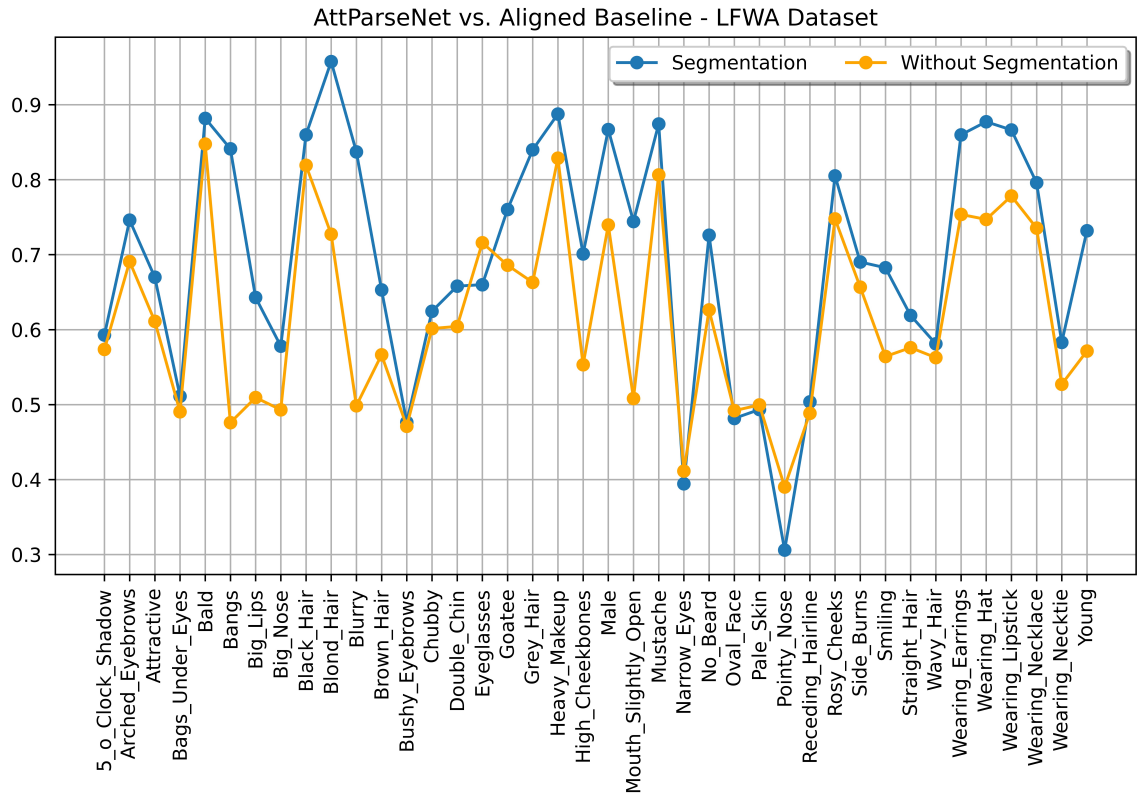


Figure 4.8: Average accuracy achieved on each facial attribute for the proposed architecture and a baseline model. AttParseNet is trained with the weak semantic segmentation task (best viewed in color).

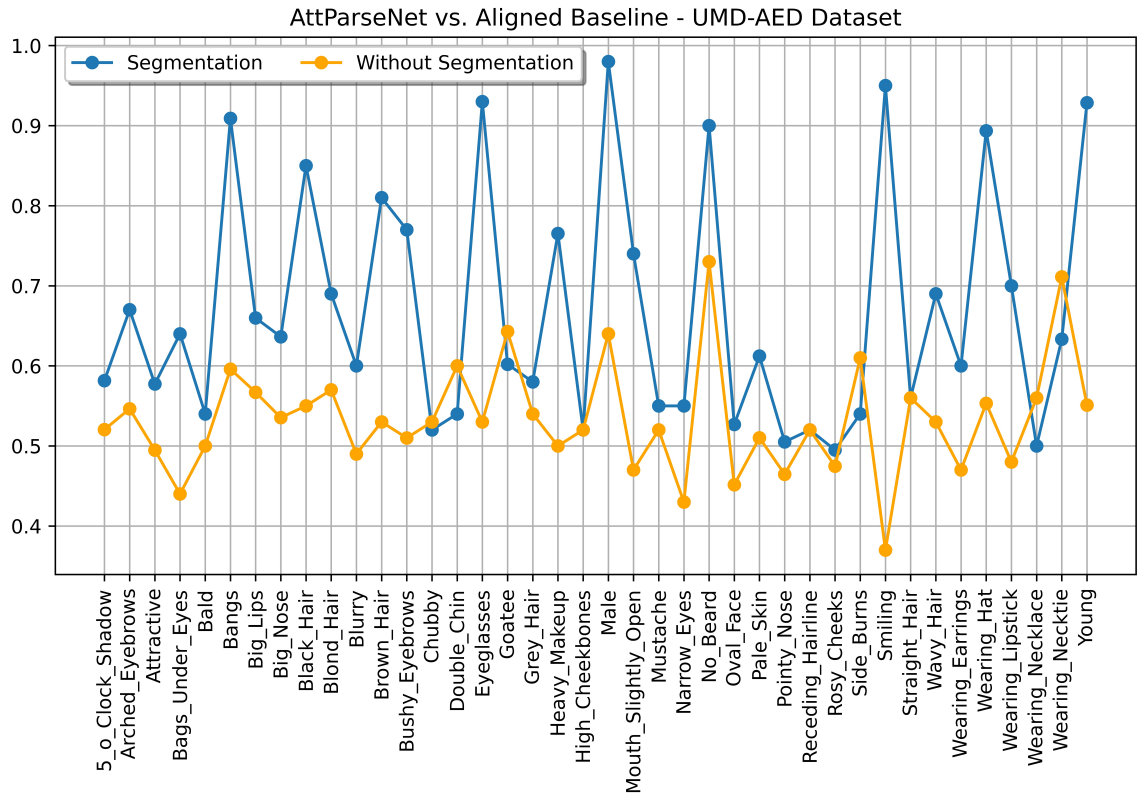


Figure 4.9: Average accuracy achieved on each facial attribute for the proposed architecture and a baseline model. AttParseNet is trained with the weak semantic segmentation task (best viewed in color).

alongside attribute classification greatly improves the performance of a base classifier. The results also suggest that enforcing features for attribute prediction coincide with the visual features which make up the target features reduces the likelihood of over-fitting, even in the presence of very few labels per attribute.

In this chapter we introduce a new method for facial attribute recognition from images, which we call AttParseNet. Our proposed method adds weakly labeled semantic segmentation of attributes as an additional level of supervision in the attribute recognition network. We also introduce a rule-based method for generating weakly labeled facial attribute segments based on landmark points. Using these weakly labeled attribute segments we are able to add a segmentation loss to the facial attribute recognition model, in addition to the attribute recognition loss. Combining these two learning tasks in a single network results in improved facial attribute recognition and generalizability of our model on unseen data.

We demonstrate the effectiveness of our method, comparing AttParseNet with the a baseline model that has the same network architecture, but is trained without the segmentation task. AttParseNet is able to take advantage of weakly labeled segmentation data to better localize and recognize facial attributes, requiring no facial landmarking at test time. In addition, there is some evidence that the semantic segmentation task has a regularization effect on the learned network weights, leading to improved model generalization to unseen data. **We emphasize that the work presented in this chapter required very little hand-labeling and no new data was collected.** Rather, we introduced a rule-based method to create weak semantic segmentation labels for added supervision in the task of attribute recognition.

Despite these improvements, issues with the existing facial attribute datasets are well documented. Architectural changes can only yield limited improvement without also improving the quality of the input data. In the next chapter we discuss known issues with available datasets and propose a novel unsupervised labeling technique.

Chapter 5

Attribute Data and Consensus Subspace Clustering

Parts of this chapter have been previously published. The work is detailed in Appendix A: 2.

5.1 Introduction

In the previous chapter we detail a novel method of attribute recognition that improves on performance metrics for attribute prediction and generalization to unseen image sets. We believe that continued improvement to the field of attribute recognition will require the collection of more high-quality attribute data. In this chapter we discuss the currently available attribute recognition datasets and provide insights into their short-comings. In addition, we present a novel method of unsupervised image labeling and clustering. Although the method achieves limited performance for images with fine-grained visual features, it is an early step toward future systems.

5.1.1 Attribute Data Problems

Despite the relatively widespread use of facial attributes for downstream tasks [20, 54, 62, 73, 167], there are few available datasets which include facial attribute labels. To the best of our knowledge the available datasets are: CelebA, LFWA, UMD-AED, UTKFace, and FairFace [49, 77, 96, 185]. The first three datasets are labeled with the same 40 attributes and contain 202,599, 13,232 and 2,800 samples respectively.

UTKFace and FairFace contain only labels relating to age and race.

One of the most prominent issues with available facial attribute datasets is label imbalance. Many datasets, such as CelebA [96], suffer from significant class imbalance, where certain attributes are underrepresented compared to others.

This imbalance can lead to biased models that perform poorly on minority classes and rely on spurious correlations between unbalanced features [49, 130]. Label imbalance can arise due to various factors, including the inherent distribution of attributes in the population, data collection biases, and annotator biases. For example, attributes like *young* or *attractive* may be over-represented in datasets that primarily consist of celebrity images, as these attributes are more common among celebrities. Similarly, annotator biases can lead to inconsistent or incomplete labeling of certain attributes, especially those that are subjective or ambiguous.

Hand et al. [49] provide attributes which are subjective and frequently mislabeled: *oval face*, *attractive*, *high cheekbones*, and *arched eyebrows*. Their work also shows that *lipstick* is inconsistently labeled. Lingenfelter et al. demonstrate significant racial and gender bias for the *big Nose* and *big lip* attributes [86, 87]. The same group also shows that nearly 7% of CelebA samples are simultaneously labeled with contradicting features. An example is a sample labeled as demonstrating both *wavy hair* and *bald*. Their analysis suggests that as many as 25% of CelebA’s attribute classes demonstrate issues which call into question their utility for downstream tasks.

To mitigate the effects of label imbalance, researchers have proposed various techniques, such as data augmentation, oversampling, and loss weighting. However, these methods often fail to address the underlying causes of imbalance and may introduce additional biases or artifacts in the training data.

The choice of attributes and the quality of annotations in facial attribute datasets can also pose challenges for attribute recognition systems. Many datasets rely on a fixed set of attributes that may not be exhaustive or relevant for all applications. For example, attributes like *attractive* or *heavy makeup* are subjective and may not be useful for tasks like identity verification or demographic analysis.

These issues motivate our next contribution, a method of unsupervised label assignment for image datasets. We avoid label biases by constructing a feature extractor which is entirely self-supervised. This technique results in extracted features correlating to the most important trends of the images seen at train time. The extracted features are clustered into groups which represent data driven visual attributes.

5.2 Unsupervised Image Labeling

Supervised learning has achieved remarkable success in various domains, including biometrics and computer vision. However, the performance of supervised methods heavily relies on the availability of large-scale, labeled datasets, which can be time-consuming, expensive, and sometimes infeasible to obtain. To address this problem, unsupervised learning methods, particularly clustering, have gained significant attention as they aim to discover meaningful patterns and structures in unlabeled data.

Traditional clustering methods, such as k-means [101], Gaussian Mixture Models (GMMs) [22, 106], and spectral clustering [114], have been applied in biometrics and computer vision. However, they often struggle to effectively group visual data. The reasons for this are two-fold. First, the image feature space is high-dimensional. This is problematic due to the curse of dimensionality [65], where the performance deteriorates as the dimensionality increases. Second, the visual features spatially correlate to one another. Clustering methods expect a flattened input, which removes the important spatial organization of the input data.

We posit that a mixture of traditional methods and deep learning has the potential to improve performance of unsupervised labeling of data. Recent advances in deep learning have been leveraged to overcome limitations of clustering on image datasets, with techniques such as autoencoders [55] and variational autoencoders (VAEs) [68] showing promising results in learning compact and meaningful representations of high-dimensional data without the supervision of classification labels.

The learned feature space from autoencoder architectures can be further reduced with matrix factorization techniques, such as Non-negative Matrix Factoriza-

tion (NMF), and a consensus of clustering results can be used as a powerful technique to enhance the robustness and stability of clustering results.

Motivated by these advancements, we propose a novel multi-step clustering approach called Consensus Subspace Clustering (CSC). CSC aims to reduce the dimensionality of input data while carefully selecting the most informative features for grouping samples into meaningful clusters. By leveraging deep learning, matrix factorization, and consensus clustering techniques, CSC captures complex patterns and relationships in high-dimensional biometric and visual data.

The main contributions of this chapter are:

- CSC utilizes a convolutional autoencoder and NMF to capture spatial relationships, identify informative features and flatten the input data.
- The flattened features are then passed into a variational autoencoder (VAE) to extract multiple representations of the flattened data.
- Consensus clustering is applied to combine clustering results from different subspaces, enhancing the stability and reliability of final cluster assignments.
- Experimental results demonstrate CSC’s competitive performance compared to state-of-the-art clustering methods in unsupervised pseudo-labeling tasks for biometric and computer vision applications.

The remainder of this chapter is organized as follows. Section 5.3 describes the proposed CSC method in detail. Section 5.4 presents the experimental setup and discusses the experimental results.

5.3 Methodology

The CSC pipeline consists of four core modules, as shown in Figure 5.1. The first module extracts features from input images using an autoencoder. The second module removes noise and unimportant features by detecting meta-features with Non-negative

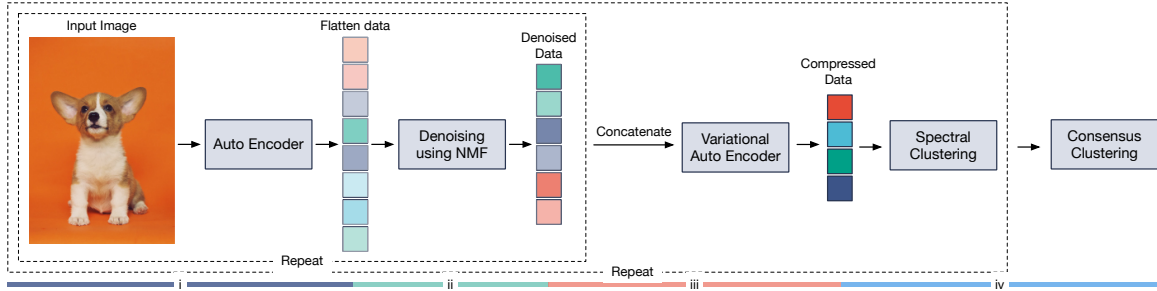


Figure 5.1: Overview of the proposed CSC pipeline. The method consists of four main modules: i) a flattening module using an autoencoder to extract features from input images, ii) a denoising module using NMF to remove unimportant features, iii) a compression module using VAE to generate a low-dimensional representation of denoised features and iv) a clustering module using spectral clustering to cluster images from their compressed representations.

Matrix Factorization (NMF) and inspecting reconstruction errors. CSC only retains features that significantly contribute to the reconstruction error, as these likely differentiate classes. These two modules are repeated to generate multiple denoised versions of the input. The third module is a Variational Autoencoder (VAE) that projects the denoised features into multiple lower-dimensional representations. The fourth module applies spectral clustering to these low-dimensional representations. All four modules are repeated to generate multiple cluster assignments per image. Finally, an ensemble approach determines the final cluster assignment for each image based on the assignments from each representation. The following subsections detail each module.

5.3.1 Feature Extraction

We scale pixel values in each image from 0 to 1 using min-max normalization. A 1-layer convolutional autoencoder then extracts 500 features from each normalized image via the bottleneck layer (Figure 5.2). Optimizing this model to generate a good, compact representation requires identifying the significant visual features. We expect that the learned feature space contains some noise, which we filter out with NMF.

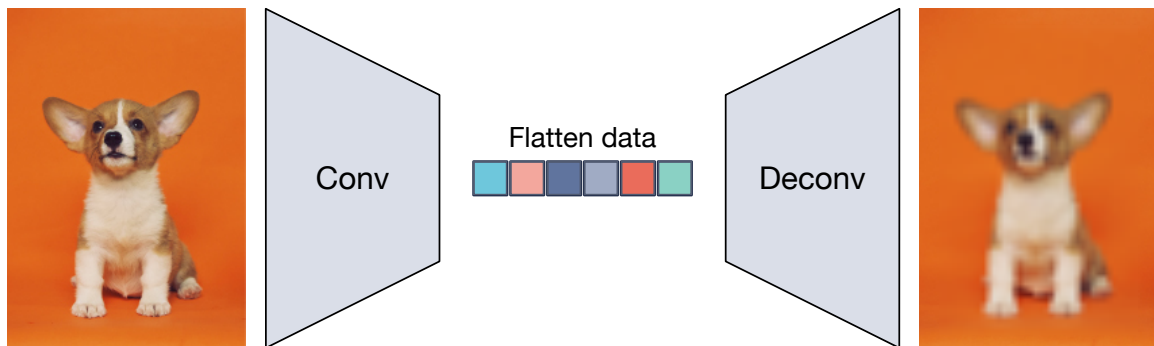


Figure 5.2: Feature extraction using our autoencoder. A 1-layer autoencoder is used to extract features from input images. The representation generated by the autoencoder has 500 dimensions.

5.3.2 Denoising Module

We expect that only a subset of the extracted features are useful for clustering images. Therefore, we filter out features unlikely to play a major role, using the workflow in Figure 5.3 based on 1-factor NMF. This module is represented by the following equation:

$$V_{m \times n} = W_{m \times k} \times H_{k \times n} + E_{m \times n}$$

In our system the latent vector from the flattening module is a vector V with dimensions $m \times n$, where m is the number of images and n is the dimensionality of the latent vector. NMF decomposes V into two matrices W and H which have dimensionality $m \times k$ and $k \times n$. Here, k represents the factor of the NMF model. The factors produced by NMF represent the most dominant trends in the input vector. E is a matrix representing the error between the original vector and the reconstructed vector.

Setting the number of factors to $k = 1$ makes fitting the model difficult for features that significantly differ between clusters - the most valuable features for clustering. By attempting to reconstruct the original matrix V , we can select the most important clustering features based on those with the highest reconstruction error [133, 163]. We sort features by their absolute error and remove the 50% with the

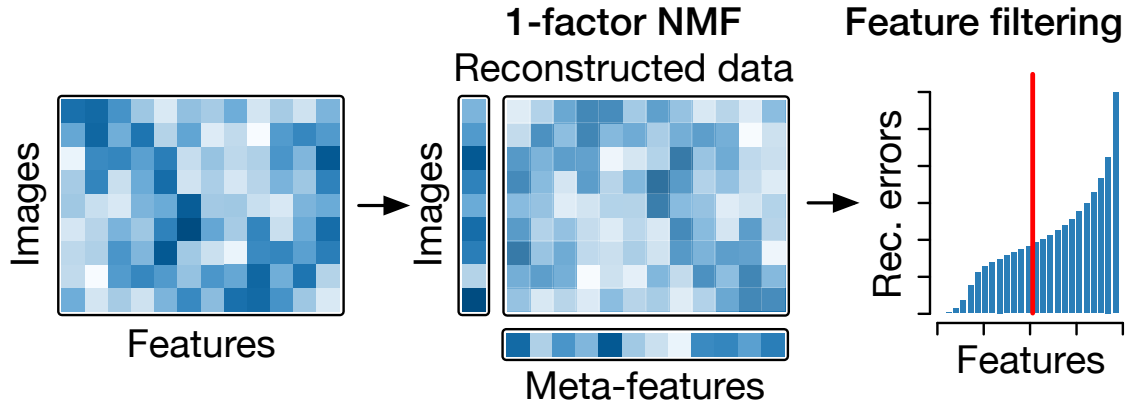


Figure 5.3: Denoising extracted features from input images using NMF. The original data matrix is decomposed into two vectors representing images and their features in 1-dimensional latent space. The error of the reconstructed data using these two vectors is used to rank each feature. Only 50% of features that have the largest error are kept for the next steps.

lowest error. Since the feature extraction and denoising modules are non-deterministic and sensitive to random factors, we repeat them ten times to obtain different denoised data versions, concatenating the results for the next step.

5.3.3 Variational Autoencoder

The previous step has removed insignificant features from the original extracted features, but the dimensions of the remaining features are still too large (2,500 features) to perform clustering efficiently. Hence, a VAE is applied to compress the significant features into a lower dimension (Figure 5.4).

The VAE architecture is similar to that of a standard autoencoder. However, rather than attempting to encode each input sample into fixed floating point features, the VAE encodes features into two vectors. These vectors represent the mean and standard deviation of a Normal distribution. This distribution is sampled to extract the latent vector. This technique results in a latent space which more smoothly transitions between classes than a traditional autoencoder.

VAEs are, however, prone to overfitting [146]. Therefore, instead of using one decoder as in a standard VAE, we use multiple decoders in our implementation to

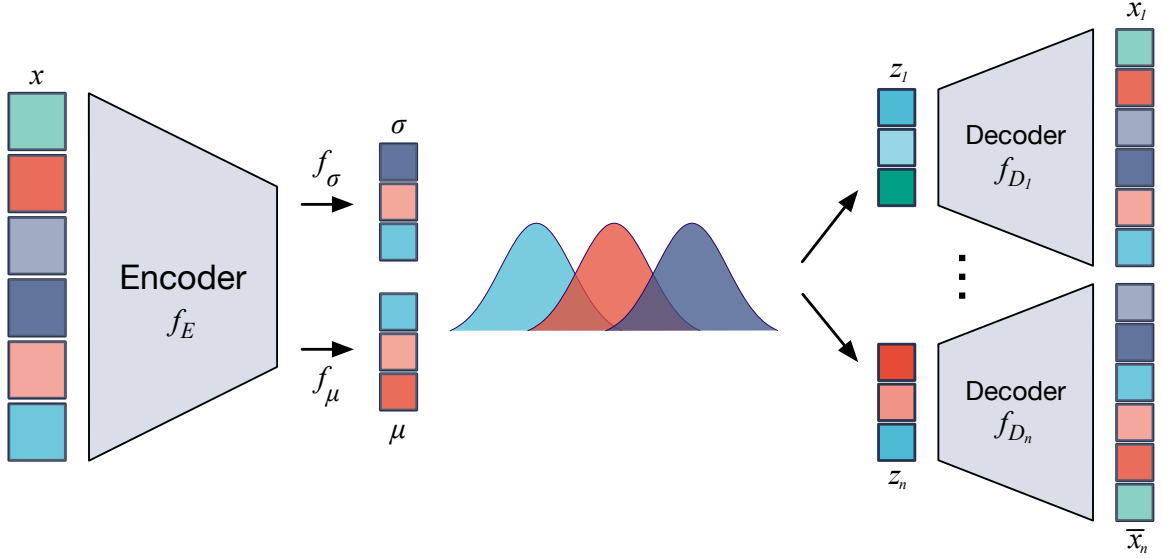


Figure 5.4: Compressing images using a VAE. Denoised images are compressed into multiple representations using a VAE. Multiple representations are obtained from one image. This is accomplished by adding different noise into the latent space and the use of multiple decoders to reconstruct the image. The representations of each image are used for clustering.

ensure that the encoder learns the generalized presentation of the input. At the end of this module we obtain three compact representations for each image by repeatedly sampling from the latent space. Output representations are gathered into 3 groups by sample number (i.e. group 1 contains representations from the first sample of an image). These groups are referred to as subspaces.

5.3.4 Basic Subspace clustering

Spectral clustering is performed on each subspace representation to form pseudo labels for the input data (i.e., each cluster represents a class). We use spectral clustering rather than k-means to better capture non-linear relationships among images.

In our pipeline, we use the K-Means adaptation of spectral clustering, proposed by Ng et al. [114], to generate pseudo labels for input images. The clustering procedure first computes the similarity matrix for all samples to use as the input graph. It then computes the symmetric and normalized Laplacian matrix (L^{sym}). Then, the

K largest eigenvectors for L^{sym} , are computed and normalized to unit length. The eigenvectors are then used to make up the columns of a matrix. Finally, the algorithm uses K-means clustering to segment the subspace into K clusters.

To select the optimal number of clusters, we run the algorithm with a different number of clusters and select the clusters that give us the best ratio r of between-sum-of-squares and total-sum-of-squares by cluster. Since the input data can be large, for each number of clusters, we sample the input multiple times and perform clustering to obtain multiple r . We take the average of all r for each k and select the optimal number of clusters K such that r is maximized.

5.3.5 Consensus Clustering

We repeat the clustering pipeline 10 times to obtain multiple cluster assignments for each image. To generate the final cluster assignment for each image, we adopt an ensemble clustering strategy called weighted-based meta-clustering (wMetaC) [127, 159].

wMetaC uses voting from each cluster assignment to determine the final clusters. First, an image-image similarity matrix is computed, with each value representing the likelihood of two images being clustered together. Next, each image is assigned a weight by summing all pairs it appears in. These similarity matrices form a cluster-cluster similarity matrix. Finally, hierarchical clustering on this matrix selects the final clusters.

5.4 Experiments and Results

To evaluate our proposed method, we compare CSC with several existing clustering methods on two different handwritten digit datasets and one general object classification dataset. Baseline methods included in our comparison are k-means, Deep Cluster [154], and Deep k-means [108]. The datasets used for experimentation are MNIST [82], USPS [57], and CIFAR-10 [71]. Widely used performance metrics are computed to compare CSC to baseline techniques and state-of-the-art methods.

5.4.1 Datasets

The datasets that we select for evaluation are USPS [57], MNIST [82], and CIFAR-10 [71]. Each of these collections are relatively small and contain low-resolution images (32x32 pixels or less). The MNIST dataset contains a total of 70,000 images of size 28x28 (60,000 images for training and 10,000 images for testing). MNIST is relatively balanced with each of the 10 classes representing close to 10% of the total population. The group with most representation makes up 11.25% and the group with least representation makes up 9%. USPS contains a total of 11,000 images with of size 16x16. Both datasets have 10 classes, which correspond with the integers ranging from 0 to 9. Each image depicts a hand-written digit. USPS is mostly balanced with the largest group representing 17% and the smallest group representing 8%. The CIFAR-10 dataset contains total of 60,000 images of size 32x32x3 (50,000 images for training and 10,000 images for testing). This dataset is balanced, with 6000 images per class. CIFAR-10 provides a much more challenging task due to significantly larger feature space and diverse class labels: airplane, automobile, bird, cat, deer, dog, frog, horse, ship and truck.

5.4.2 Methods for Comparison

Effective evaluation of CSC is achieved via comparison to state-of-the-art methods in the field. In addition, we select k-means as a baseline model. Images are flattened before being passed to k-means. K-means is run with 10 cluster centers for a maximum of 1000 iterations or until convergence. We run k-means 20 times on each dataset and select the run with best results for comparison. The selected state-of-the-art methods are Deep Cluster [154] and Deep k-means [108]. Results shown in Table 5.1 are those reported in each publication.

5.4.3 Metrics

We use Accuracy (ACC) and Normalized Mutual Information (NMI) as metrics to evaluate performance of each method. Accuracy and NMI metrics are used to be

Table 5.1: Performance of K-means, Deep Cluster, Deep K-means, and CSC on MNIST, USPS and CIFAR-10 datasets.

Method	MNIST		USPS		CIFAR-10	
	ACC	NMI	ACC	NMI	ACC	NMI
K-means	0.58	0.49	0.48	0.42	0.14	0.12
Deep Cluster	0.86	0.83	0.67	0.69	—	—
Deep K-means	0.84	0.80	0.76	0.78	—	—
CSC No Flatten	0.85	0.79	0.83	0.78	0.12	0.08
CSC No Filter	0.83	0.76	0.84	0.79	0.14	0.10
CSC No Voting	0.82	0.77	0.82	0.76	0.14	0.10
CSC	0.86	0.81	0.83	0.79	0.15	0.11

consistent with the evaluations in the original descriptions of corresponding methods included in the comparison. The metrics are calculated as follow:

$$ACC = \max_m \frac{\sum_{n=1}^N \mathbf{1}(l_i = m(c_i))}{N}$$

where $\mathbf{1}(\cdot)$ is an indicator function, l_i is the true label, c_i is the label assigned by the clustering method and $m(\cdot)$ denotes all possible one-to-one mappings between clusters.

$$NMI = \frac{I(\mathbf{l}, \mathbf{c})}{(H(\mathbf{l}) + H(\mathbf{c}))/2}$$

where \mathbf{l} denotes the ground truth labels, \mathbf{c} is the cluster assignments, $I(\cdot)$ is the mutual information metric, and $H(\cdot)$ is the entropy.

5.4.4 Results

Table 5.1 shows the Accuracy and NMI for CSC and comparison methods on the MNIST, USPS and CIFAR-10 datasets. On the MNIST task, CSC far exceeds performance of the baseline and outperforms the other methods in accuracy. Deep Cluster reports slightly better NMI for MNIST and Deep K-means outperforms CSC in both metrics on the USPS dataset. In the case of Deep Cluster, the margin of difference is very slight and shows that CSC is competitive with state-of-the-art on this task. Regarding Deep k-means, we believe that the architecture is better suited for the smaller feature space found in USPS. Each image in this dataset contains a total of only 256

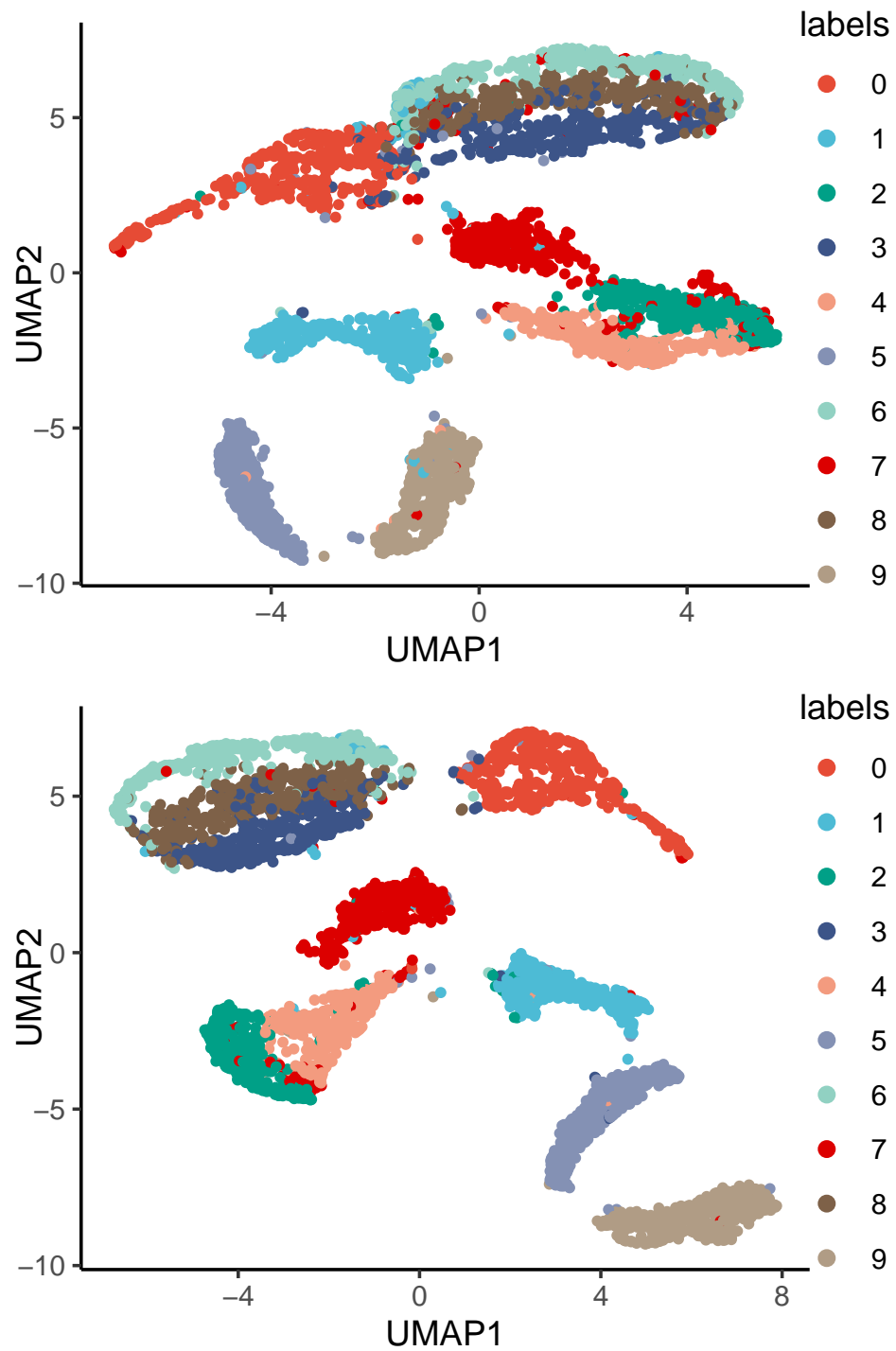


Figure 5.5: UMAP [105] visualizations of the raw USPS dataset (top) and the USPS dataset after being transformed by CSC (bottom). Each colored dot represents an input sample.

features. To reinforce this claim, we point to the method’s decreased performance on the larger MNIST and CIFAR-10 datasets. We note that the authors of Deep Cluster and Deep K-means did not evaluate their methods on the CIFAR-10 dataset.

Complete analysis of CSC requires an understanding of how each component in the pipeline effects the end performance of the model. Referencing the latter half of Table 5.1, removing the flattening module reports the least change out of all modules. However, flattening appears to become more important as the complexity of the dataset increases. Next, the filtering module is particularly important for MNIST, but less important for USPS. This is likely because the samples in USPS are mostly separated before being processed by the VAE, see Figure 5.5. Finally, voting or consensus clustering is very important for stability of clustering results. In our trials without voting, results can be extremely variable.

In this chapter, we have introduced a novel method for providing pseudo labels on arbitrary image data, which we call CSC. To the best of our knowledge we are the first to present a deep clustering method which removes inconsequential features from input data and learns multiple representations of the data to reinforce the robustness of selected cluster labels. Our experimentation shows that our work is competitive with, and in some cases, exceeds the state-of-the-art for deep clustering of image data.

Although this method contributes to the advancement of the field of unsupervised label assignment, we do not find that it is meaningful for the extraction of facial attributes. Our results suggest that CSC struggles to model fine-grained visual features. This is likely due to the simplicity of the convolution autoencoder used by the flattening module.

CSC is in essence a method for extracting arbitrary visual attributes which represent various classes in the input data. This suggests the question of which facial attributes separate highly similar individuals. Detection of such attributes will require a classifier capable of recognizing fine grained details of input images. In the next chapter we detail a novel dataset intended to further study which visual features which are most important for determining identity amongst similar individuals.

Chapter 6

DoppelVer: A Benchmark for Face Verification

Parts of this chapter have been previously published. The work is detailed in Appendix A: 3.

6.1 Introduction

The task of face recognition has received considerable attention from computer vision and pattern recognition researchers in the past 20 years. This is because face identification has significant utility in the fields of biometrics, visual search, and socially assistive technologies [4, 70]. Additionally, compute equipment capable of running increasingly powerful algorithms has become relatively cheap and widely available. Face recognition technologies have significant impact on society with a market share of \$5.69 billion worldwide in 2023 and a projected \$12.05 billion by 2028 [109].

Face recognition is separated into three well-defined steps: (1) face detection and localization, (2) extraction of features from the detected face, and (3) classification (verification or identification) [70]. The first task is to decide whether or not there are faces in an image. If there are one or more faces, then the system identifies bounding boxes for each face. The feature extraction step generates a feature vector from the localized face. This feature vector should be discriminative enough to separate images of one identity from images of other identities. Lastly, there is the classification step. This is separated into two classes of techniques: identification and verification. In

the identification scenario the system is aware of a finite number of identities and it should learn to match each image sample to one identity class. For the verification task the model is only provided with supervision in the form of a binary label which represents either same or different, and so pairs of images are compared at each step.

One might suggest that the field of face classification is reaching its maturity, citing results on the well-known benchmarks such as LFW, AgeDB, or IJB- $\{A,B,C\}$ [56, 69, 104, 110, 170]. Rather than assuming that the reported metrics are due to the techniques solving the task of visually recognizing faces, we hypothesize that the modern techniques have improved beyond the level of difficulty provided by the current benchmarks. For example, in 2015 Liu et al. published a result of 99.77% accuracy on the LFW benchmark [90]. The dataset’s evaluation protocol contains only 6000 images. This means that for nearly a decade methods have been attempting to show improvements on a method that mis-classifies only 14 images, five of which are known to be incorrectly labeled. In addition, face identification datasets are often collected with a focus on quantity, neglecting other important attributes. These problems provide the motivation for the proposed work.

This chapter introduces a new dataset – *DoppelVer* – consisting of unconstrained face images of doppelgangers – that is, individuals who look very similar and are often mistaken for each other. The purpose of DoppelVer is to challenge current SOTA facial feature extraction and face verification and identification methods. Although a number of datasets have been published to this end in the past decade, many of them are either unavailable or have been nearly solved. DoppelVer offers a specific challenge for modern face recognition methods, specifically the task of differentiating individuals who could pass for each other. To the best of our knowledge DoppelVer is the first dataset to increase face classification difficulty by increasing inter-class similarity rather than decreasing intra-class similarity.

There are a large number of datasets collected and presented for the purpose of facial feature extraction and classification. In Chapter 2 Section 2.5 we describe the major datasets that already exist for the purpose of model evaluation and benchmark-

ing. Each of the databases detailed provide an important contribution to furthering the field of face recognition. These datasets provide unconstrained images and in the cases of [110, 139, 187, 188] the sample pairs vary along specific axis which were not well represented in LFW. As mentioned previously, these datasets focus on selecting positive pairs which are visually dissimilar to one another.

DoppelVer’s goal is to expand on a dimension of challenge which has not yet been addressed. This dimension is that of visual similarity among negative samples. This yet unseen challenge will force methods to extract significantly more fine-grained, prominent features from face images. In order to achieve high performance on DoppelVer, techniques will be required to extract those features which uniquely define a given identity.

Here we detail the highlights of the DoppelVer dataset, which will be expanded upon in the remainder of this work.

- DoppelVer contains 390 unique identities, each with at least one corresponding doppelganger pair.
- We provide the unaltered source images along with cropped, aligned, and centered (CCA) images.
- There is an average of 72 CCA samples per identity, with a minimum of 11 and a maximum of 98.
- For the CCA images we provide two evaluation protocols: doppelganger and **Visual Similarity from Embeddings (ViSE)**. Under the doppelganger protocol negative samples are select images depicting an identity’s doppelganger. The ViSE protocol uses a generalized image embedding model to select negative images that are highly visually similar to the current image sample.
- Both protocols are divided into 10 cross validation splits which are distinct across identities. The doppelganger protocol’s cross validation splits are made up of 14,000 image pairs while ViSE’s splits contain 3,500 image samples.

The remainder of the chapter is organized as follows: Section 6.2 contains a more detailed description of the DoppelVer dataset including data collection, pre-processing, labeling, and the generation of the evaluation protocols. Section 6.3 provides results of our experimentation comparing the performance of SOTA facial recognition pipelines on existing benchmark datasets and DoppelVer.

6.2 Proposed Method

6.2.1 Dataset Collection

In order to construct a dataset for which negative samples are analogous to positive samples it is intuitive to begin by aggregating a list of identities which bare visual similarity to human labelers (i.e. doppelgangers). Doppelganger identity pairs were collected through labeler intuition of similar looking identities and lists of doppelgangers publicly available on the Internet. We present a large list of doppelganger identity combinations, totalling 237 pairs and 390 individuals. For each individual, 100 images were scraped from online sources. The average number of images presented in the dataset for each person is approximately 72 due to pruning of noisy samples and duplicates.

6.2.2 Data Preparation

Data preparation involved two distinct steps: (1) cropping, aligning and centering the images, and (2) hand removal of erroneous samples and duplicate images.

The first step in the data preparation is to reduce the original images into cropped, aligned, and centered images. We crop to remove information which is extraneous to the face recognition task. Alignment and centering are performed as they have been recognized as important for achieving competitive face recognition benchmark performance. Alignment involves rotating the image such that the eyes lie on a horizontal line (i.e. the same y-coordinates). The operation of centering moves the face in the frame of the image such that it appears centrally. Centering is

accomplished by repeating edge pixels along either the horizontal or vertical borders of the image. The cropping operation relies on a bounding box and centering/alignment rely on facial landmarks. We extract the bounding boxes and facial landmarks for images in DoppelVer with the MTCNN detector [182].

While processing the dataset with MTCNN, three cases may occur: (1) MTCNN does not detect a face, (2) MTCNN detects a single face, and (3) MTCNN detects multiple faces. Images where a face is not detected are pruned from the dataset. Although MTCNN returns a detected face in most images, not all detections contain the target identity or a valid face. Each detection is hand-checked for validity during the cleaning phase of pre-processing. When at least one face is detected, MTCNN returns a bounding box for the image along with five facial landmarks. The landmarks provide the detected location of the centers of the eyes, corners of the mouth, and tip of the nose.

Initially we cropped the source images to the bounding boxes predicted by MTCNN, but found that the crop was too tight. These crops often removed valuable information such as the top of head, ears and most of the neck. We expand MTCNN’s detected bounding box width and height by 50%. This produces crops which contain more contextual information. There are cases for which the detected face is near the border of the image, restricting our ability to expand the bounding box. In these cases we simply set the desired bounding box location to the border of the image.

After cropping, we align the images according to the extracted landmark locations. Our alignment rotates the images such that the detected landmark for left and right eyes have the same y-axis coordinate. During the alignment process some image information is lost due to the corners of the image rotating outside of the frame. Following the lead of the CelebA dataset, we reduce the effects of this information loss by performing same padding for any pixels that are lost due to rotation [96].

The last pre-processing step is to center the image so that the center most pixel of the image is within the bounds of the detected face. Centering is performed by computing a landmark which lies at the mid-point between the left and right eye

landmarks. Additional pixels are appended to the horizontal and vertical image borders such that the center of the face is equidistant to each border. The appended pixels are simply duplicates of the pixels which are along the border that needs to be expanded.

We remove unsatisfactory images by hand and by automatic detection. In the case of hand labeling, labelers began with the original image set collected from the internet. Their task was to pass over the images and delete any image which contained erroneous detections (e.g. not depicting the correct identity or images not containing a face). The set of images which had complete labeler agreement was accepted. The set of images which did not have agreement were re-labeled. Any remaining images which the labelers did not reach agreement on were pruned from the dataset. The images which achieved hand label agreement were passed to the automatic detection system.

The automatic detection system works by generating embeddings for each face image in the dataset with the dinov2s model [119]. dinov2s is a general purpose image embedding model, built to capture a discriminative representation of input images without finetuning. The cosine similarity is computed between all combinations of input images' embeddings to determine samples which are highly visually similar. To compute the embeddings and cosine similarities efficiently we utilize the fastdup library [158] from Visual Layer. For any image pair that has exact similarity (i.e. duplicate images), one image from the pair is pruned from the dataset. Next, we return all of the image pairs that are above a threshold of 0.92 similarity. We extract these images pairs and provide them to human labelers to find near duplicate images (i.e. images that have been horizontally flipped, color jittered, cropped slightly differently, etc.), which are removed from the dataset.

6.2.3 Protocol Generation

The DoppelVer dataset contains in total 27,967 carefully curated and processed images. The question that remains is the best way to utilize these images for assessing

and benchmarking feature extraction and face classification methods. To answer this question, we introduce two protocols for evaluation using DoppelVer: doppelganger and ViSE. Figure 6.1 provides example image pairs for each protocol in DoppelVer and Figure 6.2 shows samples from CA-LFW and CP-LFW.

Both protocols are made up of positive and negative image pairs. Positive image pairs in both protocols signify instances where both images depict the same identity. In the doppelganger protocol, negative pairs are made up of one image sample depicting the current target identity and one image sample depicting their doppelganger. In the ViSE protocol the negative pairs contain an image sample depicting an identity which does not generally appear as visually similar to the current identity, but in a one-off case is visually similar. Such similarity often arises due to comparable pose, lighting, hair style, clothing, or image background. After generating a large number of image pairs, we divide the dataset into 10 equally sized splits. Each split is divided such that images of an identity are in only a single split. Identities are divided the same in each protocol (e.g. split 0 of the doppelganger protocol depicts the same identities as split 0 of ViSE).

The doppelganger protocol is generated with our curated list of doppelganger pairs. We create the pair instances in the doppelganger protocol as follows. First, we sample 500 image combinations, without replacement, for every pair of doppelgangers and identities with themselves. After generating all pairs following this criteria we separate the samples into 10 splits based on their identities and pairs such that the same identity never shows up in multiple splits. Approximately 10 percent of the dataset is placed into each split. Finally, from each split we randomly sample 7,000 positive pairs and 7,000 negative pairs. We do this to follow the procedures laid out by LFW. This protocol has a positive label and negative label ratio of exactly 50%. It has a gender distribution of 44.96% males and 55.04% female samples respectively. Identities in each split have a relatively even representation with an average minimum contribution of 4.31%, average maximum contribution of 19.07%, and an average standard deviation between representation of 5.32%. In total the doppelganger protocol

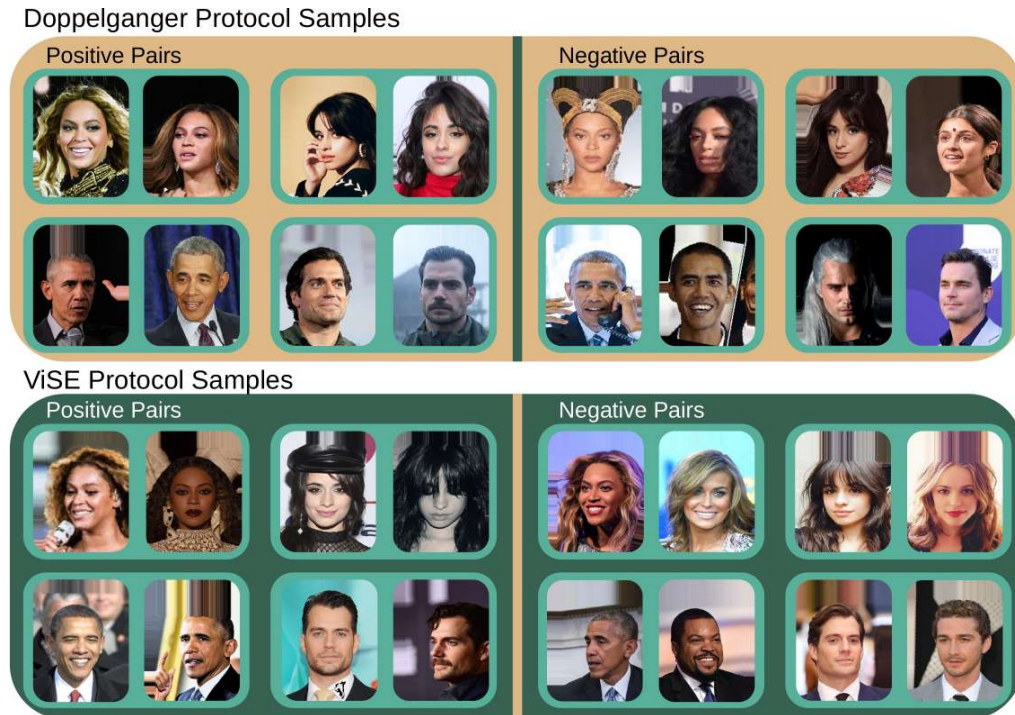


Figure 6.1: Shown above are samples from both protocols of the DoppelVer dataset – doppelganger and ViSE. We note that negative samples from the Doppelganger protocol share facial attributes while the image pairs in ViSE frequently share factors external to the face such as pose, clothing, and background.



Figure 6.2: The upper portion of this figure presents samples from the CA-LFW dataset and the lower portion contains samples from CP-LFW. The CA-LFW samples showcase differences in age while CP-LFW’s images showcase differences in pose.

has 140,000 sample pairs.

To generate the ViSE protocol we use a similar approach to the one described in the automatic detection of unsatisfactory images. We begin by generating embeddings for each image in the dataset with the dinov2s model. Next, we compute the cosine similarity between images which do not come from the same identity. We retain all image pairs that have a similarity greater than 0.80. We have found that this form of mining hard pairs image by image rather than individual by individual results in significantly more visual similarity between image pairs. Using the same identities in each split as the doppelganger protocol, we break the protocol into 10 splits with unique identities in each split. This protocol has a positive label and negative label ratio of exactly 50%. It has a gender distribution of 40.36% male and 59.64% female. Identities in each split have a relatively even representation with an average minimum contribution of 2.29%, average maximum contribution of 17.61%, and an average standard deviation between representation of 3.6%. This protocol has 35,000 verification pairs.

6.2.4 Intended Use

The DoppelVer dataset is intended to provide a new challenge for the research community developing methods in the area of facial recognition. DoppelVer has been designed to act as an evaluation dataset, not a training dataset. In the past decade the most effective methods of facial recognition have utilized large training sets such as CASIA-WebFace, MegaFace, VGGFace2, MS-Celeb-1M [14, 46, 64, 177]. These datasets contain 34.94K, 1.03M, 3.31M, 10M samples respectively. Although an aggregate of visually difficult pairs is attractive for faster convergence time, DoppelVer does not contain enough diversity to effectively and ethically train models.

We provide cross validation splits for both protocols in DoppelVer. The purpose of these splits is two-fold. First, some methods may wish to perform feature extraction prior to face classification. Such extraction methods should pre-train on external sources and infer features for each image in DoppelVer. At evaluation time final-

stage classifiers should be iteratively trained from scratch (using their pre-trained feature extraction methods) on nine splits and evaluated on the tenth. Performance should be recorded as an average across the 9 models. We refer to interaction with the dataset in this way as **View 1**. Second, methods that wish to train on external data and perform only evaluation on DoppelVer should use split 0 for algorithm development and validation of results. The model should not be exposed to data in any of the other nine splits until final evaluation. Use of the dataset in this way is called **View 2**.

Taking motivation from the LFW dataset, we suggest that researchers utilizing **View 1** report estimated mean accuracy (EM ACC) and standard error of the mean (SEM). We define these metrics in the following way:

$$\hat{\mu} = \frac{\sum_{i=1}^9 p_i}{9}, SEM = \frac{\hat{\sigma}}{\sqrt{9}}, \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^9 (p_i - \hat{\mu})^2}{9}}$$

where p_i is the percentage of correct classifications on **View 1** when using the i^{th} split for testing. $\hat{\sigma}$ is the estimate of the standard deviation. As noted by the authors of LFW, it is important that accuracy is computed with parameters and thresholds chosen independently of the test data. Researchers should not simply choose the point on a Precision-Recall curve giving the highest accuracy.

For the methods which utilize **View 2** of DoppelVer, we advocate for the use of accuracy (ACC) and area under the receiver operating characteristic curve (ROC AUC). We elect for the use of ACC and ROC AUC because of the balanced nature of classes in the Doppelganger and ViSE protocols. In addition, the correct classification of true positives is equally important to classification of true negatives.

6.3 Experiments

In this section, we highlight the challenges posed by the DoppelVer dataset as compared to other existing evaluation datasets. We detail the methods used for evaluation, the training data, and the process employed for training and testing.

6.3.1 Evaluation Model

To provide an accurate depiction of the challenge posed by DoppelVer, it is important that we evaluate DoppelVer with SOTA face recognition models. Due to ease of implementation and competitive results we have elected to utilize the techniques described by Wen et al. in SphereFace2 [169]. In particular we train the 20 layer SphereFace Network (SFNet-20), initially proposed in [92], with the following loss functions: COCO, SphereFace, CosFace, ArcFace, and SphereFace2. Following Wen et al., we equip SFNet-20 with batch normalization to facilitate model optimization. A complete implementation for training SFNet-20 with the aforementioned loss functions can be found in the OpenSphere GitHub repository [176].

6.3.2 Training and Evaluation Process

For pre-processing, we crop face images in each dataset with MTCNN, resize images to a size of 112×112 , and normalize each RGB pixel $[0, 255]$ to the range $[-1, 1]$. We trained our models on a single Nvidia Geforce RTX 3090 GPU. Each model is trained for 70,000 batches of size 512. The model weights are updated by stochastic gradient descent with a momentum of 0.9 and weight decay of 0.0005. The initial learning rate of 0.1 is reduced by a factor of 0.1 at batches 40,000; 60,000; and 70,000.

We train our dataset and protocols with VGGFace2, MS-Celeb-1M, and CASIA-WebFace [14, 46, 177]. In each run the VGGFace2 dataset was found to produce the best results on each evaluation dataset. VGGFace2 contains between 80 and 800 images for each identity making it a powerful training dataset for the face verification task. Evaluation of the trained models is performed on LFW, CA-LFW, CP-LFW, AgeDB 30, view 2 of DoppelVer’s doppelganger protocol, and view 2 of DoppelVer’s ViSE protocol. Our measured accuracy and ROC AUC are provided in Tables 6.1 and 6.2 respectively.

Table 6.1: Average accuracy of face verification for the comparison models trained with VGGFace2 and benchmarked on various datasets.

Method	LFW	CA-LFW	CP-LFW	AgeDB	Doppel.	ViSE
COCO [95]	99.08	91.25	88.48	89.40	61.14	52.53
SphereFace [92]	99.58	93.15	91.65	93.53	63.48	57.08
CosFace [160]	99.52	93.03	91.37	93.02	63.29	56.93
ArcFace [24]	99.55	93.40	91.18	92.57	63.28	57.70
SphereFace2 [169]	99.53	93.80	90.83	93.38	61.66	55.41
Average	99.45	92.93	90.70	92.38	62.57	55.93

Table 6.2: Average AUC of face verification for the comparison models trained with VGGFace2 and benchmarked on various datasets.

Method	LFW	CA-LFW	CP-LFW	AgeDB	Doppel.	ViSE
COCO [95]	99.89	96.56	93.57	96.03	65.13	50.53
SphereFace [92]	99.92	97.44	95.50	98.11	68.65	59.41
CosFace [160]	99.91	97.28	95.64	97.86	67.91	58.58
ArcFace [24]	99.89	96.99	95.46	97.53	68.15	59.79
SphereFace2 [169]	99.89	97.55	95.42	98.02	65.43	55.77
Average	99.90	97.16	95.12	97.51	67.05	56.82

6.3.3 Discussion of Results

We are satisfied with the performance achieved by the SOTA methods on the existing benchmark datasets. SOTA performance on the LFW dataset is 99.8% accuracy. Our training of SphereFace achieves an accuracy of 99.58%, mis-classifying just 25 samples. With this result we can be assured that this baseline is competitive with other SOTA methods. The best published results on the other benchmark datasets are 95.87%, 92.08%, and 98.7% accuracy on CA-LFW, CP-LFW, and AgeDB 30 respectively. Regardless of loss function, the baseline networks struggle significantly more with variations in pose than variations in age. CA-LFW and AgeDB appear to present a similar degree of difficulty to the models.

It is clear from our experiments that the doppelganger and ViSE protocols of DoppelVer are much more difficult for the classifiers than the other datasets. Results are better for the doppelganger protocol than the ViSE protocol. This result aligns with intuition. Two identities that are doppelgangers may in general share facial

attributes, but variations in clothing, hair style, lighting, and facial expression are expected when viewing a gallery of images depicting them.

On the other hand, the ViSE protocol contains image pairs which are adversarial in nature. By this we mean that the combinations of samples are those which a deep network is expected to struggle to differentiate. Although we use a different deep convolutional network to select samples which are visually similar than we do for performing facial recognition, one would expect that the visual features which are attended to by deep networks would have some similarity.

We believe that methods which will perform well on the ViSE protocol will need to extract features which are highly specific to the task of facial recognition. In addition, methods will need to not only detect relevant facial features, but discern if the features are prominent/defining to the individual's face.

In this chapter we introduce DoppelVer, a novel evaluation dataset for the tasks of facial feature extraction and face verification. DoppelVer consists of 27,967 carefully curated face images, which are used in two face verification protocols of image pairs: doppelganger and ViSE. We evaluate our methods using several SOTA methods. A near SOTA baseline model is only capable of correctly performing face verification at an accuracy of 62.57% and 55.93% in the doppelganger and ViSE protocols respectively. This indicates that despite impressive results on popular benchmark datasets, there is still work to be done in the field of facial recognition.

Chapter 7

Conclusion

This dissertation has made significant contributions to the field of facial attribute recognition, with the primary objective of enhancing performance and understanding by concentrating on the descriptive visual features of the human face.

In Chapter 3, we introduced an innovative technique for interpreting the visual features used by deep vision models when predicting facial attributes, drawing inspiration from human cognition research. This work offers valuable insights into the perceptions and decision-making processes of these models, providing a foundation for more transparent and interpretable facial attribute recognition systems.

Building upon these findings, Chapter 4 presented an improved method for facial attribute recognition that constrains deep vision models to utilize information only from the spatially relevant regions of the input image for each attribute. By enforcing this spatial prior, the proposed approach leads to better generalization, more robust predictions, and reduced susceptibility to spurious correlations in the training data.

Chapter 5 tackled the common issues found in publicly available facial attribute datasets, such as suboptimal attribute choices that are either not discriminative enough or not well-represented in the data. We introduced a novel unsupervised method to automatically discover the most visually relevant groupings of images.

Finally, Chapter 6 introduced DoppelVer, a new dataset for facial recognition composed of look-alike individuals. DoppelVer presents a unique challenge to existing face recognition systems, revealing their limitations in modeling fine-grained similarity between highly similar classes.

In conclusion, this dissertation has advanced the state-of-the-art in facial attribute recognition through a multifaceted approach encompassing model interpretability, training methodology, and analysis of representations. By focusing on human-describable attributes, this work realizes benefits in transparency, bias detection, data efficiency, and scientific understanding compared to end-to-end deep learning approaches. The work presented here lays the foundation for the development of more reliable, transparent, and equitable facial attribute recognition systems.

7.1 Future Research

Future research directions based on this work are numerous and promising. Building upon the insights gained from our study of the utility of visual data in different spatial locations, one could construct more complex alterations to facial data, such as occluding images with skin-colored regions or swapping facial parts between identities. These augmentations could be used for improving the generalizability of vision models, creating adversarial examples for unsupervised learning, and deepening our understanding of the most valuable facial attributes for downstream recognition tasks.

Another promising avenue is the further improvement of methods for automatically extracting attribute groups based on the content of the training set. Potential approaches include enhancing the CSC framework to better model fine-grained visual features and exploring the definition of attributes through relative similarity. Combining these techniques with CNNs and the visually similar identities from DoppelVer could lead to significant performance improvements in data-driven attribute discovery.

Lastly, future research should explore the insights gained from the DoppelVer dataset. Improving deep vision models to accurately classify visually similar individuals is an important step towards more robust facial recognition systems. Extending the ViSE protocol to include adversarial image pair selection and evaluation on visually dissimilar positive pairs could provide a more comprehensive assessment of facial recognition performance.

References

- [1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: large-scale machine learning on heterogeneous systems, 2015. URL: <http://tensorflow.org/>. Software available from tensorflow.org.
- [2] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018. DOI: 10.1109/ACCESS.2018.2870052.
- [3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, 9525–9536, Montréal, Canada. Curran Associates Inc., 2018.
- [4] Insaf Adjabi, Abdeldjalil Ouahabi, Amir Benzaoui, and Abdelmalik Taleb-Ahmed. Past, present, and future of face recognition: a review. *Electronics*, 9(8), 2020. ISSN: 2079-9292. DOI: 10.3390/electronics9081188. URL: <https://www.mdpi.com/2079-9292/9/8/1188>.
- [5] Séverine Affeldt, Lazhar Labiod, and Mohamed Nadif. Spectral clustering via ensemble deep autoencoder learning (sc-eda). *Pattern Recognition*, 108:107522, 2020. ISSN: 0031-3203. DOI: <https://doi.org/10.1016/j.patcog.2020.107522>. URL: <https://www.sciencedirect.com/science/article/pii/S0031320320303253>.
- [6] Mohamad Alansari, Oussama Abdul Hay, Sajid Javed, Abdulhadi Shoufan, Yahya Zweiri, and Naoufel Werghi. Ghostfacenets: lightweight face recognition model from cheap operations. *IEEE Access*, 11:35429–35446, 2023. DOI: 10.1109/ACCESS.2023.3266068.

- [7] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, and Ying Fu. Partial fc: training 10 million identities on a single machine. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1445–1449, 2021. DOI: 10.1109/ICCVW54120.2021.00166.
- [8] C. BenAbdelkader and P. Griffin. A local region-based approach to gender classification from face images. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pages 52–52, 2005. DOI: 10.1109/CVPR.2005.388.
- [9] Tamara Berg, Alexander Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV Proceedings of the 11th European conference on Computer vision: Part I*, volume 6311, pages 663–676, December 2010. DOI: 10.1007/978-3-642-15549-9_48.
- [10] Thomas Berg and Peter Belhumeur. Tom-vs-pete classifiers and identity-preserving alignment for face verification. *BMVC 2012 - Electronic Proceedings of the British Machine Vision Conference 2012*, January 2012. DOI: 10.5244/C.26.129.
- [11] G. BRADSKI. The opencv library. *Dr. Dobb's Journal of Software Tools*, 25(11):120–125, 2000. URL: <https://cir.nii.ac.jp/crid/1574231875186611456>.
- [12] Garrick Brazil, Xi Yin, and Xiaoming Liu. Illuminating pedestrians via simultaneous detection and segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4960–4969, 2017. DOI: 10.1109/ICCV.2017.530.
- [13] Xingyang Cai, Wengang Zhou, and Houqiang Li. Attribute mining for scalable 3d human action recognition. In *Proceedings of the 23rd ACM International Conference on Multimedia, MM '15*, 1075–1078, New York, NY, USA. Association for Computing Machinery, 2015. ISBN: 9781450334594. DOI: 10.1145/2733373.2806285. URL: <https://doi.org/10.1145/2733373.2806285>.
- [14] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: a dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*, pages 67–74, 2018. DOI: 10.1109/FG.2018.00020.
- [15] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XIV*, 139–156, Berlin, Heidelberg. Springer-Verlag, 2018. ISBN: 978-3-030-01263-2. DOI: 10.1007/978-3-030-01264-9_9. URL: https://doi.org/10.1007/978-3-030-01264-9_9.

- [16] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. DOI: 10.1109/WACV.2018.00097.
- [17] Kai Chen, Guiguang Ding, and Jungong Han. Attribute-based supervised deep learning model for action recognition. *Front. Comput. Sci.*, 11(2):219–229, 2017. ISSN: 2095-2228. DOI: 10.1007/s11704-016-6066-5. URL: <https://doi.org/10.1007/s11704-016-6066-5>.
- [18] Zhiyuan Chen, Annan Li, and Yunhong Wang. A temporal attentive approach for video-based pedestrian attribute recognition. In *Pattern Recognition and Computer Vision: Second Chinese Conference, PRCV 2019, Xi'an, China, November 8–11, 2019, Proceedings, Part II*, 209–220, Xi'an, China. Springer-Verlag, 2019. ISBN: 978-3-030-31722-5. DOI: 10.1007/978-3-030-31723-2_18. URL: https://doi.org/10.1007/978-3-030-31723-2_18.
- [19] Jian Cheng, Peisong Wang, Gang Li, Qinghao Hu, and Hanqing Lu. Recent advances in efficient computation of deep convolutional neural networks. *Frontiers of Information Technology and Electronic Engineering*, 19:64–77, 2018.
- [20] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, Los Alamitos, CA, USA. IEEE Computer Society, 2018. DOI: 10.1109/CVPR.2018.00916. URL: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00916>.
- [21] Garrison W. Cottrell and Janet Metcalfe. Empath: face, emotion, and gender recognition using holons. In *Proceedings of the 3rd International Conference on Neural Information Processing Systems, NIPS'90*, 564–571, Denver, Colorado. Morgan Kaufmann Publishers Inc., 1990. ISBN: 1558601848.
- [22] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977. DOI: <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>. eprint: <https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1977.tb01600.x>. URL: <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x>.
- [23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: a large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. DOI: 10.1109/CVPR.2009.5206848.

- [24] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. Arcface: additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022. DOI: 10.1109/TPAMI.2021.3087709.
- [25] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning to recognize pedestrian attribute, 2015. arXiv: 1501.00901 [cs.CV].
- [26] Yubin Deng, Ping Luo, Chen Change Loy, and Xiaoou Tang. Pedestrian attribute recognition at far distance. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, 789–792, Orlando, Florida, USA. Association for Computing Machinery, 2014. ISBN: 9781450330633. DOI: 10.1145/2647868.2654966. URL: <https://doi.org/10.1145/2647868.2654966>.
- [27] Nat Dilokthanakul, Pedro A. M. Mediano, Marta Garnelo, Matthew C. H. Lee, Hugh Salimbeni, Kai Arulkumaran, and Murray Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders, 2017. arXiv: 1611.02648 [cs.LG].
- [28] Hui Ding, Hao Zhou, Shaohua Kevin Zhou, and Rama Chellappa. A deep cascade network for unaligned face attribute classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [29] Nicolas Dupuis-Roy, Isabelle Fortin, Daniel Fiset, and Frédéric Gosselin. Uncovering gender discrimination cues in a realistic setting. *Journal of Vision*, 9(2):10–10, February 2009. ISSN: 1534-7362. DOI: 10.1167/9.2.10. eprint: https://arvojournals.org/arvo/content/_public/journal/jov/933532/jov-9-2-10.pdf. URL: <https://doi.org/10.1167/9.2.10>.
- [30] Max Ehrlich, Timothy J. Shields, Timur Almaev, and Mohamed R. Amer. Facial attributes classification using multi-task representation learning. In *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 752–760, 2016. DOI: 10.1109/CVPRW.2016.99.
- [31] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.
- [32] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Gläser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2021. DOI: 10.1109/TITS.2020.2972974.
- [33] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, 433–440, Vancouver, British Columbia, Canada. Curran Associates Inc., 2007. ISBN: 9781605603520.

- [34] Fabian Flohr and Darius Gavrila. Pedcut: an iterative framework for pedestrian segmentation combining shape models and multiple data cues. In pages 66.1–66.11, January 2013. ISBN: 1-901725-49-9. DOI: 10.5244/C.27.66.
- [35] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, 2017. DOI: 10.1109/ICCV.2017.371.
- [36] Erez Freud, Andreja Stajduhar, R. Shayna Rosenbaum, Galia Avidan, and Tzvi Ganel. The covid-19 pandemic masks the way people perceive faces. *Scientific Reports*, 10, 2020. DOI: 10.1038/s41598-020-78986-9.
- [37] Yun Fu, Guodong Guo, and Thomas S. Huang. Age synthesis and estimation via faces: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, 2010. DOI: 10.1109/TPAMI.2010.36.
- [38] Wei Gao and Haizhou Ai. Face gender classification on consumer images in a multiethnic environment. In *Proceedings of the Third International Conference on Advances in Biometrics, ICB '09*, 169–178, Alghero, Italy. Springer-Verlag, 2009. ISBN: 9783642017926. DOI: 10.1007/978-3-642-01793-3_18. URL: https://doi.org/10.1007/978-3-642-01793-3_18.
- [39] Xin Geng, Zhi-Hua Zhou, Yu Zhang, Gang Li, and Honghua Dai. Learning from facial aging patterns for automatic age estimation. In *Proceedings of the 14th ACM International Conference on Multimedia, MM '06*, 307–316, Santa Barbara, CA, USA. Association for Computing Machinery, 2006. ISBN: 1595934472. DOI: 10.1145/1180639.1180711. URL: <https://doi.org/10.1145/1180639.1180711>.
- [40] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14*, 580–587, USA. IEEE Computer Society, 2014. ISBN: 9781479951185. DOI: 10.1109/CVPR.2014.81. URL: <https://doi.org/10.1109/CVPR.2014.81>.
- [41] B. A. Golomb, D. T. Lawrence, and T. J. Sejnowski. Sexnet: a neural network identifies sex from human faces. In *Proceedings of the 3rd International Conference on Neural Information Processing Systems, NIPS'90*, page 572, Denver, Colorado. Morgan Kaufmann Publishers Inc., 1990. ISBN: 1558601848.
- [42] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), 2018. ISSN: 0360-0300. DOI: 10.1145/3236009. URL: <https://doi.org/10.1145/3236009>.
- [43] Guodong Guo, Charles R. Dyer, Yun Fu, and Thomas S. Huang. Is gender recognition affected by age? In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 2032–2039, 2009. DOI: 10.1109/ICCVW.2009.5457531.

- [44] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. Improved deep embedded clustering with local structure preservation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1753–1759, 2017. DOI: 10.24963/ijcai.2017/243. URL: <https://doi.org/10.24963/ijcai.2017/243>.
- [45] Xifeng Guo, Xinwang Liu, En Zhu, and Jianping Yin. Deep clustering with convolutional autoencoders. In Derong Liu, Shengli Xie, Yuanqing Li, Dongbin Zhao, and El-Sayed M. El-Alfy, editors, *Neural Information Processing*, pages 373–382, Cham. Springer International Publishing, 2017. ISBN: 978-3-319-70096-0.
- [46] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: a dataset and benchmark for large-scale face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 87–102, Cham. Springer International Publishing, 2016. ISBN: 978-3-319-46487-9.
- [47] Divam Gupta, Ramachandran Ramjee, Nipun Kwatra, and Muthian Sivathanu. Unsupervised clustering using pseudo-semi-supervised learning. In *International Conference on Learning Representations*, 2020. URL: <https://openreview.net/forum?id=rJlnxkSYPS>.
- [48] Manuel Günther, Andras Rozsa, and Terranee E. Boult. Affact: alignment-free facial attribute classification technique. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 90–99, 2017. DOI: 10.1109/BTAS.2017.8272686.
- [49] Emily M. Hand, Carlos Castillo, and Rama Chellappa. Doing the best we can with what we have: multi-label balancing with selective learning for attribute prediction. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI’18/IAAI’18/EAAI’18*, New Orleans, Louisiana, USA. AAAI Press, 2018. ISBN: 978-1-57735-800-8.
- [50] Emily M Hand, Carlos D Castillo, and Rama Chellappa. Predicting facial attributes in video using temporal coherence and motion-attention. In *Applications of Computer Vision (WACV), 2018 IEEE Winter Conference on*, pages 84–92. IEEE, 2018.
- [51] Emily M Hand and Rama Chellappa. Attributes for improved attributes: a multi-task network utilizing implicit and explicit relationships for facial attribute classification. In *AAAI*, pages 4068–4074, 2017.
- [52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. DOI: 10.1109/CVPR.2016.90.

- [53] Keke He, Yanwei Fu, Wuhaio Zhang, Chengjie Wang, Yu-Gang Jiang, Feiyue Huang, and Xiangyang Xue. Harnessing synthesized abstraction images to improve facial attribute recognition. In *IJCAI*, pages 733–740, 2018.
- [54] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: facial attribute editing by only changing what you want, 2018. arXiv: 1711.10678 [cs.CV].
- [55] Geoffrey E. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006.
- [56] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*, 2008.
- [57] J. J. Hull. A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. DOI: 10.1109/34.291440.
- [58] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. Massachusetts Institute of Technology, October 2007. DOI: 10.1109/ICCV.2007.4408988.
- [59] N. Jojic and Y. Caspi. Capturing image structure with probabilistic index maps. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*. Volume 1, pages I–I, June 2004.
- [60] Andrew Kae, Kihyuk Sohn, Honglak Lee, and Erik Learned-Miller. Augmenting crfs with boltzmann machine shape priors for image labeling. In June 2013. DOI: 10.1109/CVPR.2013.263.
- [61] Mahdi M Kalayeh, Boqing Gong, and Mubarak Shah. Improving facial attribute prediction using semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4227–4235. IEEE, July 2017.
- [62] Mahdi M. Kalayeh and Mubarak Shah. On symbiosis of attribute prediction and semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1620–1635, 2021. DOI: 10.1109/TPAMI.2019.2956039.
- [63] Eleni Kanasi, Srinivas Ayilavarapu, and Judith Jones. The aging population: demographics and the biology of aging. *Periodontology 2000*, 72(1):13–18, 2016. DOI: <https://doi.org/10.1111/prd.12126>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/prd.12126>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/prd.12126>.

- [64] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4873–4882, 2016. DOI: 10.1109/CVPR.2016.527.
- [65] Eamonn Keogh and Abdullah Mueen. *Curse of dimensionality*. In *Encyclopedia of Machine Learning*. Claude Sammut and Geoffrey I. Webb, editors. Springer US, Boston, MA, 2010, pages 257–258. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_192. URL: https://doi.org/10.1007/978-0-387-30164-8_192.
- [66] Fahad Shahbaz Khan, Joost van de Weijer, Rao Muhammad Anwer, Andrew D. Bagdanov, Michael Felsberg, and Jorma Laaksonen. Scale coding bag of deep features for human attribute and action recognition. *Machine Vision and Applications*, 29(1):55–71, January 2018.
- [67] Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [68] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [69] Brendan F. Klare, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark Burge, and Anil K. Jain. Pushing the frontiers of unconstrained face detection and recognition: iarpa janus benchmark a. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1939, 2015. DOI: 10.1109/CVPR.2015.7298803.
- [70] Yassin Kortli, Maher Jridi, Ayman Al Falou, and Mohamed Atri. Face recognition systems: a survey. *Sensors*, 20(2), 2020. ISSN: 1424-8220. DOI: 10.3390/s20020342. URL: <https://www.mdpi.com/1424-8220/20/2/342>.
- [71] A. Krizhevsky. Learning multiple layers of features from tiny images. In 2009.
- [72] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [73] Neeraj Kumar, Peter Belhumeur, and Shree Nayar. Facetracer: a search engine for large collections of images with faces. In *European Conference on Computer Vision*, pages 340–353. Springer, 2008.
- [74] Neeraj Kumar, Alexander Berg, Peter Belhumeur, and Shree Nayar. Attribute and simile classifiers for face verification. In *International Conference on Computer Vision*, pages 365–372. IEEE, 2009.
- [75] Neeraj Kumar, Alexander Berg, Peter N Belhumeur, and Shree Nayar. Describable visual attributes for face verification and image search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):1962–1977, 2011.

- [76] Young H Kwon and Niels da Vitoria Lobo. Age classification from facial images. *Computer vision and image understanding*, 74(1):1–21, 1999.
- [77] Kimmo Kärkkäinen and Jungseock Joo. Fairface: face attribute dataset for balanced race, gender, and age, 2019. arXiv: 1908.04913 [cs.CV].
- [78] Andreas Lanitis, Christopher J. Taylor, and Timothy F Cootes. Toward automatic simulation of aging effects on face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.
- [79] S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1265–1278, August 2005.
- [80] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.
- [81] Janet L. Leasher, Rupert R.A. Bourne, Seth R. Flaxman, Jost B. Jonas, Jill Keeffe, Kovin Naidoo, Konrad Pesudovs, Holly Price, Richard A. White, Tien Y. Wong, Serge Resnikoff, Hugh R. Taylor, and on behalf of the Vision Loss Expert Group of the Global Burden of Disease Study. Global Estimates on the Number of People Blind or Visually Impaired by Diabetic Retinopathy: A Meta-analysis From 1990 to 2010. *Diabetes Care*, 39(9):1643–1649, August 2016. ISSN: 0149-5992. DOI: 10.2337/dc15-2171. eprint: <https://diabetesjournals.org/care/article-pdf/39/9/1643/626202/dc152171.pdf>. URL: <https://doi.org/10.2337/dc15-2171>.
- [82] Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- [83] Fengfu Li, Hong Qiao, and Bo Zhang. Discriminatively boosted image clustering with fully convolutional auto-encoders. *Pattern Recognition*, 83:161–173, 2018.
- [84] Zhaoyang Li, Qianqian Wang, Zhiqiang Tao, Quanxue Gao, and Zhaohua Yang. Deep adversarial multi-view clustering network. In pages 2952–2958, August 2019. DOI: 10.24963/ijcai.2019/409.
- [85] J. Lin, H. Yang, D. Chen, M. Zeng, F. Wen, and L. Yuan. Face parsing with roi tanh-warping. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5647–5656, Los Alamitos, CA, USA. IEEE Computer Society, June 2019. DOI: 10.1109/CVPR.2019.00580. URL: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2019.00580>.
- [86] Bryson Lingenfelter, Sara R. Davis, and Emily M. Hand. A quantitative analysis of labeling issues in the celeba dataset. In George Bebis, Bo Li, Angela Yao, Yang Liu, Ye Duan, Manfred Lau, Rajiv Khadka, Ana Crisan, and Remco Chang, editors, *Advances in Visual Computing*, pages 129–141, Cham. Springer International Publishing, 2022. ISBN: 978-3-031-20713-6.

- [87] Bryson Lingenfelter and Emily M. Hand. Improving evaluation of facial attribute prediction models. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–7, 2021. DOI: 10.1109/FG52635.2021.9667077.
- [88] Zachary Chase Lipton. The mythos of model interpretability. *Queue*, 16:31–57, 2018.
- [89] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3337–3344. IEEE, 2011.
- [90] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei, and Chang Huang. Targeting Ultimate Accuracy: Face Recognition via Deep Embedding. *arXiv e-prints*:arXiv:1506.07310, arXiv:1506.07310, June 2015. DOI: 10.48550/arXiv.1506.07310. arXiv: 1506.07310 [cs.CV].
- [91] Sifei Liu, Jianping Shi, Ji Liang, and Ming-Hsuan Yang. Face parsing via recurrent propagation. *CoRR*, abs/1708.01936, 2017. arXiv: 1708.01936. URL: <http://arxiv.org/abs/1708.01936>.
- [92] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphreface: deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017. DOI: 10.1109/CVPR.2017.713.
- [93] Yanxi Liu, Yanghai Tsin, and Wen-Chieh Lin. The promise and perils of near-regular texture. *Int. J. Comput. Vision*, 62(1-2):145–159, April 2005. ISSN: 0920-5691. DOI: 10.1007/s11263-005-4639-0. URL: <http://dx.doi.org/10.1007/s11263-005-4639-0>.
- [94] Yinglu Liu, Hailin Shi, Yue Si, Hao Shen, Xiaobo Wang, and Tao Mei. A high-efficiency framework for constructing large-scale face parsing benchmark. *CoRR*, abs/1905.04830, 2019. arXiv: 1905.04830. URL: <http://arxiv.org/abs/1905.04830>.
- [95] Yu Liu, Hongyang Li, and Xiaogang Wang. Rethinking feature discrimination and polymerization for large-scale recognition. *CoRR*, abs/1710.00870, 2017. arXiv: 1710.00870. URL: <http://arxiv.org/abs/1710.00870>.
- [96] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. DOI: 10.1109/ICCV.2015.425.
- [97] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

- [98] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogério Schmidt Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. *CoRR*, abs/1611.05377, 2016. arXiv: 1611.05377. URL: <http://arxiv.org/abs/1611.05377>.
- [99] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, 4768–4777, Red Hook, NY, USA. Curran Associates Inc., 2017. ISBN: 9781510860964.
- [100] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Hierarchical face parsing via deep learning. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2480–2487, June 2012. ISBN: 978-1-4673-1226-4. DOI: 10.1109/CVPR.2012.6247963.
- [101] James B. MacQueen. Some methods for classification and analysis of multivariate observations. In 1967.
- [102] Aravindh Mahendran and Andrea Vedaldi. Salient deconvolutional networks. In volume 9910, pages 120–135, October 2016. ISBN: 978-3-319-46465-7. DOI: 10.1007/978-3-319-46466-4_8.
- [103] Daphne Maurer, Richard Le Grand, and Catherine Mondloch. The many faces of configural processing. *Trends in cognitive sciences*, 6:255–260, July 2002. DOI: 10.1016/S1364-6613(02)01903-4.
- [104] Brianna Maze, Jocelyn Adams, James A. Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K. Jain, W. Tyler Niggel, Janet Anderson, Jordan Cheney, and Patrick Grother. Iarpa janus benchmark - c: face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165, 2018. DOI: 10.1109/ICB2018.2018.00033.
- [105] Leland McInnes, John Healy, and James Melville. Umap: uniform manifold approximation and projection for dimension reduction, 2020. arXiv: 1802.03426 [stat.ML].
- [106] Geoffrey J. McLachlan, Sharon X. Lee, and Suren I. Rathnayake. Finite mixture models. *Annual Review of Statistics and Its Application*, 6(1):355–378, 2019. DOI: 10.1146/annurev-statistics-031017-100325. eprint: <https://doi.org/10.1146/annurev-statistics-031017-100325>. URL: <https://doi.org/10.1146/annurev-statistics-031017-100325>.
- [107] Tim Miller. Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.*, 267:1–38, 2019.
- [108] Maziar Moradi Fard, Thibaut Thonet, and Eric Gaussier. Deep k-means: jointly clustering with k-means and learning representations. *Pattern Recognition Letters*, 138:185–192, 2020. ISSN: 0167-8655. DOI: 10.1016/j.patrec.2020.07.028. URL: <https://www.sciencedirect.com/science/article/pii/S0167865520302749>.

- [109] MordorIntelligence, July 2023. URL: <https://www.mordorintelligence.com/industry-reports/facial-recognition-market>.
- [110] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, volume 2, page 5, 2017.
- [111] Adrian Nestor and Michael Tarr. The segmental structure of faces and its use in gender recognition. *Journal of vision*, 8:7.1–12, February 2008. DOI: 10.1167/8.7.7.
- [112] Adrian Nestor and Michael J. Tarr. Gender recognition of human faces using color. *Psychological Science*, 19:1242–1246, 2008.
- [113] Casey Newton. Facebook begins using artificial intelligence to describe photos to blind users. *theverge.com*, editor, April 2016. URL: <https://www.theverge.com/2016/4/5/11364914/facebook-automatic-alt-tags-blind-visually-impaired>. [Online; posted 5-April-2016].
- [114] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: analysis and an algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS’01, 849–856, Cambridge, MA, USA. MIT Press, 2001.
- [115] Choon-Boon Ng, Yong Haur Tay, and Bok-Min Goi. Recognizing human gender in computer vision: a survey. In volume 7458, pages 335–346, September 2012. ISBN: 978-3-642-32694-3. DOI: 10.1007/978-3-642-32695-0_31.
- [116] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. *CoRR*, abs/1505.04366, 2015. arXiv: 1505.04366. URL: <http://arxiv.org/abs/1505.04366>.
- [117] Eilidh Noyes, Josh Davis, Nikolay Petrov, Katie Gray, and Kay Ritchie. The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *Royal Society Open Science*, 8, March 2021. DOI: 10.1098/rsos.201169.
- [118] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*:685–694, 2015.
- [119] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *arXiv e-prints*:arXiv:2304.07193, arXiv:2304.07193, April 2023. DOI: 10.48550/arXiv.2304.07193. arXiv: 2304.07193 [cs.CV].

- [120] Devi Parikh and Kristen Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR 2011*, pages 1681–1688, June 2011.
- [121] Devi Parikh and Kristen Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.
- [122] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>.
- [123] Ian Pointer. Class activation mappings in pytorch. URL: <http://snappishproductions.com/blog/2018/01/03/class-activation-mapping-in-pytorch.html> (visited on 09/30/2010).
- [124] Ankit Rajpal, Khushwant Sehra, Rashika Bagri, and Pooja Sikka. Xai-fr: explainable ai-based face recognition using deep neural networks. *Wirel. Pers. Commun.*, 129(1):663–680, 2022. ISSN: 0929-6212. DOI: 10.1007/s11277-022-10127-z. URL: <https://doi.org/10.1007/s11277-022-10127-z>.
- [125] Narayanan Ramanathan and Rama Chellappa. Face verification across age progression. *IEEE Transactions on Image Processing*, 15(11):3349–3361, 2006.
- [126] Narayanan Ramanathan, Rama Chellappa, and Soma Biswas. Computational methods for modeling facial aging: a survey. *Journal of Visual Languages and Computing*, 20(3):131–144, 2009.
- [127] Yazhou Ren, Carlotta Domeniconi, Guoji Zhang, and Guoxian Yu. Weighted-object ensemble clustering: methods and analysis. English. *Knowledge and Information Systems*, 51(2):661–689, September 2017. DOI: 10.1007/s10115-016-0988-y.
- [128] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you?”: explaining the predictions of any classifier. In pages 97–101, February 2016. DOI: 10.18653/v1/N16-3020.
- [129] Andras Rozsa, Manuel Günther, Ethan M Rudd, and Terrance E Boult. Are facial attributes adversarially robust? In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pages 3121–3127. IEEE, 2016.
- [130] Ethan M Rudd, Manuel Günther, and Terrance E Boult. Moon: a mixed objective optimization network for the recognition of facial attributes. In *European Conference on Computer Vision*, pages 19–35. Springer, 2016.

- [131] Olga Russakovsky and Li Fei-Fei. Attribute learning in large-scale datasets. In Kiriakos N. Kutulakos, editor, *European Conference on Computer Vision (ECCV): Trends and Topics in Computer Vision*, pages 1–14, Berlin, Heidelberg. Springer Berlin Heidelberg, 2010.
- [132] Richard Russell. Sex, beauty, and the relative luminance of facial features. *Perception*, 32:1093–107, February 2003. DOI: 10.1068/p5101.
- [133] Farid Saberi-Movahed, Mahdi Eftekhari, and Mohammad Mohtashami. Supervised feature selection by constituting a basis for the original space of features and matrix factorization. *International Journal of Machine Learning and Cybernetics*, 11(7):1405–1421, 2020.
- [134] Pouya Samangouei and Rama Chellappa. Convolutional neural networks for attribute-based active authentication on mobile devices. In *Biometrics Theory, Applications and Systems (BTAS), 2016 IEEE 8th International Conference on*, pages 1–8. IEEE, 2016.
- [135] Pouya Samangouei, Emily Hand, Vishal M Patel, and Rama Chellappa. Active authentication using facial attributes. *Mobile Biometrics*, 3:131, 2017.
- [136] Pouya Samangouei, Vishal M Patel, and Rama Chellappa. Attribute-based continuous user authentication on mobile devices. In *Biometrics Theory, Applications and Systems (BTAS), 2015 IEEE 7th International Conference on*, pages 1–8. IEEE, 2015.
- [137] Walter J Scheirer, Neeraj Kumar, Peter N Belhumeur, and Terrance E Boulton. Multi-attribute spaces: calibration for attribute fusion and similarity search. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2933–2940. IEEE, 2012.
- [138] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128:336–359, 2019.
- [139] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. DOI: 10.1109/WACV.2016.7477558.
- [140] Gaurav Sharma, Frederic Jurie, and Cordelia Schmid. Expanded parts model for human attribute and action recognition in still images. In pages 652–659. IEEE, June 2013. DOI: 10.1109/CVPR.2013.90.
- [141] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR 2011*, pages 801–808. IEEE, June 2011. DOI: 10.1109/CVPR.2011.5995329.
- [142] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

- [143] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Hima Lakkaraju. Fooling lime and shap: adversarial attacks on post hoc explanation methods. In February 2020. DOI: 10.1145/3375627.3375830.
- [144] Brandon M. Smith, Li Zhang, Jonathan Brandt, Zhe Lin, and Jianchao Yang. Exemplar-based face parsing. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3484–3491, 2013. DOI: 10.1109/CVPR.2013.447.
- [145] Jost Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: the all convolutional net, December 2014.
- [146] Harald Steck. Autoencoders that don’t overfit towards the identity. In *NeurIPS*, 2020.
- [147] Yuyin Sun, Liefeng Bo, and Dieter Fox. Attribute based object identification. In *2013 IEEE International Conference on Robotics and Automation*, pages 2096–2103, May 2013.
- [148] David Tahmoush. Applying action attribute class validation to improve human activity recognition, 2015.
- [149] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf. Deepface: closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. DOI: 10.1109/CVPR.2014.220.
- [150] James Tanaka, Martha Kaiser, Simen Hagen, and Lara Pierce. Losing face: impaired discrimination of featural and configural information in the mouth region of an inverted face. *Attention, perception and psychophysics*, 76, January 2014. DOI: 10.3758/s13414-014-0628-0.
- [151] Nathan Thom, Andrew DeBolt, Lyssie Brown, and Emily M. Hand. Doppelver: a benchmark for face verification. In *Advances in Visual Computing: 18th International Symposium, ISVC 2023, Lake Tahoe, NV, USA, October 16–18, 2023, Proceedings, Part I*, 431–444, Berlin, Heidelberg. Springer-Verlag, 2023. ISBN: 978-3-031-47968-7. DOI: 10.1007/978-3-031-47969-4_34. URL: https://doi.org/10.1007/978-3-031-47969-4_34.
- [152] Nathan Thom and Emily M. Hand. *Facial attribute recognition: a survey*. In *Computer Vision: A Reference Guide*. Springer International Publishing, Cham, 2020, pages 1–13. ISBN: 978-3-030-03243-2. DOI: 10.1007/978-3-030-03243-2_815-1. URL: https://doi.org/10.1007/978-3-030-03243-2_815-1.
- [153] Nathan Thom, Hung Nguyen, and Emily M. Hand. Consensus subspace clustering. In *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 391–395, 2021. DOI: 10.1109/ICTAI52525.2021.00064.

- [154] Kai Tian, Shuigeng Zhou, and Jihong Guan. Deepcluster: a general clustering framework based on deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 809–825. Springer, 2017.
- [155] Yu-Ho Tseng and Shau-Shiun Jan. Combination of computer vision detection and segmentation for autonomous driving. In *2018 IEEE/ION Position, Location and Navigation Symposium (PLANS)*, pages 1047–1052, 2018. DOI: 10.1109/PLANS.2018.8373485.
- [156] Daniel A Vaquero, Rogerio S Feris, Duan Tran, Lisa Brown, Arun Hampapur, and Matthew Turk. Attribute-based people search in surveillance environments. In *Applications of Computer Vision (WACV), 2009 Workshop on*, pages 1–8. IEEE, 2009.
- [157] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, R. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. B. Blaschko, D. Weiss, B. Taskar, K. Simonyan, N. Saphra, and S. Mohamed. Understanding objects in detail with fine-grained attributes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3622–3629, June 2014.
- [158] Visual-Layer. Fastdup. <https://github.com/visual-layer/fastdup>, July 2023.
- [159] Shibiao Wan, Junil Kim, and Kyoung Won. Sharp: hyper-fast and accurate processing of single-cell rna-seq data via ensemble random projection. *Genome Research*, 30:gr.254557.119, January 2020. DOI: 10.1101/gr.254557.119.
- [160] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: large margin cosine loss for deep face recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. DOI: 10.1109/CVPR.2018.00552.
- [161] Jing Wang, Yu Cheng, and Rogerio Schmidt Feris. Walk and learn: facial attribute representation learning from egocentric video and contextual data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2295–2304, 2016.
- [162] Ruosi Wang, Jingguang Li, Huizhen Fang, Moqian Tian, and Jia Liu. Individual differences in holistic processing predict face recognition ability. *Psychological science*, 23:169–77, February 2012. DOI: 10.1177/0956797611420575.
- [163] Shiping Wang, Witold Pedrycz, Qingxin Zhu, and William Zhu. Subspace learning for unsupervised feature selection via matrix factorization. *Pattern Recognition*, 48(1):10–19, 2015.
- [164] Wei Wang, Dan Yang, Feiyu Chen, Yunsheng Pang, Sheng Huang, and Yongxin Ge. Clustering with orthogonal autoencoder. *IEEE Access*, 7:62421–62432, 2019.

- [165] Xiaoyang Wang and Qiang Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *2013 IEEE International Conference on Computer Vision*, pages 2120–2127, December 2013.
- [166] Xiaoyang Wang and Qiang Ji. Object recognition with hidden attributes. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI’16*, 3498–3504, New York, New York, USA. AAAI Press, 2016. ISBN: 9781577357704.
- [167] Zhanxiong Wang, Keke He, Yanwei Fu, Rui Feng, Yu-Gang Jiang, and Xiangyang Xue. Multi-task deep neural network for joint face recognition and facial attribute prediction. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR ’17*, 365–374, New York, NY, USA. Association for Computing Machinery, 2017. ISBN: 9781450347013. DOI: 10.1145/3078971.3078973. URL: <https://doi.org/10.1145/3078971.3078973>.
- [168] Jonathan Warrell and Simon Prince. Labelfaces: parsing facial features by multiclass labeling with an epitome prior. In pages 2481–2484, November 2009. DOI: 10.1109/ICIP.2009.5413918.
- [169] Yandong Wen, Weiyang Liu, Adrian Weller, Bhiksha Raj, and Rita Singh. Sphreface2: binary classification is all you need for deep face recognition. *CoRR*, abs/2108.01513, 2021. arXiv: 2108.01513. URL: <https://arxiv.org/abs/2108.01513>.
- [170] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K. Jain, James A. Duncan, Kristen Allen, Jordan Cheney, and Patrick Grother. Iarpa janus benchmark-b face dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–600, 2017. DOI: 10.1109/CVPRW.2017.87.
- [171] Jonathan R. Williford, Brandon B. May, and Jeffrey Byrne. Explainable face recognition. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 248–263, Cham. Springer International Publishing, 2020. ISBN: 978-3-030-58621-8.
- [172] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 529–534. IEEE, 2011.
- [173] Junyuan Xie, Ross B. Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. *CoRR*, abs/1511.06335, 2015. arXiv: 1511.06335. URL: <http://arxiv.org/abs/1511.06335>.
- [174] Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: simultaneous deep learning and clustering. In *international conference on machine learning*, pages 3861–3870. PMLR, 2017.

- [175] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International Conference on Computer Vision*, pages 1331–1338. IEEE, November 2011. DOI: 10.1109/ICCV.2011.6126386.
- [176] ydwen. Opensphere. <https://github.com/ydwen/opensphere>, July 2023.
- [177] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z. Li. Learning face representation from scratch. *CoRR*, abs/1411.7923, 2014. arXiv: 1411.7923. URL: <http://arxiv.org/abs/1411.7923>.
- [178] Kai Yu, Biao Leng, Zhang Zhang, Dangwei Li, and Kaiqi Huang. Weakly-supervised learning of mid-level features for pedestrian attribute recognition and localization. *CoRR*, abs/1611.05603:224–229, 2016.
- [179] Amir Zadeh, Tadas Baltrusaitis, and Louis-Philippe Morency. Deep constrained local models for facial landmark detection. *CoRR*, abs/1611.08657, 2016. arXiv: 1611.08657. URL: <http://arxiv.org/abs/1611.08657>.
- [180] Mohamed H. Zaki and Tarek Sayed. Using automated walking gait analysis for the identification of pedestrian attributes. *Transportation Research Part C: Emerging Technologies*, 48:16–36, 2014.
- [181] M.D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *European Conference on Computer Vision(ECCV)*, 8689:818–833, January 2013.
- [182] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016. DOI: 10.1109/LSP.2016.2603342.
- [183] Ning Zhang, Manohar Paluri, Marc’Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: pose aligned networks for deep attribute modeling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1637–1644, 2014.
- [184] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu. Attribute regularization based human action recognition. *IEEE Transactions on Information Forensics and Security*, 8(10):1600–1609, October 2013. ISSN: 1556-6021. DOI: 10.1109/TIFS.2013.2258152.
- [185] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4352–4360, 2017. DOI: 10.1109/CVPR.2017.463.

- [186] Xin Zhao, Liufang Sang, Guiguang Ding, Yuchen Guo, and Xiaoming Jin. Grouping attribute recognition for pedestrian with joint recurrent learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3177–3183. International Joint Conferences on Artificial Intelligence Organization, July 2018.
- [187] T. Zheng and W. Deng. Cross-pose LFW: A database for studying cross-pose face recognition in unconstrained environments. Technical report 18-01, Beijing University of Posts and Telecommunications, February 2018.
- [188] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age LFW: A database for studying cross-age face recognition in unconstrained environments. *CoRR*, abs/1708.08197, 2017. arXiv: 1708.08197. URL: <http://arxiv.org/abs/1708.08197>.
- [189] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*:2921–2929, 2016.
- [190] Jianqing Zhu, Shengcai Liao, Dong Yi, Zhen Lei, and Stan Z. Li. Multi-label cnn based pedestrian attribute learning for soft biometrics, 2015.
- [191] Xiaofeng Zhu, Heung-Il Suk, Seong-Whan Lee, and Dinggang Shen. Subspace regularized sparse multitask learning for multiclass neurodegenerative disease identification. *IEEE Transactions on Biomedical Engineering*, 63(3):607–618, 2016. DOI: 10.1109/TBME.2015.2466616.
- [192] Xiaofeng Zhu, Heung-Il Suk, and Dinggang Shen. A novel matrix-similarity based loss function for joint regression and classification in ad diagnosis. *NeuroImage*, 100:91–105, 2014. ISSN: 1053-8119. DOI: <https://doi.org/10.1016/j.neuroimage.2014.05.078>. URL: <https://www.sciencedirect.com/science/article/pii/S1053811914004637>.

Appendix A

List of Publications

1. Facial Attribute Recognition: A Survey

Nathan Thom and Emily M. Hand

In Computer Vision: A Reference Guide. Springer International Publishing, Cham, 2020, pages 1–13. [152]

Abstract We present a survey of attribute recognition research in the computer vision community over the past decade. Most of our attention is given to facial attributes, but attributes of objects, pedestrians, and actions are considered as well.

2. Consensus Subspace Clustering

Nathan Thom, Hung Nguyen, and Emily M. Hand

In 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), pages 391–395, 2021. [153]

Abstract One significant challenge in the field of supervised deep learning is the lack of large-scale labeled datasets for many problems. In this paper, we propose Consensus Spectral Clustering (CSC), which leverages the strengths of convolutional autoencoders and spectral clustering to provide pseudo labels for image data. This data can be used as weakly-labeled data for training and evaluating classifiers which require supervision. The primary weaknesses of previous

works lies in their inability to isolate the object of interest in an image and cluster similar images together. We address these issues by denoising input images to remove pixels which do not contain data pertinent to the target. Additionally, we introduce a voting method for label selection to improve the clustering results. Our extensive experimentation on several benchmark datasets demonstrates that the proposed CSC method achieves competitive performance with state-of-the-art methods.

3. DoppelVer: A Benchmark for Face Verification

Nathan Thom, Andrew DeBolt, Lissie Brown, and Emily M. Hand

In Advances in Visual Computing: 18th International Symposium, ISVC 2023, Lake Tahoe, NV, USA, October 16–18, 2023, Proceedings, Part I, 431–444. [151]

Abstract The field of automated face verification has become saturated in recent years, with state-of-the-art methods outperforming humans on all benchmarks. Many researchers would say that face verification is close to being a solved problem. We argue that evaluation datasets are not challenging enough, and that there is still significant room for improvement in automated face verification techniques. This paper introduces the DoppelVer dataset, a challenging face verification dataset consisting of doppelganger pairs. Doppelgangers are pairs of individuals that are extremely visually similar, oftentimes mistaken for one another. With this dataset, we introduce two challenging protocols: doppelganger and Visual Similarity from Embeddings (ViSE). The doppelganger protocol utilizes doppelganger pairs as negative verification samples. The ViSE protocol selects negative pairs by isolating image samples that are very close together in a particular embedding space. In order to demonstrate the challenge that the DoppelVer dataset poses, we evaluate a state-of-the-art face verification method on the dataset. Our experiments demonstrate that the DoppelVer dataset is significantly more challenging than its predecessors, indicating that there is still room for improvement in face verification technology.