TOWARDS INTERPRETABLE FACE RECOGNITION

By

Bangjie Yin

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science – Master of Science

2019

ABSTRACT

TOWARDS INTERPRETABLE FACE RECOGNITION

By

Bangjie Yin

Deep CNNs have been pushing the frontier of visual recognition over past years. Besides recognition accuracy, strong demands in understanding deep CNNs in the research community motivate developments of tools to dissect pre-trained models to visualize how they make predictions. Recent works further push the interpretability in the network learning stage to learn more meaningful representations. In this work, focusing on a specific area of visual recognition, we report our efforts towards interpretable face recognition. We propose a spatial activation diversity loss to learn more structured face representations. By leveraging the structure, we further design a feature activation diversity loss to push the interpretable representations to be discriminative and robust to occlusions. We demonstrate on three face recognition benchmarks that our proposed method is able to achieve the state-of-art face recognition accuracy with easily interpretable face representations.

ACKNOWLEDGEMENTS

This dissertation would not have been made possible without the help of many people.

I am very honored to have Dr. Xiaoming Liu as my advisor. His expectation and encouragement have made me achieve more than I could ever have imagined. The time we spent to discuss experiments, brainstorm, and polish papers has refined my skills in critical thinking, presentation and writing. By setting himself as an example, he has taught me what a good researcher should be like.

I am grateful for my labmates, Yousef Atoum, Xi Yin, Amin Jourabloo, Luan Tran, Garrick Brazil, Yaojie Liu, Joel Stehouwer, Shengjie Zhu, Masa Hu. The valuable comments in paper review, the willingness to help, the encouragement when I am in a bad mood, and the entertainment together have made it a very pleasant journey.

Thanks to my friends at Michigan State University, Tony, Zhongzheng, Bohao, Hieu, Lisheng, Xiaoyan, Ding, Zhiming, for their company to keep my mind refreshed.

Finally, I would like to thank my parents who have taught me to be brave, positive, and kindhearted. Thanks to my wife Jiajia for the long-term being supportive to my career and life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In the era of deep learning, one major focus in the research community has been on designing network architectures and objective functions towards discriminative feature learning He et al. (2016); Iandola et al. (2014); Lin et al. (2017); Wen et al. (2016); Liu et al. (2017b); Tran et al. (2017a). Meanwhile, given its superior even surpassing-human recognition accuracy He et al. (2015); Lu & Tang (2015), there is a strong demand from both researchers and general audiences to interpret its successes and failures Goodfellow et al. (2014); Olah et al. (2018), to understand, improve, and trust its decisions. Increased interests in visualizing CNNs lead to a set of useful tools to dissect their prediction paths to identify the important visual cues Olah et al. (2018). While it is interesting to see the visual evidences for predictions from pre-trained models, what's more interesting is to guide the learning towards better interpretability.

CNNs trained towards discriminative classification may learn filters with wide-spreading attentions – usually hard to interpret for human. Prior work even empirically demonstrate models and human attend to different image areas in visual understanding Das et al. (2017). Without design to harness interpretability, even when filters are observed to actively respond to certain local structure across several images, there is nothing preventing them to simultaneously capture a different structure; and the same structure may activate other filters too. One potential solution to address this issue is to provide annotations to learn locally activated filters and construct a structured representation from bottom-up. However, in practice, this is rarely feasible. Manual annotations are expensive to collect, difficult to define in certain tasks, and sub-optimal compared with end-to-end learned filters.

A desirable solution would keep the end-to-end training pipeline intact and encourage the interpretability with a model-agnostic design. However, in the recent interpretable CNNs Zhang et al. (2017), where filters are trained to represent object parts to make the network representation interpretable, they observe degraded recognition accuracy after introducing interpretability. While

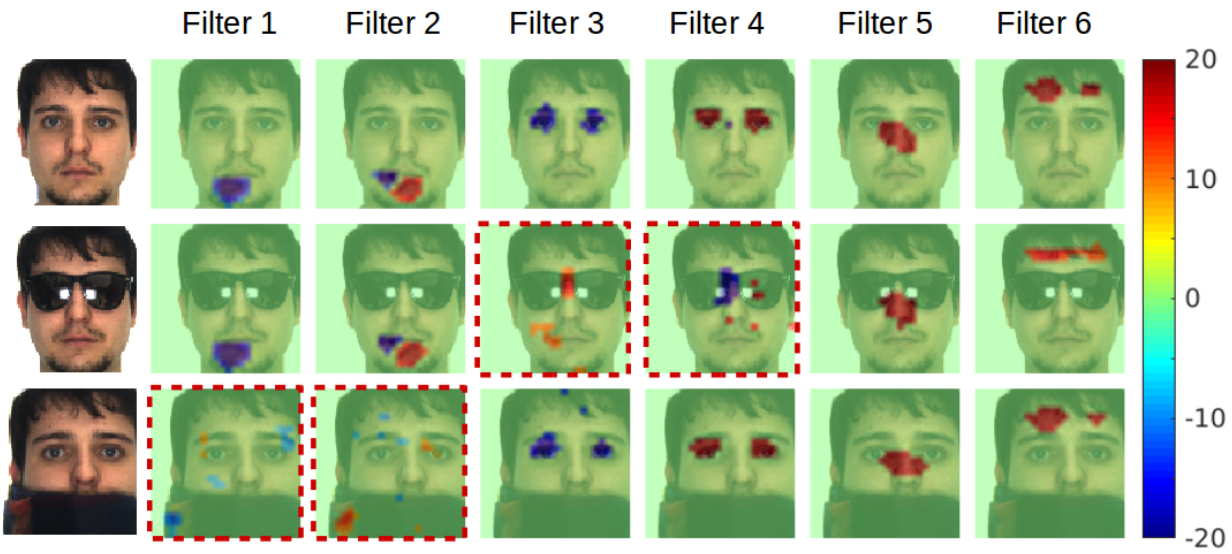


Figure 1.1: An example on the behaviors of an interpretable face recognition system: left most column is three faces of the same identity and right six columns are filter responses from six filters; each filter captures a clear and consistent semantic face part, e.g., eyes, nose, and jaw; heavy occlusions, eyeglass or scarf, alternate responses of corresponding filters and make the responses being more scattered, as shown in red bounding boxes.

the work is seminal and inspiring, this drawback largely limits its practical applicability.

In this paper, we study face recognition and strive to learn an interpretable face representation (Fig. 1.1). We define interpretability in this way that when each dimension of the representation is able to represent a face structure or a face part, the face representation is of higher interpretability. Although the concept of part-based representations has been around Li et al. (2001); Felzenszwalb et al. (2008); Berg & Belhumeur (2013); Li & Hua (2017), prior methods are not easily applicable to deep CNNs. Especially in face recognition, as far as we know, this problem is rarely addressed in the literature.

In our method, the filters are learned end-to-end from data and constrained to be locally activated with the proposed spatial activation diversity loss. We further introduce a feature activation diversity loss to better align filter responses across faces and encourage filters to capture more discriminative visual cues for face recognition, especially occluded face recognition. Compared with the interpretable CNNs from Zhang et al. Zhang et al. (2017), our final face representation does not compromise recognition accuracy, instead it achieves improved performance as well as

enhanced robustness to occlusion. We empirically evaluate our method on three face recognition benchmarks with detailed ablation studies on the proposed objective functions.

To summarize, our contributions in this paper are in three-fold: 1) we propose a spatial activation diversity loss to encourage learning interpretable face representations; 2) we introduce a feature activation diversity loss to enhance discrimination and robustness to occlusions, which promotes the practical value of interpretability; 3) we demonstrate superior interpretability, while achieving improved or similar face recognition performance on three face recognition benchmarks, compared to base CNN architectures.

CHAPTER 2

RELATED WORK

2.1 Interpretable Representation Learning

Understanding the visual recognition has a long history in computer vision Mahendran & Vedaldi (2016); Sudderth et al. (2005); Juneja et al. (2013); Singh et al. (2012); Parikh & Zitnick (2011). In early days when most models use hand-craft features, a number of research focused on how to interpret the predictions. Back then visual cues include image patches Juneja et al. (2013), body parts Yao et al. (2011), face parts Li & Hua (2017), or middle-level representations Singh et al. (2012) contingent on the tasks. For example, Vondrick et al. Vondrick et al. (2013) develop the HOGgles to visualize HOG descriptors in object detection. Since features such as SIFT Lowe (2004), LBP Ahonen et al. (2006) are extracted from image patches and serve as building blocks in the recognition pipeline, it was intuitive to describe the process from the level of patches. With the more complicated CNNs, it demands new tools to dissect its prediction. Early works include direct visualization of the filters Zeiler & Fergus (2014), deconvolutional networks to reconstruct inputs from different layers Zeiler et al. (2011), gradient-based methods to generate novel inputs that maximize certain neurons Nguyen et al. (2015), and etc. Recent efforts along this line include CAM Zhou et al. (2016) which leverages the global max pooling layer to visualize dimensions of the representation and Grad-CAM Selvaraju et al. (2016) which relaxes the constraints on the network with a general framework to visualize any convolution filters. While our method can be related to visualization of CNNs and we leverage the tools to visualize our learned filters, it is not the focus of this paper.

Visualization of CNNs is a good way to interpret the network but by itself it does not make the network more interpretable. Attention model Xu et al. (2015) has been used in image caption generation. By attention mechanism, their model can push the feature maps responding separately to each predicted caption word, which is seemingly close to our idea, but needs many labeled data

for training.

One recent work on learning a more meaningful representation is the interpretable CNNs Zhang et al. (2017). In their method, they design two losses to regularize the training of late-stage convolutional filters: one to encourage each filter to encode a distinctive object part and another to push it to respond to only one local region. AnchorNet Novotny et al. (2017) adopts the similar idea to encourage orthogonality of the filters and filter responses to keep each filter activated by a local and consistent structure. In our method, we generally extend the ideas in AnchorNet with some new aspects for face recognition in designing our spatial activation diversity loss. Another line of research in learning interpretable representations is also referred to as feature disentangling, e.g., InfoGAN Chen et al. (2016), face editing Shu et al. (2017), 3D face recognition Liu et al. (2018), and face modeling Tran & Liu (2018). They intend to factorize the latent representation to describe the inputs from different aspects, of which the direction is largely diverged from our goal in this paper.

2.2 Parts and Occlusion in Face Recognition

Face recognition is extensively studied in computer vision Learned-Miller et al. (2016); OâĂŹToole et al. (2018). Early works constructing meaningful representations for face recognition are mostly intended to improve the recognition accuracy. Some face representations are composed from face parts. The part-based models are either learned unsupervised from data Li et al. (2013) or specified by manually annotated landmarks Cao et al. (2010). Besides local parts, different face attributes are also interesting elements to build up face representations. Kumar et al. (2009) proposed to encode a face image with scores from attribute classifiers and demonstrate improved verification performance before the deep learning era. In this paper, we propose to learn meaningful part-based face representations with a deep CNN and the face part filters are learned with the carefully designed losses. We demonstrate how we leverage the interpretable representation for occlusion robust face recognition. Prior methods addressing pose variations in face recognition Li et al. (2013); Cao et al. (2010); Tran et al. (2017b); Chai et al. (2007); Yin et al. (2017); Yin & Liu

(2018) can be related since pose changes may lead to self-occlusions. However, in this work, we are more interested in more explicit situations when faces are occluded by hand, sunglasses, and other objects. Interestingly, this specific aspect is rarely studied with CNNs. Cheng et al. (2015) propose to restore occluded faces with deep auto-encoder for improved recognition accuracy. Zhou et al. (2015) argue that naively training a high capacity network with sufficient coverage in training data could achieve superior recognition performance. In our experiment, we indeed observed improved recognition accuracy to occluded faces after augmenting training data with synthetic occluded faces. However, with the proposed method, we can further improve robustness to occlusion without increasing network capacity, which highlights the merits of interpretable representation.

2.3 Occlusion Handling with CNNs

Different methods are proposed to handle occlusion with CNNs for robust object detection and recognition. Wang et al. (2017) learn an object detector by generating an occlusion mask for each object, which synthesizes harder samples for the adversarial network. In Singh & Lee, occlusion masks are utilized to enforce the network to pay attention to different parts of the objects. Ge et al. (2017) solve face detection with heavy occlusions by proposing a masked face dataset and applying it on their proposed LLE-CNNs. Despite using masked images, our occlusion robustness mainly comes from enforcing constraints for the spreadness of the feature activations and guiding the network to extract features from different parts of the face.

CHAPTER 3

PROPOSED METHOD

Our network architecture in training is shown in Fig. 3.1. From a high-level perspective, we construct a Siamese network with two branches sharing weights to learn face representations from two faces: one with synthetic occlusion and one without. We would like to learn a set of diverse filter **F**, which applies on a hyper-column descriptor Φ, consisting of feature at multiple semantic levels. The proposed Spatial Activation Diversity (SAD) loss encourages the face representation to be structured with consistent semantic meaning. Softmax loss helps encode the identity information. The input to the lower network branch is a synthetic occluded version of the above input. The proposed Feature Activation Diversity (FAD) loss requires filters to be insensitive to the occluded part, hence more robust to occlusion. At the same time, we mask out parts of the face representation sensitive to the occlusion and train to identify the input face solely based on the remaining elements. As a result, the filters respond to non-occluded parts are trained to capture more discriminative cues for identification.

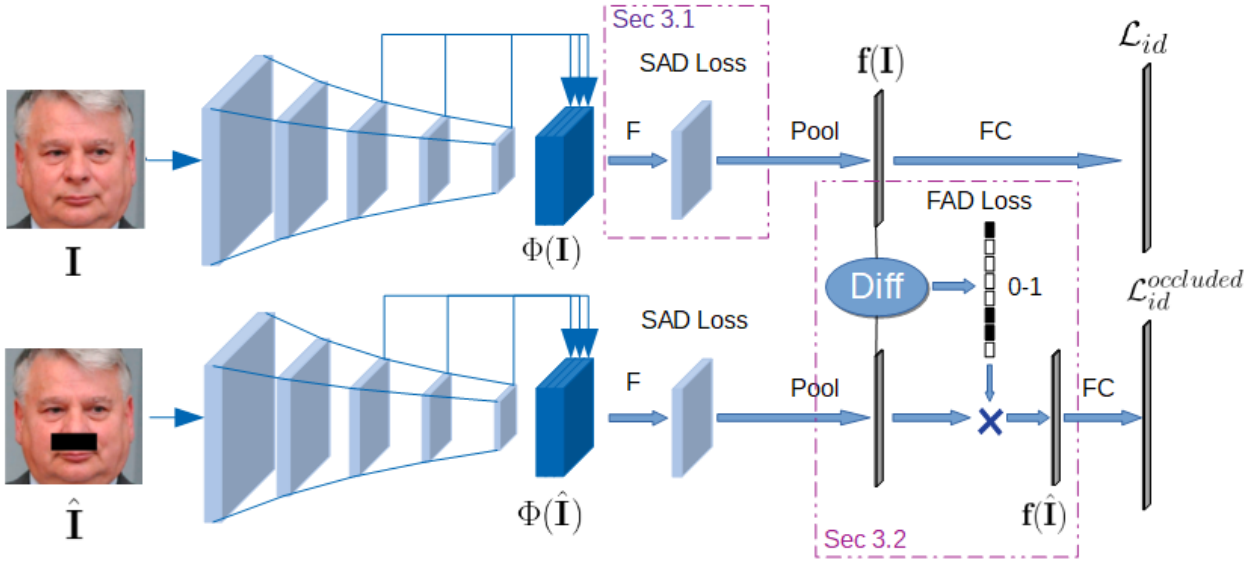


Figure 3.1: overall network architecture of the proposed method.

3.1 Spatial Activation Diversity Loss

Novotny et al. (2017) proposed a diversity loss for semantic matching by penalizing correlations among filters weights and their responses. While their idea is general enough to extend to face representation learning, in practice, their design is not directly applicable due to the prohibitively large number of identities (classes) in face recognition. Their approach also suffers from degradation in recognition accuracy. We first introduce their diversity loss and then describe our proposed modifications tailored to face recognition.

3.1.1 Spatial Activation Diversity Loss

For each of K class in the training set, Novotny et al. (2017) proposed to learn a set of diverse filters with discriminative power to distinguish an object of the category and background images. The filers $\mathbf{F}$ apply on a hypercolumns descriptor $\Phi(\mathbf{I})$, created by concatenating the filter responses of an image $\mathbf{I}$ at different convolutional layers Hariharan et al. (2015). This helps $\mathbf{F}$ to aggregate features at different semantic levels. The response map of this operation is denoted as $\psi(\mathbf{I}) = \mathbf{F} * \Phi(\mathbf{I})$.

The diversity constraint is implemented by two *diversity losses* $\mathcal{L}_{SAD}^{filter}$ and $\mathcal{L}_{SAD}^{response}$, encouraging the orthogonality of the filters and of their responses, respectively. $\mathcal{L}_{SAD}^{filter}$ makes filters orthogonal by penalizing their correlations:

$$\mathcal{L}_{SAD}^{filter}(\mathbf{F}) = \sum_{i \neq j} \left| \sum_{p} \frac{\langle \mathbf{F}_i^p, \mathbf{F}_j^p \rangle}{\|\mathbf{F}_i^p\|_F \, \|\mathbf{F}_j^p\|_F} \right|, \tag{3.1}$$

where $\mathbf{F}_i^p$ is the column of filter $\mathbf{F}_i$ at the spatial location p. Note that orthogonal filters are likely to respond to different image structures, but this is not necessarily the case. Thus, the second term $\mathcal{L}_{\text{SAD}}^{\text{response}}$ is introduced to directly decorrelate the filters' *response maps* $\psi_k(\mathbf{I})$:

$$\mathcal{L}_{SAD}^{response}(\mathbf{I}; \Phi, \mathbf{F}) = \sum_{i \neq j} \left\| \frac{\langle \psi_i, \psi_j \rangle}{\|\psi_i\|_F \|\psi_j\|_F} \right\|^2 \tag{3.2}$$

This term is further regularized by using the smoothed response maps $\psi'(\mathbf{I}) \doteq g_\sigma * (\psi(\mathbf{I}))$ in place of $\psi(\mathbf{I})$ in $\mathcal{L}_{SAD}^{response}$ loss computing. Here the channel-wise Gaussian kernel g_σ is applied to encourage filter responses to spread farther apart by dilating their activations.

3.1.2 Our Proposed Modifications

Novotny et al. (2017) learn K sets of filters, one for each of K categories. The discrimination of the features are maintained by K binary classification losses for each category vs. background images. The discriminative loss is proposed to enhance (or suppress) the maximum value in the response maps ψ_k for the positive (or negative) class. In Novotny et al. (2017), the final feature representation $\mathbf{f}$ is obtained via global max-pooling operation on ψ. This design is not applicable for face classification CNN as the number of identities K are usually prohibitively large (usually in the order of ten thousands or above).

Here, to make the feature discriminative, we only learn **one** set of filters and connect the representation $\mathbf{f}(\mathbf{I})$ directly to a K-way softmax classification:

$$\mathcal{L}_{id} = -\log(P_c(\mathbf{f}(\mathbf{I}))). \tag{3.3}$$

Here we minimize the negative log-likelihood of feature $\mathbf{f}(\mathbf{I})$ being classified to its ground-truth identity c.

Furthermore, global max-pooling could lead to unsatisfied recognition performance, as shown in Novotny et al. (2017) where they observed minor performance degradation compared to the model without their diversity loss. One empirical explanation of this performance degradation is that max-pooling has similar effect to ReLU activation which makes the response distribution biased to non-negative range $[0, +\infty)$. Hence it significantly limits the feasible learning space.

Most recent works choose to use global average pooling Yi et al. (2014); Tran et al. (2017b). However, when applying average-pooling to introduce interpretability, it does not promote desired spatially peaky distribution. Empirically, we found the learned feature response maps of average pooling failed to have strong activation in small local regions.

Here we propose to design a pooling operation that satisfies two objectives: i) promote peaky distribution to be well-cooperated with the spatial activation diversity loss; ii) maintain the statistics of the feature responses for the global average-pooling to achieve good recognition performance. Based on these considerations, we propose the operation termed **Large magnitude filtering** (LMF),

as follows:

For each channel in the feature response map, we assign d% of elements with smallest magnitude to be 0. The size of the output remains the same. The $\mathcal{L}_{SAD}^{response}$ loss is applied on the modified response map $\psi'(\mathbf{I}) \doteq g_\sigma * (\text{LMF}(\psi(\mathbf{I})))$ in place of $\psi(\mathbf{I})$ in Eqn. 3.2. Then, the conventional global average pooling is applied to LMF($\psi(\mathbf{I})$) to obtain the final representation $\mathbf{f}(\mathbf{I})$.

By removing small magnitude values from ψ_k, $\mathbf{f}$ won't be affected much after global average pooling, which favors discriminative feature learning. On the other hand, the peaks of the response maps are still well maintained, which leads to more reliable computation of the diversity loss.

3.2 Feature Activation Diversity Loss

One way to evaluate whether the diversity loss is effective is to compute the average location of the peaks within the kth response maps $\psi'_k(\mathbf{I})$ for an image set. If the average locations across K filters spread all over the face spatially, the diversity loss is well functioning and can associate each filer with a specific face area. With the SAD loss, we do observe the improved spreadness compared to the base CNN model trained without the SAD loss. Since we believe that more spreadness indicates higher interpretability, we hope to further boost the spreadness of the average peak locations across filters, i.e., elements of the learnt representation.

Motivated by the goal of learning part-based face representations, it is desirable to encourage that any local face area only affects a small subset of the filter responses. To fulfill this desire, we propose to create synthetic occlusion on local areas of a face image, and constrain on the difference between its feature response and that of the non-occluded original image. The second motivation for our proposal is to design an occlusion-robust face recognition algorithm, which, in our view, should be a natural by-product or benefit of the part-based face representation.

With this in mind, we propose a Feature Activation Diversity (FAD) Loss to encourage the network to learn filters robust to occlusions. That is, occlusion in a local region should only affect a small subset of elements within the representation. Specifically, leveraging pairs of face images $\mathbf{I}$, $\hat{\mathbf{I}}$, where $\hat{\mathbf{I}}$ is a version of $\mathbf{I}$ with a synthetically occluded region, we enforce the majority of two

feature representations, $\mathbf{f}(\mathbf{I})$ and $\mathbf{f}(\hat{\mathbf{I}})$, to be similar:

$$\mathcal{L}_{\text{FAD}}(\mathbf{I}, \hat{\mathbf{I}}) = \sum_i \left| \tau_i(\mathbf{I}, \hat{\mathbf{I}}) \left[\mathbf{f}_i(\mathbf{I}) - \mathbf{f}_i(\hat{\mathbf{I}}) \right] \right|, \tag{3.4}$$

where the feature selection mask $\tau(\mathbf{I}, \hat{\mathbf{I}})$ is defined with threshold t: $\tau_i(\mathbf{I}, \hat{\mathbf{I}}) = 1$ if $\left| \mathbf{f}_i(\mathbf{I}) - \mathbf{f}_i(\hat{\mathbf{I}}) \right| < t$, otherwise $\tau_i(\mathbf{I}, \hat{\mathbf{I}}) = 0$. There are multiple design choices for the threshold: number of elements based or value based. We evaluate and discuss these choices in the experiments.

We also would like to correctly classify occluded images using just subset of feature elements, which is insensitive to occlusion. Hence, the softmax identity loss in the occlusion branch is applied to the masked feature:

$$\mathcal{L}_{id}^{occluded} = -\log(P_c(\tau(\mathbf{I}, \hat{\mathbf{I}}) \odot \mathbf{f}(\hat{\mathbf{I}}))). \tag{3.5}$$

By sharing the classifier's weights between two branches, this classifier is learned to be more robust to occlusion. It also leads to a better representation as filters respond to non-occluded parts need to be more discriminative.

3.3 Implementation Details

Our proposed method is model agnostic. To demonstrate this, we apply the proposed SAD and FAD losses to two different network architectures: one inspired by the widely used CASIA-Net Yi et al. (2014); Tran et al. (2018), the other based on ResNet50 He et al. (2016), both of which are popular in face recognition community. The structure of the CASIA-Net is shown in Tab. 3.3. ResNet50 contains a lot of layers, we put its structure in Appendix. And the We add HC-descriptor-related blocks for our SAD loss learning. Conv33, conv44, conv54 layers are used to construct the HC descriptor via conv upsampling layers. We set the feature dimension $N^f = 320$. As for ResNet50, we take the modified version in Deng et al. (2018), where $N^f = 512$. We also construct the HC descriptor by using 3 layers at different resolutions. To speed up the training, we reuse the pretrained feature extraction network shared by Tran et al. (2018) and Deng et al. (2018). All new weights are randomly initialized using a truncated normal distribution with std of 0.02. The entire network is jointly trained using SGD optimizer at an initial learning rate of 0.001 and a momentum of 0.9. The learning rate is divided by 10 for twice when the training loss is stabled.

Table 3.1: The structures of our network architecture.

Layer	Input	Filter/Stride	Output Size
conv11	Image	$3 \times 3/1$	$96 \times 96 \times 32$
conv12	conv11	$3 \times 3/1$	$96 \times 96 \times 64$
conv21	conv12	$3 \times 3/2$	$48 \times 48 \times 64$
conv22	conv21	$3 \times 3/1$	$48 \times 48 \times 64$
conv23	conv22	$3 \times 3/1$	$48 \times 48 \times 128$
conv31	conv23	$3 \times 3/2$	$24 \times 24 \times 128$
conv32	conv32	$3 \times 3/1$	$24 \times 24 \times 96$
conv33	conv32	$3 \times 3/1$	$24 \times 24 \times 192$
conv41	conv33	$3 \times 3/2$	$12 \times 12 \times 192$
conv42	conv41	$3 \times 3/1$	$12 \times 12 \times 128$
conv43	conv42	$3 \times 3/1$	$12 \times 12 \times 256$
conv51	conv43	$3 \times 3/2$	$6 \times 6 \times 256$
conv52	conv51	$3 \times 3/1$	$6 \times 6 \times 160$
conv53	conv52	$3 \times 3/1$	$6 \times 6 \times N^f$
conv43-U	conv43	upsampling	$24 \times 24 \times 256$
conv44	conv43-U	$1 \times 1/1$	$24 \times 24 \times 192$
conv53-U	conv53	upsampling	$24 \times 24 \times 320$
conv54	conv53-U	$1 \times 1/1$	$24 \times 24 \times 192$
Φ (HC)	conv33,44,54	$3 \times 3/1$	$24 \times 24 \times 576$
Ψ	Φ	$3 \times 3/1$	$24 \times 24 \times N^f$
AvgPool	Ψ	$24 \times 24/1$	$1 \times 1 \times N^f$

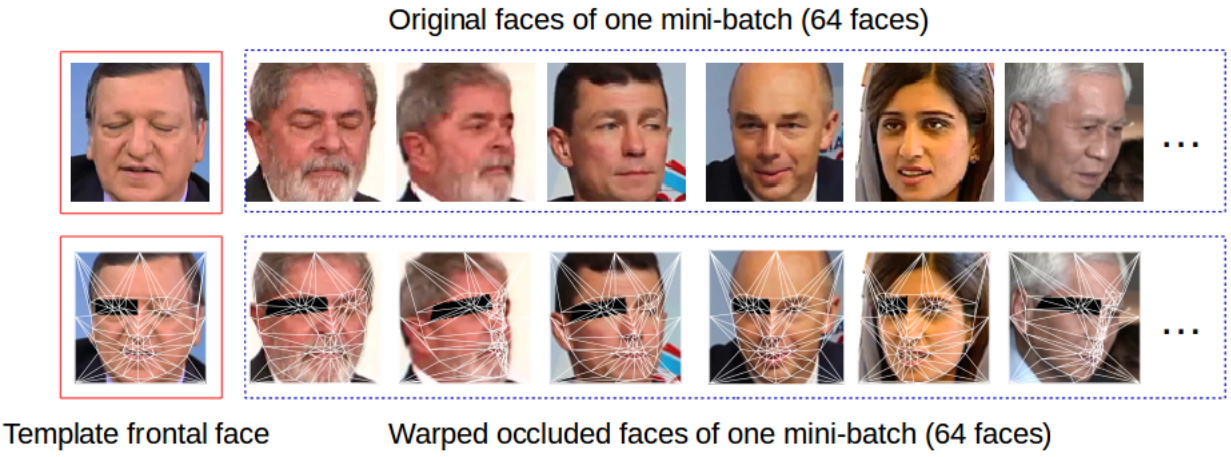


Figure 3.2: With barycentric coordinates, we warp the vertices of the template face mask to each image within the 64-image mini-batch.

For FAD, the feature mask τ can be computed per image pair $\mathbf{I}$ and $\hat{\mathbf{I}}$. However, to obtain a more reliable feature mask, we opt to compute τ using multiple image pairs sharing the same *physical* occluded mask, i.e., $\tau_i(\{\mathbf{I}, \hat{\mathbf{I}}\}_{j=1}^{N}) = 1$ if $\frac{1}{N}\sum_{j=1}^{N} \left|\mathbf{f}_i(\mathbf{I}_j) - \mathbf{f}_i(\hat{\mathbf{I}}_j)\right| < t$, otherwise 0.

To provide the same *physical* mask to images in a batch regardless their poses, we first define a frontal face template with 142 triangles created by 68 facial landmarks. A 32×12 rectangle, which is randomly placed and may cover the face area, such as eye, nose and mouth, is selected as a normalized mask. Each of the rectangle's four vertices can be represented by the barycentric coordinate w.r.t. the triangle enclosing the vertex. For each image within a mini-batch, corresponding four vertices of a quadrilateral can be found via the same barycentric coordinates. This quadrilateral denotes the location of a warped mask of that image. An example of this mask warping process is shown in Fig. 3.2.

CHAPTER 4

EXPERIMENTAL RESULTS

4.1 Experimental Settings

4.1.1 Introduction

The following sections provide ablation studied, qualitative and quantitative evaluation. Firstly, to further analyze the influence of parameters setting in our model, ablation studies are conducted. We set the different thresholds in feature activation diversity loss to explore the face recognition performance, and then compare the performance among models trained with different occlusion types and dynamic occlusion window sizes. Besides, turn off one of the diversity losses also help us to understand the effect of the proposed two loss functions. Secondly, to better illustrate the results of our method, we present qualitative visualization of the learned representations, response maps, etc. Lastly, we compare the face recognition performance on three benchmark datasets: IJB-A Klare et al. (2015), IJB-C Brianna Maze et al. (2018) and AR face Martinez (1998).

4.1.2 Database

For CASIA-Net, we take CASIA-WebFace databases Yi et al. (2014) as the training databases. For ReNet50, we use MS-Celeb-1M database Guo et al. (2016) for training, and IJB-A Klare et al. (2015), IJB-C Brianna Maze et al. (2018) and AR face Martinez (1998) for testing. CASIA-WebFace contains $493,456$ images of $10,575$ subjects. MS-Celeb-1M includes $1M$ images of $100k$ subjects. Since it contains many labeling noise, we use a cleaned version of MS-Celeb-1M Guo et al. (2016). In our experiments, we evaluate IJB-A in three different scenarios, i.e., original faces, synthetic occlusion and natural occlusion faces. For synthetic occlusion, we randomly generate a warped occluded area for each testing image, the same as what we did in training. IJB-C extends IJB-A, also is a video-based face database with $3,134$ images and $117,542$ frames from videos of

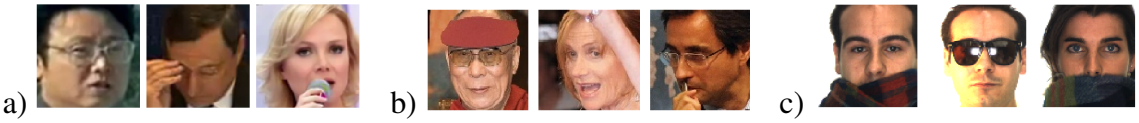


Figure 4.1: Example face from (a) IJB-A, (b) IJB-C and (c) AR face databases. The occlusions include scarf, eyeglass, hands, etc.

3,531 subjects. One unique property of IJB-C is its label on fine-grained occlusion area. Thus, we use IJB-C to evaluate occlusion-robust face recognition, using testing images with at least one occluded area. AR face is another natural occlusion face database, with ~ 4K faces of 126 subjects. We only use AR faces with natural occlusions, including wearing glass and scarfs. Some examples of IJB-A, IJB-C and AR databases are in Fig. 4.1. Following the setting in Deng et al. (2018), all training and test images are processed and resized to 112×112. Note that, all ablation and qualitative evaluations use the CASIA-Net-based model, and the quantitative evaluations use both models.

4.2 Ablation Study

4.2.1 Different Thresholds

As mentioned in Sec. 3.2, the threshold t for Eqn. 3.4 denotes the number of elements in two N^f-dim features that the FAD loss encourages their similarity. To study the effect of t to face recognition, we train different models with $t = 130, 260, 320$. The first three rows in Tab. 4.2.1 show the comparison on all three variants of IJB-A dataset. When forcing all elements of $\mathbf{f}(\mathbf{I})$ and $\mathbf{f}(\hat{\mathbf{I}})$ to be the same ($t = N^f = 320$), the performance significantly drops on all three sets. In this case, the feature representation of the non-occluded face is negatively affected as being completely pushed toward a representation of the occluded one. While the model with $t = 130$ has similar performance with model $t = 260$, we will use the latter for the rest of the paper, due to the observation that the latter model affects less filters, push other filter responses away from any local occlusions, and subsequently enhances the spreadness of the average response locations.

Moreover, in our Feature Activation Diversity Loss (FAD), the feature mask is computed by

Table 4.1: Ablation study on IJB-A database.

Method	IJB-A		Manual Occlusion		Natural Occlusion	
Metric (%)	@FAR=.01	@Rank-1	@FAR=.01	@Rank-1	@FAR=.01	@Rank-1
BlaS(t = 130)	79.0 ± 1.6	89.5 ± 0.8	76.1 ± 1.7	88.0 ± 1.4	66.2 ± 4.0	73.0 ± 3.3
BlaS(t = 260)	79.2 ± 1.8	89.4 ± 0.8	76.1 ± 1.4	88.0 ± 1.2	66.5 ± 6.4	72.3 ± 2.8
BlaS(t = 320)	74.6 ± 2.4	88.9 ± 1.3	71.8 ± 3.1	87.5 ± 1.6	61.0 ± 6.5	71.6 ± 3.2
GauD(t = 260)	**79.3** ± 2.0	**89.9** ± 1.0	**76.2** ± 2.4	**88.6** ± 1.1	**66.8** ± 3.5	**73.2** ± 3.3
SAD only	78.1 ± 1.8	88.1 ± 1.1	66.6 ± 5.6	81.2 ± 1.9	64.2 ± 6.9	71.0 ± 3.3
FAD only	76.7 ± 2.0	88.1 ± 1.1	75.2 ± 2.4	85.1 ± 1.2	66.5 ± 6.4	72.3 ± 2.8

Table 4.2: Comparison of selected filter numbers.

Method	Forehead	Eye_L	Nose	cheek	Mouth
Metric ($n_{sel} \setminus n_{total}$)					
GauD(t = 260)	60 \ 320	60 \ 320	60 \ 320	60 \ 320	60 \ 320
BlaS(fixed)	1 \ 320	59 \ 320	98 \ 320	41 \ 320	82 \ 320

thresholding on the averaged feature difference. There are also multiple design choices for the thresholding operation, such as number of element based or value based thresholding. In our paper, we explored the first choice: minimizing the difference of t, out of 320, elements with smaller averaged difference values. Under this setting, the number of filters enforced to be similar by FAD loss are fixed regardless the occlusion location. Intuitively, dynamic selected number of filters can be more natural, as different occlusion areas might cover different regions, some of which contains more discriminative feature (i.e eye, mouth area) than the others (cheek, forehead). To further evaluate this effect, we explore the latter thresholding options: setting a fixed threshold value to select different number of filters. As shown in Table 4.2,the forehead only has one filter been selected, which means covering forehead won't affect the face recognition, since forehead contain little identity information. However, for eye, nose and mouth, there are 80 filter on average been affected. Because they are the most dicriminative parts of the face.

4.2.2 Different Occlusions and Dynamic Window Size

In FAD loss, we use the warped black window as the synthetic occlusion. It is important to introduce another type of occlusion to see the effects on face recognition. Thus, we use Gaussian noise to replace the black color in the occlusion window. Further, we employ a dynamic window size by randomly generating a value from $[12, 32]$ for both the window height and width. The face recognition results on IJB-A are shown in Tab. 4.2.1, where 'BlaS' means black window with static sizes, while 'GauD' means Gaussian noise window with dynamic sizes. From the results, it is interesting to find that the performance is slightly better than 'BlaS'. Comparing to black window, Gaussian noise contains more diverse information.

4.2.3 Spatial vs. Feature Diversity Loss

Since we propose two different diversity losses in our model, it is important to evaluate the effects of two losses on face recognition performance respectively. As shown in Tabs 4.2.1, we can train our models using either diversity loss, or both of them. We observe that, while the SAD loss performs reasonably well on general IJB-A, it suffers for data with occlusions, being synthetic or natural. Alternatively, using only the FAD loss can improve the performance on the two occlusion datasets. Finally, using both losses, the row of 'BlaS($t = 260$)', improves upon both models with only one loss.

4.3 Qualitative Evaluation

4.3.1 Spreadness of Average Locations of Filter Response

Given an input face image, our model computes $\psi'(\mathbf{I})$, the 320 feature maps of size 24×24, where the average pooling of one map is one element of the final 320-d feature representation. Each feature map contains both the positive and negative response values, which are distributed at different spatial areas of the face. We select the locations of both the highest value for positive response and the lowest value for negative response as the peak response locations. To better

illustrate the spatial distribution of peak locations, we randomly select $1,000$ testing images and calculate the weighted average location for each filter, with three notes. One is that there are two types of locations, for the highest (positive) and lowest (negative) responses respectively. The other is that, since the filters are responsive to semantic facial components, their 2D spatial locations may change with pose. To compensate that, we warp the peak location in an arbitrary-view face to a canonical frontal-view face, by its barycentric coordinates w.r.t. the triangle enclosing it. Similar to Fig. 3.2, we use 68 estimated landmarks Liu et al. (2017a); Jourabloo & Liu (2017) and control points on the image boundary to define the triangular mesh. Finally, the weight of each location is determined by the magnitude of its peak response.

With that, the average locations for all feature maps (or filters) are shown in Fig. 4.2. To compare the visualization results between our models and CNN base model, we compute $\bar{d} = \frac{1}{N^f}\sum_i^{N^f}\left|c_i - \frac{1}{N^f}\sum_i^{N^f} c_i\right|$ to quantify the average locations spreadness, where c_i denotes the (x, y) coordinates of the ith average location. For both the positive and negative peak response, we take the mean of their $\bar{d}$. As stated in Fig. 4.2, our model with SAD loss enlarge the spreadness of the average locations. Further, our model with both losses continue to push the filter responses apart from each other. This demonstrates that indeed our model is able to push filters to attach to diverse face areas, while all filters doesn't attach to specific facial part, results in average locations stay near the image center. In addition, we compute the standard deviation for each filter's peak location from $1,000$ images. From Fig. 4.2, we can observe that the base model has larger standard deviations than SAD only model or our model, which means our model can better concentrate on a local part than the base model.

In above analysis, we set the LMF rate d to be 95.83%. It is worthy to ablate the impact of the rate d. We train models with different d of 0%, 75%, 87.5% and 95.83%. Since before average pooling the feature map is of $24 * 24$, the last 3 percentages mean that we remove 24×18, 24×21 and 24×23 responses respectively for each model and 0% denotes the base model. Tab. 4.3 compares the average of standard deviations of peak locations across 320 filters. Note the values of the best model $(12.9/13.4)$ equals to the average color of Fig. 4.2(c). When we use the larger LMF rate, the

Table 4.3: Compare standard deviations of peaks with varying d.

LMF (d%)	0	75.00	87.50	95.83
std(pos./neg.)	25.7/25.7	14.7/14.4	13.5/14.0	12.9/13.4

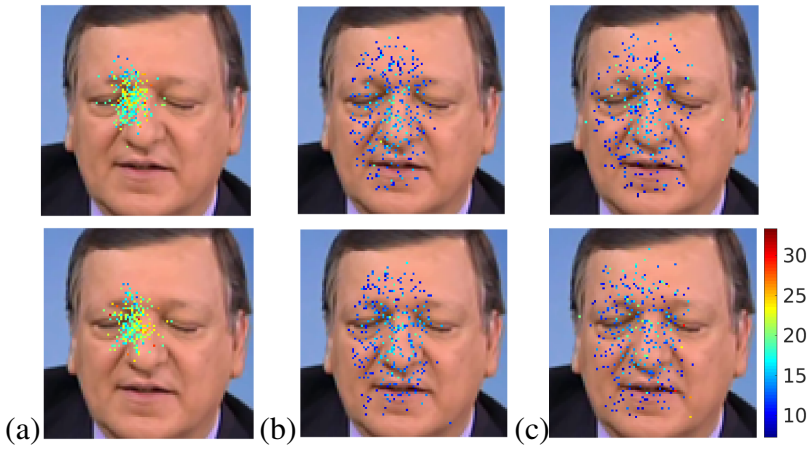
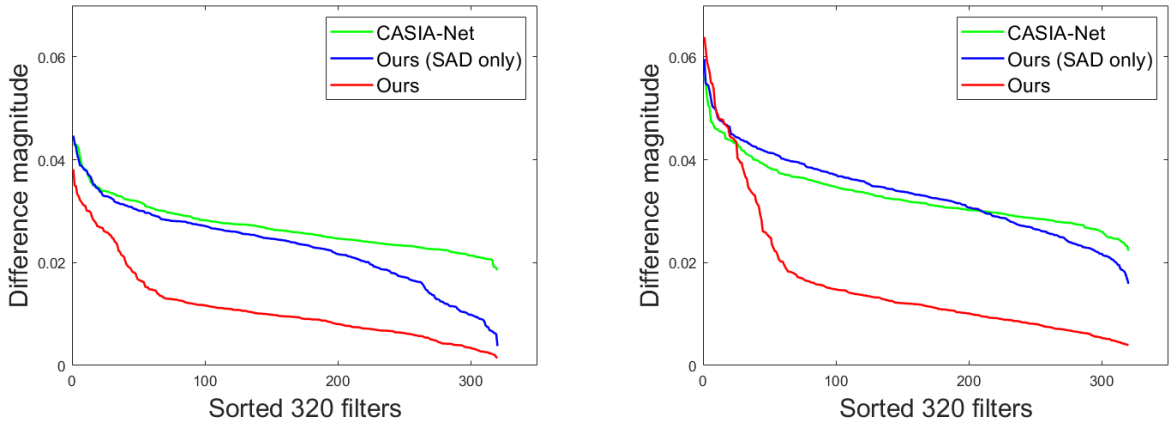


Figure 4.2: The average locations of positive (top) and negative (bottom) peak responses of 320 filters for three models: (a) base CNN model ($\bar{d} = 6.9$), (b) our (SAD only, $\bar{d} = 17.1$), and (c) our model ($\bar{d} = 18.7$), where $\bar{d}$ quantifies the average locations spreadness. The color on each location denotes the standard deviation of peak locations. The face size is 96×96.

model tends to be more concentrated onto a local facial part. For this reason, we set $d = 95.83\%$.

4.3.2 Mean Feature Difference Comparison

Both of our losses promotes part-based feature learning, which leads to occlusion robustness. Especially, in FAD, we directly minimize the difference in a portion of representation of face with and without occlusion. We now study the effect of our loss on faces with occlusion. Firstly, we randomly select $1,000$ test faces in different poses and generate the synthetic occlusion. After that, we calculate the mean of feature difference of each filter on both original and occluded faces based on different models. Fig. 4.3 (a) and (b) illustrates the sorted feature difference of three models at two different occlusion parts, eye and nose, respectively. Compare to the base CNN (trained with $\mathcal{L}_{id}$), both of our losses have smaller magnitude of differences. Diversity properties of SAD loss could help to reduce the feature change on occlusion, even without directly minimizing this



(a) Eye part occlusion (b) Nose part occlusion

Figure 4.3: Mean of feature difference on two occluded parts.

difference. FDA loss further enhances robustness by only letting the occlusion alternate a small portion of the representation, keeping the remaining elements invariant to the occluded part.

4.3.3 Visualization on Feature Difference Vectors

Fig. 4.2 demonstrates that each of our filter spatially corresponds to a face location. Here we further study the relation of these average locations and semantic meaning on input images. In Fig. 4.4, we visualize the magnitude of changes of each filter response due to five different occlusions. We observe the locations of points with large feature difference are around the occluded face area, which means our learned filters are indeed sensitive to various facial areas. Further, the magnitude of the feature difference can be vary with different occlusions. The maximum feature difference can be as high as 0.7 with occlusion in eye or mouth, meanwhile this number is only 0.3 in less critical area, e.g., forehead.

4.3.4 Filter Response Visualization

Fig. 4.5 visualizes the feature responses of some filters on different subjects' faces. From the heat maps, we can see how each filter is attached to a specific semantic location on the faces, independent to either identities or poses. This is especially impressive for faces with varying poses,

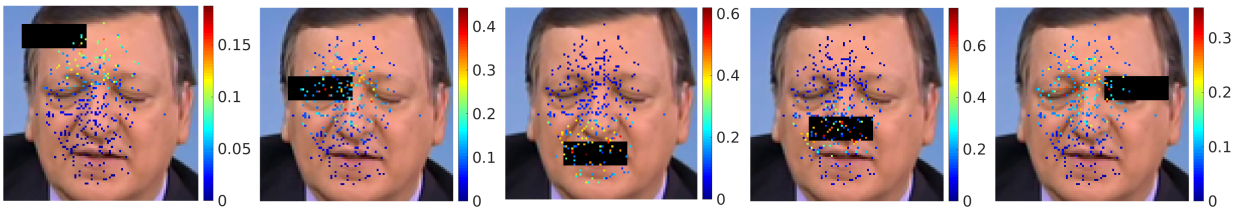


Figure 4.4: The correspondence between feature difference magnitude and occlusion locations. Best viewed electronically.

in that despite no pose prior is used in training, the filter can always respond to the semantically equivalent local part.

4.4 Quantitative Evaluation on Benchmark

Although we have shown some results of standard deviations of peaks in Fig. 4.2, we still want to further investigate the response concentration properties between our model and base model, which we will give some histograms to illustrate. To show that our method is model agnostic, we use two different base CNN models, CASIA-Net and ResNet50. Our proposed method and the respective base model only differs in the loss functions. E.g., both our CASIA-Net-based model and base CASIA-Net model use the same network architecture as Tab. 3.3. We test on two types of datasets: the generic in-the-wild faces and occlusion faces.

4.4.1 Standard Deviation of Peaks

Fig. 4.2 shows the average of the peak locations of 320 filter responses, and also use the different color to show the vary standard deviations of peak locations for each filter across 1,000 images. But it is hard to tell which model has the smallest standard deviation, therefore We can also compute the histograms of standard deviations across 320 filters, for both the positive responses and negative responses, respectively. As shown in Fig. 4.6, our model can generate filter responses whose locations have smaller standard deviations than CNN base model, i.e., our filter can be more concentrated on one specific location of the face. Note that the SAD loss only model also reduces the standard deviation over the CNN base model, our model can have a slightly smaller standard

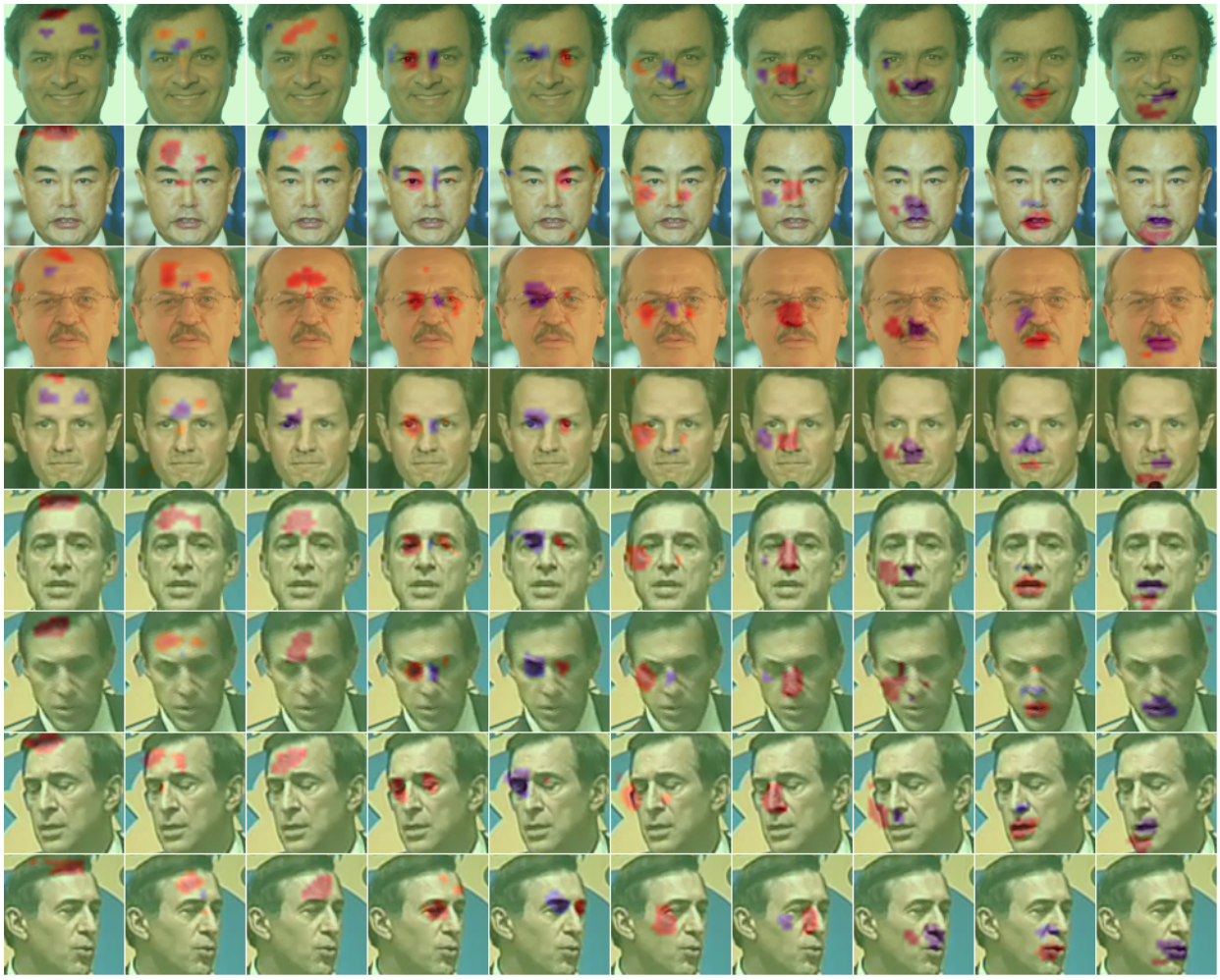


Figure 4.5: Visualization of filter response “heat maps" of 10 different filters on faces from different subjects (top 4 rows) and the same subject (bottom 4 rows). The positive and negative responses are shown as two colors within each image. Note the high consistency of response locations across subjects and across poses.

deviations than SAD loss only model.

4.4.2 Generic in-the-wild faces

As shown in Tabs. 4.4.2, 4.4.3, when comparing to the base CASIA-Net model, our CASIA-Net-based model with two losses achieves the superior performance. The same superiority is demonstrated w.r.t. CASIA-Net with data augmentation, which shows that the gain is caused by the novel loss function design. For the deeper ResNet50 structure, our proposed model achieves

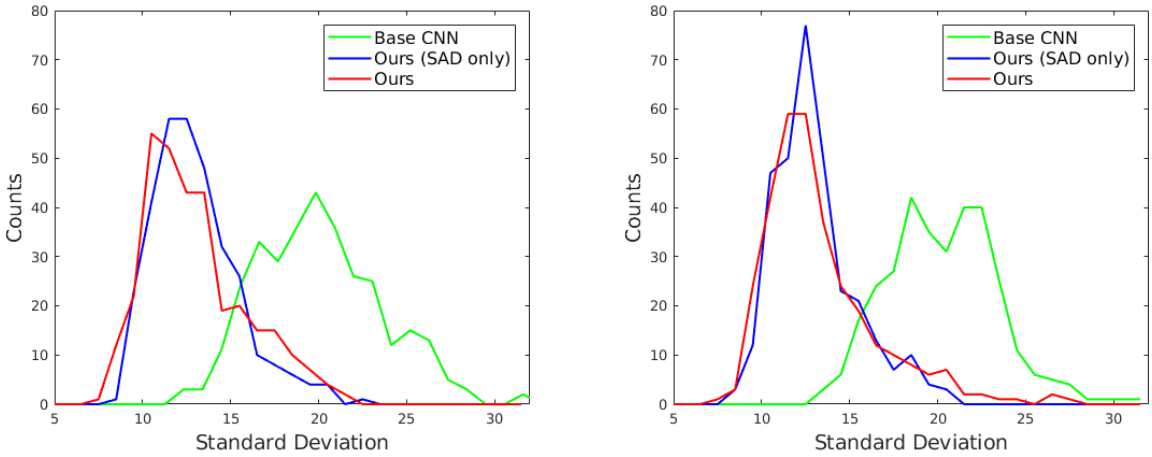


Figure 4.6: Histograms of standard deviations of peak locations for positive (left) and negative (right) responses.

.

Table 4.4: Comparison on IJB-A database.

Method ↓	Verification		Identification	
Metric (%) →	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
DR-GAN Tran et al. (2018)	79.9 ± 1.6	56.2 ± 7.2	88.7 ± 1.1	95.0 ± 0.8
CASIA-Net	74.3 ± 2.8	49.0 ± 7.4	86.6 ± 2.0	94.2 ± 0.9
Ours (CASIA-Net)	79.3 ± 2.0	60.2 ± 5.5	89.9 ± 1.0	95.6 ± 0.6
FaceID-GAN Shen1 et al. (2018)	87.6 ± 1.1	69.2 ± 2.7	–	–
VGGFace2 Cao et al. (2018)	93.9 ± 1.3	85.1 ± 3.0	**96.1** ± 0.6	**98.2** ± 0.4
PRFaceCao2 et al. (2018)	94.4 ± 0.9	86.8 ± 1.5	92.4 ± 1.6	96.2 ± 1.0
ResNet50 He et al. (2016)	**94.8** ± 0.6	86.0 ± 2.6	94.1 ± 0.8	96.1 ± 0.6
Ours (ResNet50)	94.6 ± 0.8	**87.9** ± 1.0	93.7 ± 0.9	96.0 ± 0.5

similar performance as the base model, and both outperform the models with CASIA-Net as the base. Even comparing to state-of-art methods, the performance of our ResNet50-based model is still competitive. It is worthy note that this is the first time that a reasonably interpretable representation is able to demonstrate competitive state-of-the-art recognition performance on a widely used benchmark, e.g., IJB-A.

Table 4.5: Comparison on IJB-C database.

Method ↓	Verification		Identification	
Metric (%) →	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
DR-GAN Tran et al. (2018)	88.2	73.6	74.0	84.2
CASIA-Net	87.1	72.9	74.1	83.5
Ours (CASIA-Net)	89.2	75.6	77.6	86.1
VGGFace2 Cao et al. (2018)	95.0	90.0	89.8	**93.9**
Mn-v Xie & Zisserman (2018)	**96.5**	92.0	–	–
AIM Zhao et al. (2018)	96.2	**93.5**	–	–
ResNet50 He et al. (2016)	95.9	93.2	**90.5**	93.2
Ours (ResNet50)	95.8	93.2	90.3	93.2

4.4.3 Occlusion faces

We test our models and base models on multiple occlusion face datasets. The synthetic occlusion of IJB-A, the natural occlusion of IJB-A, and the natural occlusion of IJB-C have $500/25,795$, $466/12,703$, and $3,329/78,522$ images/subjects, respectively. As shown in Tabs. 4.4.3, 4.4.3, 4.4.3, the performance improvement on the occlusion datasets are more substantial than the generic IJB-A database, which shows the advantage of interpretable representations in handling occlusions.

For AR faces, we select all 810 images with eyeglasses and scarfs occlusions, from which $6,000$ same-person and $6,000$ different-person pairs are randomly selected. We compute the representations of an image pair and its cosine distance.

As shown in Fig. 4.7, the Equal Error Rates of CASIA-Net, ours (CASIA-Net), ResNet50 and ours (ResNet50) are 21.6%, 16.2%, 4.2% and 3.9%, respectively. We observe that our model based on CASIA-Net achieves the superior performance comparing to the CASIA-Net base model. And as for the state-of-art ResNet50 model, we can still observe the performance improvement of our model to the ResNet50 base model.

Table 4.6: Comparison on IJB-A database with synthetic occlusions.

Dataset	IJB-A synthetic occlusion			
Method ↓	Verification		Identification	
Metric (%) →	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
DR-GAN Tran et al. (2018)	61.9 ± 4.7	35.8 ± 4.3	80.0 ± 1.1	91.4 ± 0.8
CASIA-Net	61.8 ± 5.5	39.1 ± 7.8	79.6 ± 2.1	91.4 ± 1.2
Ours (CASIA-Net)	76.2 ± 2.4	55.5 ± 5.7	88.6 ± 1.1	95.0 ± 0.7
ResNet50 He et al. (2016)	93.0 ± 0.7	80.9 ± 4.7	92.8 ± 0.9	95.5 ± 0.8
Ours (ResNet50)	**94.2** ± 0.6	**87.5** ± 1.5	**93.4** ± 0.7	**95.8** ± 0.4

Table 4.7: Comparison on IJB-A database with natural occlusions.

Dataset	IJB-A natural occlusion			
Method ↓	Verification		Identification	
Metric (%) →	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
DR-GAN Tran et al. (2018)	64.7 ± 4.1	41.8 ± 6.4	70.8 ± 3.6	81.7 ± 2.9
CASIA-Net	64.4 ± 6.1	40.7 ± 6.8	71.3 ± 3.5	81.6 ± 2.5
Ours (CASIA-Net)	66.8 ± 3.4	48.3 ± 5.5	73.2 ± 2.5	82.3 ± 3.3
ResNet50 He et al. (2016)	**86.0** ± 1.8	64.3 ± 7.7	79.8 ± 4.2	84.9 ± 3.1
Ours (ResNet50)	**86.0** ± 1.6	**72.6** ± 5.0	**80.0** ± 3.2	**85.0** ± 3.1

Table 4.8: Comparison on IJB-C database with natural occlusions.

Dataset	IJB-C natural occlusion			
Method ↓	Verification		Identification	
Metric (%) →	@FAR=.01	@FAR=.001	@Rank-1	@Rank-5
DR-GAN Tran et al. (2018)	82.4	66.1	70.8	82.8
CASIA-Net	83.3	67.0	72.1	83.3
Ours (CASIA-Net)	83.8	69.3	74.5	83.6
ResNet50 He et al. (2016)	93.1	89.0	**87.5**	**91.0**
Ours (ResNet50)	**93.4**	**89.8**	87.4	90.7

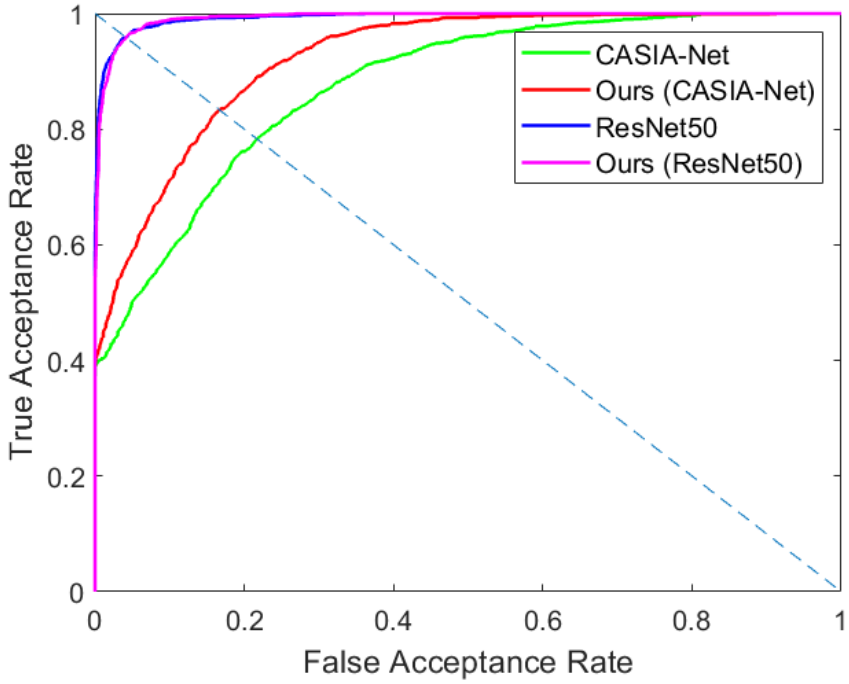


Figure 4.7: ROC curves of different models on AR database.

.

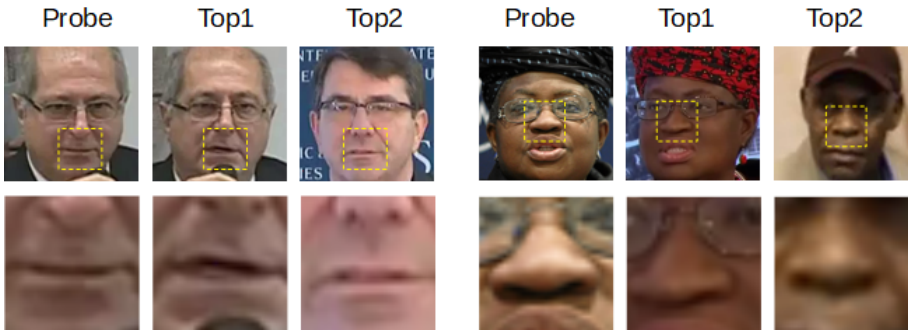


Figure 4.8: Partial face retrieval with mouth (left), and nose (right).

4.5 Other Applications

4.5.1 Partial face retrieval

In addition to the interpretable face recognition, another novel potential application of our method is partial face retrieval. Assuming we are interested in retrieving images with similar mouth, we can define“mouth filters" base on filters’ average peak location with our models, as in Fig. 4.9.

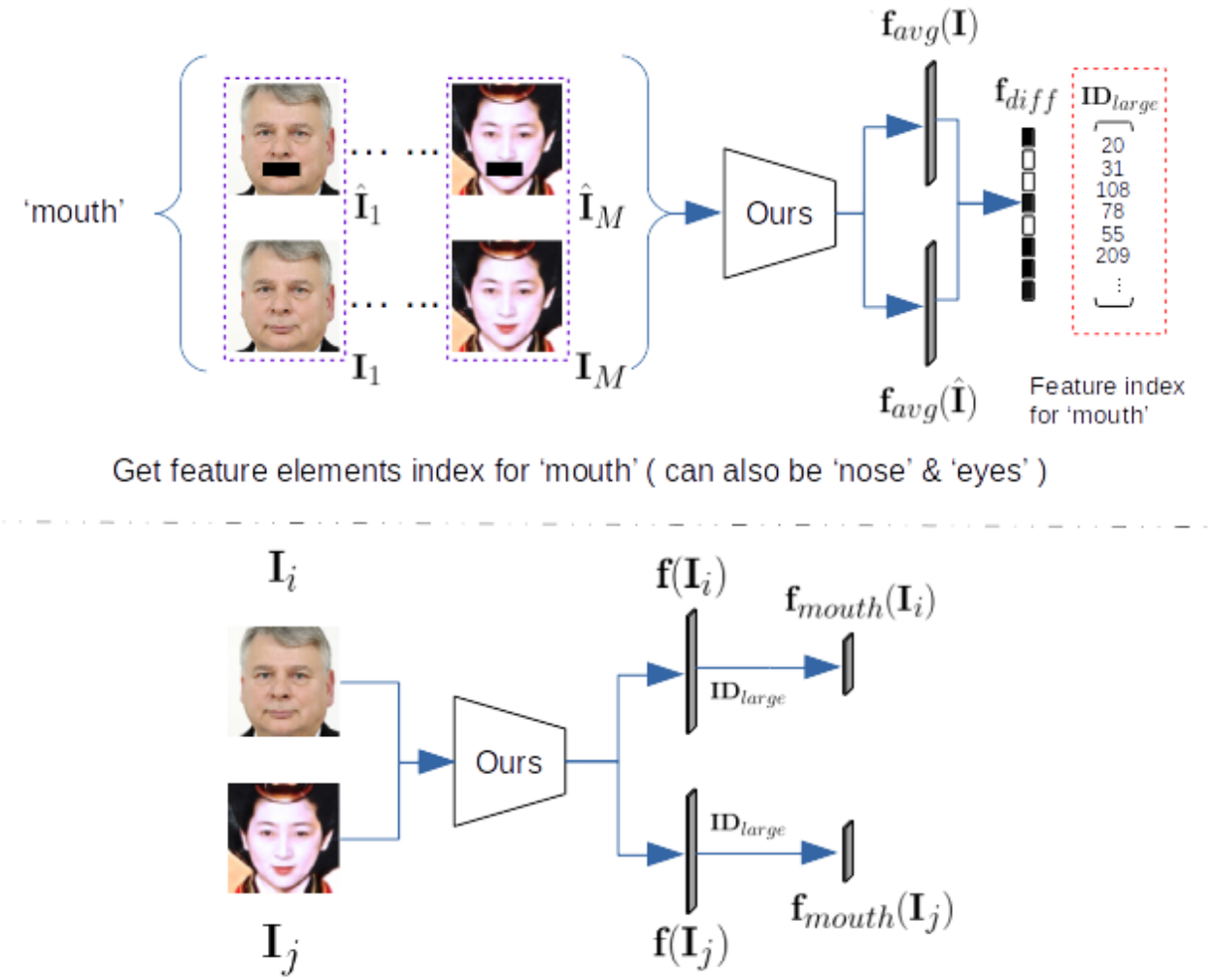


Figure 4.9: The overall framework of partial face retrieval.

Assume we have M original and occluded face pairs as the input, for all the pairs, we compute their average feature difference $\mathbf{f}_{diff} = |\mathbf{f}_{avg}(\mathbf{I}) - \mathbf{f}_{avg}(\hat{\mathbf{I}})|$, where $\mathbf{f}_{avg} = \frac{\sum_{i=1}^{M} \mathbf{f}_i}{M}$. Then we find the indexes of top $N_f - t$ elements in $\mathbf{f}_{diff}$, which we denote as $\mathbf{ID}_{large}$. After that, we can use $\mathbf{ID}_{large}$ to filter the 'mouth' related feature elements for each testing face $\mathbf{I}_i$. By giving a probe face $\mathbf{I}_i$ and the gallery faces $\mathbf{I}_j$, $j \in \{1, ..., L\}$, our model conducts the features $\mathbf{f}(\mathbf{I}_i)$ and $\mathbf{f}(\mathbf{I}_j)$. For each probe-gallery pair, the 'mouth' related features, $\mathbf{f}_{mouth}(\mathbf{I}_i)$ and $\mathbf{f}_{mouth}(\mathbf{I}_j)$, will be computed. Through applying cosine distance on those two part-based features, we can retrieve the most similar $\mathbf{I}_j$ to $\mathbf{I}_i$.

For experimental demonstration, we select one pair of images from a subset of 150 identities from IJB-A test set, to create a set of 300 images in total. Using different facial parts of each image as a query, our accuracy of retrieving the remaining image of the same subject as the top 1 result are 71%, 58% and 69% for eyes, mouth, and nose respectively. Results are visualized in Fig. 4.8, we can retrieve facial parts that are not from the same identity but visually very similar to the query part.

4.5.2 Occlusion detection

Face occluded area detection is another interesting task that we can explore. As observed from Fig. 1.1, filter responses will be weak and scattered if there exists a heavy occlusion on the region to which the filter responding. This observation can be leveraged to unsupervisely detect the existed occlusion areas. Fig. 4.10 describes our approach of occlusion detection. For each filter, its visibility is defined by three criterias: (i) distance to the average peak location, (ii) the feature activation spreadness, (iii) inverse peak value. More specifically, we would like to have a detailed discussion here about this approach.

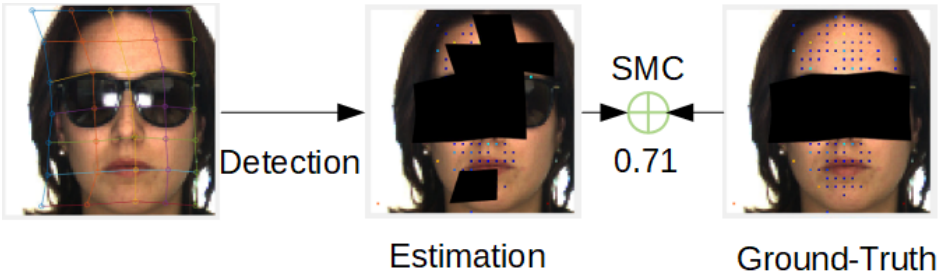


Figure 4.10: The framework of occlusion detection on AR database.

.

Assume we have N occluded images just like the leftmost face shown in Fig. 4.10. Then we create a 6×4 grid on the facial part for each occluded face. Since manually labeling those grids will be inefficiently, we firstly select a frontal face and create such grid for it, then we use the barycentric coordinates to warp the vertices of the grid to other faces within the N occluded images set. Once we get the grids, their ground truth labels can be given. When looking at the rightmost

face given by Fig. 4.10, the ground truth label is constructed as a binary form string, 1 denotes existing occlusions within a square of the grids, while 0 defines the non-occluded square. For example, the ground truth label of the face in Fig. 4.10 is "000000001111111100000000".

After obtaining labels for all the occluded faces, we should design our approach to detect the occlusions. Before that, we pair the occluded face image with another twin image, which has the similar properties except having the heavy occlusions. By utilizing the twin images, we can train a two-class classifier for the visibility. Because we have defined the criterias for its visibility of each filter, it is worthy to discuss the detailed formulations of those criterias.

Firstly, distance to the average peak location is a meaningful way to measure the visibility. Our previous experiments have shown that the standard deviations among peaks will be small by applying our two loss functions. If there is a heavy occlusion, the locations of the peak response of the filters could be scattered. We can compute a average location for each filter across N non-occluded images and then for both non-occluded images and occluded images sets, we can calculate the average distances to the average peak location for each filter. Ideally, the one computed on the occluded images will be larger.

Secondly, we can also evaluate the difference of the feature activation spreadness. In our assumption, smaller activation spreadness means stronger interpretability. We observe that the feature response of a filter for non-occluded face is concentrated on a local part, in other words, its area will be small. As for occluded face, the heavy occlusion will push the filters respond to a scatter area. Based on this observation, we compute the average area of each filter for the two sets of images. By comparing the average areas, we can get some knowledge about the difference between occluded and non-occluded images.

Thirdly, except for using the spreadness of the feature response, we now explore the property of response strength. Through looking at the value of peaks of each filters, we find that the filters tend to respond stronger to non-occluded faces than the occluded faces. To quantitatively evaluate this response strength, we select the average peak values for each filter across N images for both occluded and non-occluded sets.

The summation of the normalized scores of three criterias' output will determine the filter visibility. For each region in the 6×4 grid, the binary decision of the region's visibility is decided by majority votes from all filters it contains. One note, our output is also a binary string. As shown in Fig. 4.10, the middle face illustrates the estimated detection results, its predicted label is "001001111110111000000100". Using Simple Matching Coefficient (SMC) metrics, we can compute the coefficient of the sample image to be 0.71. On $N = 810$ faces with occlusion of AR dataset, our method achieves the averaged SMC score of 0.58.

CHAPTER 5

CONCLUSIONS

In this paper, we present our efforts towards interpretable face recognition. Our grand goal is to learn from data a structured face representation where each dimension activates on a consistent semantic face part and captures its identity information. We propose two novel losses to encourage both spatial activation diversity and feature activation diversity in the final-stage convolutional filters and the face representation. We empirically demonstrate the proposed method can lead to more locally constrained individual filter responses and overall widely-spreading filters distribution. A by-product of the harnessed interpretability is improved robustness to occlusions in face recognition.

CHAPTER 6

RESNET50

6.1 The network structure of our modified ResNet50

Table 6.1: The structures of the modified ResNet50.

Layer	Input	Filter/Stride	Output Size
conv11	Image	$3 \times 3/2$	$56 \times 56 \times 64$
MaxPool	conv11	$3 \times 3/2$	$56 \times 56 \times 64$
conv21	MaxPool	$1 \times 1/1$	$56 \times 56 \times 64$
conv22	conv21	$3 \times 3/1$	$56 \times 56 \times 64$
conv23	conv22	$1 \times 1/1$	$56 \times 56 \times 256$
conv24	conv23	$1 \times 1/1$	$56 \times 56 \times 64$
conv25	conv24	$3 \times 3/1$	$56 \times 56 \times 64$
conv26	conv25	$1 \times 1/1$	$56 \times 56 \times 256$
conv27	conv26	$1 \times 1/1$	$56 \times 56 \times 64$
conv28	conv27	$3 \times 3/1$	$56 \times 56 \times 64$
conv29	conv28	$1 \times 1/1$	$56 \times 56 \times 256$
conv31	conv29	$1 \times 1/2$	$28 \times 28 \times 128$
conv32	conv32	$3 \times 3/1$	$28 \times 28 \times 128$
conv33	conv32	$1 \times 1/1$	$28 \times 28 \times 512$
conv34	conv33	$1 \times 1/1$	$28 \times 28 \times 128$
conv35	conv34	$3 \times 3/1$	$28 \times 28 \times 128$
conv36	conv35	$1 \times 1/1$	$28 \times 28 \times 512$
conv37	conv36	$1 \times 1/1$	$28 \times 28 \times 128$
conv38	conv37	$3 \times 3/1$	$28 \times 28 \times 128$
conv39	conv38	$1 \times 1/1$	$28 \times 28 \times 512$
conv310	conv39	$1 \times 1/1$	$28 \times 28 \times 128$
conv311	conv310	$3 \times 3/1$	$28 \times 28 \times 128$
conv312	conv311	$1 \times 1/1$	$28 \times 28 \times 512$
conv41	conv312	$1 \times 1/2$	$14 \times 14 \times 256$
conv42	conv41	$3 \times 3/1$	$14 \times 14 \times 256$
conv43	conv42	$1 \times 1/1$	$14 \times 14 \times 1024$
conv44	conv43	$1 \times 1/1$	$14 \times 14 \times 256$
conv45	conv44	$3 \times 3/1$	$14 \times 14 \times 256$
conv46	conv45	$1 \times 1/1$	$14 \times 14 \times 1024$
conv47	conv46	$1 \times 1/1$	$14 \times 14 \times 256$
conv48	conv47	$3 \times 3/1$	$14 \times 14 \times 256$
conv49	conv48	$1 \times 1/1$	$14 \times 14 \times 1024$
conv410	conv49	$1 \times 1/1$	$14 \times 14 \times 256$
conv411	conv410	$3 \times 3/1$	$14 \times 14 \times 256$
conv412	conv411	$1 \times 1/1$	$14 \times 14 \times 1024$
conv413	conv412	$1 \times 1/1$	$14 \times 14 \times 256$
conv414	conv413	$3 \times 3/1$	$14 \times 14 \times 256$
conv415	conv414	$1 \times 1/1$	$14 \times 14 \times 1024$
conv416	conv415	$1 \times 1/1$	$14 \times 14 \times 256$
conv417	conv416	$3 \times 3/1$	$14 \times 14 \times 256$
conv418	conv417	$1 \times 1/1$	$14 \times 14 \times 1024$
conv51	conv418	$1 \times 1/2$	$7 \times 7 \times 512$
conv52	conv51	$3 \times 3/1$	$7 \times 7 \times 512$
conv53	conv52	$1 \times 1/1$	$7 \times 7 \times 2048$
conv54	conv53	$1 \times 1/2$	$7 \times 7 \times 512$
conv55	conv54	$3 \times 3/1$	$7 \times 7 \times 512$
conv56	conv55	$1 \times 1/1$	$7 \times 7 \times 2048$
conv57	conv56	$1 \times 1/2$	$7 \times 7 \times 512$
conv58	conv57	$3 \times 3/1$	$7 \times 7 \times 512$
conv59	conv58	$1 \times 1/1$	$7 \times 7 \times N^f$
conv418-U	conv418	upsampling	$28 \times 28 \times 1024$
conv419	conv43-U	$1 \times 1/1$	$28 \times 28 \times 512$
conv59-U	conv59	upsampling	$28 \times 28 \times 512$
conv510	conv510-U	$1 \times 1/1$	$28 \times 28 \times 512$
Φ (HC)	conv312,419,510	$3 \times 3/1$	$28 \times 28 \times 1536$
Ψ	Φ	$3 \times 3/1$	$28 \times 28 \times N^f$
AvgPool	Ψ	$28 \times 28/1$	$1 \times 1 \times N^f$

As shown in Tab.6.1, for ResNet50, the image input size will change to 112×112. And the final dimention of the feature representaion will also be $N^f = 512$. And we take the same block

setting as described in Deng et al. (2018), which is shown in Fig. 6.1. This more advanced residual unit, which has a BN-Conv-BN-PReLu-Conv-BN structure, has been proved to be efficient in face reocognition.

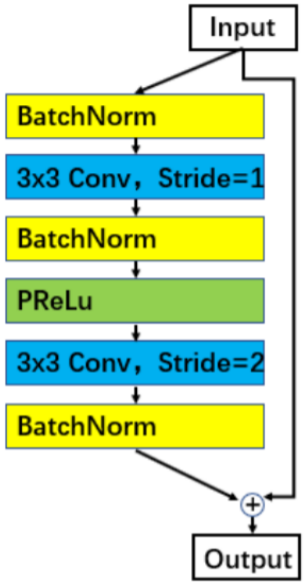


Figure 6.1: The block setting.

.

BIBLIOGRAPHY

BIBLIOGRAPHY

Ahonen, Timo, Abdenour Hadid & Matti Pietikainen. 2006. Face description with local binary patterns: Application to face recognition. *TPAMI* .

Berg, Thomas & Peter N Belhumeur. 2013. Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *Cvpr*, .

Brianna Maze, Jocelyn Adams, Nathan Kalka James A. Duncan, Charles Otto Tim Miller, W. Tyler Niggel Anil K. Jain, Jordan Cheney Janet Anderson & Patrick Grother. 2018. IARPA Janus Benchmark-C: Face dataset and protocol. In *Icb*, .

Cao, Qiong, Li Shen, Weidi Xie, Omkar M. Parkhi & Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *Fg*, .

Cao, Zhimin, Qi Yin, Xiaoou Tang & Jian Sun. 2010. Face recognition with learning-based descriptor. In *Cvpr*, .

Cao2, Kaidi, Yu Rong1, Cheng Li, Xiaoou Tang & Chen Change Loy. 2018. Pose-robust face recognition via deep residual equivariant mapping. In *Cvpr*, .

Chai, Xiujuan, Shiguang Shan, Xilin Chen & Wen Gao. 2007. Locally linear regression for pose-invariant face recognition. *TIP* .

Chen, Xi, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever & Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Nips*, .

Cheng, Lele, Jinjun Wang, Yihong Gong & Qiqi Hou. 2015. Robust deep auto-encoder for occluded face recognition. In *Icm*, .

Das, Abhishek, Harsh Agrawal, Larry Zitnick, Devi Parikh & Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *CVIU* .

Deng, Jiankang, Jia Guo & Stefanos Zafeiriou. 2018. Arcface: Additive angular margin loss for deep face recognition. In *arxiv preprint arxiv:1801.07698*, .

Felzenszwalb, Pedro, David McAllester & Deva Ramanan. 2008. A discriminatively trained, multiscale, deformable part model. In *Cvpr*, .

Ge, Shiming, Jia Li, Qiting Ye1 & Zhao Luo1. 2017. Detecting masked faces in the wild with lle-cnns. In *Cvpr*, .

Goodfellow, Ian J, Jonathon Shlens & Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* .

Guo, Yandong, Lei Zhang, Yuxiao Hu, Xiaodong He & Jianfeng Gao. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Eccv*, .

Hariharan, Bharath, Pablo Arbeláez, Ross Girshick & Jitendra Malik. 2015. Hypercolumns for object segmentation and fine-grained localization. In *Cvpr*, .

He, Kaiming, Xiangyu Zhang, Shaoqing Ren & Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Iccv*, .

He, Kaiming, Xiangyu Zhang, Shaoqing Ren & Jian Sun. 2016. Deep residual learning for image recognition. In *Cvpr*, .

Iandola, Forrest, Matt Moskewicz, Sergey Karayev, Ross Girshick, Trevor Darrell & Kurt Keutzer. 2014. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869* .

Jourabloo, Amin & Xiaoming Liu. 2017. Pose-invariant face alignment via CNN-based dense 3D model fitting. *IJCV* .

Juneja, Mayank, Andrea Vedaldi, CV Jawahar & Andrew Zisserman. 2013. Blocks that shout: Distinctive parts for scene classification. In *Cvpr*, .

Klare, Brendan F, Ben Klein, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah & Anil K Jain. 2015. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark-A. In *Cvpr*, .

Kumar, Neeraj, Alexander C Berg, Peter N Belhumeur & Shree K Nayar. 2009. Attribute and simile classifiers for face verification. In *Iccv*, .

Learned-Miller, Erik, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li & Gang Hua. 2016. Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis*, .

Li, Haoxiang & Gang Hua. 2017. Probabilistic elastic part model: a pose-invariant representation for real-world face verification. *TPAMI* .

Li, Haoxiang, Gang Hua, Zhe Lin, Jonathan Brandt & Jianchao Yang. 2013. Probabilistic elastic matching for pose variant face verification. In *Cvpr*, .

Li, Stan Z, Xin Wen Hou, Hong Jiang Zhang & Qian Sheng Cheng. 2001. Learning spatially localized, parts-based representation. In *Cvpr*, .

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He & Piotr Dollár. 2017. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002* .

Liu, Feng, Dan Zeng, Qijun Zhao & Xiaoming Liu. 2018. Disentangling features in 3D face shapes for joint face reconstruction and recognition. In *Cvpr*, .

Liu, Yaojie, Amin Jourabloo, William Ren & Xiaoming Liu. 2017a. Dense face alignment. In *Iccv workshop*, .

Liu, Yu, Hongyang Li & Xiaogang Wang. 2017b. Learning deep features via congenerous cosine loss for person recognition. *arXiv preprint arXiv:1702.06890* .

Lowe, David G. 2004. Distinctive image features from scale-invariant keypoints. *IJCV* .

Lu, Chaochao & Xiaoou Tang. 2015. Surpassing human-level face verification performance on LFW with gaussianface. In *Aaai*, .

Mahendran, Aravindh & Andrea Vedaldi. 2016. Visualizing deep convolutional neural networks using natural pre-images. *IJCV* .

Martinez, Aleix M. 1998. The AR face database. *CVC Technical Report24* .

Nguyen, Anh, Jason Yosinski & Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Cvpr*, .

Novotny, David, Diane Larlus & Andrea Vedaldi. 2017. Anchornet: A weakly supervised network to learn geometry-sensitive features for semantic matching. In *Cvpr*, .

Olah, Chris, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye & Alexander Mordvintsev. 2018. The building blocks of interpretability. *Distill* .

OâĂŹToole, Alice J., Carlos D. Castillo, Connor J. Parde, Matthew Q. Hill & Rama Chellappa. 2018. Face space representations in deep convolutional neural networks. *Trends in cognitive sciences* .

Parikh, Devi & C Zitnick. 2011. Human-debugging of machines. *NIPS WCSSWC* .

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh & Dhruv Batra. 2016. Grad-cam: Visual explanations from deep networks via gradient-based localization. *https://arxiv. org/abs/1610.02391 v3* .

Shen1, Yujun, Ping Luo1, Junjie Yan, Xiaogang Wang & Xiaoou Tang. 2018. Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis. In *Cvpr*, .

Shu, Zhixin, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman & Dimitris Samaras. 2017. Neural face editing with intrinsic image disentangling. *arXiv preprint arXiv:1704.04131* .

Singh, Krishna Kumar & Yong Jae Lee. ???? Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Iccv*, .

Singh, Saurabh, Abhinav Gupta & Alexei A Efros. 2012. Unsupervised discovery of mid-level discriminative patches. In *Eccv*, .

Sudderth, Erik B, Antonio Torralba, William T Freeman & Alan S Willsky. 2005. Learning hierarchical models of scenes, objects, and parts. In *Iccv*, .

Tran, Luan & Xiaoming Liu. 2018. Nonlinear 3D face morphable model. In *Cvpr*, .

Tran, Luan, Xiaoming Liu, Jiayu Zhou & Rong Jin. 2017a. Missing modalities imputation via cascaded residual autoencoder. In *Cvpr*, .

Tran, Luan, Xi Yin & Xiaoming Liu. 2017b. Disentangled representation learning GAN for pose-invariant face recognition. In *Cvpr*, .

Tran, Luan, Xi Yin & Xiaoming Liu. 2018. Representation learning by rotating your faces. *TPAMI* .

Vondrick, Carl, Aditya Khosla, Tomasz Malisiewicz & Antonio Torralba. 2013. Hoggles: Visualizing object detection features. In *Iccv*, .

Wang, Xiaolong, Abhinav Shrivastava & Abhinav Gupta. 2017. A-fast-rcnn: Hard positive generation via adversary for object detection. In *Cvpr*, .

Wen, Yandong, Kaipeng Zhang, Zhifeng Li & Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *Eccv*, .

Xie, Weidi & Andrew Zisserman. 2018. Multicolumn networks for face recognition. In *arxiv preprint arxiv:1807.09192*, .

Xu, Kelvin, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel & Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Icml*, .

Yao, Bangpeng, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas & Li Fei-Fei. 2011. Human action recognition by learning bases of action attributes and parts. In *Iccv*, .

Yi, Dong, Zhen Lei, Shengcai Liao & Stan Z Li. 2014. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* .

Yin, Xi & Xiaoming Liu. 2018. Multi-task convolutional neural network for pose-invariant face recognition. *IEEE Transactions on Image Processing* .

Yin, Xi, Xiang Yu, Kihyuk Sohn, Xiaoming Liu & Manmohan Chandraker. 2017. Towards large-pose face frontalization in the wild. In *Iccv*, .

Zeiler, Matthew D & Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Eccv*, .

Zeiler, Matthew D, Graham W Taylor & Rob Fergus. 2011. Adaptive deconvolutional networks for mid and high level feature learning. In *Iccv*, .

Zhang, Quanshi, Ying Nian Wu & Song-Chun Zhu. 2017. Interpretable convolutional neural networks. *arXiv preprint arXiv:1710.00935* .

Zhao, Jian, Yu Cheng, Yi Cheng, Yang Yang, Haochong Lan, Fang Zhao, Lin Xiong, Yan Xu, Jianshu Li, Sugiri Pranata, Shengmei Shen, Junliang Xing, Hengzhu Liu, Shuicheng Yan1 & Jiashi Feng. 2018. Look across elapse: Disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. In *arxiv preprint arxiv:1809.00338*, .

Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva & Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Cvpr*, .

Zhou, Erjin, Zhimin Cao & Qi Yin. 2015. Naive-deep face recognition: Touching the limit of LFW benchmark or not? *arXiv preprint arXiv:1501.04690* .