CQA-Face: Contrastive Quality-aware Attentions for Face Recognition

Qiangchang Wang,[1] **Guodong Guo,**[1,2,3*]

[1]Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, USA
[2]Institute of Deep Learning, Baidu Research, Beijing, China
[3]National Engineering Laboratory for Deep Learning Technology and Application, Beijing, China
qiangchang.wang@gmail.com, guoguodong01@baidu.com

Abstract

Few existing face recognition (FR) models take local representations into account. Although some works achieved this by extracting features on cropped parts around face landmarks, landmark detection may be inaccurate or even fail in some extreme cases. Recently, without relying on landmarks, attention-based networks can focus on useful parts automatically. However, there are two issues: 1) It is noticed that these approaches focus on few facial parts, while missing other potentially discriminative regions. This can cause performance drops when emphasized facial parts are invisible under heavy occlusions (e.g. face masks) or large pose variations; 2) Different facial parts may appear at various quality caused by occlusion, blur, or illumination changes. In this paper, we propose contrastive quality-aware attentions, called CQA-Face, to address these two issues. First, a Contrastive Attention Learning (CAL) module is proposed, pushing models to explore comprehensive facial parts. Consequently, more useful parts can help identification if some facial parts are invisible. Second, a Quality-Aware Network (QAN) is developed to emphasize important regions and suppress noisy parts in a global scope. Thus, our CQA-Face model is developed by integrating the CAL with QAN, which extracts diverse quality-aware local representations. It outperforms the state-of-the-art methods on several benchmarks, demonstrating its effectiveness and usefulness.

Introduction

Face recognition (FR) has many practical applications. Previous works can be divided into two categories: global-based approaches and local-based methods. The former learns features on global face images (Deng et al. 2019; Cao et al. 2020). However, they rarely consider representations on local patches to improve discrimination of face features. Three positive pairs are shown in Fig. 1 (a)&(f), the holistic faces change dramatically by blur changes in (1), pose variations in (2), and age gaps in (3). However, some local patches remain similar which can help the verification. For example, the similar eyes in (1) or the similar noes in (2)&(3).

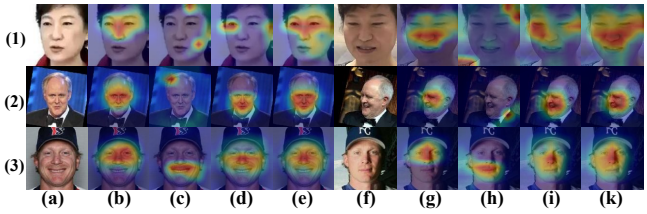

Figure 1: Effects of the CQA-Face model on three positive pairs. **(a)&(f)**: Input faces are affected by blur changes **(1)**, pose variations **(2)**, and aging **(3)**. **(b-d)&(g-i)**: 1st, 2nd, and 3rd local branches of local CNNs where different local patches are located in each local branch. **(e)&(k)**: Discriminative local patches are emphasized in local CNNs. For a good illustration, only three local branches are used to show class activation maps (CAMs) (Zhou et al. 2016).

There are mainly two groups of methods to extract discriminative local representations: landmark-based and attention-based methods. Landmark-based works detected face landmarks first and then extract local features on cropped regions centered around landmarks (Sun, Wang, and Tang 2014a; Ding and Tao 2017; Kang, Kim, and Kim 2018). However, these methods largely depend on accuracy of landmark detection and may even suffer from detection failure under dramatic pose variations or heavy occlusions. For example, due to the ongoing outbreak of the COVID-19, people are wearing face masks. Because some facial parts (e.g. noses or mouths) are invisible, landmark detection may be inaccurate or failed. Besides, even if face landmarks are detected, the cropped parts may include masks which would inevitably deteriorate the extracted face features. Without relying on landmarks, some models employ attention modules to automatically locate useful parts (Wang and Guo 2019; Kang et al. 2019). However, no mechanism is designed to select the local patches effectively, which may miss some important facial parts, and thus limit the performance.

FR is challenging mainly because of two reasons: 1) High intra-class variations. Faces from the same subject may appear at different poses or occlusion levels. In such a case, if only few facial parts are located, the extracted features may not be sufficient when these parts are invisible under occlusions (e.g. face masks) or pose variations. 2) Small inter-

*Corresponding author.
†This work was done when the first author was an intern at IDL, Baidu Research.

class differences. Faces from different subjects may have similar local appearances, especially considering a large number of subjects. The performance would decline if the few emphasized facial parts across different subjects look similar. To alleviate the above issues, it is necessary to capture local representations that are as rich as possible. In such a way, more potentially useful facial parts can contribute to FR if the few emphasized parts are invisible or remain similar across different subjects. To achieve this goal, a contrastive attention learning (CAL) module is devised to encourage the diversity among different attention maps. Consequently, it is ensured that varying local patches are well-explored across face images and diverse discriminative facial parts are emphasized.

Local representations can provide discriminative information for FR. However, concatenating local representations directly without considering relations between different facial parts may not be optimal, as discussed in previous paragraph. To address the limitations of separated facial parts, the structural correlation of facial parts is built to boost the discriminative ability of local representations. Besides, it is observed that different facial parts have various quality under occlusion, blur, or pose variations, as illustrated in Fig. 1 (b-d)&(g-i). The performance would be deteriorated if concatenating these local features directly. To address these two issues, we devise a quality-aware network (QAN). It introduces a part-level quality-aware module to explore the relation between an individual facial part and the rest parts by using graph convolutional networks. In such a manner, it has two benefits. First, each local part itself and its relations with other parts are considered simultaneously. This encourages each part-level feature to utilize information from other parts, making them more discriminative. Second, it can estimate the quality of located facial parts, emphasizing informative parts and suppressing noisy ones.

By putting the CAL and QAN together, a CQA-Face model is developed to learn rich quality-aware local representations. Our main contributions are three-fold:

1. A contrastive attention learning (CAL) is designed to ensure the localization of comprehensive facial parts, especially the less discriminative but still useful facial parts;

2. A region-level quality-aware network (QAN) is proposed to generate quality scores for each facial part by exploring its relation with the rest, emphasizing important facial parts and suppressing noisy ones;

3. By combining the CAL and QAN, the CQA-Face model is developed, which outperforms the state-of-the-art methods on many challenging datasets.

Related Works

Related works about face recognition and quality-aware networks are reviewed briefly.

Face Recognition

Most existing FR works trained networks on global face images (Deng et al. 2019; Wang et al. 2019; Cao et al. 2020). However, they may suffer from performance drops when tackling faces taken in challenging cases. Several approaches (Sun, Wang, and Tang 2014b; Sun et al. 2014; Xie, Shen, and Zisserman 2018; Ding and Tao 2017) trained models on cropped parts around face landmarks. However, landmark detection may be unreliable in many hard cases.

Rather than relying on landmarks, recent attention-based networks weighed varying local patches adaptively. (Wang and Guo 2019) was an early effort to apply attentions to automatically emphasize important local patches and suppress noisy regions. (Kang et al. 2019) used attention scores to capture the unique local pair relations. However, multiple attention maps may have similar responses around few facial parts. To address this problem, (Wang et al. 2020a) maximized pairwise attention distances. However, these approaches typically ignore the mining of facial structure patterns. Different from these, our CQA-Face model learns quality scores for different facial parts, highlighting discriminative parts and suppressing useless parts.

Quality-aware Networks

A quality-aware network was proposed in (Liu, Yan, and Ouyang 2017) to learn a quality score automatically for each sample in a set. Features and quality scores for all samples in a set are fused to output set-level features. A dependency-aware attention control network was presented in (Liu et al. 2018) to fully exploit correlations among orderless images within a set. An aggregation network was proposed in (Gong, Shi, and Jain 2019) to fuse representations of frames in a video based on quality and context information. A region-based quality estimation network aggregated complementary region-based features in a sequence (Song et al. 2017). An attention-aware method was introduced in (Rao, Lu, and Zhou 2017) for video face recognition where deep reinforcement learning is used to discard the uninformative frames and focus on the important ones. However, these approaches aim at set-level or video-level FR problems. A FR network was proposed in (Liu and Tan 2021) which gives a quality score along with a feature vector. This is the first work to generate an explicit quantitative quality score for a face image. In contrast, our CQA-Face model is capable of solving the image-based FR task by learning a quality score for each facial part.

Methods

The overall framework of our approach is shown in Fig. 2, mainly composed of the stem CNN, Contrastive Attention Learning (CAL), and Quality-Aware Networks (QAN). The stem CNN extracts high-level feature maps. Because HSNet-61 model (Wang and Guo 2019) has an excellent generalization ability, it is used as the default stem CNN.

If without a proper signal, multiple attention maps tend to focus on few facial parts, while neglecting other important parts. Motivated by this observation, the CAL is devised to guide multiple local branches to extract diverse local representations based on high-level feature maps.

However, since comprehensive image details are explored thoroughly across an image, it is possible that some distracted regions (e.g. face masks) may also be represented. To alleviate this, the QAN is proposed to learn quality scores among different local patches from a global scope. Finally,

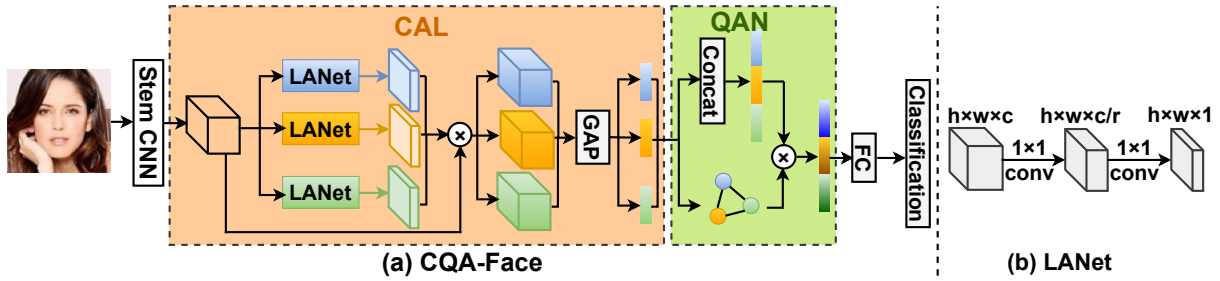

Figure 2: **(a) CQA-Face**: The CAL pushes models to explore comprehensive local representations. The QAN learns a quality score for each local patch in a global scope. GAP and FC refer to the global average pooling and fully-connected layers, respectively. Notice that only three local branches are used for illustration. **(b) LANet**: The spatial attention in (Wang and Guo 2019) is used where h, w, and c is the height, width, and number of channels. r is the reduction ratio.

diverse quality-aware local representations are concatenated for face matching. Details are illustrated as follows.

Contrastive Attention Learning

Existing attention-based methods do not perform localization of less discriminative but still useful ones. If models focus on few facial parts, they would suffer from performance drop if facial parts are invisible or remain similar across different subjects. To address this, a contrastive attention learning (CAL) mechanism is proposed to encourage diversity in multiple attention maps. Consequently, learned models can retrieve comprehensive local patches across face images.

Suppose that $X \in R^{\mathrm{h}\times\mathrm{w}\times\mathrm{c}}$ denote the input of the CAL, where h, w, and c refer to the height, width, and number of channels, respectively. Next, it is expected to learn diverse attention maps in different local branches, i.e. $[M_1, M_2, ..., M_\mathrm{b}]$, where b is the number of local branches. As shown in Fig. 2 (b), the spatial attention (i.e. LANet) in (Wang and Guo 2019) is used to weigh different regions, assigning high weights to important parts and small weights to useless ones. There are two convolutional layers in LANet in which the first layer uses the ReLU function and the second layer adopts the Sigmoid function.

For the i^{th} local branch, the output O_i is calculated by the product of the attention map M_i and input X as follows:

$$O_i = X \circ M_i, \tag{1}$$

where $\circ$ denotes Hadamard product.

However, if without a proper guidance, multiple attention maps $[O_1, O_2, ..., O_\mathrm{b}]$ tend to have similar responses around the same facial parts, limiting the representational ability.

To address this issue, we propose a contrastive attention learning (CAL) using a divergence loss as follows:

$$L_{\mathrm{CAL}} = \frac{2}{b(b-1)} \sum_{i=1}^{b} \sum_{j=i+1}^{b} max(0, -t + exp^{-(\mathrm{M_i}-\mathrm{M_j})^2/\sigma}), \tag{2}$$

where M_i and M_j mean attention maps in the i^{th} and j^{th} local branches, respectively. t is a hyper-parameter margin. σ is a positive value to control the shape of the Gaussian function.

The loss L_{CAL} measures the correlation between M_i and M_j. When M_i and M_j are similar, L_{CAL} tends to be large, pushing M_i and M_j to be dissimilar. In such a way, M_i and M_j would have responses around different facial parts. Comprehensive facial parts are well-explored across face images. Consequently, models can rely on more useful parts if some parts are invisible for faces with occlusions or pose variations.

After that, a global average pooling (GAP) layer is used to pool feature maps $[O_1, O_2, ..., O_\mathrm{b}]$ in different local branches, generating local representations in each local branch $[P_1, P_2, ..., P_\mathrm{b}]$.

It should be mentioned that the diverse learning proposed in (Wang et al. 2020a) can also extract diverse local patches. The divergence loss is defined as follows:

$$L_{\mathrm{D}} = \frac{2}{b(b-1)} \sum_{i=1}^{b} \sum_{j=i+1}^{b} max(0, t - dist(M_i, M_j)), \tag{3}$$

where $dist(M_\mathrm{i}, M_\mathrm{j})$ is the Euclidean distance between M_i and M_j. As a result, varying attention maps are encouraged to be different from each other, and thus focus on different facial parts. Compared with L_{D} (Wang et al. 2020a), our proposed L_{CAL} is more flexible to measure distances of two attention maps. Experimental results also verify the superiority of our L_{CAL}.

Quality-aware Networks

Local patches appear at facial images can be divided into three categories: Few important local patches (e.g. the nose, mouth, both eyes, or mouth) which can significantly contribute to FR; Some unimportant regions, like the cheek or forehead; Distracted background or occlusions which may be represented. This is because these parts are close to important parts (e.g. eyes) in profile or masked faces. It is highly desired that discriminative parts should be emphasized and distracted regions should be suppressed.

The attention module automatically determines which regions should be highlighted. However, few existing approaches design a mechanism to refine attention responses. Consequently, some distracted regions may be extracted, such as noisy background or face masks, which deteriorate the performance. This issue is especially serious when the

CAL is used, because every face detail is explored to extract comprehensive features. For example, for masked face matching, because face masks are very close to discriminative parts, noisy information tends to be represented.

Therefore, there are two issues: Local patches may exhibit varying quality scores; It is necessary to infer the relations between one facial part and others because the relations provide important clues to refine attention maps. To alleviate these issues, a quality-aware network (QAN) is designed. Specifically, it is implemented by a graph convolutional network where relationships between different local patches are captured and quality scores are learned simultaneously.

Since each local branch generates a local patch, the number of local patches is b. Assume that $\mathcal{G}(\mathcal{V}, \xi)$ is the constructed graph with b nodes where $\mathcal{V} = [P_1, P_2, ..., P_{\mathrm{b}}]$ and the edge ξ models the relation between two patches. The adjacent matrix $A \in R^{\mathrm{b}\times\mathrm{b}}$ represents pairwise relations of local patches.

Inspired by the work (Wang and Gupta 2018), pairwise relations between every two local patches are represented as follows:

$$e(P_{\mathrm{i}}, P_{\mathrm{j}}) = \phi(P_{\mathrm{i}})^{\top}\varphi(P_{\mathrm{j}}), \tag{4}$$

where ϕ and φ represent two different projects of local features. More specifically, $\phi = W^{\phi}P_{\mathrm{i}}$ and $\varphi = W^{\varphi}P_{\mathrm{j}}$ where $W^{\phi}, W^{\varphi} \in R^{\mathrm{c/r}\times\mathrm{c}}$. r is a dimension reduction ratio to reduce the number of parameters. W^{ϕ} and W^{φ} are implemented by two 1×1 convolutional layers followed by batch normalization (BN) and ReLU operations. In such a manner, the transformations allow models to learn correlations between different local patches adaptively.

The element $A_{i,j}$ in adjacency matrix A represents the relation between P_i and P_j. For the i^{th} feature node, pairwise relations with all the nodes are stacked in a fixed order to obtain a relation vector $A_i = [A(i,:), A(:,i)] \in R^{2\mathrm{b}}$. Since the relations are stacked into a vector with a fixed order, spatial information is also represented in the relation vector A_i.

Each P_i represents original local features, while A_i denotes the structural relations. They complement and reinforce each other but in different embedding spaces. Therefore, we combine them in their respective embedding space and jointly learn an importance S_i of feature node P_i by the following formulation:

$$S_i = \theta([\mu(P_i), \nu(A_i)]), \tag{5}$$

where μ and ν represent projection functions for P_i and A_i, respectively. Specifically, they consist of a spatial 1×1 convolutional layer followed by BN and ReLU operations. Then, an embedding function is conducted to mine rich information from them for inferring quality scores through two 1×1 convolutional layers.

For all nodes, we have a quality score list $S = [S_1, S_2, ..., S_b]$. We normalize the quality scores by the Sigmoid function to obtain normalized quality scores $\hat{S} = [\hat{S_1}, \hat{S_2}, ..., \hat{S_b}]$ as follows:

$$\hat{S_i} = \frac{1}{1 + e^{-\mathrm{S_i}}}. \tag{6}$$

Then, each local representation is multiplied by its corresponding quality score as follows:

$$\hat{P_i} = \hat{S_i} \cdot P_i. \tag{7}$$

For an image, we first concatenate local representations $[\hat{P_1}, \hat{P_2}, ..., \hat{P_b}]$ from b local branches, which are followed by a FC layer with 512 units before the classification layer.

Overall Loss

L_{Cls} is the CosFace loss (Wang et al. 2018), guiding models to learn discriminative features. CosFace loss is formulated as follows:

$$L_{\mathrm{Cls}} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(\cos(\theta_{y_i,i})-m)}}{e^{s(\cos(\theta_{y_i,i})-m)} + \sum_{j\neq y_i} e^{s\cos\theta_{j,i}}}, \tag{8}$$

where N represents the number of samples. m is the cosine margin to maximize the decision margin in the angular space. The sample x_i is normalized and re-scaled to s, belonging to the y_i class.

The overall loss is defined as follows:

$$L_{\mathrm{Overall}} = L_{\mathrm{Cls}} + \lambda L_{\mathrm{CAL}}, \tag{9}$$

where L_{CAL} encourages multiple attention maps to locate diverse facial parts, learning comprehensive local representations. λ is a coefficient to control the balance in these two losses.

Experiments

Data

Training and test faces are aligned via MTCNN (Zhang et al. 2016). VGGFace2 (Cao et al. 2018) and MS-Celeb-1M (Guo et al. 2016) are used as the training data. The former has 3.14M faces of 8,631 identities which is the default training data. The MS1MV2 version (Deng et al. 2019) contains 5.8M images of 85K subjects by removing noisy labels in the original MS-Celeb-1M. In testing, we evaluate our models on various challenging tasks: 1) Cross-quality face matching on IJB-A quality (Guo and Zhang 2018); 2) Cross-pose face matching on CPLFW (Zheng and Deng 2018) and CFP-FP (Sengupta et al. 2016); 3) Cross-age face matching on CALFW (Zheng, Deng, and Hu 2017); 4) General face matching on LFW (Huang et al. 2008). For more details about test datasets, please refer to the related references.

We also consider masked face matching which is important under the global outbreak of the COVID-19. Real-world masked face recognition dataset (RMFRD) and masked LFW (MLFW) dataset in (Wang et al. 2020b) are used. The RMFRD contains 1,945 masked and 80,577 normal faces from 403 common identities, with two sub-tasks designed. First, **Masked2Normal (M2N)** means matching between masked faces and normal faces, which has verification and identification protocols. In verification, each masked face is matched with every normal face, generating 394K positive pairs and 156M negative pairs. Performances are reported when FAR$=0.1$, 0.01, and 0.001, respectively. In identification, the gallery set consists of features which are computed by averaging features of normal faces within the same

Table 1: Effectiveness of different modules. L_D refers to the diverse learning in (Wang et al. 2020a).

CAL	QAN	L_D	LFW	CALFW	CPLFW
			99.35	92.55	89.55
√			99.47	92.90	89.85
	√	√	99.33	93.20	89.82
√	√		**99.52**	**93.37**	**90.20**

subject. Every masked face is employed as the probe face and Rank-1 accuracy is reported. Second, **Masked2Masked (M2M)** refers to matching between each masked face and every other masked faces. There are 3.76M negative pairs and 20K positive pairs. TAR (True Accept Rate) values are reported when FAR= 0.1, 0.01, and 0.001, respectively. In **MLFW**, a method was used to generate synthetic masked faces for face images in LFW (Wang et al. 2020b). Because some faces are failed to wear masks, there are 13,175 masked faces from 5,749 subjects. Consequently, the verification protocol has 5,955 pairs in MLFW. We measure the performance by the accuracy and TAR values when FAR= 0.1, 0.01, and 0.001, respectively.

Implementation Details

CosFace loss (Wang et al. 2018) is used. The number of warm-up epochs is 2. The batch-size is set to 256. During training on VGGFace2, learning rate starts at 0.03 and is divided by 10 at the 7^{th} and 10^{th} epochs, respectively. The learning rate is set to 1e-4 at the 12^{th} epoch. Training stops at the 12^{th} epoch. During training on MS1MV2, learning rate is 0.03 and is divided by 10 at the 13^{th} and 19^{th} epochs, respectively. It is set to 1e-4 at the 23^{rd} epoch. Training stops at the 24^{th} epoch. The ResNet-100 used as the stem CNN.

After comparative experiments, the number of local branches (b) is set to 4. The σ and t in Eqn. (2) are set to 0.01 and 0.2, respectively. The λ in Eqn. (9) is 0.5.

Ablation Study

In ablation study, experiments are conducted, as shown in Tables 1 and 2 where a tuple (-, -, -) refers to results on LFW, CALFW, and CPLFW, respectively.

Effects of the proposed components. The proposed CQA-Face model mainly consists of two components: the CAL and QAN. Their performances are shown in Table 1.

Without an explicit guidance, multiple attention maps tend to focus on only few facial parts and miss other important regions. As shown in Fig. 3 (b)&(g), attention-based methods can learn some important face parts automatically. However, since there is not an explicit signal, some important regions are ignored. To alleviate this problem, the CAL is proposed to locate diverse local patches by encouraging distances between each two attention maps. For face pairs from the same subject, it is possible that some local patches are distinctive across large age gaps. This is the reason why the accuracy is boosted from 92.55% to 92.90% on CALFW. For profile faces, some facial parts are occluded due to viewpoint changes. This would deteriorate the performance if only few facial parts are located. In contrast,

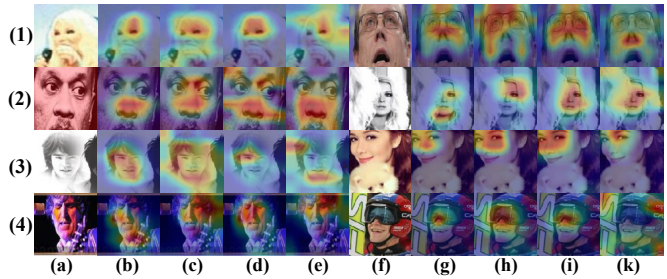

Figure 3: Effects of the CQA-Face model on several challenging faces. **(a)&(f):** Input faces which are affected by illumination changes, occlusions, or pose variations where MTCNN fails to detect landmarks. **(b)&(g):** Without an explicit signal, multiple attention maps emphasize few face parts, while missing some important parts. **(c)&(h):** The CAL can locate rich local patches. However, different located regions may exhibit different discriminative ability, like background and snow goggles. **(d)&(i):** The QAN can emphasize important face parts and suppress noisy regions. **(e)&(k):** Located regions if the L_D in Eqn. (3) is used.

the CAL can generate diverse local patches. Consequently, other important regions can help the matching besides occluded ones, leading to a moderate improvement on CPLFW which is from 89.55% to 89.85%. Since rich local representations are mined across face images, the performance on LFW is slightly increased from 99.35% to 99.47%. Besides, we also qualitatively show localization ability of diverse local patches in Fig. 1 (b-d)&(g-i). Guided by the CAL, diverse discriminative local patches are located.

While the CAL generates diverse local patches by exploring every image detail, this may raise two questions: Noisy information may be extracted, like face masks; Different facial parts may appear at various quality under occlusion, blur, or illumination changes. For instances, since background is very close to discriminative parts for profile faces, noisy information tends to be represented. The QAN alleviates these issues by learning a quality score for each face part automatically from a global view. As shown in Fig. 4, discriminative local patches are assigned with large quality scores, while less important regions have small quality scores. Consequently, discriminative parts are highlighted and useless regions are suppressed. Therefore, the accuracy is increased from (99.47%, 92.90%, 89.85%) to (99.52%, 93.37%, 90.20%). Second, some background information may be represented under the guidance of the CAL, the QAN can assign low weights to distracted regions and large weights to discriminative parts, which well prepares our models for pose changes. The large accuracy gain (0.35%) on CPLFW verifies the effectiveness of the QAN.

As shown in Fig. 3 (b)&(g), attention-based methods can learn some important face parts automatically. However, since there is not an explicit signal, some important regions are ignored. This would inevitably deteriorate the performance if only some face parts are visible under pose variations or occlusions. The CAL alleviates this issue by en-

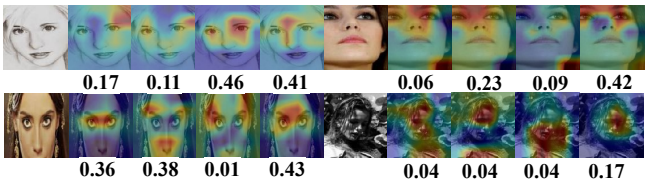

Figure 4: Quality scores learned by the QAN on 4 local branches.

larging pairwise distances about each two attention maps. As demonstrated in (c)&(h), rich local regions are located under the guidance of the CAL. However, it can be observed that some noisy regions are emphasized, like background, or snow goggles. The QAN can adaptively assign weights to different regions through learning from a global perspective. As observed in (d)&(i), some important regions are emphasized and noisy regions are suppressed. As a result, more discriminative features are extracted to improve the performance.

Comparison between CAL and diverse learning and dropout. The diverse learning in (Wang et al. 2020a) encourages the dissimilarities among multiple attention maps by the L_D in Eqn. (3). Our proposed CAL is compared with the L_D in Table 1. If the L_D is used, its performance is (99.33%, 93.20%, 89.82%). In contrast, our CAL is able to achieve a higher performance of (99.52%, 93.37%, 90.20%). This is because our CAL adopts a more flexible divergence loss where the σ controls the variation around its mean value.

Besides quantitative results in Table 1, we run a experiment by adding only L_D into our baseline, achieving (99.42%, 92.67%, 89.68%), which is inferior than only adding CAL (99.47%, 92.90%, 89.85%). Besides, as shown in Fig. 3, qualitative results between L_D (e&k) and CAL (d&i) also demonstrate the superiority. This is because L_D uses Euclidean distance to calculate the similarity of two attention maps. However, it is sensitive to scales due to characteristics of Euclidean distances. This makes it less effective under the scenario of large scale variations in attention maps where some noise may have high responses. In contrast, CAL takes advantages of Gaussian function and the diversity penalization is affected by the Gaussian distance, preventing more penalization variances and achieving better performance than Euclidean distance (Gong, Zhong, and Hu 2019). Furthermore, we thoroughly explore different values of σ which can control the Gaussian distances flexibly. Three local branches locate different local patches in Fig. 1 (b-d)&(g-i).

If dropout is used instead of CAL, it achieves (99.42%, 92.72%, 89.52%). Although dropout can encourage the learning of more parts, this signal is too weak to achieve diversity. In contrast, our CAL enforces attention maps away from each other, locating diverse local patches.

Different number of local branches b. Different number of local branches (i.e. 2, 4, 8, and 16) are compared in Table 2. If b is set to 2, only few facial parts are located, while missing some important local patches. b is improved to 4, boosting the results from (99.32%, 93.15%, 90.02%) to (99.52%, 93.37%, 90.20%). However, if the b is 8, some unnecessary information may be represented. Consequently, the result is dropped to (99.45%, 93.13%, 90.00%). This issue is becoming more serious when b is 16, which achieves the result of (99.40%, 92.80%, 89.78%).

Table 2: Performance comparison (%) of different hyper-parameters: Varying number of local branches (b) and values of λ Eqn. (9), and t and σ in Eqn. (2).

b	λ	σ	t	LFW	CALFW	CPLFW
2	0.50	0.01	0.2	99.32	93.15	90.02
4	0.50	0.01	0.2	**99.52**	**93.37**	**90.20**
8	0.50	0.01	0.2	99.45	93.13	90.00
16	0.50	0.01	0.2	99.40	92.80	89.78
4	**5.00**	0.01	0.2	99.43	93.10	**90.22**
4	**1.00**	0.01	0.2	99.42	93.18	90.00
4	**0.50**	0.01	0.2	**99.52**	**93.37**	90.20
4	**0.10**	0.01	0.2	99.47	92.90	89.85
4	**0.05**	0.01	0.2	99.40	92.88	90.18
4	0.50	**0.001**	0.2	99.40	93.17	**90.22**
4	0.50	**0.005**	0.2	**99.55**	93.23	90.07
4	0.50	**0.010**	0.2	99.52	**93.37**	90.20
4	0.50	**0.050**	0.2	99.40	93.27	89.92
4	0.50	**0.100**	0.2	99.47	93.20	90.12
4	0.50	0.01	**0.6**	99.35	93.18	90.07
4	0.50	0.01	**0.4**	99.35	93.07	90.17
4	0.50	0.01	**0.2**	**99.52**	**93.37**	**90.20**
4	0.50	0.01	**0.0**	99.45	93.18	90.17

Different values of λ. As illustrated in Table 2, we compare the performance of different values of λ in Eqn. (9): 5, 1, 0.5, 0.1, and 0.05. If we reduce the value of λ from 5, 1 to 0.5, the overall accuracy is boosted to (99.52%, 93.37%, 90.20%). This is because some noisy information is extracted under the strong guidance of the ACL if λ is too large. On the other hand, if λ is decreased from 0.5 to 0.1 and 0.05, the performance declines. This means that the signal of the ACL is too weak to miss some important features. Therefore, λ is assigned to 0.5, which obtains a good trade-off between the L_{Cls} and L_{CAL}.

Different values of σ. The σ controls the shape of the Gaussian function. When σ becomes larger, more variances are allowed around the mean; As σ becomes smaller, the less variances allow. It is observed that the highest accuracy is achieved when $\sigma = 0.01$.

Varying values of t. We investigate varying values of the hyper-parameter margin t in Eqn. (2) in Table 2: 0.6, 0.4, 0.2, and 0. It shows that the best margin $t = 0.2$ which achieves the best accuracy.

Experiments on Masked Face Matching

The most recent publicly available models are compared using the M2N, M2M, and MLFW protocols, respectively, in Table 3. For a fair comparison, CQA-Face adopts ResNet-100 as the stem CNN which is used in both ArcFace (Deng

Table 3: Performance comparison (%) of our CQA-Face models with two public models on masked face matching

Methods	M2N				M2M			MLFW			
	FAR= 0.1	FAR= 0.01	FAR= 0.001	Rank-1	FAR= 0.1	FAR= 0.01	FAR= 0.001	FAR= 0.1	FAR= 0.01	FAR= 0.001	Acc.
ArcFace (Deng et al. 2019)	31.53	8.54	1.45	15.73	29.22	7.25	2.06	44.92	12.85	1.81	69.25
CurricularFace (Huang et al. 2020)	33.43	13.21	4.42	17.38	31.55	9.79	2.95	59.24	23.45	3.62	74.61
CQA-Face	**59.40**	**34.22**	**16.72**	**40.46**	**57.84**	**34.93**	**17.75**	**89.33**	**86.68**	**84.23**	**92.78**

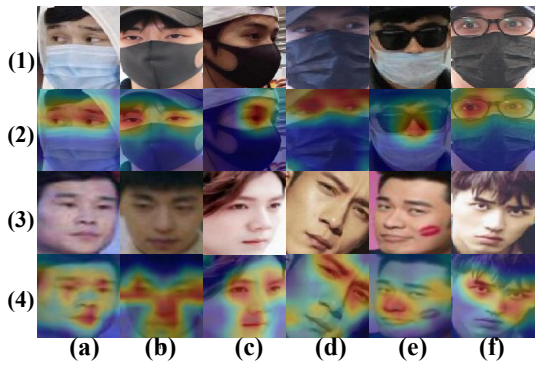

Figure 5: Qualitative illustration of the CQA-Face model on masked face matching. **(1)**&**(3)**: Input faces with blur changes in **(a)**&**(b)**, pose variations in **(c)**&**(d)**, and heavy occlusions in **(e)**&**(f)**). **(2)**&**(4)** show CAMs on these faces.

et al. 2019) and CurricularFace (Huang et al. 2020).

M2N refers to matching between masked and normal faces. For a clear presentation, a tuple (a, b, c, d) is used to denote the performance where a, b, c is the TAR values when FAR=0.1, 0.01, 0.001, respectively. d refers to the Rank-1 accuracy. Because both masked and normal faces are captured in the real-world scenarios, this is a very challenging task. Since ArcFace and CurricularFace are trained on global faces without considering locating discriminative local representations, the performance is significantly decreased on this task. Specifically, ArcFace obtains the accuracy of (31.53%, 8.54%, 1.45%, 15.73%) and CurricularFace achieves the performance of (33.43%, 13.21%, 4.42%, 17.38%). However, due to its localization ability of diverse local patches under the supervision of the CAL and its part-level quality-aware capacity among multiple local patches with the QAN, our CQA-Face model boosts the performance to (59.40%, 34.22%, 16.72%, 40.46%). Although our model uses an inferior CosFace, it still outperforms ArcFace and CurricularFace, showing the superiority.

The M2M protocol refers to the verification between masked and masked faces. For clarity, a tuple (a, b, c) is used to denote the performance where a, b, c is the TAR values when FAR=0.1, 0.01, 0.001, respectively. Since both faces in gallery and probe sets are masked faces where some facial clues are occluded in M2M, thus M2M is more challenging than M2N. Consequently, global-based models (i.e. ArcFace, CurricularFace) have inferior performances, with results of (29.22%, 7.25%, 2.06%) and (31.55%, 9.79%, 2.95%), respectively. Compared with them, our CAL module can locate rich local representations. Consequently, non-occluded facial parts can contribute to the FR even if some face clues are invisible. Besides, the QAN learns quality scores for multiple local patches in a global scope, emphasizing important local patches and suppressing noisy regions. This is especially important for masked face matching where face masks cover some important face regions. In such a manner, the QAN assigns low weights to unimportant located parts. Therefore, our CQA-Face achieves the highest overall performance of (57.84%, 34.93%, 17.75%).

MLFW means the verification protocol on masked LFW. A tuple (a, b, c, d) is used where a, b, and c represent TAR values on FAR=0.1, 0.01, and 0.001, respectively, and d denotes the accuracy. As illustrated in Table 3, although ArcFace (Deng et al. 2019) model* achieves 99.77% on the original LFW, its accuracy is reduced significantly to (44.92%, 12.85%, 1.81%, 69.25%). Similarly, although CurricularFace (Huang et al. 2020) obtains the 99.80% on the original LFW, it achieves the accuracy of (59.24%, 23.45%, 3.62%, 74.61%) on MLFW. Our CQA-Face model improves over them, achieving (89.33%, 86.68%, 84.23%, 92.78%).

As shown in Fig. 5, global faces change significantly under different challenging factors, like blur changes in (a)&(b), pose variations in (c)&(d), and heavy occlusions in (e)&(f). However, our CQA-Face model can locate discriminative parts effectively. As demonstrated in the above analysis, our CQA-Face model is more robust to face masks, which shows great potentials for recognizing masked faces under the pandemic of the COVID-19.

Experiments on Cross-quality Face Matching

In some real-world applications, the captured face images may be of low-quality, which are matched with enrolled high-quality faces. Here cross-quality face matching is conducted to simulate the above scenarios. Several public models are compared on IJB-A quality in Table 4. TAR values are reported when FAR=0.01 and 0.001, respectively.

Several global-based approaches employed full face images as the training input: Center loss (Wen et al. 2016), SphereFace (Liu et al. 2017), and ArcFace (Deng et al. 2019). However, they fail to extract discriminative local patches, which lead to sub-optimal representations. Differently, LS-CNN (Wang and Guo 2019) used spatial attention modules to focus on important local parts, and boosted the

*https://github.com/deepinsight/insightface/wiki/Model-Zoo

Table 4: Performance comparison (%) of our CQA-Face model with the state-of-the-art on different face matching tasks.

Methods	IJB-A quality		CPLFW	CFP-FP	CALFW	LFW
	FAR= 0.01	FAR= 0.001				
Center loss (Wen et al. 2016)	52.5	31.3	77.48	-	85.48	99.13
SphereFace (Liu et al. 2017)	54.8	39.6	81.40	-	90.30	99.42
LS-CNN (Wang and Guo 2019)	87.5	75.5	88.03	97.17	92.00	99.52
MV-Softmax (Wang et al. 2019)	-	-	89.69	95.70	95.63	99.79
ArcFace (Deng et al. 2019)	68.6	65.7	92.08	98.37	95.45	**99.83**
HPDA (Wang et al. 2020a)	87.6	80.3	92.35	-	95.90	99.80
DBM (Cao et al. 2020)	-	-	92.63	-	96.08	99.78
EQFace (Liu and Tan 2021)	-	-	92.60	98.34	95.98	99.82
CQA-Face	**91.1**	**86.4**	**93.00**	**98.49**	**96.12**	**99.83**

performance to (87.5%, 75.5%). Furthermore, HPDA model (Wang et al. 2020a) enlarged pairwise attention distances, emphasizing rich facial parts to make models robust to blur changes. It obtains the performance of (87.6%, 80.3%). Like ArcFace (Deng et al. 2019) which uses ResNet-100 as the stem, CQA-Face boosts the performance from (68.6%, 65.7%) to (91.1%, 86.4%).

Experiments on Cross-pose Face Matching

CPLFW and CFP-FP datasets are used to evaluate the performance where a tuple (a, b) is used to show the results.

Since some face clues are missed in profile view, it is inevitable that global-based methods (i.e. Center loss (Wen et al. 2016), SphereFace (Liu et al. 2017), and MV-Softmax (Wang et al. 2019)) suffer from performance drops. In contrast, LS-CNN (Wang and Guo 2019) locates discriminative local patches by spatial attentions. Thus, it achieves competitive performances on these datasets, which shows great potentials of local descriptions in this task.

However, some important local patches may be missed. HPDA (Wang et al. 2020a) alleviates this issue by maximizing pairwise attention Euclidean distances, and thus outperforms LS-CNN, which is (92.35%, -) versus (88.03%, 97.17%). $-$ denotes that the result is unreported. Meanwhile, CQA-Face adopts ResNet-100 as stem CNN, but further increases the overall performance to (93%, 98.49%). Our explanation is that the CAL enlarges pairwise attention distances, and thus make models locate diverse facial parts. Non-occluded parts can contribute to FR if the occluded parts are invisible, making models robust to pose variations. And also, the QAN learns a quality score for each local representation in a global scope, assigning high weights to important parts and low weights to noisy parts. Although EQFace (Liu and Tan 2021) learns a quality score for a whole facial image and achieves (92.60%, 98.34%), it fails to explore quality scores for different facial parts. This is important for cross-pose face matching because different facial parts have different importance.

Experiments on Cross-age Face Matching and LFW dataset

We investigate the performance on cross-age face matching. As shown in Table 4, our CQA-Face achieves the best result on CALFW (96.12%). Finally, our CQA-Face model obtains the competitive result on LFW, achieving 99.83% accuracy.

Conclusions

We have proposed a new method, called CQA-Face model, which is mainly composed of contrastive attention learning (CAL) and quality-aware network (QAN). First, the CAL encourages the diversity among different attention maps. In such a manner, it is guaranteed that comprehensive facial regions are explored. Second, the QAN learns quality scores for each local representations from a global scope. Our CQA-Face model outperforms the state-of-the-art methods on several challenging tasks, illustrating the importance and usefulness of our proposed approach.

Acknowledgement

Part of the work was supported by an NSF CITeR grant.

References

Cao, D.; Zhu, X.; Huang, X.; Guo, J.; and Lei, Z. 2020. Domain Balancing: Face Recognition on Long-Tailed Domains. *arXiv preprint arXiv:2003.13791*.

Cao, Q.; Shen, L.; Xie, W.; Parkhi, O. M.; and Zisserman, A. 2018. Vggface2: A dataset for recognising faces across pose and age. In *FG*, 67–74. IEEE.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Ding, C.; and Tao, D. 2017. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Gong, S.; Shi, Y.; and Jain, A. K. 2019. Recurrent embedding aggregation network for video face recognition. *arXiv preprint arXiv:1904.12019*.

Gong, Z.; Zhong, P.; and Hu, W. 2019. Diversity in machine learning. *IEEE Access*, 7: 64323–64350.

Guo, G.; and Zhang, N. 2018. What Is the Challenge for Deep Learning in Unconstrained Face Recognition? In *FG*, 436–442. IEEE.

Guo, Y.; Zhang, L.; Hu, Y.; He, X.; and Gao, J. 2016. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 87–102. Springer.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

Huang, Y.; Wang, Y.; Tai, Y.; Liu, X.; Shen, P.; Li, S.; Li, J.; and Huang, F. 2020. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5901–5910.

Kang, B.-N.; Kim, Y.; Jun, B.; and Kim, D. 2019. Attentional Feature-Pair Relation Networks for Accurate Face Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 5472–5481.

Kang, B.-N.; Kim, Y.; and Kim, D. 2018. Pairwise Relational Networks for Face Recognition. *arXiv preprint arXiv:1808.04976*.

Liu, R.; and Tan, W. 2021. EQFace: A Simple Explicit Quality Network for Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1482–1490.

Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; and Song, L. 2017. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 212–220.

Liu, X.; Vijaya Kumar, B.; Yang, C.; Tang, Q.; and You, J. 2018. Dependency-aware Attention Control for Unconstrained Face Recognition with Image Sets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 548–565.

Liu, Y.; Yan, J.; and Ouyang, W. 2017. Quality aware network for set to set recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5790–5799.

Rao, Y.; Lu, J.; and Zhou, J. 2017. Attention-aware deep reinforcement learning for video face recognition. In *Proceedings of the IEEE international conference on computer vision*, 3931–3940.

Sengupta, S.; Chen, J.-C.; Castillo, C.; Patel, V. M.; Chellappa, R.; and Jacobs, D. W. 2016. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1–9. IEEE.

Song, G.; Leng, B.; Liu, Y.; Hetang, C.; and Cai, S. 2017. Region-based quality estimation network for large-scale person re-identification. *arXiv preprint arXiv:1711.08766*.

Sun, Y.; Chen, Y.; Wang, X.; and Tang, X. 2014. Deep learning face representation by joint identification-verification. In *Advances in neural information processing systems*, 1988–1996.

Sun, Y.; Wang, X.; and Tang, X. 2014a. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1891–1898.

Sun, Y.; Wang, X.; and Tang, X. 2014b. Deep learning face representation from predicting 10,000 classes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1891–1898.

Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; and Liu, W. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5265–5274.

Wang, Q.; and Guo, G. 2019. LS-CNN: Characterizing local patches at multiple scales for face recognition. *IEEE Transactions on Information Forensics and Security*, 15: 1640–1653.

Wang, Q.; Wu, T.; Zheng, H.; and Guo, G. 2020a. Hierarchical Pyramid Diverse Attention Networks for Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8326–8335.

Wang, X.; and Gupta, A. 2018. Videos as space-time region graphs. In *ECCV*, 399–417.

Wang, X.; Zhang, S.; Wang, S.; Fu, T.; Shi, H.; and Mei, T. 2019. Mis-classified Vector Guided Softmax Loss for Face Recognition. *arXiv preprint arXiv:1912.00833*.

Wang, Z.; Wang, G.; Huang, B.; Xiong, Z.; Hong, Q.; Wu, H.; Yi, P.; Jiang, K.; Wang, N.; Pei, Y.; et al. 2020b. Masked face recognition dataset and application. *arXiv preprint arXiv:2003.09093*.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, 499–515. Springer.

Xie, W.; Shen, L.; and Zisserman, A. 2018. Comparator networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 782–797.

Zhang, K.; Zhang, Z.; Li, Z.; and Qiao, Y. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10): 1499–1503.

Zheng, T.; and Deng, W. 2018. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep*, 5.

Zheng, T.; Deng, W.; and Hu, J. 2017. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2921–2929.