Learning from the Web: Webly Supervised Meta-Learning for Masked Face Recognition

Wenbo Zheng [1,3] Lan Yan [3,4] Fei-Yue Wang [3,4] Chao Gou [2] *
[1] School of Software Engineering, Xi'an Jiaotong University
[2] School of Intelligent Systems Engineering, Sun Yat-sen University
[3] The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences
[4] School of Artificial Intelligence, University of Chinese Academy of Sciences
zwb2017@stu.xjtu.edu.cn;yanlan2017@ia.ac.cn;feiyue.wang@ia.ac.cn;gouchao@mail.sysu.edu.cn

Abstract

Mask wearing has been considered as an effective measure to prevent the spread of COVID-19 during the current pandemic. However, most advanced face recognition approaches are not adequate for masked face recognition, particularly in dealing with the issue of training through the datasets covering only a limited number of images with ground-truth labels. In this work, we propose to learn from the large scale of web images and corresponding tags without any manual annotations along with limited fully annotated datasets. In particular, inspired by the recent success of webly supervised learning in deep neural networks, we capitalize on readily-available web images with noisy annotations to learn a robust representation for masked faces. Besides, except for the conventional spatial representation learning, we propose to leverage the power of frequency domain to capture the local representative information of unoccluded facial parts. This approach learns robust feature embeddings derived from our feature fusion architecture to make joint and full use of information from both spatial and frequency domains. Experimental results on seven benchmarks show that the proposed approach significantly improves the performance compared with other state-of-the-art methods.

1. Introduction

Masked face recognition is a special kind of occluded face recognition [42, 50, 53]. Different from other types of face recognition, this task focuses on recognizing people wearing face masks via their face images [41], and examples are shown in Figure 1. During the ongoing outbreak of coronavirus disease 2019 (COVID-19), almost everyone wears a mask [46]. On the other hand, for masked face recognition, conventional facial recognition technology is ineffective in many cases, such as community access control, face access control, facial attendance, facial security checks at train stations, etc. Therefore, it is very *urgent* and essential to design a robust and effective model for masked face recognition.

Figure 1: Illustration of Masked Face Recognition Task. The left samples (image and its label) are from fully annotated datasets [41]. The right samples (image and its tags) are from weakly annotated datasets .

Although, these datasets cover a significant number of images (e.g., about 1M in Megaface and 200K in CelebA), creating a larger dataset with image-label pairs is extremely difficult and labor-intensive [15]. Especially, considering the current global epidemic, it may be impossible to establish large-scale manual labeling. Moreover, it is gen-

*Chao Gou is the corresponding author.

Figure 2: The Open-World Setting of Our Paper. We focus on the learning of robust masked face recognition model using clean images with ground-truth labels, and update this learning by utilizing web images and its noisy associated tags. During this process, the latent space is learned and tested by images and tags from our web datasets.

erally feasible to have only a limited number of users to annotate training images, which may lead to a biased model [29, 30]. Hence, while these datasets provide a convenient modeling assumption, they are very restrictive considering the enormous amount of detailed descriptions that a human can compose. Accordingly, although trained models show excellent performance on benchmark datasets for masked face recognition task, applying such models in the open-world setting is unlikely to show satisfactory generalization.

Streams of images with noisy labels are readily available in datasets [47] as well as in nearly infinite numbers on the web. Developing a practical system for masked face recognition, considering a large number of web images, is more likely to be robust. However, inefficient utilization of weakly annotated images may increase ambiguity and degrade performance.

Motivated by this observation, we pose an essential question in this paper: *Can a large number of web images with noisy annotations be leveraged upon with a fully annotated dataset of images to learn a better model for masked face recognition?* Figure 2 shows an illustration of this scenario.

In this work, we study how to judiciously utilize web images to develop a strong masked face recognition system. We propose a novel framework that can augment knowledge through an effective masked face recognition model with weakly supervised web data. Our approach outperforms previous approaches significantly in masked face recognition.

In masked face images shown in Figure 1, even though there are a little noise and shadow, humans are very good at recognizing the subjects. Why can human beings identify these subjects quickly and accurately with very little direct supervision or none at all? Probably because human beings can use the experience to learn [35], while the network cannot. And isn't this one of the mechanisms of meta-learning [34]? *Therefore, why don't we use the principle of meta-learning to build a network to capture the relation of subjects and its masked face images?*

Furthermore, under the occlusion by masks, the accuracy of deep-network-based face recognition models would be degraded[7, 19]. This may be because the existing network is based on the down-sampling operation, which would cause redundant and salient information loss [22]. However, the frequency domain analysis method (e.g., discrete cosine transform (DCT)) arises the redundant and salient information of the image [26]. Thus, *how to apply frequency domain analysis to make up for the deficiency of existing deep networks?*

On the other hand, inspired by the success of the residual network [12] and its variants [48, 27], which are both fusion of multiple shallow networks, we intend to investigate network feature fusion strategies to overcome the aforementioned problem. However, it has not been considered in existing work about network feature fusion for jointing spatial domain with frequency domain information. *How to design spatial-frequency fusion to make the full of and joint use of spatial domain and frequency domain information?*

To tackle all the aforementioned problems, in this paper, we propose a novel meta-learning based model for masked face recognition. In this work, we attempt to utilize web images annotated with noisy tags for improving the model trained using a dataset of images and labels. We build a two-branch relation network via meta-learning. First, we use the embedding approach to do feature extraction of training images. In this process, we design the spatial domain network and the frequency domain network, and propose a feature fusion approach to get salient information of masked faces via combining the learning of our two networks. Then, to compare the features, we design a relation model that determines if they are from matching categories or not. Experimental results show that our model performs better than similar works, and has strong robustness. On the side, we propose a two-stage approach that learns the image-label representation. In Stage I, we use a supervised formulation that leverages the available clean image-label pairs from a dataset. In Stage II, we utilize weakly-annotated image-

tags pairs from the web (e.g., Google Photo) to update the previously learned image representation, which allows us to transfer knowledge from thousands of freely available weakly annotated images to develop a better masked face recognition system. We address a novel and practical problem in this paper that how to exploit large-scale web data for learning an effective masked face recognition model without requiring a large amount of human-crafted training data. Towards solving this problem, we make the following main contributions:

⊛ We propose a webly-supervised approach utilizing web image collection with associated noisy tags, and a clean dataset containing images and their labels for learning robust masked face recognition model. Experimental results show that the proposed approach has strong robustness and superiorities.

⊛ We propose an effective and novel frequency domain network for masked face recognition.

⊛ We design spatial-frequency fusion architecture to embed frequency features into the spatial network to enhance the recognition performance.

2. Webly Supervised Meta-Learning Approach

In this section, we first describe the problem definition. Then, based on our definition, we propose network-based representation learning for masked face recognition. Finally, we present our proposed strategy to incorporate the tags in the framework to learn improved representation learning.

2.1. Problem Definition

We consider the problem of masked face recognition as meta-learning- and webly-supervised-learning-based (i.e., webly supervised meta-learning based) few-shot classification. Before elaborating on the proposed network, we briefly present the notations and prior knowledge.

Learning to Learn The ability to learn new classes is crucial to the development of real-world artificial intelligence systems. In this paper, we focus on the few-shot learning problem. In this setting of our meta-learning, there are three datasets: a training set, a support set, and a testing set. The support set and testing set share the same label space, but the training set has its own label space that is disjoint with support/testing set.

We can, in principle, train a classifier to assign a class label $\hat{y}$ to each sample $\hat{x}$ in the test set while we only use the support set. However, in most cases, the performance of such a classifier is usually not excellent because of the lack of the labeled samples in the support set. Therefore, we use the meta-learning on the training set to transfer the extracted knowledge to the support set. It aims to perform the few-shot learning on the support set better and classify the test set more successfully. The training procedure of our approach is shown in Figure 3.

We propose novel matching networks[36, 34] to solve the problem of adversarial image classification. Suppose there are m labeled samples for each of n unique classes in support set. We select randomly n classes from the training set with m labeled samples from each of the n classes to conduct the sample set $\mathfrak{D}^S = \{(x_i, y_i)\}^z{}_{i=1} (z = m \times n)$, and we select the remaining samples to conduct the query set $\mathfrak{D}^Q = \{(x_j, y_j)\}^v{}_{j=1}$. This sample/query set split is designed to simulate the support/test set that will be encountered at test time.

2.2. Network Representation Learning

Considering that (1) existing neural networks mainly operate in the spatial domain, and (2) the downsampling operations of these neural networks remove both redundant and salient information obliviously, which results in accuracy degradation, we propose a novel matching networks based on relational network [9, 23] to solve the problem of masked face recognition. First, we meta-learn a transferable feature exaction model through the designed *spatial-frequency fusion network*, which considers both spatial and frequency domain sampling of images. The well-learned features of the query samples in the support set are then fed into the non-linear distance metric to learn the similarity scores. Further, we conduct a few-shot classification based on these scores.

Meta-Learning Based Classifier: As illustrated in Figure 4, our matching network consists of two branches: a **feature extraction model** and a **relation model** during the training of our network.

❶ **Meta-Learning Based Feature Extraction:** As illustrated in Figure 4, our spatial-fresquency fusion network consists of three parts: spatial domain network, frequency domain network, and spatial-frequency fusion architecture.

① **Spatial Domain Network:** We use the ResNet-50 network [12] to deals with traditional convolution, pooling and activation neurons for input images, and get the spatial domain features of these images.

② **Frequency Domain Network:** We design the frequency domain network utilizing ResNet-50 with discrete cosine transform (DCT) to process additional frequency feature representations of the same image. Firstly, the input image, the shape of which is $H \times W \times C$, where $C = 3$ and the height and width of the image is denoted as H and W respectively, is transformed to the YCbCr color space and converted to the frequency domain using DCT. Then, the two-dimensional DCT coefficients at the same frequency are grouped into one channel to form three-dimensional DCT cubes. Since the JPEG compression standard [31, 33] uses 8×8 DCT transformation on the YCbCr color space, we group the components of the same frequency in all the

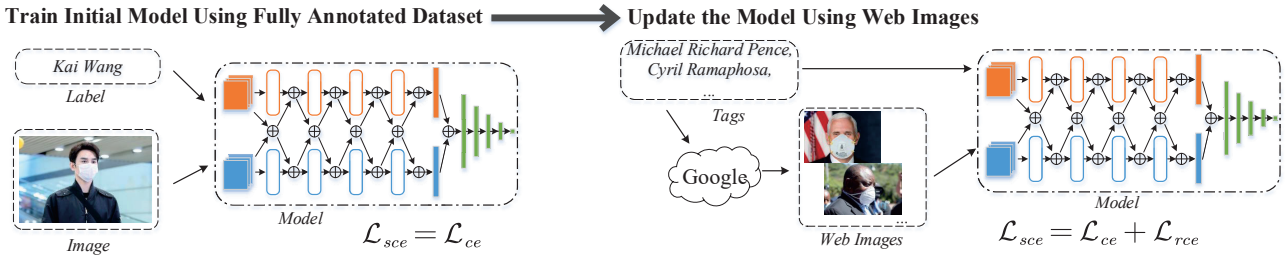


Figure 3: A Brief Illustration of the Proposed Framework. We use the image-label pairs from the clean dataset and image-tag pairs from the web to learn the masked face recognition model. Firstly, we use the image from the clean and their labels for learning our model. Then, we update our model using web images and their tags.

Figure 4: Our Relation Network. Left: the framework of our relation network. It contains two modules: a feature extraction model and a relation model. The feature extraction model $f_{network}$ produces feature maps to represent feature extraction function. Our $f_{network}$ includes two kinds of networks: spatial domain network and frequency domain network. The relation model $J_{relation}(\cdot)$ represents the similarity between sample and query; Right: the architecture of our relation network. Firstly, we preprocess the input image. Secondly, for the image whose size is $448 \times 448 \times 3$, we use our designed frequency domain network to get the frequency domain features. Similarly, we use spatial domain network to get spatial domain features. We meta-learn a transferable feature exaction model through the designed *spatial-frequency fusion mechanism*, where we use the solid circle denotes this operation. By this way, our networks considers both spatial and frequency domain sampling of images. Thirdly, the well-learned features of the query samples in the support set are then fed into the non-linear distance metric to learn the similarity scores. Finally, we conduct a few-shot classification based on these scores.

8×8 blocks into one channel, maintaining their spatial relations at each frequency. Thus, each of the Y, Cb, and Cr components provides $8 \times 8 = 64$ channels, one for each frequency, with a total of 192 channels in the frequency domain. At this point, the preliminary feature shape becomes $\frac{H}{8} \times \frac{W}{8} \times 64C$, which maintains the same input data size. Since the input feature maps in the frequency domain are smaller in the H and W dimensions but larger in the C dimension than the spatial-domain counterpart, we skip the input layer of a conventional ResNet-50, which is usually a stride-2 convolution. We skip the max-pooling operator as well. Then we adjust the channel size of the next layer to match the number of channels in the frequency domain. This way, we minimally adjust the existing ResNet-50 to accept the frequency-domain features as input.

All in all, the spatial domain network and the frequency domain network are both composed of a sequence of residual blocks. Each residual block contains two branches: identity mapping and residual branch. The corresponding function is given as,

$$\begin{aligned} \mathbf{x}_{t+1}^{spatial} &= \text{Res}_t(\mathbf{x}_t^{spatial}) + \mathbf{x}_t^{spatial} \\ \mathbf{x}_{t+1}^{frequency} &= \text{DCT - Res}_t(\mathbf{x}_t^{frequency}) + \mathbf{x}_t^{frequency} \end{aligned} \tag{1}$$

where $\mathbf{x}_t^{spatial}$ and $\mathbf{x}_t^{frequency}$ denote the input of the t-th residual block of the spatial domain network and the frequency domain network, respectively. $\text{Res}_t(\cdot)$ and $\text{DCT - Res}_t(\cdot)$ are transition functions of the spatial domain network and the frequency domain network respectively, corresponding to the residual branch composed of a few stacked layers.

③ **Spatial-Frequency Fusion Architecture:** We aim to jointly map spatial domain features and frequency domain features to a unitary feature space. This combined features

are called spatial-frequency features. A feature fusion architecture is proposed in order to use frequency features to spatial domain features so as to make each network extract complementary features. Our fusion architecture is shown in Figure 4 and is formed by assembling each residual branch of the spatial domain network and the frequency domain network: Sum the inputs of each residual branch, and add the sum to the output of each residual branch as the input of each subsequent residual branch, respectively. It is formulated as below,

$$\begin{aligned}
&\mathbf{x}_{t+1}^{spatial} = (\text{Res}_t(\mathbf{x}_t^{spatial}) + \mathbf{x}_t^{spatial}) + \\
&(\mathbf{x}_t^{frequency} + \mathbf{x}_t^{spatial}) \\
&\mathbf{x}_{t+1}^{frequency} = (\text{DCT - Res}_t(\mathbf{x}_t^{frequency}) + \mathbf{x}_t^{frequency}) + \\
&(\mathbf{x}_t^{frequency} + \mathbf{x}_t^{spatial})
\end{aligned} \tag{2}$$

❷ **Meta-Learning Based Relation Model:** We further propose non-linear distance relation model to learn to compare the sample features in a few-shot classification.

Suppose sample x_j in the query set $\mathfrak{D}^Q$ and sample x_i in the sample set $\mathfrak{D}^S$, we define the function $f_{network}$ which represents feature extraction function using network to produce feature maps $f_{network}(x_j)$ and $f_{network}(x_i)$. The feature maps are combined using the function $C_{network}$. In this work, we assume the $C_{network}(\cdot,\cdot)$ to be concatenation of corresponding feature maps in depth. The combined feature map of the sample and query is used as the relation model $J_{relation}(\cdot)$ to get a scalar in range of 0 to 1 representing the similarity between x_i and x_j, which is called relation score. Suppose we have one labeled sample for each of n unique classes, our model can generate n relation scores $Judge_{i,j}$ for the relation between one query input x_j and training sample set examples x_i:

$$\begin{aligned}
&Judge_{i,j} = J_{relation}(C_{network}(f_{network}(x_i), f_{network}(x_j))) \\
&i = 1, 2, \cdots, n
\end{aligned} \tag{3}$$

Furthermore, we can do the operation of the element-wise sum over our feature extraction model outputs of all samples from each training class to form this class's feature map. And this pooled class-level feature map is concatenated with the feature map of the test samples as above.

2.3. Training with Noisy Web Data

In this work, we try to utilize image-tag pairs from the web for improving joint embeddings trained using a clean dataset with image-label pairs. We aim to learn a good representation of spatial-frequency fusion that ideally ignores the data-dependent noise and generalizes well. The utilization of web data effectively increases the sample size used for training our model and can be considered as implicit data augmentation. On the other hand, due to the outbreak of the coronavirus disease 2019 (COVID-19) epidemic, it is impossible for a large number of workers to label data. The exploitation of web data also frees up a lot of human resources so that these can be used to save the lives of patients. In particular, we propose a two-stage approach to train image representation. In the first stage, we leverage the available clean image-label pairs from a dataset to learn an aligned representation. In the second stage, we adapt the model trained in the first stage with noisy web data.

2.3.1 Stage I: Initial training

We leverage image-label pairs from an annotated dataset to learn network-based representation. To this end, we use the **symmetric cross entropy**, which provides its effectiveness against various types and rates of label noise.

$$\mathcal{L}_{sce} = \tau \times \mathcal{L}_{ce} + \upsilon \times \mathcal{L}_{rce} \tag{4}$$

where τ and υ are two hyperparameters, $\mathcal{L}_{ce}$ means a standard cross-entropy loss[49], and $\mathcal{L}_{rce}$ means reverse cross-entropy loss [40]. Details about $\mathcal{L}_{sce}$ are shown in Ref[40].

In Eq. 4, τ and υ are predefined weights for different losses. In the first training stage, the reverse cross-entropy loss is not used ($\tau = 1$ and $\upsilon = 0$) while in the second stage, both losses are used ($\tau = 1$ and $\upsilon = 1$).

2.3.2 Noisy Web Data

We use Google Photo API[1] to retrieve web images via inputting tags from the labels from MS-Celeb-1M[2], MS1MV3[2], and Celeb500K[5]. We then query and retrieve around 200 images per query, together with their tags. In this way, we collect about $3,000,000$ masked face images with tags. We do not collect more than 5 images from a single owner. Furthermore, we also collect about $2,000,000$ original face images with tags. We have developed a mask wearing software based on Dlib library[3] to perform mask wearing automatically. This software is then used to wear masks on original face images. Based on this software, we also can get masked face images from the web. All in all, we collect about $5,000,000$ masked face images with tags in total.

Debias in Web Image-set We use the "Nationality" attribute of FreeBase[4] celebrities to directly select Asians and Indians. For Caucasians and Africans, Face++ API[5] is used to estimate race. The identity will be accepted only if its most images are estimated as the same race. Otherwise, it will be abandoned. To avoid the negative effects caused by

[1] https://developers.google.com/photos
[2] http://trillionpairs.deepglint.com/overview
[3] http://dlib.net/
[4] https://developers.google.com/freebase/data
[5] https://www.faceplusplus.com/

the biased Face++ tool, we manually check some images with low confidence scores from Face++. Finally, we construct our web image sets, including four subsets, namely Caucasian, Asian, Indian, and African.

2.3.3 Stage II: Model Adaptation with Web Data

After Stage I converges, we have the representation of an image with a learned model. In Stage II, we utilize weakly-annotated image-tags pairs from our web data to update the previously learned network. This enables us to transfer knowledge from thousands of freely available weakly annotated images in learning the representation. We utilize a smaller learning rate in Stage II, as the network achieves competitive performance after Stage I and tuning the representation network with a high learning rate from weakly-annotated data may lead to catastrophic forgetting.

As web data is very prone to label noise, we find it is hard to learn good representation for our task in many cases. Hence, in Stage II, we adopt a curriculum learning-based strategy [1, 39, 14] in training. Curriculum learning allows the model to learn from easier instances first so they can be used as building blocks to learn more complex ones, which leads to a better performance in the final task. It has been shown in many previous works that appropriate curriculum strategies guide the learner towards better local minima [24, 40]. Our idea is to gradually inject difficult information to the learner such that in the early stages of training, the network is presented with images related to frequently occurring features in the clean training set. Images related to rarely occurring concepts are presented at a later stage. Since the network trained in Stage I is more likely to have learned well about frequently occurring features, label noise is less likely to affect the network.

3. Experiments

In this section, we firstly introduce our settings of the experiment. Then we conduct detailed ablation study over the vital modules of our model. Further, we show evaluation on the different types of testing datasets and compare to state-of-the-art methods.

3.1. Experimental Settings

In this subsection, we describe the used datasets and the implementation details.

Dataset. In our training phase, we use Real-world Masked Face Recognition Dataset (RMFRD) [41] as the clean dataset, and use our collected *Web Image-set* as web data for Stage II. In our testing phase, our testing datasets consist of Simulated Masked Face Recognition Dataset (SMFRD) [41], CFP [28], YTF [45], MegaFace [16], IJB-B [44], IJB-C [25] and FMA-3D [38]. For compared methods, we also use RMFRD and Web Image-set for training, and use the same seven datasets for testing.

Training and Testing Settings The classic pipeline in meta-learning is first to train a model on a set of base classes and then to evaluate it on a different set of novel classes (each set of classes is split into train and validation subsets) [34]. For our experiments, we use this protocol. In this paper, we resize the images from all datasets to $448 \times 448 \times 3$. For all methods consisting of ours and compared methods, in the training process, we randomly choose 800 face images of the 80 subjects for training subsets, and the remaining of these 80 subjects are used for validation subsets; We randomly choose 10 times as per the above strategy and take the average recognition performance for comparison. For widely comparison, we use the unrestricted verification protocol on the SMFRD, FMA-3D and YTF datasets; we use the frontal-profile (FP) protocol on the CFP dataset; we test on both verification and identification protocols of MegaFace. Specifically, for face identification (Id.), the Cumulative Matching Characteristics (CMC) curves are adopted to evaluate the Rank-1 accuracy, and also report the precision, recall, F_1-Measure. For face verification (Veri.), the Receiver Operating Characteristic (ROC) curves at different false alarm rates are adopted; we test on both verification and identification protocols of the two benchmarks: IJB-B and IJB-C. These comparison principles are same to the previous works [17].

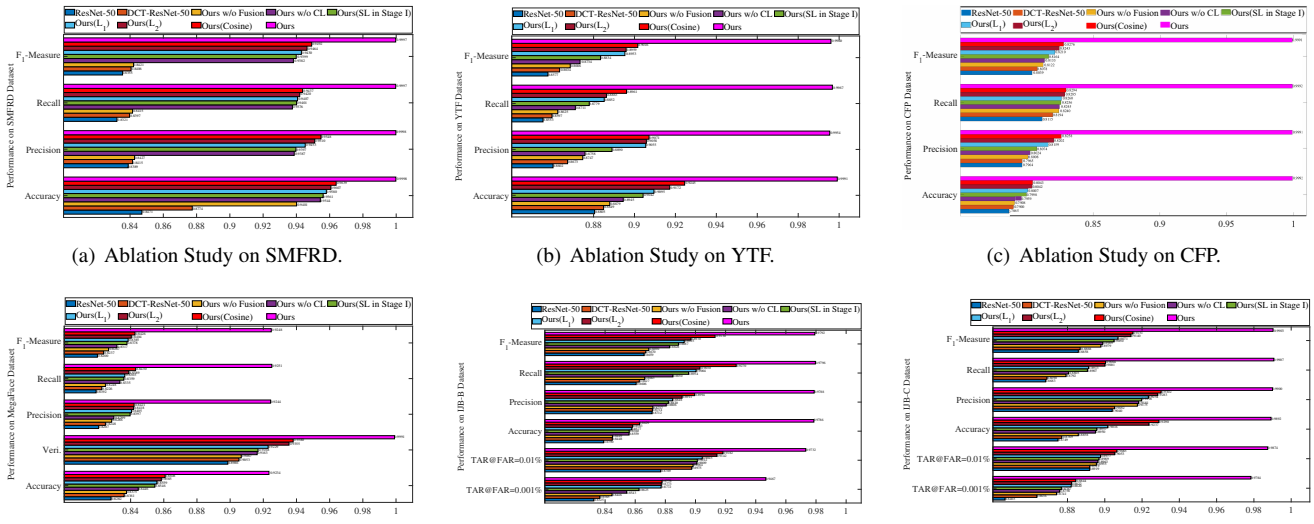
Our Network Settings. We continue training Stage I for an initial 120 epochs. Then we start updating the learned model in Stage I with web images in Stage II for another 120 epochs. Our network architecture is shown in Figure 4. Our network consists of two parts:

❶ **Feature Extraction Model:** We employ the ResNet-50 architecture [12] for learning the feature exaction model. When meta-learn the transferable feature exaction, we use Adam optimizer [18] with a learning rate of 0.001 and a decay for every 40 epochs. We totally train 1000 epochs and adopt the semi-hard mining strategy [11] when the loss starts to converge.

❷ **Relation Model:** We use the 4-layer network architecture. Taking two feature maps from the spatial domain network and the frequency domain network, respectively, as input, we take the concatenation of these features. Then, we apply the fully connected layer to change into 8192-dimensional vector. Finally, we use three fully-connected layers to have 1024, 8 and 1 outputs respectively, followed by a loss function to get the final similarity scores. Other network settings are similar to our feature extraction model.

3.2. Ablation Study

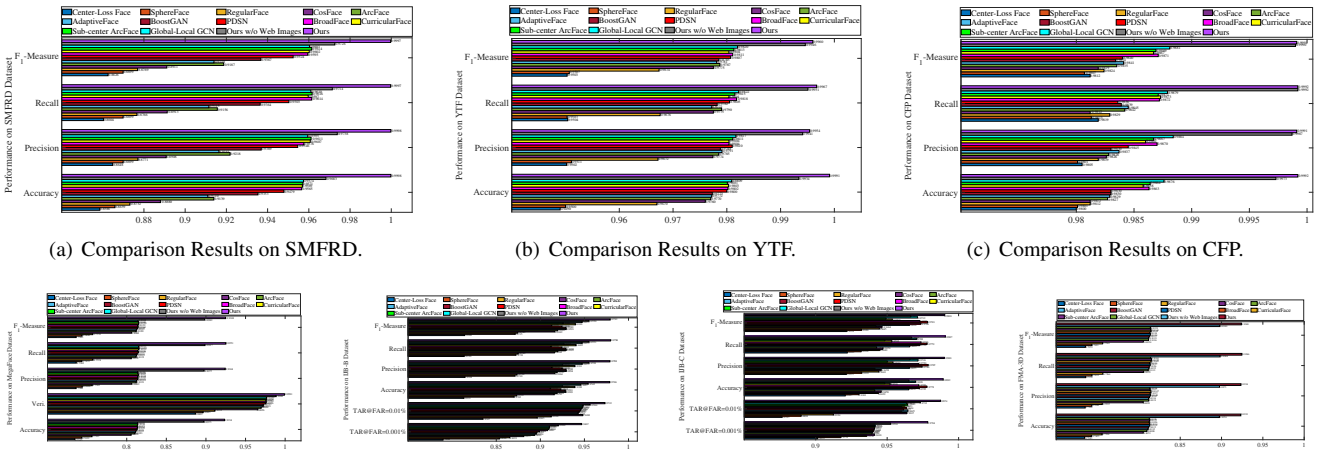
In order to verify the reasonableness and effectiveness of each part of our network, we design the ablation experiment. In Figure 5, "**Ours w/o Fusion**" means a variant of



(a) Ablation Study on SMFRD. (b) Ablation Study on YTF. (c) Ablation Study on CFP.

(d) Ablation Study on MegaFace. (e) Ablation Study on IJB-B. (f) Ablation Study on IJB-C.

Figure 5: The Results of Ablation Study.



(a) Comparison Results on SMFRD. (b) Comparison Results on YTF. (c) Comparison Results on CFP.

(d) Comparison Results on MegaFace. (e) Comparison Results on IJB-B. (f) Comparison Results on IJB-C. (g) Comparison Results on FMA-3D.

Figure 6: Comparison Results.

Ours, which removes spatial-frequency fusion architecture; "**ResNet-50**" means a variant of Ours, which only using the spatial domain network; "**DCT-ResNet-50**" means a variant of Ours, which only using the frequency domain network. "**Ours w/o CL**" means a variant of Ours, which removes curriculum learning-based strategy [1] during training. "**Ours (SL in Stage I)**" means a variant of Ours, which use the symmetric cross entropy loss [40] during the training of Stage I. "**Ours**(L_1)" means a variant of Ours, which using L_1 loss function [10, 8] and not using our loss funtion. "**Ours**(L_2)" means a variant of Ours, which using L_2 loss function [10, 8] and not using our loss function. "**Ours(Cosine)**" means a variant of Ours, which using cosine similarly function [10, 8] and not using our loss function. We analyze the following two aspects:

Compared with "Ours w/o Fusion" From Figure 5, "Ours w/o Fusion" is 0.0927 and 0.0621 higher than "ResNet-50", and "DCT-ResNet-50", on the SMFRD dataset, in terms of accuracy, respectively. *As we can see, "Ours w/o Fusion" is better than two single networks. These suggest the importance of making joint use of spatial domain and frequency domain information.*

Compared with "Ours" From Figure 5, "Ours w/o CL" has lower performance than "Ours". This shows

that the curriculum learning-based strategy is effective for our model. Similarly, "Ours(SL in Stage I)", "Ours(L_1)","Ours(L_2)" and "Ours(Cosine)" have lower performance than "Ours". This shows that the settings and utility pattern of symmetric cross entropy loss are effective for our model. *As we can see, "Ours" is better than others. These suggest making full use of spatial domain and frequency domain information helps us to improve the task of masked face recognition.*

From the above, we get the conclusion in the following two aspects:

(1) *It is apparent that the design of our spatial-frequency fusion architecture improves the task of masked face recognition.*

(2) *It is manifest that the design of the frequency domain network is better than the spatial domain network. This suggests that the design of frequency domain network is more effective.*

Moreover, by analyzing the results in Figure 5 on other datasets, we can get similar conclusions.

3.3. Comparison with State-of-The-Art Methods

In this subsection, we compare the state-of-the-art approaches with our model. "**Ours w/o Web Images**" means a variant of Ours, which only using the clean dataset and without web images.

Baseline We compare ours with the state-of-the-art approaches, including Center-Loss Face[43], SphereFace[21], RegularFace[52], CosFace[37], ArcFace[4], AdaptiveFace[20], BoostGAN[6], PDSN [32], BroadFace[17], CurricularFace[13], Sub-center ArcFace[3], Global-Local GCN[51]. The results are shown in Figure 6.

Effect of Proposed Webly Supervised Training. For evaluating the impact of our approach, we compare results reported in row-"Ours w/o Web Images" and row-"Ours". Our method utilizes the same loss functions and features used in row-"Ours w/o Web Images" for a fair comparison. From Figure 6, we observe that the proposed approach improves performance consistently in all the cases. For instance, our method is 0.0057, 0.0013, 0.0016, 0.0014 higher than "Ours w/o Web Images", in terms of accuracy, precision, recall, F_1-measure, on the YTF dataset, respectively. *It is evident that using webly supervised training can enhance the effectiveness of our approach.*

Effect of Our Approach. From Figure 6, it is evident that our approach is better than others. For example, our method is 0.0501, 0.0491, 0.0321, 0.0231, 0.0221, 0.0219, 0.0218, 0.0191, 0.0189, 0.0188, 0.019, and 0.0182 higher than Center-Loss Face, SphereFace, RegularFace, CosFace, ArcFace, AdaptiveFace, BoostGAN, PDSN, BroadFace, CurricularFace, Sub-center ArcFace, and Global-Local GCN, in terms of accuracy, on the YTF dataset, respectively;our method is 0.0457, 0.0453, 0.0286, 0.0186, 0.0173, 0.0179, 0.0175, 0.0153, 0.0150, 0.0157, 0.0147, and 0.0141 higher than Center-Loss Face, SphereFace, RegularFace, CosFace, ArcFace, AdaptiveFace, BoostGAN, PDSN, BroadFace, CurricularFace, Sub-center ArcFace, and Global-Local GCN, in terms of F_1-measure, on the YTF dataset, respectively. *From above, our approach is more effective and robust than the state-of-the-arts approaches on these seven benchmark datasets.*

3.4. Discussion on the Generalization Ability

The data distributions in different masked face scenarios could be different from that during model development. To explore the generalization ability of the proposed method, we firstly use the RMFRD datasets as sketch datasets to train our model. We evaluate it with cross-database testing on the SMFRD dataset. Then, we use the SMFRD datasets as sketch datasets to train our model. We evaluate it with cross-database testing on the RMFRD dataset. We run ten times following the above strategy in this discussion. In results, the recognition accuracy comparisons of testing on these datasets are shown in Figure 7. *This experiment indicates that the proposed method could achieve good recognition performance in such a challenging scenario.*

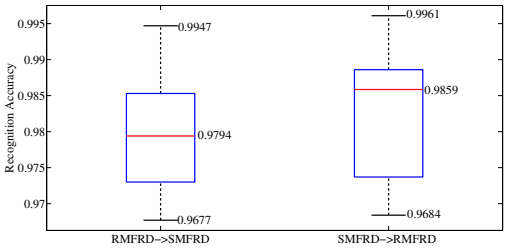


Figure 7: Cross-Database Recognition Accuracy of the Proposed Method.

4. Conclusion

In this work, we show how to leverage large scale of web images with tags to augment knowledge for masked face recognition models with limited labeled data. We attempt to address the challenge by proposing a two-stage approach that can augment knowledge through an effective masked face recognition model with weakly supervised web data. Extensive experiments demonstrate that our approach significantly improves the performance in the masked face recognition task in seven benchmark datasets.

Acknowledgments This work is supported in part by National Key R&D Program of China (2020YFB1600400), in part by Key Research and Development Program of Guangzhou (202007050002).

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 41–48, New York, NY, USA, 2009. ACM. 6, 7

[2] J. Cao, Y. Li, and Z. Zhang. Celeb-500k: A large training dataset for face recognition. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2406–2410, 2018. 5

[3] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2020. 8

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 8

[5] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 5

[6] Q. Duan and L. Zhang. Look more into occlusion: Realistic face frontalization and recognition with boostgan. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15, 2020. 8

[7] S. Ge, C. Li, S. Zhao, and D. Zeng. Occluded face recognition in the wild by identity-diversity inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2020. 2

[8] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016. 7

[9] J. Guan, Z. Lu, T. Xiang, A. Li, A. Zhao, and J. Wen. Zero and few shot learning with semantic feature synthesis and competitive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 3

[10] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011. 7

[11] Ben Harwood, Vijay Kumar B G, Gustavo Carneiro, Ian Reid, and Tom Drummond. Smart mining for deep metric learning. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 3, 6

[13] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 8

[14] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G Hauptmann. Self-paced curriculum learning. In *AAAI*, volume 2, page 6, 2015. 6

[15] B. Jin, M. V. O. Segovia, and S. Süsstrunk. Webly supervised semantic segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1705–1714, July 2017. 1

[16] Ira Kemelmacher-Shlizerman, Steven M Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4873–4882, 2016. 6

[17] Yonghyun Kim, Wonpyo Park, and Jongju Shin. Broadface: Looking at tens of thousands of people at once for face recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2020. 6, 8

[18] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, page arXiv:1412.6980, Dec 2014. 6

[19] S. Kumar and S. K. Singh. Occluded thermal face recognition using bag of cnn (bocnn). *IEEE Signal Processing Letters*, 27:975–979, 2020. 2

[20] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z. Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 8

[21] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 8

[22] Zhenhua Liu, Jizheng Xu, Xiulian Peng, and Ruiqin Xiong. Frequency-domain dynamic pruning for convolutional neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1043–1053. Curran Associates, Inc., 2018. 2

[23] Puneet Mangla, Nupur Kumari, Abhishek Sinha, Mayank Singh, Balaji Krishnamurthy, and Vineeth N Balasubramanian. Charting the right manifold: Manifold mixup for few-shot learning. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 3

[24] T. Matiisen, A. Oliver, T. Cohen, and J. Schulman. Teacher-student curriculum learning. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–9, 2019. 6

[25] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. Iarpa janus benchmark - c: Face dataset and protocol. In *2018 International Conference on Biometrics (ICB)*, pages 158–165, 2018. 6

[26] P. Meer, E. S. Baugher, and A. Rosenfeld. Frequency domain analysis and synthesis of image pyramid generating kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(4):512–522, 1987. 2

[27] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[28] S. Sengupta, J. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. 6

[29] T. Shen, G. Lin, C. Shen, and I. Reid. Bootstrapping the performance of webly supervised semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1363–1371, June 2018. 2

[30] Yunhang Shen, Rongrong Ji, Zhiwei Chen, Xiaopeng Hong, Feng Zheng, Jianzhuang Liu, Mingliang Xu, and Qi Tian. Noise-aware fully webly supervised object detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, June 2020. 2

[31] S. Shimizu and T. Suzuki. Flexibly-tunable bitcube-based perceptual encryption within jpeg compression. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2702–2706, 2020. 3

[32] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 8

[33] Mengdi Sun, Xiaohai He, Shuhua Xiong, Chao Ren, and Xinglong Li. Reduction of jpeg compression artifacts based on dct coefficients prediction. *Neurocomputing*, 384:335 – 345, 2020. 3

[34] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018. 2, 3, 6

[35] Flood Sung, Li Zhang, Tao Xiang, Timothy M. Hospedales, and Yongxin Yang. Learning to learn: Meta-critic networks for sample efficient learning. *CoRR*, abs/1706.09529, 2017. 2

[36] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3637–3645, USA, 2016. Curran Associates Inc. 3

[37] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 8

[38] Jun Wang, Yinglu Liu, Yibo Hu, Hailin Shi, and Tao Mei. Facex-zoo: A pytorch toolbox for face recognition. *arXix preprint arXiv:2101.04407*, 2021. 6

[39] Weixun Wang, Tianpei Yang, Yong Liu, Jianye Hao, Xiaotian Hao, Yujing Hu, Yingfeng Chen, Changjie Fan, and Yang Gao. From few to more: Large-scale dynamic multiagent curriculum learning. In *AAAI*, pages 7293–7300, 2020. 6

[40] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 5, 6, 7

[41] Zhongyuan Wang, Guangcheng Wang, Baojin Huang, Zhangyang Xiong, Qi Hong, Hao Wu, Peng Yi, Kui Jiang, Nanxi Wang, Yingjiao Pei, et al. Masked face recognition dataset and application. *arXiv preprint arXiv:2003.09093*, 2020. 1, 6

[42] X. Wei, C. Li, Z. Lei, D. Yi, and S. Z. Li. Dynamic image-to-class warping for occluded face recognition. *IEEE Transactions on Information Forensics and Security*, 9(12):2035–2050, 2014. 1

[43] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 499–515, Cham, 2016. Springer International Publishing. 8

[44] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother. Iarpa janus benchmark-b face dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–600, 2017. 6

[45] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, 2011. 6

[46] Colin J. Worby and Hsiao-Han Chang. Face mask use in the general population and optimal resource allocation during the covid-19 pandemic. *Nature Communications*, 11(1):4049, Aug 2020. 1

[47] Zhonghua Wu, Qingyi Tao, Guosheng Lin, and Jianfei Cai. Exploring bottom-up and top-down cues with attentive learning for webly supervised object detection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, June 2020. 2

[48] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995, 2017. 2

[49] Y. Yuan, G. Xun, F. Ma, Y. Wang, N. Du, K. Jia, L. Su, and A. Zhang. Muvan: A multi-view attention network for multivariate temporal data. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 717–726, Nov 2018. 5

[50] W. Zhang, S. Shan, X. Chen, and W. Gao. Local gabor binary patterns based on kullback–leibler divergence for partially occluded face recognition. *IEEE Signal Processing Letters*, 14(11):875–878, 2007. 1

[51] Yaobin Zhang, Weihong Deng, Mei Wang, Jiani Hu, Xian Li, Dongyue Zhao, and Dongchao Wen. Global-local gcn: Large-scale label noise cleansing for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 8

[52] Kai Zhao, Jingyi Xu, and Ming-Ming Cheng. Regularface: Deep face recognition via exclusive regularization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 8

[53] Wenbo Zheng, Chao Gou, and Fei-Yue Wang. A novel approach inspired by optic nerve characteristics for few-shot occluded face recognition. *Neurocomputing*, 376:25 – 41, 2020. 1