Two-Stream Prototype Learning Network for Few-Shot Face Recognition Under Occlusions

Xingyu Yang, Mengya Han, *Graduate Student Member, IEEE*, Yong Luo, *Member, IEEE*, Han Hu, and Yonggang Wen, *Fellow, IEEE*

***Abstract*—Few-shot face recognition under occlusion (FSFRO) aims to recognize novel subjects given only a few, probably occluded face images, and it is challenging and common in real-world scenarios. Unknown occlusions may deteriorate the class prototypes, while an occluded image in the support set may be critical for recognition if the query image is occluded. This motivates us to propose a novel Two-stream Prototype Learning Network (TSPLN) for FSFR under occlusions by simultaneously considering the quality of support images and their relevance to the query image. Specifically, we design a two-stream architecture, which mainly consists of a support-centered stream and query-centered stream, to learn the optimal class prototypes. The former stream is to reduce the negative impact of occluded images on the prototype. This is achieved by exploring the similarities between different images in the support set. In the query-centered stream, we exploit the relevance between the query and support set based on feature alignment (FA). We conduct extensive experiments on two popular datasets: CASIA-WebFace and RMFRD. The experimental results show that our proposed method achieves the state-of-the-art performance for occluded face recognition in the few-shot setting.**

***Index Terms*—Face recognition under occlusions, few-shot learning, two-stream, feature alignment, similarity.**

I. Introduction

FACE recognition under occlusion is a challenging face recognition task. In the past few years, deep learning approaches have achieved significant advances in recognizing occluded face images, mainly attributing to the following factors: large-scale and high-quality image galleries and limited variants of occluded scenes [1]. For example, the occluded face dataset is synthesized using occluders such as sunglasses and scarves. However, such occlusion patterns are quite different from those in the real world application. In practice, face images also suffer from self-occlusion (Non-frontal pose), facial accessory occlusion (masked face image) [2], extreme illumination (Part of face highlighted), and low resolution. Besides, In real-world applications, it is common that only a few labeled face images are available for a novel subject. Therefore, it is desirable to design a method for *few-shot face recognition under occlusion (FSFRO)*.

There are two main challenges in FSFRO: a novel subject face has only a few annotated images and the face may be occluded in the image. It should be noted that FSFRO differs significantly from the single sample per person (SSPP) face recognition problem [3]. For SSPP, all subjects at the test time are seen during training phases, and only one labeled image of the subject is available during training. In contrast, for FSFRO, we aims to train a model that can learn to recognize unseen (novel) subjects where only a few images of each novel subject are available.

Current face recognition approaches usually deal with the occlusion issue by extracting features that are robust to occlusion scenes. If a large number of training data or augmented data are provided, we can obtain the occlusion-robust face feature representations by employing advanced deep CNN architectures [4], [5], [6]. However, it is challenging to collect a large number of occlusion data and augmented data cannot be adapted to general diverse types of occlusion. It is extremely nontrivial to obtain features that are robust to occlusion in the few-shot setting. Some works [7], [8], [9], [10], [11] try to learn robust features by employing the feature fusion strategy, but the learning for novel subjects given only a few and probably occluded face images is not investigated.

To deal with this issue, we can simply apply existing few-shot image classification approaches to face recognition, where a subject is considered as a category. As shown in Fig. 1, each subject has three support images, where some faces have non-frontal poses, or are occluded by glasses or masks. For FSFRO, we consider the impact of occluded images on the feature space as *class-level feature jitters* [12]. Since the number of labeled face images for each novel subject is small (usually from 1 to 5), the *class-level feature jitter* to the prototypes caused by the occlusion in one image can be significant. Most of the existing

Manuscript received 15 July 2022; revised 11 December 2022 and 9 February 2023; accepted 17 February 2023. Date of publication 6 March 2023; date of current version 8 May 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2021YFC3300200, in part by the Special Fund of Hubei Luojia Laboratory under Grant 220100014, and in part by the National Natural Science Foundation of China under Grant 62276195. The Guest Editor coordinating the review of this manuscript and approving it for publication was Prof. Jun Wu. *(Corresponding author: Yong Luo.)*

Xingyu Yang, Mengya Han, and Yong Luo are with the School of Computer Science, National Engineering Research Center for Multimedia Software, Wuhan University, and Hubei Luojia Laboratory, Wuhan, Hubei 430072, China (e-mail: yangxingyu2021@whu.edu.cn; myhan1996@whu.edu.cn; yluo180@gmail.com).

Han Hu is with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: hhu@bit.edu.cn).

Yonggang Wen is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: ygwen@ntu.edu.sg).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMM.2023.3253054, provided by the authors.

Digital Object Identifier 10.1109/TMM.2023.3253054

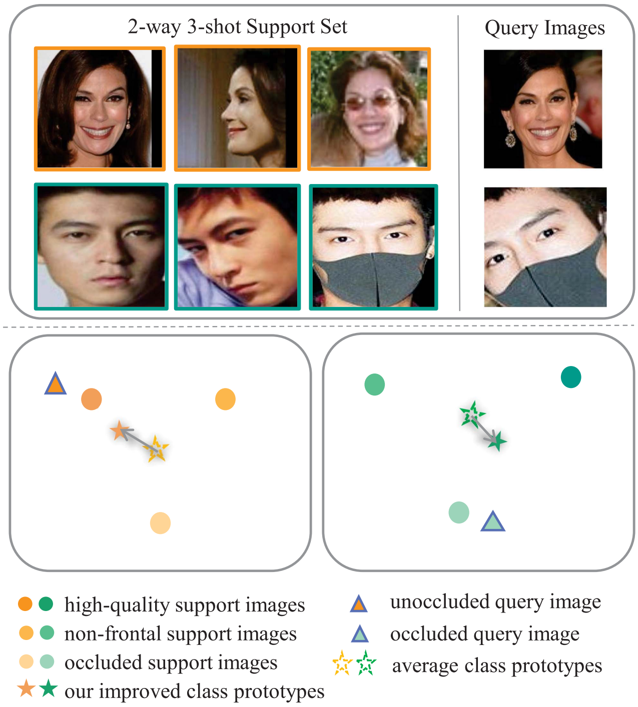

Fig. 1. An illustration of the few-shot face recognition under occlusion task. The face images of the same person contain standard high-quality images, images under natural occlusion, and images under artificial occlusion. If the query face is unoccluded, higher-quality support images will be assigned with larger weights in the learning of class/subject prototype, which will be away from the occluded support images due to their low-quality (as shown in the bottom left subfigure). If the query face is occluded, the significance of occluded support images will increase and the learned prototype will be close to the occluded support images (as shown in the bottom right subfigure).

approaches minimize the intra-class gap to learn robust prototype [12], [13], and these approaches can be regarded as extracting features robust to occlusion scene. A major drawback of these approaches is that they only consider the relationship either between images within the support set [12], [14] or between the query and support images [15], [16].

This motivates us to design a novel framework termed **T**wo-**S**tream **P**rototype **L**earning **N**etwork (**TSPLN**) to learn optimal class prototypes by simultaneously exploring both kinds of relationships. The main idea is to learn adaptive weights for different support images by considering both their quality and relevance to the query, where the weights are induced using middle-level features since they can better adapt to novel classes due to their high transferable ability [17], [18], [19]. The proposed TSPLN is based on a meta-learning paradigm and mainly consists of a **support-centered stream** and **query-centered stream**. Specifically, for the support-centered stream, we adopt a pre-trained transferable similarity relation network (TSRN) to obtain the similarities between support images in the same class (e.g., for a N-way K-shot task, we obtain N similarity matrices of size $K \times K$). Then low-quality (such as occluded) support images are given lower weights when constructing class prototypes due to their low similarities to normal support images. In regard to the query-centered stream, we propose a middle-level feature alignment module, which aims to match the query and support images, and larger alignment scores indicate higher relevance to the query image. In addition, only utilizing the middle-level features to calculate the score may lead to mismatches between a query and some support images belonging to different classes. For example, if the query image is labeled "*Justin Bieber*," but it is very similar to a support image labeled "Troye Sivan". Although the two images have different labels, they may have a high alignment score. This issue is alleviated by computing the similarities between query and support images using their high-level prediction features and enforce the alignment scores to be consistent with these similarities. Finally, we measure the distance between the query image and prototypes and choose the class corresponding to the prototype with the shortest distance as the recognition result.

Our main contributions can be summarized as follows:

- We propose a general two-stream framework termed to learn class prototypes in the few-shot setting, where the correlations between support images, and their relationships to the query are simultaneously exploited.
- We introduce a similarity relation network [20] to evaluate the significance of different support images in our FSFRO problem.
- We design a middle-level feature map alignment module, and consistency between the alignment scores and high-level similarities is explored to improve the learning of class prototypes.

Extensive experiments on two popular datasets show that our proposed method achieves the state-of-the-art performance for occluded face recognition in the few-shot setting. Specifically, we obtain a significant 5.33% improvement in the 3-shot setting. Furthermore, ablation experiments verify the effectiveness of each component.

II. Related Work

In this section, we briefly describe the existing approaches related to our work in terms of both face recognition under occlusions and few-shot classification.

A. *Face Recognition Under Occlusions*

Face recognition under occlusion (FRO) is the task of recognizing faces that are occluded by uncertain objects, such as sunglasses and masks. The existing general CNN models for face recognition may have poor performance in recognizing faces with occlusion. Thus some approaches are specially designed for FRO, and they can be divided into three categories [2]: extract features robust to the occlusion scene, exclude the occluded region, and recover the occluded face.

In the early works, subspace learning is applied to feature extraction [21]. Later, the work of sparse representation classifiers (SRC) [22] interprets occlusions as a linear combination of training samples plus sparse error. Deep learning has been introduced to extract occlusion-robust features in recent years. For example, DDRC [23] linearly encodes features based on a dictionary containing in-depth features and occlusion patterns of test face images to reduce the impact of occlusion. The exclude occluded region approach ignores the features of the occluded part [24], [25], [26]. It is common to crop out multiple patches from the whole face, aiming to use the unoccluded part of the face for recognition. PDSN [27] learns a mask dictionary from the high-level convolutional feature differences of a pair of occluded-unoccluded face images. FROM [1] can predict feature masks end-to-end without relying on additional occlusion detectors, utilizing a mask decoder. Recovery of occlusion approaches recovers the whole face from the occluded face as a substitute image for recognition. One is a reconstruction technique [28], [29], [30] based on the relationship between the occluded and unoccluded face, and the other is an image inpainting technique [31], [32], [33] focusing on repairing the occluded part. These approaches usually require large amounts of labeled training data, which may be not available in practice.

To the best of our knowledge, there are quite few works that explore few-shot face recognition under occlusion. S^{2}P^{2}FR [9] adopts adaptive fusion of multiple features (structural element feature, connected-granule labeling feature, etc.) and artificially synthesized occluded images to address the challenge of occlusion and limited number of training images, respectively. A few-shot learning method based on Siamese Network [34] is proposed in [7] to overcome the image discrepancy of occlusion. However, both of them are not particularly in line with the standard few-shot learning setting, which adopts a setting similar to the SSPP problem. In contrast, many few-shot classification approaches explore the problem caused by occlusions in images.

B. Few-Shot Classification

Due to the data-hungry property of deep learning, few-shot learning has attracted increasing attention. Few-shot learning aims to learn a model that generalizes well to novel classes, relying on only a few clean, labeled samples per class. Each of the novel classes is unseen during training. According to how they use prior knowledge, few-shot learning methods can be roughly categorized as data-based, model-based, and algorithm-based [35].

This work focuses on metric learning approaches in model-based few-shot learning [36], [37], [38], [39], [40], [41]. The core idea of the metric learning approach is to find the nearest category to the query sample features in the feature space by similarity measure. A pioneering work is prototypical Network [37], which calculates the Euclidean distance between the query features and the mean feature vector of each suppport class (namely prototype), belonging to the fixed distance metric. Afterward, in the Relation Network [38], the similarity between two samples can be learned and predicted by a nonlinear CNN classifier. In the ATL-Net [39], the authors introduce an attention mechanism into an episode, allowing the network to perceive important local regions adaptively based on the context [42] of the task in an episode. Recently, RENet [40] has learned an end-to-end relational embedding by combining two kinds of relationships: the auto-correlation relationship within a single image and the cross-correlation attention between two images, which boosted model performance. Besides, there exist some approaches that study the problem of occlusions in images. For example, in the work of IFSM [12], the authors argue that the outliers (including occluded images) have a significant impact on average class prototypes. A hybrid GNN [43] composed of an instance GNN and a prototype GNN, is designed to overcome the problem caused by badly sampled images. In [14], in order to minimize the negative impacts of low-quality samples, the authors propose to generate reference class prototypes based on the whole based class images before episode training.

Some other approaches that explore feature relationships [44], [45], [46] have achieved tremendous progress in few-shot learning. However, existing works do not combine the feature relationships of images within the support set and the relationship between the query set and the support set images. More detailed differences between these works and our method are discussed in the supplementary material.

III. Preliminary

In this paper, we aim to address the problem of few-shot face recognition under occlusion (FSFRO), where each subject is considered as a category. The challenges caused by occluded image samples for few-shot classification can be overcomed by modeling adaptive feature weights from the perspectives of the support set and query set.

In our setting, the dataset consists of non-overlapping base classes $\mathcal{C}_{base}$ and novel classes $\mathcal{C}_{novel}$, where $\mathcal{C}_{base} \cap \mathcal{C}_{novel} = \emptyset$, and each of them has a large number of image samples. The base classes serve as the meta-training dataset $\mathcal{D}_{train}$, and the novel classes serve as the meta-testing dataset $\mathcal{D}_{test}$. Following the standard few-shot learning setting, the model learns from task instances with an episodic-training based meta-learning strategy [47]. The training and testing tasks are sampled from the base and novel classes, respectively. A sampled N-way K-shot training/testing episodic task is defined as $\{\mathcal{S}_N, \mathcal{Q}_N\}$, where $\mathcal{S}_N$ represents the support set and $\mathcal{Q}_N$ represents the query set. In one episode, we randomly sample N classes $\mathcal{C}_N$ from the base/novel classes and then sample K samples from each of these N classes to form the support set and Q samples to form the query set. The training and test tasks have the same formulation but disjoint label spaces because they are sampled from classes that do not overlap. In the support set, $K \ll M$ (M is the number of samples per class), simulating cases where only a few samples per novel class are available during testing.

$$\begin{aligned}\mathcal{S}_N &= \{(x_i^s, y_i^s)\,|y_i^s \in \mathcal{C}_N, i \in [1, N \times K]\}\\ \mathcal{Q}_N &= \left\{\left(x_j^q, y_j^q\right)|y_j^q \in \mathcal{C}_N, j \in [1, N \times Q]\right\}\end{aligned} \tag{1}$$

We denote the feature extractor as $\boldsymbol{f}$ and the classifier as ϕ. During the training phase, for a given query image x_j^q, the classifier takes the feature embedding $\boldsymbol{f}(x_j^q)$ as input and learns a mapping $y_j^q = \phi(\mathcal{S}_N, x_j^q)$. Once trained, given a testing task $\{\mathcal{S}_N, \mathcal{Q}_N\}$ sampled from a novel class, the model is expected to adapt both parts to the novel classes, classifying x_j^q in the testing task to one of the N novel classes.

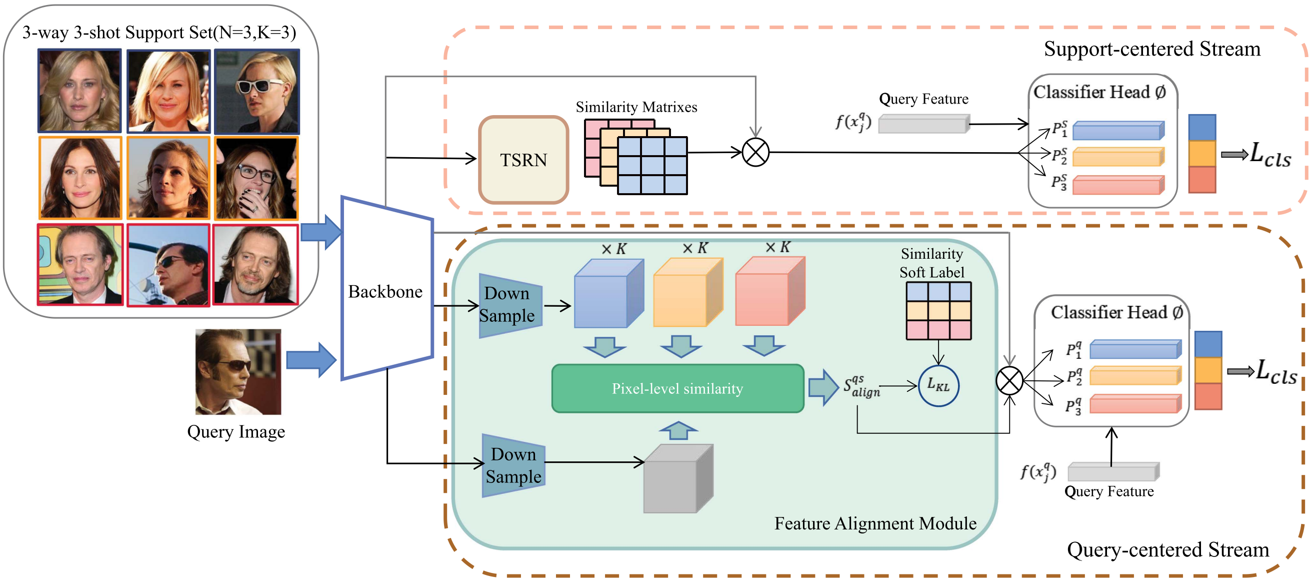

Fig. 2. An overview of our proposed Two-Stream Prototype Learning Network (TSPLN). We train TSPLN using the episode training strategy. Firstly, the features extracted by the shared backbone are fed into two streams. In the support-centered stream, we fine-tune a pre-trained transferable similarity relation network and obtain intra-class similarity matrices of images from the same class in the support set. For example, for the 3-way 3-shot support set, we will obtain three similarity matrices size of 3 × 3. The matrix indicates the similarity relationships of the intra-class images and provides the weight information of each image when constructing the class prototype. In the query-centered stream, we down-sample the original middle-layer feature map to select certain features. Then the feature alignment module aligns the sampled middle-level features of query images and support images. The alignment scores are used as image weights for class prototypes in the query-centered stream. Besides, we use the cosine similarity between high-level features as soft labels to aid the learning of the alignment module, and improve the reliability of the learned relevance to the query. Finally, we apply the prototype network metric to classify the query images based on the respective class prototypes of the two streams.

IV. Method

The proposed **T**wo-**S**tream **P**rototype **L**earning **N**etwork (**TSPLN**) for FSFRO mainly consists of a support-centered stream and query-centered stream. In the following, we first briefly describe the overall architecture in Section IV-A. Then we present the technical details of the two streams of TSPLN in Sections IV-B and IV-C, respectively, and finally depict our training objectives in Section IV-D.

A. Architecture Overview

For the problem of FSFRO, we argue that when the query image has occlusion, occluded support images may play a critical role in learning optimal class prototypes. Therefore, the information required to build the class prototype should also come from the relationship between the query set and the support set, in addition to the relationship within the support set. Motivated by these, we propose the novel two-stream framework termed to learn class prototypes in a few-shot setting by simultaneously exploring both kinds of relationships.

Fig. 2 shows the overall framework of our approach, where TSPLN learns two sets of class prototypes from the support-centered stream and the query-centered stream, respectively. The two stream share the same backbone and will be optimized jointly. We then apply the prototype network metric to classify the query images according to the respective class prototypes of the two streams. In one episode, the sampled support set $\mathcal{S}_N$ and query set $\mathcal{Q}_N$ are fed into TSPLN, and the pre-trained backbone extracts deep features for support and query images into common embedding space. In the support-centered stream, we obtain similarities between images from the same class in the support set based on a pre-trained transferable similarity relation network (TSRN). Then the intra-class similarity can be utilized as the weight of the support image for inducing the class prototype. As a result, those low-quality (such as occluded) support images are given lower weights due to their low similarity to other normal support images. The support-centered prototype can effectively mitigate the negative impact of the occluded face images. The query-centered stream feeds the extracted middle-level feature maps to the feature alignment module, which matches the middle-level features between query and support images. Then the alignment scores are used as sample weights for class prototypes in the query-centered stream. In particular, The alignment module is jointly optimized with similarity soft labels of high-level features and prototype-based classification loss. The class prototypes obtained from the query-centered stream are more relevant to the query images of the current task and are query-specific. This helps the network to better match the query images with the support images. Overall, the class prototypes of the two streams are complementary to each other, and more details are depicted as follows.

B. Support-Centered Stream

In few-shot learning, the total number of annotated support set images for a N-way K-shot task is small (only $N \times K$).

Therefore, the *class-level feature jitters* from low-quality occluded images will likely have significant negative impact on the feature space for classification. To minimize this effect, an intuitive and straightforward method is to make the features of occluded images contribute less in the subsequent construction of class prototypes. The face images in Fig. 1 show that the appearance difference between occluded and unoccluded face images is relatively significant. Their similarity can naturally reflect the extent of the difference (e.g., the more significant the difference, the lower the similarity) and provide a guide for using image features. Hence, we keep similarities of the low-quality occluded face images to other unoccluded images within the same class consistent with its contribution weights. We adopt a pre-trained transferable similarity relation network (TSRN) to the support-centered stream, which can output the semantic similarity between the input images. A detailed description of TSRN can be found in the supplementary material.

Prototype Learning: Given an N-way K-shot support set, we feed K images $x_i^s (i = 1, \ldots, K)$ belonging to class $C_n (n = 1, \ldots, N)$ into TSRN, and we will get N semantic similarity matrixes $M_{K \times K}$ for each class. When forming the prototype of one class, high-quality support images are given larger weights due to their high similarity to each other, while occluded low-quality support images are given lower weights due to their low similarity to other images. The weight of each support image w_i^n comes from the average similarity between this image x_i^s and other images in the same class C_n:

$$w_i^n = \frac{1}{K} \sum_{j=1}^{K} M_{i,j}. \tag{2}$$

After that, we can calculate the class prototype P_n^s of the learned support-centered stream for each class C_n instead of simply averaging the features as a prototype, i.e.,

$$P_n^s = \sum_{i=1}^{K} w_i^n \boldsymbol{f}\left(x_i^s\right), y_i^s \in C_n, \tag{3}$$

where w_i^n represents the average similarity weight between the i^{th} image and other images from category C_n in the support set. In the end, we measure the Euclidean distance between the query image to each class prototype, and the class with the closest distance will be the classification result.

C. Query-Centered Stream

The support-centered stream only relies on the limited information within the support set to construct class prototypes. However, we cannot know the diverse query images in advance. Thus, it is difficult for the support-centered prototype to be optimal consistently when facing different query images (e.g., for FSFRO, the query image likely happens to be an occluded image.) At this time, the occluded images in the support set will play a critical role in the classification task, and the weight should not be reduced. To this end, a query-centered stream aims to learn the relationship between query images and support images, select those support image features that are more relevant to query images and make them account for a larger proportion of class prototypes. In addition, considering that the middle-level features of the image are class-agnostic, the middle layer of the backbone is equivalent to an implicit memory block, which stores the knowledge that can be shared and transferred between the base classes and the novel classes. Therefore, we design a module that aligns the middle-level features of query images and support set images.

Feature Alignment (FA): takes the feature maps of query and support images at the third layer in the backbone as inputs. The down-sample block plays a role of preprocessing for middle features. It makes the middle-level features compact and discriminative by mainly utilizing 2D convolution operators, and obtains updated feature maps $\{\boldsymbol{f}_{down}^i\}_{NK}$ and $\{\boldsymbol{f}_{down}^j\}_{NQ}$ corresponding to the support set and query set. Then, as shown in Fig. 2, given the updated middle-level features, we define the set of alignment score S_{align}^{qs} between all query images and all support images, where the alignment score S_{align}^{ji} for each query image x_j^q and support image x_i^s is the pixel-wise similarity of the updated features after normalization, i.e.,

$$S_{align}^{ji} = \sum_{m,n=1}^{HW} \frac{p_m^T p_n}{\|p_m\| \|p_n\|}, \tag{4}$$

where p_m and p_n denote a pixel in the feature map $\boldsymbol{f}_{down}^j$ and $\boldsymbol{f}_{down}^i$ respectively. For a query set with NQ query samples, there will be $NQ \times NK$ alignment scores S_{align}^{qs}. However, we cannot completely rely on this alignment score since the query image may be misclassified due to the specific background, hair color, etc, which may lead to high alignment scores for the query and some support images belonging to non-query categories. Hence, we optimize the FA module using the distance loss of the class prototype and a Kullback–Leibler divergence loss by (5) to enforce the cosine similarity of the high-level feature map to be consistent with the alignment score. Then the similarities can be regarded as soft labels.

$$\mathcal{L}_{kl} = KL\left(\cos\left(\boldsymbol{f}_{high}^q, \boldsymbol{f}_{high}^s\right), S_{align}^{qs}\right), \tag{5}$$

where $\boldsymbol{f}_{high}^q$ and $\boldsymbol{f}_{high}^s$ represent the high-level features for the query set and support set.

In the end, similar to the support-centered stream classification process, the alignment scores provide weight information instead of similarity matrixes when constructing class prototypes. We calculate the class prototype P_n^q of the learned query-centered stream based on weight information for each class C_n as follows:

$$P_n^q = \sum_{i=1}^{K} S_{align}^{qi} \boldsymbol{f}\left(x_i^s\right), y_i^s \in C_n, \tag{6}$$

where S_{align}^{qi} represents the alignment score between the query set and the i^{th} image from category C_n within the support set. The query images are categorized into the class to which the closest prototype belongs.

D. Training Objectives

During training, we calculate the distance between the query image and all class prototypes in each stream, and then use the softmax function to obtain the classification probability $\hat{p}$, and finally use the cross-entropy function to classify. The classification loss of $\mathcal{L}_{cls}^{sc}$ and $\mathcal{L}_{cls}^{qc}$ are as follows:

$$\mathcal{L}_{cls}^{sc} = -\sum_{i=1}^{Q}\sum_{j=1}^{N}\mathbb{I}(y_i == j)\log \hat{p}_{i,j}^{s},$$

$$\mathcal{L}_{cls}^{qc} = -\sum_{i=1}^{Q}\sum_{j=1}^{N}\mathbb{I}(y_i == j)\log \hat{p}_{i,j}^{q}, \tag{7}$$

where y_i is the label and $\mathbb{I}(\cdot)$ is an indicator function: $\mathbb{I}(\cdot) = 1$ when $\cdot$ is true and 0 otherwise. The overall loss function in (8) combines the prototype classification loss $\mathcal{L}_{cls}^{qc}$ for a support-centered stream, $\mathcal{L}_{cls}^{sc}$ for a query-centered stream, and the additional $\mathcal{L}_{kl}$ when optimizing alignment scores.

$$\mathcal{L} = \mathcal{L}_{cls}^{qc} + \mathcal{L}_{cls}^{sc} + \lambda\mathcal{L}_{kl}, \tag{8}$$

where $\lambda \geq 0$ is a trade-off hyper-parameter.

V. Experiments

In this section, we mainly present the comparison with the state-of-the-art approaches, and the ablation experiments. More experimental results of our proposed method can be found in the supplementary material.

A. Datasets

We evaluate our approach on two publicly available and widely used face datasets: CASIA-WebFace [48] and RMFRD [49], each of which contains many subjects. Both datasets are designed for general supervised face recognition. Therefore, we utilize a part of the subjects in the experiments and resplit them for few-shot learning.

B. Comparisons

1) Baselines: We compared the proposed method with the state-of-the-art few-shot methods, including algorithm-based and metric-based. MAML [50] is a typical algorithm-based method that learns suitable initialization parameters, enabling the model to find the gradient descent direction on new tasks quickly. Prototype Network [37], Relation Network [38], DN4 [51], ATL-Net [39], ADM _ KL [36], RENet [40], DeepBDC [52] are all metric-based methods. These methods have achieved improvements in terms of feature embedding or metrics and have achieved an impressive performance. Besides, ArcFace [53] is designed to obtain discriminative depth features for the face recognition task. We add the ArcFace loss function to the typical metric-based prototype network [37] as a new baseline, named "ArcFace+Prototype Network," which is re-implemented based on prototype network.

2) Performance Comparisons: The comparison results on the RMFRD and CASIA-WebFace datasets are summarized in Tables I and II, respectively. From the experimental results, we can observed that: 1) our TSPLN achieves state-of-the-art performance on both datasets under the 5-way 3-shot and 5-shot settings. In the 3-shot setting, compared with the strongest baseline RENet [40], TSPLN improves accuracy by 3.19% and 5.33% on the RMFRD and CASIA-WebFace datasets, respectively. In the 5-shot setting, TSPLN also improves the accuracy by 1.84% and 3.46% on the RMFRD and CASIA-WebFace datasets, respectively; 2) our method dramatically outperforms the baselines that directly introduce the ArcFace loss into a regular few-shot learning method, indicating that ArcFace loss fails to help solve the FSFRO problem. In contrast, our proposed approach can address the challenges of few-shot and occlusions of images in a general schema without special consideration of face distance metrics; 3) the results in Tables I and II show that the proposed method obtains more improvement in the 3-shot setting. We hypothesize few-shot occluded face recognition as *class-level feature jitters* in few-shot classification. When the number of face images is smaller, *class-level feature jitters* are more likely to occur. Therefore, the baselines achieve the worst performance in the 3-shot setting compared with the 5-shot setting. In contrast, the proposed method degrades less as the number of shots decreases

TABLE I
Comparison Results on the RMFRD Dataset Under 5-Way 3-Shot and 5-Way 5-Shot Few-Shot Settings in Terms of Accuracy (%)

Methods	Backbone	RMFRD	
		3-shot	5-shot
MAML [50]	ResNet18	63.67	71.93
Prototype Network [37]	ResNet50	62.20	65.47
ArcFace+Prototype Network	ResNet50	63.13	64.13
Relation Network [38]	Conv64	60.13	68.33
DN4 [51]	Conv64	63.47	66.93
ATL-Net [39]	Conv64	66.80	71.40
ADM_KL [36]	Conv64	70.80	77.33
RENet [40]	ResNet50	81.81	85.53
DeepBDC [52]	ResNet50	79.89	83.52
TSPLN(Ours)	ResNet50	**85.00**	**87.37**

TABLE II
Comparison Results on the CASIA-WebFace Dataset Under 5-Way 3-Shot and 5-Way 5-Shot Few-Shot Settings in Terms of Accuracy (%)

Methods	Backbone	CASIA-WebFace	
		3-shot	5-shot
MAML [50]	ResNet18	67.60	74.33
Prototype Network [37]	ResNet50	64.20	73.33
ArcFace+Prototype Network	ResNet50	60.60	75.60
Relation Network [38]	Conv64	68.66	71.40
DN4 [51]	Conv64	50.20	62.73
ATL-Net [39]	Conv64	54.53	68.53
ADM_KL [36]	Conv64	69.67	79.47
RENet [40]	ResNet50	80.87	86.03
DeepBDC [52]	ResNet50	78.67	81.04
TSPLN(Ours)	ResNet50	**86.20**	**89.49**

TABLE III
ABLATION STUDY OF THE PROPOSED TWO-STREAM ARCHITECTURE ON RMFRD AND CASIA-WEBFACE DATASETS UNDER 5-WAY 3-SHOT AND 5-WAY 5-SHOT FEW-SHOT SETTINGS IN TERMS OF ACCURACY (%)

Methods	RMFRD		CASIA-WebFace	
	3-shot	5-shot	3-shot	5-shot
Support-centered Prototype	81.87	85.27	83.73	89.13
Query-centered Prototype	77.20	78.13	78.67	79.40
TSPLN(Both)	**85.00**	**87.37**	**86.20**	**89.49**

(e.g., 2.37% on RMFRD). This indicates that our method performs better for the few-shot occlusion problem. Overall, our proposed approach is effective in real-world scenarios where novel subjects have fewer face images and rich natural occlusions in the images.

C. Ablation Studies

1) Two-Stream Architecture: The proposed framework is a two-stream architecture, including a support-centered stream and a query-centered stream. To verify the effectiveness of two-stream architecture, we conduct ablation studies under 5-way 3-shot and 5-shot settings. The ablation analysis results are tabulated in Table III. According to the results in Table III, the single-stream obtains the worst performance, and the support-centered stream together with the query-centered stream, brings the best performance, reflecting that the two streams could cooperate for further performance improvement. Note that only the query-centered stream in the model degrades the recognition performance significantly (about 9%). This is because the query-centered stream is equivalent to an unsupervised feature alignment. Thus, only considering the relevance of the support images to the query image, we cannot obtain a clear inter-class boundary from the query images without labels. When the model contains only a support-centered stream, the model performance drops between $[0.36\%, 3.13\%]$. Under the two-stream architecture, our method achieves the best performance on both datasets in all settings. These improvements are derived from two aspects: the support-centered stream can reduce the negative impact of occluded images on the prototype, and the query-centered stream can explore the relevance between the query and support images. The joint learning of support-centered and query-centered streams is beneficial for learning more optimal class prototypes.

TABLE IV
ABLATION STUDY OF TSRN IN THE SUPPORT-CENTERED STREAM ON RMFRD AND CASIA-WEBFACE UNDER 5-WAY 3-SHOT AND 5-WAY 5-SHOT SETTINGS IN TERMS OF ACCURACY (%)

Methods	RMFRD		CASISA-WebFace	
	3-shot	5-shot	3-shot	5-shot
Prototype Network [37]	62.20	65.47	64.20	73.33
Support-centered Prototype	81.87	85.27	83.73	89.13

TABLE V
ABLATION STUDY OF THE FEATURE ALIGNMENT MODULE IN THE QUERY-CENTERED STREAM ON THE CASIA-WEBFACE AND RMFRD DATASETS UNDER THE 5-WAY 5-SHOT SETTING IN TERMS OF ACCURACY (%). THE FEATURE ALIGNMENT MODULE CONTAINS TWO SUB-MODULES: DOWN-SAMPLE AND SIMILARITY SOFT LABEL

Query-centered Prototype		CASIA-WebFace	RMFRD
Down sample	Soft label	5-way 5-shot	
		70.93	73.27
√		78.33	77.47
	√	75.73	74.67
√	√	**79.40**	**78.13**

2) Effectiveness of TSRN: To verify the effectiveness of the designed TSRN, we remove the pre-trained TSRN in the support-centered stream to construct a model which directly utilizes the extracted features to classify. The results are given in Table IV. Without the TSRN, the model turns into a prototype network baseline [37] and obtains an accuracy of 73.33% on the CASIA-WebFace dataset under the 5-shot setting. The proposed method TSPLN achieves a significant improvement of 15.8% on the CASIA-WebFace dataset under the same setting. This indicates that the critical component, TSRN, contributes significantly to the boost. The TSRN can output similarity matrices between images that provide more proper prototype weights.

3) Feature Alignment: The query-centered stream mainly depends on the feature alignment module, which contains two sub-modules: down-sample and similarity soft label. To evaluate the effectiveness of the feature alignment module, we conduct ablation experiments on both datasets under the 5-shot setting. The baseline directly utilizes middle-level features for feature alignment. The results are summarized in Table V. For the CASIA-WebFace dataset, compared to the baseline, the down-sample sub-module improves performance by 7.4%, and the consistent supervision of similarity soft labels brings an improvement of 4.8%. These improvements show that the proposed down-sample benefits the optimal selection of features, and the consistency supervision signals from similarity soft labels can also try to avoid misleading alignment scores and help correctly classify query images. Moreover, the entire feature alignment module combined with down-sample and soft labels achieves the best performance, outperforming the baseline by a large margin.

VI. CONCLUSION

In this paper, we propose a two-stream framework to learn optimal class prototypes for the FSFRO problem. The proposed TSPLN can learn adaptive weights for different support images by simultaneously considering their quality and correlations with query images. To achieve this goal, we introduce a pre-trained transferable similarity relation network to the support-centered stream to reduce the negative impact of occluded images for class prototype learning. We also design an alignment module for middle-level features, which can focus on features more relevant to the query image in constructing class prototypes. Moreover, the consistency constraints with similarities between high-level features make the learned prototypes more reliable. From the comprehensive experiments on two popular datasets (RMFRD and CASIA-WebFace), we mainly conclude that: 1) the two-stream framework is beneficial for learning complementary class prototypes; 2) TSPLN can learn from a few novel subjects with potentially unknown occlusions. A drawback of our method is that the local patterns of the occluded regions are not explored. Therefore, in the future, we intend to evaluate the significance of different images in a more fine-grained manner, and apply the proposed method to more recognition problems under occlusions in addition to faces.

Acknowledgment

The authors would like to thank the handling editor and all the anonymous reviewers for their constructive comments.

References

[1] H. Qiu, D. Gong, Z. Li, W. Liu, and D. Tao, "End2End occluded face recognition by masking corrupted features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6939–6952, Oct. 2022.

[2] D. Zeng, R. Veldhuis, and L. Spreeuwers, "A survey of face recognition techniques under occlusion," *IET Biometrics*, vol. 10, no. 6, pp. 581–606, 2021.

[3] Y.-F. Yu, D.-Q. Dai, C.-X. Ren, and K.-K. Huang, "Discriminative multiscale sparse coding for single-sample face recognition with occlusion," *Pattern Recognit.*, vol. 66, pp. 302–312, 2017.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 7–9, 2015.

[6] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2818–2826.

[7] A. Holkar, R. Walambe, and K. Kotecha, "Few-shot learning for face recognition in the presence of image discrepancies for limited multi-class datasets," *Image Vis. Comput.*, vol. 120, 2022, Art. no. 104420.

[8] W. Zheng, L. Yan, F.-Y. Wang, and C. Gou, "Learning from the web: Webly supervised meta-learning for masked face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4304–4313.

[9] W. Zheng, C. Gou, and F.-Y. Wang, "A novel approach inspired by optic nerve characteristics for few-shot occluded face recognition," *Neurocomputing*, vol. 376, pp. 25–41, 2020.

[10] X. Lan, Q. Hu, and J. Cheng, "ATF: An alternating training framework for weakly supervised face alignment," *IEEE Trans. Multimedia*, early access, Apr. 05, 2022, doi: 10.1109/TMM.2022.3164798.

[11] Y. Zhaoet al., "ChildPredictor: A child face prediction framework with disentangled learning," *IEEE Trans. Multimedia*, early access, Apr. 05, 2022, doi: 10.1109/TMM.2022.3164785.

[12] C. Cai, M. Yuan, and T. Lu, "IFSM: An iterative feature selection mechanism for few-shot image classification," in *Proc. IEEE 25th Int. Conf. Pattern Recognit.*, 2021, pp. 9429–9436.

[13] J. Liu, L. Song, and Y. Qin, "Prototype rectification for few-shot learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 741–756.

[14] M. Yuan et al., "Learning class-level prototypes for few-shot learning," 2021, *arXiv:2108.11072*.

[15] C. Doersch, A. Gupta, and A. Zisserman, "Crosstransformers: Spatially-aware few-shot transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21981–21993.

[16] T. Perrett, A. Masullo, T. Burghardt, M. Mirmehdi, and D. Damen, "Temporal-relational crosstransformers for few-shot action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 475–484.

[17] M. Salzmann et al., "Learning transferable adversarial perturbations," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 13950–13962.

[18] S. Huang and D. Tao, "All you need is a good representation: A multi-level and classifier-centric representation for few-shot learning," 2019, *arXiv:1911.12476*.

[19] Z. Wang, B. Du, and Y. Guo, "Domain adaptation with neural embedding matching," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 7, pp. 2387–2397, Jul. 2020.

[20] J. Chen, L. Niu, L. Liu, and L. Zhang, "Weak-shot fine-grained classification via similarity transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 7306–7318.

[21] V. Štruc, S. Dobrišek, and N. Pavešić, "Confidence weighted subspace projection techniques for robust face recognition in the presence of partial occlusions," in *Proc. IEEE 20th Int. Conf. Pattern Recognit.*, 2010, pp. 1334–1338.

[22] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[23] F. Cen and G. Wang, "Dictionary representation of deep features for occlusion-robust face recognition," *IEEE Access*, vol. 7, pp. 26595–26605, 2019.

[24] L. He, H. Li, Q. Zhang, and Z. Sun, "Dynamic feature learning for partial face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7054–7063.

[25] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 773–782.

[26] P. Wang et al., "Quality-aware part models for occluded person re-identification," *IEEE Trans. Multimedia*, early access, Mar. 07, 2022, doi: 10.1109/TMM.2022.3156282.

[27] L. Song, D. Gong, Z. Li, C. Liu, and W. Liu, "Occlusion robust face recognition based on mask learning with pairwise differential siamese network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 773–782.

[28] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust LSTM-autoencoders for face de-occlusion in the wild," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 778–790, Feb. 2018.

[29] L. Cheng, J. Wang, Y. Gong, and Q. Hou, "Robust deep auto-encoder for occluded face recognition," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1099–1102.

[30] N. Zhang, N. Liu, J. Han, K. Wan, and L. Shao, "Face de-occlusion with deep cascade guidance learning," *IEEE Trans. Multimedia*, early access, Mar. 08, 2022, doi: 10.1109/TMM.2022.3157036.

[31] G. Liu et al., "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 85–100.

[32] Y.-A. Chen, W.-C. Chen, C.-P. Wei, and Y.-C. F. Wang, "Occlusion-aware face inpainting via generative adversarial networks," in *Proc. IEEE Int. Conf. Image Process.*, 2017, pp. 1202–1206.

[33] B. Hu, Z. Zheng, P. Liu, W. Yang, and M. Ren, "Unsupervised eyeglasses removal in the wild," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4373–4385, Sep. 2021.

[34] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a" siamese" time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1993, pp. 737–744.

[35] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, 2020.

[36] W. Li et al., "Asymmetric distribution measure for few-shot learning," in *Proc. 29th Int. Joint Conf. Artifi. Intell.*, C. Bessiere, Ed., 2020, pp. 2957–2963.

[37] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4080–4090.

[38] F. Sung et al., "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1199–1208.

[39] C. Dong, W. Li, J. Huo, Z. Gu, and Y. Gao, "Learning task-aware local representations for few-shot learning," in *Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell.*, 2021, pp. 716–722.

[40] D. Kang, H. Kwon, J. Min, and M. Cho, "Relational embedding for few-shot classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 8802–8813.

[41] H. Zhang, H. Li, and P. Koniusz, "Multi-level second-order few-shot learning," *IEEE Trans. Multimedia*, early access, Jan. 13, 2022, doi: 10.1109/TMM.2022.3142955.

[42] M. Lan, J. Zhang, and Z. Wang, "Coherence-aware context aggregator for fast video object segmentation," *Pattern Recognit.*, vol. 136, 2023, Art. no. 109214.

[43] T. Yu, S. He, Y.-Z. Song, and T. Xiang, "Hybrid graph neural networks for few-shot learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 3179–3187.

[44] X. Zhong, C. Gu, M. Ye, W. Huang, and C.-W. Lin, "Graph complemented latent representation for few-shot image classification," *IEEE Trans. Multimedia*, early access, Jan. 11, 2022, doi: 10.1109/TMM.2022.3141886.

[45] X. Zhu et al., "Few-shot action recognition with prototype-centered attentive learning," in *Proc. 32nd Brit. Mach. Vis. Conf.*, Nov. 22–25, 2021, p. 249.

[46] K. Guo, C. Shen, B. Hu, M. Hu, and X. Kui, "RSNet: Relation separation network for few-shot similar class recognition," *IEEE Trans. Multimedia*, early access, Apr. 19, 2022, doi: 10.1109/TMM.2022.3168146.

[47] T. M. Hospedales, A. Antoniou, P. Micaelli, and A. J. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, 2022.

[48] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.

[49] Z. Wang et al., "Masked face recognition dataset and application," 2020, *arXiv:2003.09093*.

[50] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[51] W. Li et al., "Revisiting local descriptor based image-to-class measure for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7260–7268.

[52] J. Xie, F. Long, J. Lv, Q. Wang, and P. Li, "Joint distribution matters: Deep brownian distance covariance for few-shot classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 7972–7981.

[53] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4685–4694.

Xingyu Yang received the B.S. degree in computer science from Northwestern Polytechnical University, Xi'an, China. She is currently working toward the M.S. degree with the School of Computer Science, Wuhan University, Wuhan, China. Her research interests include computer vision and machine learning.

Mengya Han (Graduate Student Member, IEEE) received the B.S. degree in computer science from Bengbu University, Bengbu, China, and the M.S degree with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China. She is currently working toward the Ph.D. degree with the School of Computer Science, Wuhan University, Wuhan, China. Her research interests include computer vision and machine learning.

Yong Luo (Member, IEEE) received the B.E. degree in computer science from Northwestern Polytechnical University, Xi'an, China, and the D.Sc. degree with the School of Electronics Engineering and Computer Science, Peking University, Beijing, China. He is currently a Professor with the School of Computer Science, Wuhan University, Wuhan, China. He has authored or coauthored more than 60 papers in top journals and prestigious conferences, including IEEE T-PAMI, IEEE T-NNLS, IEEE T-IP, IEEE T-KDE, IEEE T-MM, ICCV, WWW, IJCAI, and AAAI. His research interests include machine learning and data mining with applications to visual information understanding and analysis. He is serving on Editorial Board of IEEE T-MM. He was the recipient of the IEEE Globecom 2016 Best Paper Award, and was nominated as the IJCAI 2017 Distinguished Best Paper Award. He was also the recipient of the IEEE ICME 2019 and IEEE VCIP 2019 Best Paper Awards.

Han Hu received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2007 and 2012, respectively. He is currently a Professor with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China. His research interests include multimedia networking, edge intelligence, and space-air-ground integrated network. He was the recipeint of several academic awards, including the Best Paper Award of the IEEE TCSVT 2019, Best Paper Award of the IEEE Multimedia Magazine 2015, and Best Paper Award of the IEEE Globecom 2013. He was an Associate Editor for IEEE TMM and Ad Hoc Networks, and a TPC Member of Infocom, ACM MM, AAAI, and IJCAI.

Yonggang Wen (Fellow, IEEE) received the Ph.D. degree in electrical engineering and computer science (minor in Western Literature) from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2008. He is currently a Professor of computer science and engineering with Nanyang Technological University (NTU), Singapore. Since 2018, he has been the Associate Dean (Research) with the College of Engineering, NTU. He was the Acting Director of Nanyang Technopreneurship Centre, NTU from 2017 to 2019, and the Assistant Chair (Innovation) of School of Computer Science and Engineering, NTU from 2016 to 2018. His research interests include cloud computing, green data center, distributed machine learning, blockchain, Big Data analytics, multimedia network, and mobile computing. He was the recipient of multiple journal best papers awards, including IEEE Transactions on Circuits and Systems for Video Technology (2019), IEEE Multimedia (2015), and and several best paper awards from international conferences, including 2020 IEEE VCIP, 2016 IEEE Globecom, 2016 IEEE Infocom MuSIC Workshop, 2015 EAI/ICST Chinacom, 2014 IEEE WCSP, 2013 IEEE Globecom and 2012 IEEE EUC. He was the recipient of 2016 IEEE ComSoc MMTC Distinguished Leadership Award. He is a Fellow of IEEE.