Self-Supervised Learning With Data-Efficient Supervised Fine-Tuning for Crowd Counting

Rui Wang, Yixue Hao, *Member, IEEE*, Long Hu, Jincai Chen, Min Chen, *Fellow, IEEE*, and Di Wu, *Senior Member, IEEE*

***Abstract*—Due to the expensive and laborious annotations of labeled data required by fully-supervised learning in the crowd counting task, it is desirable to explore a method to reduce the labeling burden. There exists a large number of unlabeled images in the wild that can be easily obtained compared to labeled datasets. Based on the characteristics of consistent spatial transformation with the annotations of heads and image, this paper proposes a self-supervised learning framework with unlabeled and limited labeled data for pre-training and fine-tuning crowd counting model (SSL-FT). It includes an online network and a target network that receive the same images but are randomly processed by two defined augmentation transformations. We leverage unlabeled data to pre-train the online network based on a self-supervised loss and small-scale labeled data to transfer the model to a specific domain based on a fully-supervised loss. We demonstrate the effectiveness of the SSL-FT on four public datasets including ShanghaiTech PartA, PartB, UCF-QNRF and WorldExpo'10 utilizing a classical counting model. Experimental results show that our approach performs better than state-of-art semi-supervised methods.**

***Index Terms*—Crowd counting, augmentation transformation, self-supervised learning, self-supervised loss.**

I. Introduction

CROWD counting is to infer the number of people in images or videos. As a basic computer vision task, it has drawn increasing attention recently because of its significance in practical applications, e.g., crowd management, traffic control, emergency evacuation, urban planning [1], [2], [3], [4].

To realize crowd counting, many deep-learning-based methods are proposed with promising performance, most of which mainly can be divided into detection-based approaches [5], [6] and density map-based approaches [7], [8]. The key idea of the former method is to employ bounding boxes to locate the position of the head or body in an image, while the latter aims to generate density maps and sum up them to obtain the counts. Meanwhile, some point-based counting methods are also explored to predict a set of points to represent heads in an image [9], [10]. Considering that the crowd location is not included in the evaluation criteria of the counting model, the point-level labels are actually redundant. To avoid over-labeling, some methods [11], [12] are also proposed to directly generate the number of people and only utilize the count-level labels without the labels of head locations. Though remarkable counting performance has been achieved by the above fully-supervised methods, they all require a large amount of point-level or count-level annotations to guide network learning. In many real applications, however, access to plentiful labeled data is costly and time-consuming, especially when the scenario is very crowded with heavy occlusion.

To this end, exploring a not-so-supervised learning approach is necessary by leveraging limited labeled data as well as unlabeled data. Compared to labeled data, there is a large number of wild images without annotations that can be obtained and utilized to enhance the performance [13] [14]. For unlabeled data, semi-supervised methods [15] commonly generate pseudo-labels by using limited labeled data. However, the unreliable pseudo-labels and the unsatisfying generalization ability are the main drawbacks of these approaches. On the other hand, self-supervised learning without any annotations is a potential technique, which involves the formulation of contrastive loss from the hidden semantics of images not relying on ground truth to update the counting model [16], [17]. Unsurprisingly, the network model trained only using unlabeled data shows weaker performance than semi-supervised methods.

To obtain an accurate counting result without adding additional annotation workload, in our work, we propose a self-supervised learning method with unlabeled data to pre-train the model and fine-tune the model using limited labeled data (SSL-FT). Unlike existing semi-supervised and self-supervised methods, SSL-FT leverages both limited labeled data as well as extensive unlabeled data to train the counting model in a self-supervised manner with an overlapping-consistency strategy, as shown in Fig. 1. Specifically, for an image and its

Manuscript received 30 June 2022; revised 30 October 2022 and 6 February 2023; accepted 17 February 2023. Date of publication 1 March 2023; date of current version 8 May 2023. This work was supported by National Natural Science Foundation of China (NSFC) under Grants 62176101, 62272178, U1911201, and U2001209. The work of Yixue Hao was supported by the Hubei Natural Science Foundation under Grant 2021CFB050. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Jun Wu. *(Corresponding authors: Long Hu; Jincai Chen; Min Chen.)*

Rui Wang, Yixue Hao, and Long Hu are with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: ruiwang2020@hust.edu.cn; yixuehao@hust.edu.cn; hulong@hust.edu.cn).

Jincai Chen is with the Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: jccchen@hust.edu.cn).

Min Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510640, China, and also with the Pazhou Laboratory, Guangzhou 510640, China (e-mail: minchen@ieee.org).

Di Wu is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong 510006, China, and also with the Guangdong Key Laboratory of Big Data Analysis and Processing, Guangdong 510006, China (e-mail: wudi27@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TMM.2023.3251106

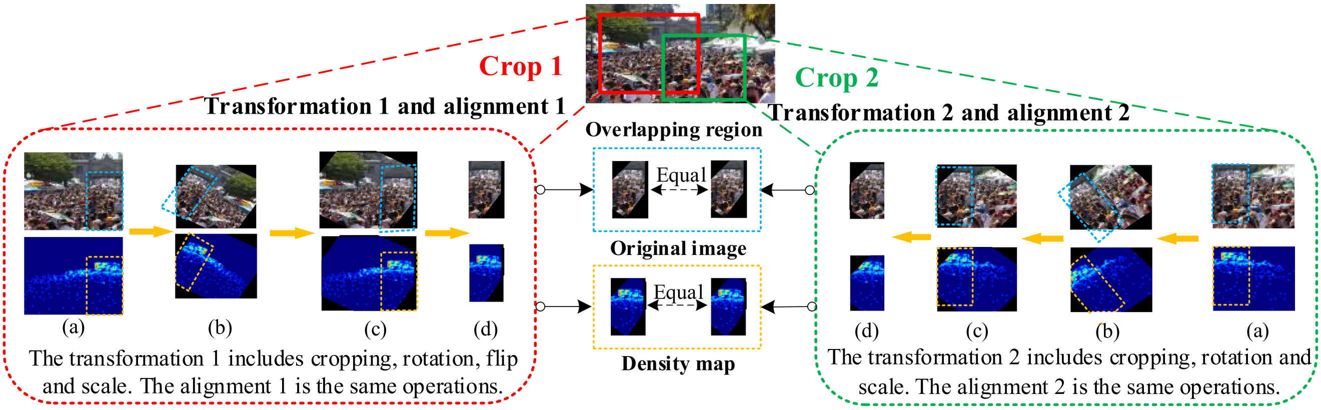
Crop 1
Transformation 1 and alignment 1
Crop 2
Transformation 2 and alignment 2
Overlapping region
Equal
Original image
Equal
Density map
(a)
(b)
(c)
(d)
(d)
(c)
(b)
(a)
The transformation 1 includes cropping, rotation, flip and scale. The alignment 1 is the same operations.
The transformation 2 includes cropping, rotation and scale. The alignment 2 is the same operations.

Fig. 1. The presentation of input image processed by different transformations. (a) Crop the image with a certain size. (b) Scale and perform affine transformation for the image. (c) Align and scale the image. (d) Crop the overlapping region of the two transformed images.

associated density map, when the same image is processed by both transformations, their density maps also change accordingly. Therefore, the density maps of the overlapping areas of the two images are regarded as ground truth to each other, and the gap between the two density maps is named as a self-supervised loss. Our counting model is first trained with a self-supervised of the density maps of the same cropped image after going through different transformations based on self-supervised training. And then we utilize limited labeled data to fine-tune the trained model in a fully-supervised manner. In this way, our SSL-FT addresses the domain-specific shortcomings of limited labeled data while making full use of unlabeled data. There are two networks in the SSL-FT including an online network and a target network. The online network and the target network have exactly the same network structure, which are initialized with random parameters. During model training, they receive cropped patches from the same image processed by different transformations. The update of the online network relies on the output of the target network, and the update of the target network uses the weighting of the previous parameters and the parameters of the current online network. Finally, we use dilated convolutional neural networks for understanding the highly congested scene (CSRNet) [18] due to its strong scalability as the online network and target network to perform experiments based on the SSL-FT framework. In summary, the main contributions are as follows:

- We propose a self-supervised method with the unlabeled dataset to pre-train the counting model, and labeled data to guide the domain adaptation, improving the adaptive performance of the model.
- We design a series of augmentation transformations and alignment schemes that are suitable for labeled and unlabeled images for density map-based models, increasing the variety of datasets.
- Extensive experiments show that SSL-FT achieves superior performance on four challenging datasets, proving the effectiveness and superiority of our method.

The remainder of this paper is organized as follows: Section II presents related works with the crowd counting and self-supervised learning. Section III elaborates on the technical details of our SSL-FT. Section IV performs extensive experiments to demonstrate the effectiveness of the proposed method. Finally, Section V concludes the whole paper.

II. Related Work

In recent years, deep learning has achieved a series of breakthroughs in crowd counting tasks [19], [20]. In this section, we present semi-supervised methods for crowd counting and the achievements of self-supervised learning.

A. Semi-Supervised Methods for Counting

Semi-supervised learning is a commonly used method in the counting task with minimal labeled samples [21], [22], [23], [24], [25]. Liu et al. utilized an unlabeled dataset by ranking cropped images at different scales in a multi-network task [13]. Meng et al. adopted the teacher-student framework to solve the problem of noise supervision of unlabeled data [12]. Yu et al. used the continuity among video frames to reconstruct the density map and guided the training based on the associated loss among frames [26]. Liu et al. designed a segmentation map predictor for unlabeled data and leveraged a threshold method to judge whether there is a head annotation existing on the pixel [11]. Vishwanath et al. designed a Gaussian Processes-based iterative learning framework to count people [23]. Gao et al. extracted abundant relations and structural information and employed partial orders from the latent feature spaces to reduce the estimation error on crowd counting [24]. These methods reduce the burden of manually labeling data but their performances are also influenced by inaccurate pseudo-labels and weak semantic information.

B. Self-Supervised Learning

Self-supervised learning means that the dataset does not provide extra information as supervised signals to train the model, which only takes advantage of the properties of images [27]. It is generally implemented based on contrastive learning [28], which refers to making two kinds of data become similar or

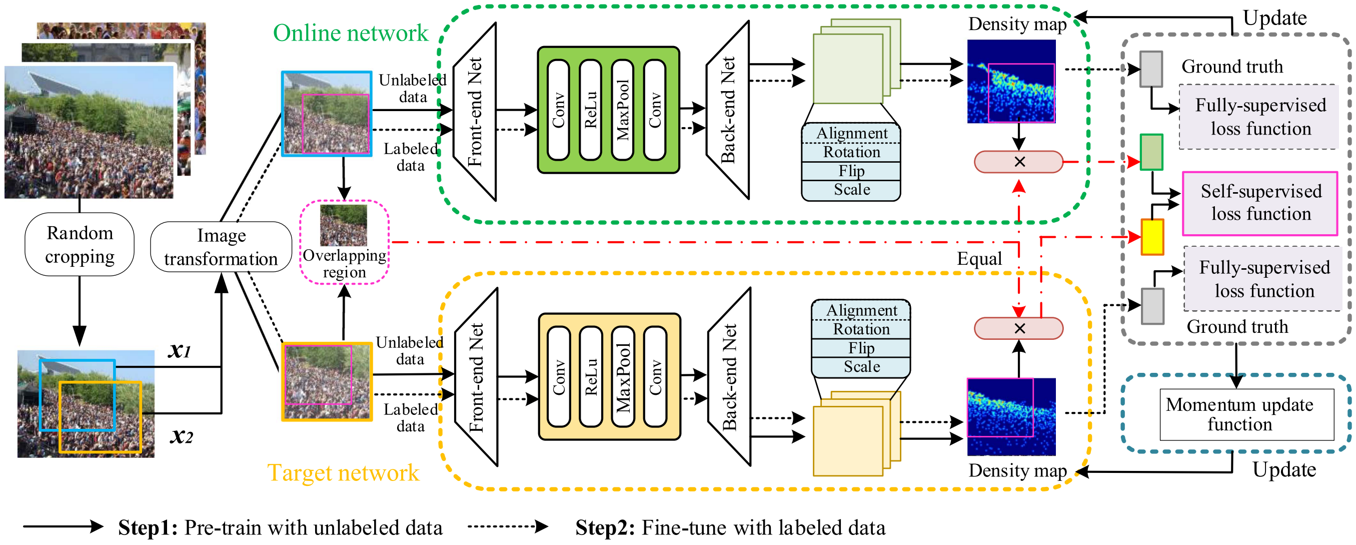

Fig. 2. Overview of our SSL-FT framework. The SSL-FT framework consists of two steps. In Step 1, the model is pre-trained with unlabeled data, and the online network is updated with a self-supervised loss function. In Step 2, the model is fine-tuned with labeled data, and the online network is updated with a self-supervised loss function and a fully-supervised loss function. The target network is updated with the momentum function both in Step 1 and Step 2.

different to train the model via learning latent code. MoCo randomly augments the data from two views, calculates the cosine value of positive and negative examples, and employs a double-tower structure to update the encoder parameters in a momentum way [16]. SimCLR adds negative examples to the loss computation [17]. BYOL improves the performance of the target network by reducing the distance between different views of the same image [29]. MoBY integrates the two methods of MoCo and BYOL and uses positive and negative examples to compare and update the prediction results based on the online network and target network [30]. In terms of crowd counting, Deepak et al. designed a completely self-supervised paradigm based on density regression, exploiting the idea that natural populations obey a power-law distribution [31]. Duan et al. proposed S4-Crowd to formulate self-supervised loss to simulate crowd scale and illumination changes [32]. Instead of other counting works based on self-supervised learning, we aim to utilize the basic consistency characteristic between image transformation and head location to perform self-supervised crowd counting using large-scale unlabeled images and limited labeled images.

III. Method

In this section, we briefly introduce a preliminary and overview of the architecture design of our SSL-FT. Then, we formulate the transformation and alignment scheme based on contrastive learning. Finally, we present the details of the loss function and the process of training and testing.

A. Preliminary

In our work, the typical density map-based regression method is determined as the supporting path to choose counting network. To obtain the continuous labels about the head position, we use the Gaussian kernel to transform the pixel-level annotations to block-level values based on convolution operations. Density maps are generated via geometry-adaptive kernels to avoid the influence of perspective deformation [33]. For an image with labels, the density map is calculated as

$$\mathcal{D}(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x), \tag{1}$$

$$\sigma_i = \beta \overline{d}_i, \tag{2}$$

where x represents a pixel in the image, x_i represents the pixel of the head position, N represents the number of heads, and G_{σ_i} represents a Gaussian kernel with variance of σ_i. The value of σ_i is adaptively determined by the average distance $\overline{d}_i$ between x_i and its nearest head point and the parameter β.

B. Overview of SSL-FT

To make full use of a large amount of unlabeled data and limited labeled data, as shown in Fig. 2, our SSL-FT contains two steps: (1) The density map-based model is first pre-trained with unlabeled data, and the supervision information is derived from the self-supervision of the same image processed by different transformations. (2) Limited domain-specific labeled data is used to fine-tune the model to achieve domain adaptation in a supervised manner.

In the SSL-FT framework, there are two-stream networks consisting of an online network and a target network. The architecture and the number of parameters of the target network are exactly the same as the online network. The training of SSL-FT is divided into two stages. First, the unlabeled images through two kinds of random affine transformations are input into the online network and the target network, and the two networks output different feature maps. The parameters including operations and magnitudes of the affine transformation participate in the inverse transformation of the feature map. And then the contrastive loss of the overlapping region of the feature maps corresponding to the two transformation maps is measured. This contrastive loss is used to perform gradient update for the online network θ_{online}. Meanwhile, the target network θ_{target} is updated by the momentum function, which is expressed as follows:

$$\theta_{target} = \varphi \cdot \theta_{target} + (1 - \varphi) \cdot \theta_{online}, \tag{3}$$

$$\varphi = 1 - \frac{(1 - \varphi_0) \cdot \cos(1 + \pi \cdot k/K)}{2}, \tag{4}$$

where φ is the step size of the momentum update, φ_0 is the initial value, K is the total number of iterations of the training and k is the index of the current training iteration. It is noted that the function avoids network vibration and performance degradation caused by violent updates. And the model is fine-tuned in dynamic uncertain changes through the adjustment of periodic momentum update parameters by the cosine function, making the counting network break through the local optimum. After the network is pre-trained with unlabeled data, minimal labeled data from a specific dataset is used to fine-tune the online network and target network. Different from the pre-training stage, the loss function not only includes the distance of the density maps from overlapping regions between two different transformed images, but also the distance between the estimated map and the ground truth. It boosts the online network to realize the domain adaptation on a specific domain. The update of the target network is still performed with the momentum update function according to (3). In the stage of model inference, the online network predicts the number of people for an input image and the target network is not allowed to participate in the inference stage. In addition, the SSL-FT framework can be applied to strong supervision models, e.g., density map-based regression and point-based regression methods, and weak supervision models of the count-level method.

C. Transformation and Alignment

A series of affine methods and basic transformations are applied to feature maps based on SSL-FT. In this paper, images are transformed according to the following orders:

1) Randomly crop the original image with a size of 90% to 130% of the original image size. It is required to include at least one person in the cropped patch.
2) Scale the cropped images to a uniform size. It is determined by the scale of the target scenario generally varied from 256×256 to 512×512.
3) Randomly flip the image with 50% probability. The flip direction includes vertical flip and horizontal flip.
4) Randomly rotate the image by a certain angle along with the center of the image generally varied from -30 degrees to 30 degrees.
5) Convert the image to a grayscale image with an executing probability of 0.2, and then process the image with the gaussian blur method with an executing probability of 0.2.
6) Apply gaussian noise with the mean and standard deviation of 0 and 0.5, and salt and pepper noise with the proportion of $1e-6$ performed on the image.
7) Adjust the brightness, contrast, saturation, and hue of the image with values of 0.4, 0.4, 0.4 and 0.2, and with an executing probability of 0.6.

The 1) and 2) generate diverse sizes of patches in the images to fit real application scenarios. The 3) and 4) leverage the characteristic of consistency between different transformations to generate supervised information. And the 5), 6), and 7) make augmentations for the image without changing the head locations. The listed series of transformations can be formulated as follows:

$$I' = T(I), \tag{5}$$

where $T(\cdot)$ represents the above-mentioned transformation from 1) to 7), I and I' are the images before and after the transformation, and the parameters of the affine transformation and scale transformation for each image are recorded. And then the transformed images are input into the counting network to obtain the density map M. To match the M with another density map to obtain the supervised information, the inverse transformation is performed on M to obtain the original perspectives, which is expressed as follows:

$$M' = T'(M), \tag{6}$$

where $T'(\cdot)$ represents the inverse transformations, including rotation, flip and scale operations, and the transformations in 5) to 7) are not required to perform an inverse transformation. And the parameters of inverse operations are opposite to the original transformation. M and M' represent the feature maps before and after inverse transformation. For example, a series of transformations in T from 1) to 4) for image I is expressed as crop size with 500×500, resizing to 512×512, vertical flip and rotating 10 degrees. And a series of inverse transformation operations in T' for density map M' is expressed as rotating -10 degrees, vertical flip and resizing to 500×500. For the same image processed by two different transformations of T_1 and T_2, two density maps are obtained through the online network and the target network, represented as M_1 and M_2 respectively. Then the two images are inversely transformed using T_1' and T_2' to obtain M_1' and M_2'. And then the original image and the transformation sequence are used to align the two density maps, and the overlapping area is

$$m = M_1' \cap M_2'. \tag{7}$$

In the image, m is represented as the position coordinates of the upper left and lower right corners of the overlapping regions of M_1' and M_2', which is calculated by comparing the positions they occupy in the original image. Thus, after the image is transformed, the feature maps output by the online network and the target network are aligned to find the overlapping area. The signals for the mutual supervision of these two images are obtained. In the crowd counting task, the pixel of head location strictly follows the one-to-one correspondence in the affine transformation. For different transformed images, it is a brand new image for the counting network. Therefore, through transformation and alignment, it is guaranteed that the pseudo-label can be obtained.

Meanwhile, the extensiveness of the data is expanded through augmentation operations, which enhances the robustness of the model.

D. Loss Function

When the network is trained with unlabeled data, only the information of the transformation and alignment operations is used to achieve self-supervision. The self-supervised loss function for unlabeled data is expressed as

$$\mathcal{L}_{self}(M_1, M_2) = \sum \|M_1'(m) - M_2'(m)\|^2, \tag{8}$$

where $M_1'(m)$ and $M_2'(m)$ represent the feature maps of the aligned overlapping regions output by the online network and target network. When the network is trained on minimal labeled data, in addition to the self-supervised loss, the loss between the transformed image and the ground truth is also included, which is expressed as:

$$\mathcal{L}_{fully} = \sum \|M - G\|^2, \tag{9}$$

where G represents the ground truth of the density map corresponding to the image. Therefore, for the update of the online network, the loss function is expressed as:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{self}^1 + \lambda_1 \cdot \mathcal{L}_{self}^2 + \lambda_2 \cdot \mathcal{L}_{fully}, \tag{10}$$

where λ_1 and λ_2 are the balance parameters of the two loss values and λ_2 is 0 when the model is trained by unlabeled data. To make full use of existing data participating in model training and introduce more information, two cropped images from the same image transformed by different operations are simultaneously input to the online network and the target network, and the cross-computation comparison loss is used to update the network. There are two self-supervised loss values in the training phase, which are $\mathcal{L}_{self}^1(M_1, M_2)$ and $\mathcal{L}_{self}^2(M_2, M_1)$. In $\mathcal{L}_{self}^1(M_1, M_2)$, M_1 is generated by online network and M_2 is generated by target network, and in $\mathcal{L}_{self}^2(M_2, M_1)$, M_1 is generated by target network and M_2 is generated by online network. In this way, the limited labeled data contribute more information to the model training.

IV. Experiments

In this section, we evaluate the performance of the proposed SSL-FT framework on public datasets and compare it with state-of-the-art methods.

A. Experimental Setting

Network structure: We leverage a typical model CSRNet with the simple structure and optimization goal [18] as the counting network of the SSL-FT framework. We choose a dilated convolutional network with a dilation rate equipped with channel numbers of 512, 512, 512, 256, 128, and 64.

Compared methods: We illustrate the details of compared semi-supervised methods with our method as follows:

- *Mean Teacher (MT)* [21]: The method consists of a teacher network and a student network, both of which use the original image and the augmented image to estimate the density map.
- *Unsupervised Data Augmentation (UDA)* [22]: The method utilizes the advanced noise augmentation methods to participate in semi-supervised learning.
- *Interpolation Consistency Training (ICT)* [25]: This is a semi-supervised method, and the model learning is guided by maintaining consistency in the transformation.
- *L2R* [13]: The method utilizes a large number of unlabeled images and implements a novel counting method by ranking the number of people with the cropped images.
- *Gaussian Process-based (GP)* [23]: The method aims to achieve counting using small-scale labeled images and large-scale unlabeled data based on the Gaussian process.
- *Inter-Relationship-Aware Self-Training (IRAST)* [11]: The method is to keep consistent with the crowd distribution of segmentation maps for different scales.
- *Spatial Uncertainty-Aware (SUA)* [12]: This is a semi-supervised model based on a teacher-student framework, both of which estimate the spatial uncertainty map.
- *S4-Crowd* [32]: This is a semi-supervised learning framework for crowd counting. The two supervised loss functions are proposed to simulate the variations among images and generate fine-grained pseudo-labels.
- *S^{2}FPR* [24]: The method extracts the structural information from the latent feature space. And unlabeled data is employed to enhance the representation ability.

Datasets: We evaluate our method on four popular datasets including ShanghaiTech part A (SHA), ShanghaiTech part B (SHB) [33], WorldExpo'10 (WE) [38] and UCF-QNRF [39]. We randomly take out a proportion of labeled data in the training dataset of these four datasets and evaluate the model on the complete testing dataset.

Implementation details: When transforming the image, we use the horizontal flip direction instead of the vertical flip. Other transformation parameters are set as mentioned above. We employ the training dataset of SHA removing labels to complete the pre-training task on CSRNet as unlabeled samples, and then only select 25% of the datasets in the target domain including SHB, WE and UCF-QNRF dataset for fine-tuning the model. The four datasets are processed by Gaussian kernel with the fixed kernel size of 15×15 and σ of 4 to obtain the ground truth of the labeled dataset [33]. In particular, when making comparisons on the SHA dataset, we use the SHB dataset without their labels to pre-train the counting model and then the labeled SHA dataset is used to fine-tune the model. Compared with other semi-supervised models, we adopt external images not belonging to the target dataset, which improves the robustness of the model. We adopt Adam optimizer with a learning rate and weight decay rate of $1e-4$. The initial value of the momentum update parameter φ is 0.99. The cropped size for each image is 512×512 on all datasets. The λ_1 and λ_2 are set as 1 and 1 when there are labeled images.

Evaluation protocol: We use common evaluation criteria, mean absolute error (MAE) and mean squared error (MSE),

TABLE I
EVALUATIONS ON FOUR CROWD COUNTING DATASETS COMPARED WITH THE FULLY-SUPERVISED METHODS AND SEMI-SUPERVISED METHODS (* REPRESENTS THE PERCENTAGES OF LABELED DATA ON SHA, SHB, WE AND UCF-QNRF ARE 30%, 30%, 28% AND 60%)

Methods		Labeled data	SHA		SHB		WE		UCF-QNRF	
			MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Fully-supervised	MCNN [32]	100%	110.2	173.2	26.4	41.3	11.6	-	277	426
	CMTL [33]	100%	101.3	152.4	20.0	31.1	-	-	252	514
	Switch-CNN [34]	100%	90.4	135.0	21.6	33.4	9.4	-	228	445
	CP-CNN [35]	100%	73.6	106.4	20.1	30.1	8.9	-	-	-
	ACSCP [36]	100%	75.7	102.7	17.2	27.4	7.5	-	-	-
	CSRNet [18]	100%	68.2	115.0	10.6	16.0	8.6	-	119.2	211.4
Semi-supervised	MT [21]	50%	88.2	151.1	15.9	25.7	-	-	147.2	249.6
	MT [21]	*	94.5	156.1	15.6	24.5	14.1	-	145.5	250.3
	UDA [22]	*	93.8	157.2	15.7	24.1	14.2	-	144.7	255.9
	ICT [37]	*	92.5	156.8	15.4	23.8	14.9	-	144.9	250
	L2R [13]	50%	86.5	148.2	16.8	25.1	-	-	145.1	256.1
	L2R [13]	*	90.3	153.5	15.6	24.4	13.9	-	148.9	249.8
	GP [23]	25%	91.0	149.0	-	-	-	-	147.0	226.0
	IRAST [11]	*	86.9	148.9	14.7	22.9	11.1	-	135.6	233.4
	SUA [12]	50%	-	-	14.1	20.6	-	-	-	-
	S4-Crowd [31]	25%	84.5	148.2	12.9	21.6	-	-	140.5	224.8
	S^2FPR [24]	25%	93.5	148.4	-	-	-	-	146.9	237.2
Our method	**Our baseline**	25%	86.0	137.9	15.0	28.1	**10.0**	-	**132.5**	**210.2**
	Our SSL-FT	25%	**82.1**	**132.9**	**11.5**	**20.2**	12.1	-	151.0	259.1

to evaluate counting performance:

$$MAE = \frac{1}{N}\sum_{i}^{N} \|\hat{g}_i - g_i\|, \tag{11}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i}^{N} \|\hat{g}_i - g_i\|^2}, \tag{12}$$

where N is the number of images, and $\hat{g}$ and g are the predicted count and real count in the image.

B. Counting Results

We compare our method with fully-supervised and semi-supervised methods, as shown in Table I. Our baseline method is directly trained with minimal labeled data without going through the pre-training stage on unlabeled data. Our SSL-FT means that we first pre-train the model with unlabeled data not belonging to the target dataset, and then fine-tune it with domain-specific labeled datasets. Since we adopt CSRNet as the network architecture of the SSL-FT framework, our method is weaker than the fully-supervised method CSRNet. However, the performance is stronger than some fully supervised models such as CMTL [34] and Switch-CNN [35] on some datasets. For semi-supervised methods, it can be found that our method significantly outperforms them on both MAE and MSE on the four datasets, although we use less labeled data than these methods by 5% or even 35%. For three semi-supervised crowd counting models, our SSL-FT is better than these three models including GP, IRAST and S4-Crowd on both SHA and SHB datasets, better than S4-Crowd by 2.4 and 15.4 in MAE and MSE, and our baseline also performs better on WE and UCF-QNRF datasets than these three methods. The MAE and MSE of our method on UCF-QNRF are 132.5 and 210.2. Compared with the S^2FPR, SSL-FT performs better on SHA and UCF-QNRF datasets when using the same amount of labeled images.

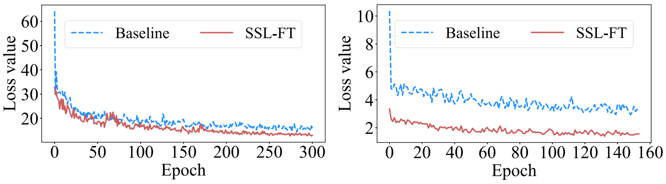

Fig. 3. The comparisons of convergence speed between baseline model and SSL-FT model. The loss values include self-supervised loss value and fully-supervised loss value.

Meanwhile, we observe that our SSL-FT shows the best performance on SHA and SHB, while our baseline performs better on WE and UCF-QNRF. In our experiment, the number of labeled data of WE and UCF-QNRF are 845 and 300. It is concluded that when there are many labeled images on the training dataset, the pre-trained model has little effect on improving the overall performance of the model. The results are consistent with the conclusions in [40] and [41]. In addition, the pre-training dataset SHB has a gap with the fine-tuning dataset UCF-QNRF and WE in the crowd distribution. As a result, the pre-trained model is not trained well to learn the region of interest that is consistent with the target domain. When the model is transferred to the target scenario, the performance of the model is decreased. Therefore, SSL-FT is helpful in applications with small-scale labeled datasets. When there are more labeled samples, the effect of pre-training is diminished. In this case, we can just leverage the baseline method to train the model and conduct inference.

C. Ablation Studies

Effect of pre-training for convergence rate: We compare the convergence rate of the baseline method and the SSL-FT method on the SHA dataset and SHB dataset, as shown in Fig. 3. It can be seen that the initial loss value of SSL-FT is lower than

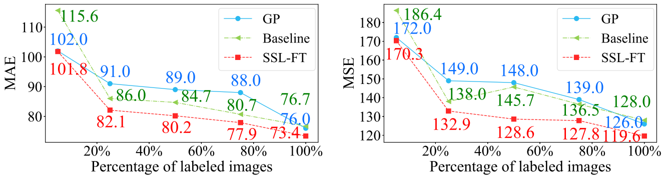

Fig. 4. The effect of the percentage of labeled images. We evaluate our SSL-FT method and GP method [23] on the labeled SHA dataset with 5%, 25%, 50%, 75% and 100% of labeled data.

TABLE II
THE EFFECT OF THE NUMBER OF UNLABELED IMAGES. THERE ARE 50, 100, 200, 300 AND 400 UNLABELED IMAGES IN SHB USED TO PRE-TRAIN THE MODEL AND FINE-TUNE WITH 75 IMAGES IN SHA

The number of unlabeled images	MAE	MSE
50	80.01	128.29
100	**76.30**	**125.48**
200	87.97	139.30
300	89.98	138.97
400	82.12	132.92

the baseline, and during 300 epochs, its loss value keeps decreasing on the SHA dataset. And the same phenomenon appears on the SHB dataset. Therefore, it can be concluded that the pre-trained model using SSL-FT converges faster than the baseline method, which is beneficial in scenarios where the collection of labeled data is difficult. And the convergence rate depends on the size of the dataset and the transferring ability of the pre-trained model. Therefore, the computational resources consumed in the pre-training stage can be compensated by reducing the number of training epochs in the fine-tuning stage, showing the energy efficiency of our method are not weaker than other semi-supervised methods.

Effect of the percentage of labeled images: We evaluate the effect of the proportion of labeled data, as shown in Fig. 4. The percentages of 5%, 25%, 50%, 75% and 100% of the labeled data are selected to train the baseline model and the SSL-FT model. We compare the MAE and MSE of baseline, the SSL-FT method and the GP method [23] which has the same configurations with SSL-FT on the testing dataset. It can be found that when the proportion of labeled data is only 5%, the performance of SSL-FT and GP is similar but the baseline model is weaker than theirs. When the percentages of labeled data are 25%, 50% and 75%, the performance gap between the SSL-FT and GP increases, and our baseline and SSL-FT method are far superior to the GP method. For the model trained on 75% of labeled data, the MAE and MSE of SSL-FT are 77.9 and 127.8, which are reducer by 10 than the GP method. When these models are trained on the fully labeled dataset, the performance gap of the models is further narrowed. This proves that our SSL-FT method is robust on labeled datasets in different settings, which can be leveraged to generalize better performance in the real world with diverse scenarios.

Effect of the percentage of unlabeled images: We validate the effect of the number of unlabeled data from 50 to 400, and fix the number of labeled images to 75. According to the results shown in Table II, it can be found that the performance of the model varies greatly, which may be related to the matching degree of scenarios between unlabeled data and labeled images. When there are more unlabeled images participating in the stage of pre-training, the distribution between the source domain and the target domain has a bad effect on the model training. When the number of unlabeled images is 100, the MAE and MSE reach better performance with 76.30 and 125.48. If the distributions of unlabeled data and labeled data are quite different, the pre-training stage for the model may weaken the robustness of the model.

TABLE III
THE EFFECT OF THE TYPE OF PRE-TRAINING DATASET. THE UNLABELED SHA, SHB, UCF-QNRF AND WE DATASET ARE USED AS THE PRE-TRAINING DATASET TO TRAIN THE MODEL AND FINE-TUNE ON THE 75 IMAGES IN SHA

The type of pretrained dataset	MAE	MSE
SHA	**80.78**	132.38
SHB	82.12	132.92
UCF-QNRF	85.19	**129.15**
WE	92.01	139.02

TABLE IV
THE EFFECT OF CONTRAST MOMENTUM ON THE 75 IMAGES IN SHA

Contrast momentum (φ)	MAE	MSE
0.99	**82.12**	132.92
0.95	83.45	**129.66**
0.90	83.91	134.99
0.85	86.97	133.14
0.80	85.48	132.63

Effect of the type of pre-training dataset: We evaluate the effect of the type of pre-training dataset with unlabeled SHA, SHB, UCF-QNRF and WE dataset. The model is fine-tuned with the left-labeled SHA dataset and the results are shown in Table III. It can be observed that when using 225 unlabeled images in SHA as the pre-training dataset, the fine-tuned model reaches the best performance with the MAE and MSE of 80.78 and 132.38 compared with other models pre-trained on SHB, UCF-QNRF and WE datasets. And a large distribution gap between SHA and WE according to the results weaker than 11.23 and 6.64 of MAE and MSE on WE dataset. It is concluded that we need to select a suitable unlabeled dataset matching the target domain as much as possible to improve counting performance.

Effect of the contrast momentum: We evaluate the effect of the step size of contrast momentum with 25% labeled data of the SHA dataset, as shown in Table IV. When φ is 0.99, MAE and MSE are 82.12 and 132.92 to achieve better performance. However, when we set φ less than 0.90, the counting performances degrade greatly. As the contrast momentum becomes smaller, it proves that the model retains less of the parameters recorded from the previous step, and the performance of the model tends to deteriorate. Generally, φ should maintain a relatively large value to fully retain the previous parameters to ensure steady updates of the model.

Effect of the type of transformation operation: We explore the effect of multiple operations in the transformation orders with 25% labeled data of SHA dataset, and the results are shown in

TABLE V
THE EFFECT OF TRANSFORMATION TYPE ON THE 75 IMAGES IN SHA

The type of transformation	MAE	MSE
w/o Rotation	86.98	140.22
w/o Brightness	86.93	138.33
w/o Gaussian and salt-and-pepper noises	85.64	134.90
w/ All (T)	**82.12**	**132.92**

TABLE VI
THE EFFECT OF SELF-SUPERVISED LOSS FUNCTION ON THE 75 IMAGES IN SHA

Loss function	MAE	MSE
w/o $\mathcal{L}_{self}$	85.09	138.96
w/o $\mathcal{L}^1_{self}$	83.95	133.60
w/o $\mathcal{L}^2_{self}$	95.31	139.04
w/ All ($\mathcal{L}$)	**82.12**	**132.92**

Table V. The model performs worse when transformations including rotation, brightness, and gaussian and salt noises are removed. Specifically, if there is a lack of rotation transformation, the MAE and MSE are 86.98 and 140.22. It is also observed that gaussian and salt-and-pepper noises have less effect on performance improvement. Therefore, we need to perform reasonable augmentation transformation and choose appropriate parameters for the images.

Effect of the self-supervised loss function: In the stage of the fine-tuning model with 25% labeled data, we study the effect of the self-supervised loss $\mathcal{L}_{self}$, as shown in Table VI. The $\mathcal{L}_{self}$ means that the two transformed images for the same image are both input to the online network and the target network simultaneously, and then the self-supervised loss values of the overlapping area are alternately calculated. When we do not use self-supervised loss or use only one of the self-supervised loss values for transformed images, the performance is weaker than using two alternating loss values. This indicates that the self-supervised loss also has a large effect when the model is trained on labeled data. The supervised loss can be further applied in the fully-supervised methods as an augmentation strategy to improve the counting accuracy.

V. CONCLUSION

This paper proposes a self-supervised learning framework SSL-FT utilizing unlabeled data and minimal labeled data for crowd counting. First, the counting model is pre-trained by unlabeled data. The model is then fine-tuned on the labeled data to realize the domain adaptation in a specific dataset. We present transformation and alignment schemes during self-supervised learning. Extensive experiments on four public datasets demonstrate that our method achieves state-of-the-art performance. SSL-FT is expected to be widely used with low labeling costs while ensuring counting performance in the real world.

REFERENCES

[1] M. K. K. Reddy, M. Rochan, Y. Lu, and Y. Wang, “AdaCrowd: Unlabeled scene adaptation for crowd counting,” *IEEE Trans. Multimedia*, vol. 24, pp. 1008–1019, 2022.

[2] M. Wang, H. Cai, X. Han, J. Zhou, and M. Gong, “STNet: Scale tree network with multi-level auxiliator for crowd counting,” *IEEE Trans. Multimedia*, 2022, early access, Jan. 13, 2022, doi: 10.1109/TMM.2022.3142398.

[3] X. Jiang et al., “Density-aware multi-task learning for crowd counting,” *IEEE Trans. Multimedia*, vol. 23, pp. 443–453, 2021.

[4] C. Zhao et al., “Deep fusion feature representation learning with hard mining center-triplet loss for person re-identification,” *IEEE Trans. Multimedia*, vol. 22, pp. 3180–3195, 2020.

[5] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, “Density map regression guided detection network for RGB-D crowd counting and localization,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1821–1830.

[6] S. Hou et al., “Improved instance discrimination and feature compactness for end-to-end person search,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2079–2090, Apr. 2022.

[7] Z. Ma, X. Wei, X. Hong, and Y. Gong, “Bayesian loss for crowd count estimation with point supervision,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6141–6150.

[8] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, and A. Hauptmann, “Learning spatial awareness to improve crowd counting,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6151–6160.

[9] Q. Song et al., “Rethinking counting and localization in crowds: A purely point-based framework,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3345–3354.

[10] H. Liu, Q. Zhao, Y. Ma, and F. Dai, “Bipartite matching for crowd counting with point supervision,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 860–866.

[11] Y. Liu, L. Liu, P. Wang, P. Zhang, and Y. Lei, “Semi-supervised crowd counting via self-training on surrogate tasks,” in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 242–259.

[12] Y. Meng et al., “Spatial uncertainty-aware semi-supervised crowd counting,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15529–15539.

[13] X. Liu, J. Van De Weijer, and A. D. Bagdanov, “Leveraging unlabeled data for crowd counting by learning to rank,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7661–7669.

[14] Q. Wang, J. Gao, W. Lin, and Y. Yuan, “Learning from synthetic data for crowd counting in the wild,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8190–8199.

[15] Z. Zhao, M. Shi, X. Zhao, and L. Li, “Active crowd counting with limited supervision,” in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 565–581.

[16] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual representation learning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9726–9735.

[17] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1597–1607.

[18] Y. Li, X. Zhang, and D. Chen, “CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1091–1100.

[19] Q. Wang, J. Gao, W. Lin, and X. Li, “NWPU-crowd: A large-scale benchmark for crowd counting and localization,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 6, pp. 2141–2149, Jun. 2021.

[20] X. Jiang et al., “Attention scaling for crowd counting,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4705–4714.

[21] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1195–1204.

[22] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, “Unsupervised data augmentation for consistency training,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6256–6268.

[23] V. A. Sindagi, R. Yasarla, D. S. Babu, R. V. Babu, and V. M. Patel, “Learning to count in the crowd from limited labeled data,” in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 212–229.

[24] J. Gao et al., “S^{2}FPR: Crowd counting via self-supervised coarse to fine feature pyramid ranking,” 2022, *arXiv:2201.04819*.

[25] V. Verma et al., “Interpolation consistency training for semi-supervised learning,” *Neural Netw.*, vol. 145, pp. 90–106, 2022.

[26] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, “Uncertainty-aware self-ensembling model for semi-supervised 3D left atrium segmentation,” in *Proc. 22nd Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2019, pp. 605–613.

[27] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, “S4L: Self-supervised semi-supervised learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1476–1485.

[28] J. Li, P. Zhou, C. Xiong, and S. Hoi, "Prototypical contrastive learning of unsupervised representations," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=KmykpuSrjcq

[29] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[30] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9992–10002.

[31] D. Babu Sam et al., "Completely self-supervised crowd counting via distribution matching," in *Proc. 17th Eur. Conf. Comput. Vis.*, 2022, pp. 186–204.

[32] H. Duan and Y. Guan, "S4-crowd: Semi-supervised learning with self-supervised regularisation for crowd counting," 2021, *arXiv:2108.13969*.

[33] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 589–597.

[34] V. A. Sindagi and V. M. Patel, "CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *Proc. IEEE 14th Int. Conf. Adv. Video Signal Based Surveill.*, 2017, pp. 1–6.

[35] D. Babu Sam, S. Surya, and R. Venkatesh Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4031–4039.

[36] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1879–1888.

[37] Z. Shen et al., "Crowd counting via adversarial cross-scale consistency pursuit," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5245–5254.

[38] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 833–841.

[39] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 532–546.

[40] K. He, R. Girshick, and P. Dollár, "Rethinking imageNet pre-training," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4917–4926.

[41] B. Zoph et al., "Rethinking pre-training and self-training," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 3833–3845.

Rui Wang received the bachelor's degree in computer science and technology from Lanzhou University, Lanzhou, China. He is currently working toward the Ph.D degree with the Embedded and Pervasive Computing Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. Her research interests include computer vision and emotion recognition. ruiwang2020@hust.edu.cn

Yixue Hao (Member, IEEE) received the Ph.D degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, in 2017. He is currently an Associate Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. His research interests include 5G network, Internet of Things, edge computing, edge caching, and cognitive computing. yixuehao@hust.edu.cn

Long Hu is currently an Assistant Professor with the School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China. From August 2015 to April 2017, he was a Visiting Student with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, USA. His research interests include the Internet of Things, software defined networking, caching, 5G, body area networks, body sensor networks, and mobile cloud computing. hulong@hust.edu.cn

Jincai Chen received the Ph.D. degree majoring in computer architecture from Xi'an Jiaotong University, Xi'an, China, in 2000. He is currently a Professor. From 2001 to 2003, he was engaged in postdoctoral research with the National Laboratory of Storage Systems, Huazhong University of Science and Technology, Wuhan, China. During 2011–2012, he was a Visiting Researcher with the University of California, Santa Cruz, Santa Cruz, CA, USA. In recent years, he has taken charge of four projects of the National Natural Science Foundation of China, participated in 1 "973" project and four "863" projects. He has authored or coauthored more than 80 papers, and authorized more than 30 invention patents (including two international invention patents). His main research directions are storage theory and technology, Big Data, machine learning, image processing, and affective computing. jcchen@hust.edu.cn.

Min Chen (Fellow, IEEE) is currently a Full Professor with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. He is also the Director of Embedded and Pervasive Computing Lab, Huazhong University of Science and Technology (HUST), Wuhan, China. He was an Assistant Professor with the School of Computer Science and Engineering, Seoul National University, Seoul, South Korea, before he joined HUST. His Google Scholar Citations reached more than 39,000 with an H-index of 93. His top paper was cited more than 4,090 times. From 2018 to 2022, he was selected as Highly Cited Researcher. He was the recipient of IEEE Communications Society Fred W. Ellersick Prize in 2017, IEEE Jack Neubauer Memorial Award in 2019, and IEEE ComSoc APB Oustanding Paper Award in 2022. He is a Fellow of IET. He is the Founding Chair of IEEE Computer Society Special Technical Communities on Big Data. He is the Chair of IEEE Globecom 2022 eHealth Symposium. minchen@ieee.org

Di Wu (Senior Member, IEEE) received the B.S. degree from the University of Science and Technology of China, Hefei, China, in 2000, the M.S. degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2003, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2007. From 2007 to 2009, he was a Postdoctoral Researcher with the Department of Computer Science and Engineering, Polytechnic Institute of New York University, Brooklyn, NY, USA, advised by Prof. K. W. Ross. He is currently a Professor and the Associate Dean of the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China. His research interests include edge/cloud computing, multimedia communication, Internet measurement, and network security. He was the recipient of the IEEE INFOCOM 2009 Best Paper Award and IEEE Jack Neubauer Memorial Award. wudi27@mail.sysu.edu.cn