Resolution-Aware Knowledge Distillation for Efficient Inference

Zhanxiang Feng, *Member, IEEE*, Jianhuang Lai, *Senior Member, IEEE*, and Xiaohua Xie

Abstract— **Minimizing the computation complexity is essential for the popularization of deep networks in practical applications. Nowadays, most researches attempt to accelerate deep networks by designing new network structure or compressing the network parameters. Meanwhile, transfer learning techniques such as knowledge distillation are utilized to keep the performance of deep models. In this paper, we focus on accelerating deep models and relieving the computation burden by using low-resolution (LR) images as inputs while maintaining competitive performance, which is rarely researched in the current literature. Deep networks may encounter serious performance degradation when using LR inputs because many details are unavailable from LR images. Besides, the existing approaches may fail to learn discriminative features for LR images because of the dramatic appearance variations between LR and high-resolution (HR) images. To tackle with the above problems, we propose a resolution-aware knowledge distillation (RKD) framework to narrow the cross-resolution variations by transferring knowledge from HR domain to LR domain. The proposed framework consists of a HR teacher network and a LR student network. First, we introduce a discriminator and propose an adversarial learning strategy to shrink the variations between inputs with changing resolution. Then we design a cross-resolution knowledge distillation (CRKD) loss to train discriminative student network by exploiting the knowledge of the teacher network. The CRKD loss is consisted of a resolution-aware distillation loss, a pair-wise constraint, and a maximum mean discrepancy loss. Experimental results on person re-identification, image classification, face recognition, and defect segmentation tasks demonstrate that RKD outperforms traditional knowledge distillation method by achieving better performance with lower computation complexities. Furthermore, CRKD surpasses the state-of-the-art knowledge distillation methods in transferring knowledge across different resolutions under RKD framework, especially when coping with large resolution differences.**

Index Terms— **Knowledge distillation, deep learning, cross-resolution discrepancy, adversarial learning.**

Manuscript received October 11, 2020; revised May 3, 2021 and June 16, 2021; accepted July 26, 2021. Date of publication August 4, 2021; date of current version August 9, 2021. This work was supported by the NSFC under Grant 61902444, Grant 62076258, and Grant 61672544. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhu Li. *(Corresponding author: Jianhuang Lai.)*

Zhanxiang Feng is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China (e-mail: fengzhx7@mail.sysu.edu.cn).

Jianhuang Lai is with the School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou 510006, China, also with Guangzhou Xinhua University, Guangzhou 510006, China, and also with the Guangdong Key Laboratory of Information Security Technology, Guangzhou 510006, China (e-mail: stsljh@mail.sysu.edu.cn).

Xiaohua Xie is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China, and also with the Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Sun Yat-sen University, Guangzhou 510006, China (e-mail: xiexiaoh6@mail.sysu.edu.cn).

Digital Object Identifier 10.1109/TIP.2021.3101158

I. Introduction

DEEP learning has attracted widespread research attention in recent years. Because of the emergence of large-scale benchmarks and the superior learning ability, deep networks have achieved remarkable breakthroughs over a variety of computer vision and machine learning topics, including image classification [1]–[3], object segmentation [4], [5], face recognition [6], [7], person re-identification [8]–[10], etc. Nowadays, the researchers design cumbersome structures to achieve good performance, making deep neural networks high computation consumptions [11]–[13]. A prominent problem of deep networks lies in the balance between the performance and the computation complexities. Deeper networks achieve higher performance. However, the computation cost of deep networks grows dramatically along with the increasing number of network layers, which restricts the applications of deep networks towards realistic environments where computation resources are limited.

Recently, knowledge distillation [14] is proposed to obtain a better tradeoff between the performance and complexity. The knowledge distillation technique focuses on transferring the priors of the cumbersome teacher network to the compact student network while generating discriminant features. Knowledge distillation is effective for reducing the computation burden of deep networks by manipulating the structure of student networks and compressing the network parameters. Meanwhile, the student network generates robust features by exploiting the knowledge of the teacher network. Therefore, the student network is computational efficient without losing too much discriminative power. However, knowledge distillation is inconvenient for practical applications in unknown environments. We have to design different simplified network structures for changing environments with varying computing resources. Besides, traditional knowledge distillation techniques only consider accelerating deep networks by compressing network parameters. The resolution of the input for the neural network, another important factor which significantly affects the robustness and the executive speed of a deep model, is largely ignored by current literatures.

Intuitively, we can extract stronger features and get more information from HR images rather than from LR images. Compared with LR images, HR images contain shaper and clearer edges and provide wider receptive fields for the same network. Nevertheless, the computation cost of deep networks increases quadratically with the resolution of the input. When the resolution of the input becomes 4 times as the original resolution, the computation complexities of a network become approximately 16 times and the executive speed is much slower. The above phenomenon can be considered in another perspective, that is we can accelerate a deep model by using LR images as inputs instead of HR images. The ideal situation is that deep networks can extract robust representations using LR inputs. However, extracting features from LR inputs may lead to a serious performance degradation for deep networks because of the lack in image details. When the cross-resolution variations become too large, the existing techniques such as knowledge distillation and transfer learning may fail to shrink the gaps between LR and HR images. To summarize, adopting LR input leads to fast feature extraction process and low computation consumptions whereas using HR input is beneficial for learning discriminative features with more information. Yet this contradiction is largely ignored by the existing researches, and few studies have made efforts to adopt LR inputs to improve the computation efficiency of deep networks while preserving the network performance.

To tackle with the contradiction between efficiency and accuracy, we propose a resolution-aware knowledge distillation (RKD) framework to accelerate deep networks while generating representative features. RKD is designed to improve the overall performance of deep networks by distilling knowledge from HR domain to LR domain. We accelerate the executive speed of deep models by extracting features using LR inputs. Besides, our method exploits the semantic knowledge from HR domain to enhance the discriminative power of features from LR images. The proposed framework is composed of a HR teacher network and a LR student network, and is implemented in a two-step manner. First, we train the teacher network using HR inputs to explore knowledge in the HR domain. Then, we introduce a discriminator and employ an adversarial learning manner between the teacher network and the student network to minimize the cross-resolution margin. Finally, we propose a cross-resolution knowledge distillation loss (CRKD) to distill knowledge from the HR teacher model to the LR student model. Notably, the CRKD consists of a resolution-aware distillation loss, a pari-wise constraint, and a maximum mean discrepancy loss. The benefits of the RKD method are as follows. First, the RKD technique can largely lessen the computation burden of a network and expedite the feature extraction process through generating features from LR images. Second, the RKD method is valuable for extracting stronger features from LR images by inferring useful information from high-resolution images. Third, the RKD focuses on transferring knowledge across networks with inputs of changing resolutions, which is an under-study topic. Finally, compared with the state-of-the-art methods, the RKD approach provides a more reliable and flexible solution to shrink the margin across images with large resolution variations. Because the computation efficiency of a deep network is highly correlated to the resolution, we can easily adjust the resolution of the inputs for the student network according to the computing power available in realistic environments. Experiments are conducted on multiple popular computer vision tasks including person re-identification, image classification, and face recognition. Experimental results verify the effectiveness of RKD in improving the overall performance of deep networks. RKD achieves almost the same performance as the teacher network with only 7% to 35% computation complexities. Moreover, RKD significantly improves the discriminative power of deep features from LR inputs and remarkably outperforms the state-of-the-art methods, especially when dealing with very low resolution images.

In summary, we make the following contributions.

- We propose a novel RKD framework to flexibly accelerate deep networks while maintaining the recognition performance.
- We introduce a discriminator and employ the adversarial learning strategy to narrow the gap between the features from different resolution domains.
- We propose the CRKD loss to transfer knowledge from the HR teacher model to the LR student model.

II. Related Work

A. Compression and Network Acceleration

With the development of deep learning theory, deep networks have achieved significant breakthroughs for many computer vision tasks. As the neural network becomes deeper and deeper, the high computation complexity becomes the bottleneck for the applications of deep learning methods. Researchers have made great efforts to compress the parameters of deep models and accelerate the feature extraction process in recent years. Howard *et al.* [15] propose to reduce redundant parameters and build lightweight deep networks using an efficient structure named MobileNet which trades off between latency and accuracy by replacing traditional convolution layers with depth-wise separable convolutions. Zhang *et al.* [16] propose to adapt mobile applications with very limited computing power using a computation-efficient architecture named ShuffleNet which achieves fast execution and accurate prediction by employing point-wise group convolution and channel shuffle. Wu *et al.* [17] introduce the Quantized CNN framework to accelerate feature extraction and reduce the storage consumptions on memory overhead of CNN models. Han *et al.* [18] and Lin *et al.* [19] propose a pruning method to accelerate deep networks by removing the neurons which have low influence on the output features.

Some researchers have tried to accelerate deep networks by using LR images. Dong *et al.* [20] and Feng *et al.* [21] propose to accelerate the executive speed of super-resolution models by extracting features and implementing SR on LR images. Chen *et al.* [22] propose the Wavelet-like Auto-Encoder (WAE) which decomposes the input image into two low-resolution channels to accelerate deep networks, of which one carries low-frequency information and the other carries high-frequency information correspondingly. The deep network is accelerated by extracting features from the low-frequency channel and using a lightweight model to extract features from the high-frequency channel. Besides, some object detection [23] and de-noising methods [24], [25] down-sample the images to accelerate the feature extraction process. Zhang *et al.* [25] propose the FFDNet based on pixel-shuffle technique which works on down-sampled sub-images for fast and efficient CNN based image denoising. The major difference between RKD and the above methods lies in that RKD utilizes the prior knowledge from the HR teacher network to improve the performance of the LR student network while the knowledge in HR domain is not used by the above methods. Moreover, RKD can directly resize the images to the desired resolution to achieve varying computation complexities while most approaches such as WAE has to modify the structure of the encoders and decoders to adapt different resolutions. Notably, the super-resolution (SR) methods are also useful for improving the performance of deep networks using LR images. The resolution distillation method is different from the SR methods in the following aspects. First, the ambition of the SR method is to enhance the quality of the LR inputs while the resolution distillation method is designed to accelerate deep networks while keeping high performance. Second, the SR methods will bring additional computational consumption, which is contradicted to the goal of the resolution distillation approach.

B. Knowledge Distillation

The above researches tend to accelerate deep networks by simplifying or improving the network structure. Recently, some studies consider exploiting the prior knowledge to compress the parameters of deep networks. Hinton *et al.* [14] propose the pioneering knowledge distillation approach to compress network parameters by transferring the knowledge from a cumbersome model to a small model. Since then, knowledge distillation has attracted increasing research attention [26]–[31]. The existing studies generally employ distillation loss between the teacher outputs and the student outputs to force the student network to imitate the behaviors of the teacher network. Romero *et al.* [32] propose the FitNets to train thin and deep networks and guide the training process of the wide and shallower networks by exploiting intermediate-level hints from the teacher hidden layers. Zagoruyko *et al.* [33] integrate the attention mechanism into knowledge distillation and show that imitating the attention information dramatically improves the performance of the student network. Heo *et al.* [34] propose to discover samples supporting a decision boundary by employing an adversarial attack so as to transfer information that is closely related to the decision boundary. Peng *et al.* [27] propose to exploit the correlation information and design a correlation congruence based knowledge distillation framework to transfer not only the instance level information but also the correlation knowledge between different instances. Liu *et al.* [35] introduce a novel holistic distillation approach to train a compact segmentation network by distilling the structured holistic knowledge. Wu *et al.* [29] propose a multi-teacher adaptive similarity distillation framework to transfer knowledge from multiple teacher networks to a compact student network without using the training data from the source benchmarks. Tian *et al.* [36] propose a contrastive learning method to transfer structural knowledge from the teacher network and capture more significant information in the representations of teacher network. Park *et al.* [37] introduce the dubbed relational knowledge distillation approach which transfers mutual relations of data examples by imposing distance-wise and angle-wise distillation losses that penalize structural differences in relations. Wang *et al.* [38] integrate the mixup and active learning techniques into a knowledge distillation framework to transfer knowledge from a black-box teacher model and train a high-performing student neural network in a data-efficient manner. Yuan *et al.* [39] develop a teacher-free knowledge distillation framework which achieves competitive performance without using a stronger teacher model. A student model is optimized from a manually-designed regularization distribution.

The knowledge distillation technique provides an effective solution to reduce the number of parameters of a deep network while avoiding serious degradation in performance, which is essential for lightening the storage and computation burden of deep networks. However, the resolution of the input for a network, another factor that influences the computation burden and robustness of a deep network, is scarcely discussed in the existing distillation methods. When the resolution of the input decreases, the computation cost of a neural network drops significantly. Nevertheless, the performance of the deep network will be unsatisfactory when using LR inputs. To tackle with the above problem, we propose a resolution-aware knowledge distillation approach to transfer knowledge from HR domain to LR domain. The resolution-aware distillation technique explores a reliable solution to accelerate deep networks using LR inputs while maintaining high performance, which is complementary for the current literature. Moreover, our approach is flexible to trade off the efficiency and accuracy of a deep network by simply changing the resolution of the input images, making it adaptive to practical applications.

III. Proposed Methodology

A. Notation

In this paper, we denote the training set and the corresponding ground-truth labels as $\chi = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ and $Y = \{y_1, y_2, \ldots, y_N\}$, where N denotes the number of training samples. Through operation such as resizing, we can obtain the target LR and HR training samples, which can be denoted as $\chi^L = \{\boldsymbol{x}_1^l, \boldsymbol{x}_2^l, \ldots, \boldsymbol{x}_N^l\}$ and $\chi^H = \{\boldsymbol{x}_1^h, \boldsymbol{x}_2^h, \ldots, \boldsymbol{x}_N^h\}$. During training, we refer the parameters of the teacher network and the student network as $\boldsymbol{W}_t$ and $\boldsymbol{W}_s$, and the feature extraction process of the teacher network and the student network as $f(\boldsymbol{x}^h; \boldsymbol{W}_t)$ and $f(\boldsymbol{x}^l; \boldsymbol{W}_s)$ correspondingly.

B. Revisit Knowledge Distillation

We will first revisit the traditional knowledge distillation technique in this section. Figure 1 shows the framework of traditional knowledge distillation method. The knowledge distillation framework is composed of a teacher network and a student network, and the teacher network contains more parameters than the student network. Generally, the knowledge distillation method trains the teacher network in advance. During the distillation process, a distillation loss is employed to improve the performance of the student model by transferring knowledge from the pre-trained cumbersome teacher model to the compact student model. In most cases, the distillation loss is implemented between the teacher outputs and the student outputs, which are extracted from the same training

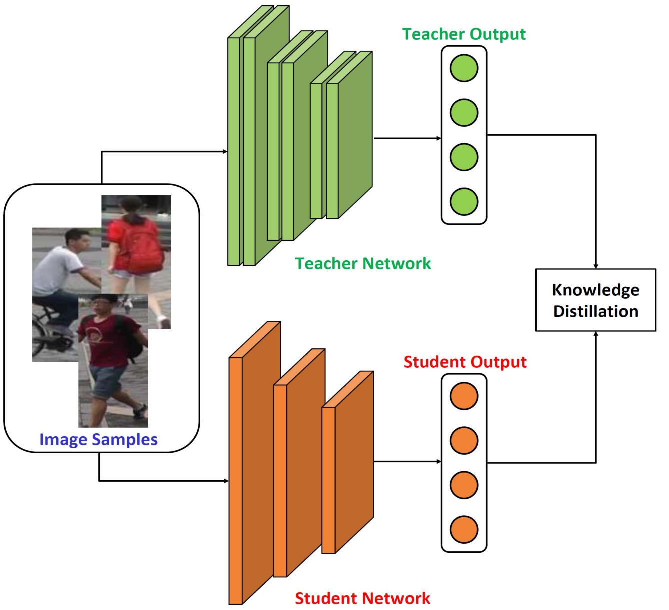


Fig. 1. Traditional knowledge distillation framework. The student network is cheaper than the teacher network in terms of storage and computation consumptions. Knowledge is transferred from the teacher network to enhance the generalization ability of the student network.

inputs. Because the student network is compact, the executive consumption is reduced, and the student network remains competitive with the guidance of the teacher prior.

C. Resolution-Aware Knowledge Distillation

In this paper, we focus on accelerating deep networks by reducing the resolution of the input images. Obviously, LR inputs are lack of high-frequency information, which is harmful for extracting discriminative representations. To address this issue, we propose a novel resolution-aware knowledge distillation framework to overcome the cross-resolution appearance discrepancies between inputs with different resolutions and prevent serious performance degradation when extracting features from LR inputs. The proposed approach manages to distill valuable knowledge from HR domain to LR domain, which is under-study in current literatures.

Figure 2 demonstrates the resolution-aware knowledge distillation (RKD) framework. The proposed framework is composed of two networks, namely the HR teacher network and the LR student network. In this paper, we impose the same network structure for both the teacher network and the student network to pay more attention on distilling knowledge across changing resolutions. A discriminator is integrated between the teacher output and the student output, and the adversarial learning is introduced to minimize the gap between features from different resolutions. Furthermore, we propose to implement a cross-resolution knowledge distillation loss to force the student network with LR inputs to mimic the activations of the teacher network and generate robust representations regardless of the resolution variations.

RKD is different from traditional knowledge distillation (KD) in the following aspects. First, the RKD framework adopts different inputs for the teacher network and the student network, whereas the KD approach uses the same inputs for different networks. We use HR inputs for the teacher network to learn powerful features and adopt LR inputs for the student network to enhance the computation efficiency. Second, RKD proposes to reduce the computation complexities of deep networks by extracting features from LR inputs and enhance the generalization ability of the LR features by distilling knowledge from HR features, whereas KD focuses on exploiting the prior knowledge to compress the network parameters while preserving good performance. Finally, RKD is more flexible than KD for practical applications with changing computation resources. For RKD, we can change the resolution of the inputs to obtain different computation complexities, whereas for KD we have to design a new structure to change the computation efficiency.

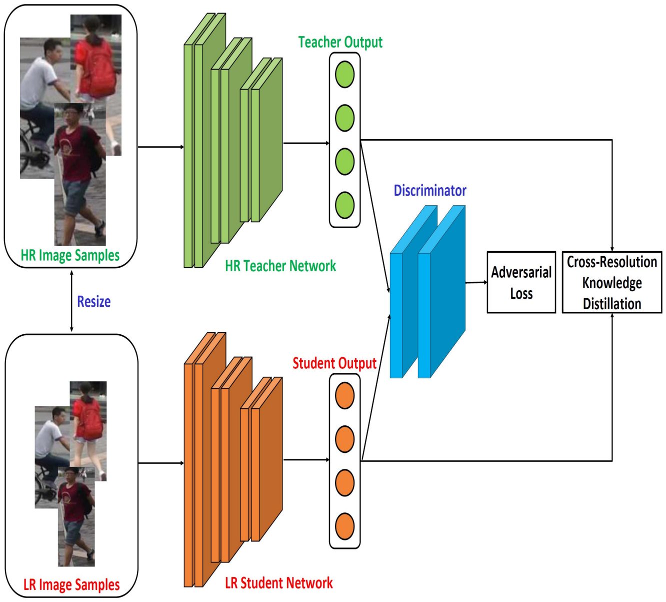


Fig. 2. The proposed resolution-aware knowledge distillation framework. The teacher network and student network are implemented using inputs with different resolutions. Knowledge are transferred from HR to LR domain to improve the performance of the student network. Furthermore, an adversarial loss is implemented to narrow the margin between different resolutions.

D. Training Teacher Network

The RKD approach is implemented in a two-step manner. The first step is to train a well-performed HR teacher network which provides reliable guidance for the LR student network. We first resize the training samples to obtain HR training set χ^H, and then employ the Softmax loss [40] to train a discriminative teacher model. The output z_t of the teacher network can be computed by:

$$z_t = f(\boldsymbol{x}^h; \boldsymbol{W}_t). \tag{1}$$

Finally, the loss function of the teacher network can be formulated as:

$$\mathcal{L}_t = \mathcal{L}_{CE}(y, z_t), \tag{2}$$

where $\mathcal{L}_{CE}$ refers to the cross-entropy loss [14].

E. Adversarial Learning

The appearance discrepancies will be very large when the resolution gaps between LR and HR images are too large.

TABLE I

DETAILS OF DISCRIMINATOR

Layer name	Input size	Parameters	Output size
Input	B	$B * 128$	128
Output	128	$128 * 2$	2

Compared with HR images, LR images are lack of discriminative details, which leads to the degradation of recognition performance of the student model. Therefore, the student network may achieve poor performance when adopting very low resolution images as inputs. To address the above issue, we propose to impose the adversarial learning technique to shrink the margin between the features of the teacher network and the student network. Note that some knowledge distillation methods [41], [42] also adopt the adversarial learning technique to accelerate the training process of knowledge distillation. Nevertheless, the purpose of the adversarial learning technique in RKD is to narrow the gaps between features from different resolution domains, which is different to the GAN-based KD methods. Particularly, we design a discriminator to decide whether the features are from HR images or LR images. The discriminator is composed by 2 fully-connected layers, with 0 indicating LR image and 1 indicating HR image. The structure of the discriminator is demonstrated in Figure 2, and the details are shown in Table I. An adversarial loss is proposed to enforce a min-max game between the student network and the discriminator, where B is the dimension of the output backbone feature. The student network tries to fool the discriminator by learning features similar with the teacher network while the discriminator tries to distinguish the features from HR and LR images. Denote the discriminator as D, features from the teacher network and the student network as $f(\boldsymbol{x}^h; \boldsymbol{W}_t)$ and $f(\boldsymbol{x}^l; W_s))$, the adversarial loss for the discriminator can be written as:

$$\mathcal{L}^d_{adv} = E_{\boldsymbol{x}^h}[log(D(f(\boldsymbol{x}^h; \boldsymbol{W}_t)))] + E_{\boldsymbol{x}^l}[log(1 - D(f(\boldsymbol{x}^l; \boldsymbol{W}_s)))], \quad (3)$$

For student network, the adversarial loss can be written as:

$$\mathcal{L}^s_{adv} = E_{\boldsymbol{x}^l}[log(D(f(\boldsymbol{x}^l; \boldsymbol{W}_s)))]. \quad (4)$$

The training process of adversarial learning is similar to traditional GAN-based methods. First, we extract features from the student network and the teacher network. Then both features are used to update the parameters of the discriminator using $\mathcal{L}^d_{adv}$. Finally, the adversarial loss is applied to the student network to generate features similar to the teacher network.

F. Cross-Resolution Knowledge Distillation

Despite the adversarial learning, we also try to transfer the knowledge from the pre-trained teacher model to the student model to narrow the cross-resolution disparities by implementing constraints between the teacher output and the student output. Notably, the teacher network is fixed after the first step. During training, we use image pairs $(\boldsymbol{x}^h, \boldsymbol{x}^l)$ as inputs and obtain the outputs $(f(\boldsymbol{x}^h; W_t), f(\boldsymbol{x}^l; W_s))$ from the teacher network and the student network correspondingly. Furthermore, we implement a cross-resolution knowledge distillation (CRKD) loss between the teacher output and the student output to enforce a strong constraint over features extracted across variant resolutions. The CRKD loss consists of three parts, of which one is the resolution-aware distillation (RD) loss, one is the pair-wise constraint (PC), and the remainder is the maximum mean discrepancy (MMD) loss.

Resolution-Aware Distillation Loss: The features of the teacher network and student network are resolution-aware, which is related to different resolution domains. The ambition of the RKD framework is to shrink the cross-resolution margin and improve the generalization ability of the student network so that features from LR inputs are congruent with those from HR inputs. Therefore, we propose a resolution-aware distillation loss (RD) to narrow the cross-resolution gaps by transferring resolution-aware knowledge from HR domain to LR domain. The RD loss ensures that the teacher network and the student network produce similar features with consistent distribution across changing resolutions. Particularly, the RD loss forces the student network to mimic the activations of the teacher network by using the KL divergence [43] loss, and the loss function $\mathcal{L}_{RD}$ can be written as:

$$\mathcal{L}_{RD} = \frac{1}{N}\sum_{i=1}^{N} T^2 \mathcal{L}_{KL}(\sigma(\frac{z_t}{T}), \sigma(\frac{z_s}{T})), \quad (5)$$

where z_t and z_s are corresponding to the outputs of the teacher network and the student network, $\sigma(.)$ refers to the softmax function, and T is the temperature hyperparameter to smooth the distillation loss. The KL divergence loss can be written as:

$$\mathcal{L}_{KL}(P(z_t), P(z_s)) = \sum P(z_t)log(\frac{P(z_t)}{P(z_s)}), \quad (6)$$

where $P(z_t) = \sigma(\frac{z_t}{T})$ and $P(z_s) = \sigma(\frac{z_s}{T})$.

Pair-Wise Constraint: The RD loss employs an indirect solution to make the student network predict features with similar distribution as the teacher network regardless of the resolution variations. Despite the RD loss, we also enforce a pair-wise constraint to transfer knowledge directly in the feature space so that the features remain the same for inputs from both LR and HR domains. Specially, we propose a pair-wise constraint (PC) to force the student model to imitate the teacher model directly in the feature space using L1 constraint. The formula of the pair-wise constraint is as follows:

$$\mathcal{L}_{PC} = \frac{1}{N}\sum_{i=1}^{N} |f(\boldsymbol{x}^h; \boldsymbol{W}_t) - f(\boldsymbol{x}^l; \boldsymbol{W}_s)|. \quad (7)$$

Maximum Mean Discrepancy Loss: Eventually, we employ the Maximum Mean Discrepancy (MMD) loss to tackle with the cross-resolution problem because MMD has been proven effective for shrinking the discrepancies in changing domains. Given the corresponding LR image features $\{z^l_i\}, i = 1, 2, \ldots, N$ and HR image features $\{z^h_j\}, j = 1, 2, \ldots, N$, the MMD loss can be formulated as:

$$\mathcal{L}_{MMD} = \| \frac{1}{N}\sum_{i=1}^{N} \phi(z^l_i) - \frac{1}{N}\sum_{j=1}^{N} \phi(z^h_j) \|^2_2, \quad (8)$$

TABLE II

COMPARISON RESULTS BETWEEN KD AND RKD FOR PERSON RE-IDENTIFICATION

Method	Resolution	Market-1501				DukeMTMC				RegDB				FLOPs	Time (s)
		r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP		
ResNet50$_t$	384 × 128	93.9	98.8	99.3	84.7	85.7	94.7	96.4	73.6	71.7	86.7	91.7	68.9	6.1×10^9	0.0037
ResNet18	384 × 128	91.9	98.2	99.0	79.9	83.3	93.8	95.2	69.0	57.5	80.1	88.3	55.8	3.0×10^9	0.0018
ResNet18$_s$+KD		93.3	98.4	99.0	83.3	84.7	94.7	96.1	72.9	66.4	82.5	90.7	66.1		
ResNet50	128 × 128	90.8	98.2	98.9	78.4	83.3	93.3	95.6	69.2	62.7	83.0	89.2	62.5	2.1×10^9	0.0017
ResNet50$_s$+RKD		**93.4**	**98.6**	**99.1**	**83.7**	**85.8**	**95.1**	**96.2**	**73.0**	**69.4**	**85.5**	**91.4**	**67.9**		

where $\phi(.)$ denotes the projection function. Derived from [44], the MMD loss can be written as:

$$\mathcal{L}_{MMD} = \| \frac{1}{N^2} \sum_{i,j=1}^{N} k(z_i^l, z_j^l) - \frac{2}{N^2} \sum_{i,j=1}^{N} k(z_i^l, z_j^h) + \frac{1}{N^2} \sum_{i,j=1}^{N} k(z_i^h, z_j^h) \|_2^2, \quad (9)$$

where $k(.)$ denotes the kernel function. Particularly, we adopt the Gaussian kernel to measure the distance between the LR features and HR features, which can be formulated as:

$$k(z_i^l, z_j^h) = e^{-\|(z_i^l - z_j^h)\|_2^2 / 2\sigma^2} \quad (10)$$

Eventually, the overall function of the cross-resolution knowledge distillation loss can be formulated as:

$$\mathcal{L}_{CRKD} = (1-\alpha)\mathcal{L}_{CE}(y, z_s) + \alpha\mathcal{L}_{RD} + \beta\mathcal{L}_{PC} + \gamma\mathcal{L}_{MMD}, \quad (11)$$

where α, β, and γ are the hyperparameters for balancing the classification loss, the resolution-aware distillation loss, the pair-wise Euclidean constraint, and the maximum mean discrepancy loss.

IV. EXPERIMENT

In this paper, we conduct extensive experiments on multiple popular computer vision tasks including person re-identification, image classification, face recognition, and defect segmentation to demonstrate the effectiveness of the RKD framework and the proposed CRKD loss. The evaluated statistics include the recognition performance, the computation complexities, and the execution time. Besides, we compare RKD with traditional KD to determine which framework is more effective in balancing the complexity and performance using the same distillation loss (CRKD). We also compare CRKD with the stat-of-the-art distillation functions under the RKD framework to prove the superiority of the proposed approach in distilling knowledge across different resolution domains. Eventually, we conduct ablation experiments to demonstrate that every component of the RKD approach is valuable for improving the performance of deep networks using LR inputs.

A. Person Re-Identification

Datasets: We conduct experiments on two large-scale re-id benchmarks, namely Market-1501 [45] and DukeMTMC [46], and a visible-infrared re-id dataset named RegDB [47] to evaluate the effects of the RKD approach. Particularly, Market-1501 contains 32,668 annotated bounding boxes of 1,501 identities from six cameras in an open system, and DukeMTMC contains 36,411 images of 1,812 people captured from 8 camera views, among which 408 people are from one camera view. The RegDB contains 8,240 images of 412 identities from a dual-camera system. We evaluate the effects of the proposed approach using the standard testing protocols on the Market-1501 [48], DukeMTMC [49], and RegDB [50] datasets.

Implementation Details: For the RKD framework, we adopt ABD-Net [49] based on ResNet50 as the baseline network to extract features for both the teacher network and the student network. The resolution of the input images for the teacher network is 384 × 128 pixels and that of the student network varies from 32 × 32 pixels to 128 × 128 pixels. For the KD framework, we use the same structure as the baseline of the cumbersome teacher network and use ABD-Net based on ResNet18 as the baseline of the student network. Both the teacher network and the student network adopt images with 384 × 128 pixels as inputs. Note that ABD-Net employs tricks such as attention [51] and feature orthogonality, proving that RKD can be easily adapted to any trick. This study is implemented using the Pytorch framework. During the pre-training stage, the parameters of the teacher network are optimized using Adam optimizer with an initial learning rate of 3×10^{-4}, which decreases by half every 20 epochs. The pre-training of the teacher network ends when the epoch number reaches 80. After pre-training, the teacher network is fixed and the student network is optimized using an Adam optimizer with the initial learning rate of 3×10^{-4} and decreases by half every 20 epochs. The training of the student network ends after 80 epoches. The hyperparameters are set to be $\alpha = 0.5$, $\beta = 0.001$, $T = 8$, and $\gamma = 1$.

Table II shows the comparison results between the RKD and KD frameworks on Market-1501, DukeMTMC, and RegDB datasets, where subscript t refers to the teacher network, subscript s refers to the student network, and FLOPs denotes the computation complexities. Apparently, both compressing the network parameters and reducing the resolution of inputs result in performance degradation. Compared with the teacher model, training a shallow network (ResNet18) encounters a degradation of 2%/2.4%/14.2% in rank-1 accuracy on Market-1501/DukeMTMC/RegDB dataset. Similarly, training ResNet50 with LR inputs (128 × 128) leads to a decrease of 3.1%/2.4%/9.0% in rank-1 accuracy on Market-1501/DukeMTMC/RegDB dataset. Notably, the performance degradation is more serious on RegDB because of the cross-modality visual variations. The experimental results show that the RKD method is effective for improving the performance of the LR student network. RKD improves the rank-1/mAP accuracies of the student network with a margin of 3.6%/5.3% on Market-1501, 2.5%/3.8% on DukeMTMC, and 9.3%/7.9%

TABLE III

EXPERIMENTAL RESULTS FOR DIFFERENT DISTILLATION LOSSES UNDER VARYING RESOLUTIONS

Resolution	Method	Market-1501				DukeMTMC				RegDB			
		r1	r10	r20	mAP	r1	r10	r20	mAP	r1	r10	r20	mAP
384×128	ResNet50$_t$	93.9	98.8	99.3	84.7	85.7	94.7	96.4	73.6	71.7	86.7	91.7	68.9
128×128	ResNet50	90.8	98.2	98.9	78.4	83.3	93.3	95.6	69.2	60.1	79.8	87.5	59.6
	ResNet50$_s$+Hinton	92.3	98.2	98.9	80.3	84.0	94.0	95.8	69.2	64.7	83.6	90.3	64.3
	ResNet50$_s$+Fitnet	93.3	98.6	99.2	82.5	85.2	94.9	96.2	70.4	65.7	83.4	89.2	64.3
	ResNet50$_s$+PKT	92.6	98.3	98.9	81.4	83.7	93.9	95.8	67.4	67.7	85.0	91.2	66.5
	ResNet50$_s$+AT	93.1	98.2	98.8	82.0	84.5	94.3	95.8	70.5	62.9	80.2	89.8	63.9
	ResNet50$_s$+SP	93.1	98.3	98.8	82.3	84.9	94.3	95.7	70.2	68.4	85.3	91.4	67.8
	ResNet50$_s$+AB	93.3	98.3	98.8	82.2	84.6	94.5	95.8	70.6	66.7	80.5	89.5	66.1
	ResNet50$_s$+Relation KD	93.3	98.3	99.1	81.8	83.9	94.0	95.7	69.6	64.3	81.7	89.1	64.4
	ResNet50$_s$+CC	93.1	98.5	98.9	82.2	84.7	94.4	96.0	71.6	63.8	80.3	87.8	63.8
	ResNet50$_s$+CRKD	**93.4**	**98.6**	**99.2**	**83.7**	**85.8**	**95.1**	**96.2**	**73.0**	**69.4**	**85.5**	**91.4**	**67.9**
64×64	ResNet50	82.4	95.5	96.9	61.3	70.6	86.4	89.6	49.4	44.0	65.3	76.3	44.1
	ResNet50$_s$+Hinton	84.7	96.1	97.5	64.2	73.4	88.0	90.8	53.0	49.6	67.8	76.4	50.5
	ResNet50$_s$+Fitnet	86.2	96.7	97.9	67.8	75.6	89.3	92.2	57.7	53.4	74.6	83.8	55.2
	ResNet50$_s$+PKT	85.7	96.8	98.0	66.4	74.0	88.2	90.8	53.4	53.5	73.9	81.8	52.7
	ResNet50$_s$+AT	86.4	96.6	97.8	67.6	75.6	88.9	91.4	56.3	51.4	69.9	80.6	51.6
	ResNet50$_s$+SP	86.2	96.6	97.9	68.2	76.3	89.3	91.7	51.6	55.6	74.4	83.0	55.1
	ResNet50$_s$+AB	86.7	96.4	97.7	68.4	77.5	90.0	92.0	59.0	58.1	75.4	83.2	57.9
	ResNet50$_s$+Relation KD	86.6	96.8	98.0	68.4	74.6	89.0	91.6	54.7	53.7	73.8	83.4	54.0
	ResNet50$_s$+CC	86.4	96.8	97.9	67.9	76.6	90.2	92.9	58.9	51.6	72.3	82.4	52.3
	ResNet50$_s$+CRKD	**87.7**	**96.8**	**98.1**	**68.6**	**78.6**	**90.8**	**92.7**	**60.7**	**58.6**	**77.1**	**86.2**	**58.1**
32×32	ResNet50	43.5	76.0	83.6	25.3	38.4	64.4	71.7	22.1	25.6	43.3	54.5	28.3
	ResNet50$_s$+Hinton	51.3	80.1	85.9	28.6	39.9	65.7	73.0	23.2	28.6	44.8	56.0	30.8
	ResNet50$_s$+Fitnet	60.2	87.5	92.0	37.8	48.5	71.9	77.5	29.3	29.9	47.7	58.8	30.9
	ResNet50$_s$+PKT	57.5	86.1	90.8	35.9	43.5	67.6	75.0	24.9	28.2	47.3	57.3	29.2
	ResNet50$_s$+AT	55.3	83.1	88.2	33.5	49.6	74.5	80.9	31.4	27.6	44.4	55.5	29.6
	ResNet50$_s$+SP	63.3	89.0	93.7	41.0	46.4	70.7	76.4	28.0	30.7	49.4	61.9	33.7
	ResNet50$_s$+AB	58.8	85.6	90.7	36.4	48.7	74.7	79.1	29.1	32.6	53.1	63.5	35.1
	ResNet50$_s$+Relation KD	65.6	89.0	92.4	40.3	53.7	75.4	81.5	33.5	32.4	50.8	61.7	34.2
	ResNet50$_s$+CC	66.4	89.8	93.4	43.2	54.0	77.5	82.5	33.7	29.9	47.6	59.3	31.5
	ResNet50$_s$+CRKD	**69.7**	**91.5**	**94.6**	**46.9**	**55.7**	**77.8**	**83.4**	**35.9**	**35.2**	**54.6**	**65.6**	**36.6**

on RegDB. Furthermore, RKD outperforms KD in balancing the computation complexities and recognition accuracies. RKD surpasses KD by 1.1%/3% in rank-1 accuracy on DukeMTMC/RegDB dataset with a lower computation cost (2.1×10^9 FLOPs V.S. 3.0×10^9 FLOPs). Meanwhile, the running speed of the student model is very closed for the RKD and KD approaches. Both RKD and KD approaches accelerate the speed of feature extraction process by nearly 2 times. Finally, RKD is competitive against the teacher network with a degradation of only 0.5%/2.3% in rank-1 accuracy on Market-1501/RegDB dataset while reporting a reduction of more than 65% computation complexities and 50% execution time. Notably, RKD outperforms the teacher network in terms of rank-1 accuracy on DukeMTMC dataset. Consequently, RKD significantly improves the efficiency of deep networks while preserving the discriminative power, which is important for applications with limited computation resources.

Despite comparing the RKD framework with the KD framework, we also conduct experiments to investigate the effects of RKD framework when dealing with resolutions varying from 32×32 to 128×128. Besides, we compare the CRKD with other knowledge distillation functions under resolution distillation framework. Table III shows the experimental results. Obviously, the performance of deep model degrades more dramatically when using lower resolution images because of lacking sharp details. Take DukeMTMC dataset into consideration, the performance of the baseline student model trained by images of 128×128 pixels is lower than the teacher model by a margin of 2.4%/4.4% in terms of rank-1/mAP. When using inputs of 64×64 resolution, the performance gap is expanded to 15.1%/24.2 in rank-1/mAP. When the resolution gap becomes very large, that is, when using inputs of 32×32 pixels, the performance degradation is 47.3%/51.5%, which is unacceptable for applications. The above results show that although employing LR inputs is advantage for reducing the computation complexities, the performance may degrade seriously, which prevents the applications of deep networks adopting LR inputs. Experiments demonstrate that the proposed approach is effective in dealing with the above problem. Table III show that RKD significantly improves the performance of deep networks when extracting features from LR images. Notably, the proposed approach achieves higher promotions when using lower resolution inputs. Take Market-1501 benchmark into consideration, RKD achieve an improvements of 2.6%/5.3% in rank-1/mAP when using inputs of 128×128, an improvements of 5.3%/7.3% in rank-1/mAP when using inputs of 64×64, and an inspiring improvement of 26.2%/21.6% in rank-1/mAP when using inputs of 32×32. Finally, we compare CRKD loss with other state-of-the-art distillation losses including Hinton [14], Fitnet [32], PKT [52], AT [33], SP [53], AB [34], Relation KD [37], and CC [27] under the cross-resolution distillation framework to prove the superiority of our method for transferring knowledge from HR domain to LR domain. CRKD achieves the best recognition performance for all involved datasets and resolutions. Particularly, CRKD is more superior when transferring knowledge to the lower resolution networks. Take Market-1501 into consideration, CRKD outperforms the best compared models (AB loss) by 0.1% in rank-1 accuracy when dealing with images of 128 × 128 pixels. The performance disparity becomes 1% for images of 64 × 64 pixels and 3.3% for images of 32 × 32 pixels.

B. Image Classification

Dataset: We conduct experiments on CIFAR-100 [54] to show the effectiveness of RKD in improving the performance of features extracted from LR images and the superiority of CRKD against the state-of-the-art methods on image classification. The CIFAR-100 dataset consists of 60,000 images from 100 classes. Each class contains 600 images, among which 500 images are used for training and the rest images are used for testing. The images in CIFAR-100 dataset are natural colored images with 32 × 32 pixels, which can be considered LR inputs for deep networks. For the teacher network, the training images are resized to 64 × 64 pixels to achieve better classification accuracy. For the student network, we adopt images of 32 × 32 pixels and 16 × 16 pixels as inputs to get faster inference speed. Although the interpolation process does not introduce new information, the receptive fields of HR images will be larger than LR images, which is crucial for extracting discriminant features. Our method is also useful for improving the performance of deep networks by distilling knowledge from networks trained by interpolated images. We compare the proposed approach with other distillation methods including Hinton [14], Fitnet [32], PKT [52], AT [33], SP [53], AB [34], Relation KD [37], and CC [27].

Implementation Details: We adopt ResNet50 [55] as the baseline network for both the teacher network and the student network of the RKD framework to extract backbone features. The distillation process is optimized in a SGD optimizer with an initial learning rate of 1×10^{-1}, which decreases by 80% at epoch 30, 60 and 80, and the batch size is set to be 128. The training process ends when the epoch number reaches 100. We set the hyperparameters as $\alpha = 0.5$, $\beta = 1$, $T = 8$, and $\gamma = 1$.

Table IV shows the experimental results on CIFAR-100 benchmark. We can see that the baseline network encounters a degradation of 3.98%/2.17% in rank-1/mAP when the resolution of inputs changes from 64 × 64 to 32 × 32. Furthermore, when using inputs with 16 × 16 pixels, the performance of the baseline network is poor and the degradation reaches 15.3%/8.25% in rank-1/mAP. The above observation illustrates that the performance of deep networks drops dramatically when using very low-resolution inputs, which is an important reason why HR inputs are required for deep networks. The experimental results also prove that RKD is valuable for improving the performance of LR student networks by inferring knowledge from HR domain to LR domain. The proposed approach improves the performance of the baseline network with a margin of 4.74%/2.79% in terms of rank-1/mAP for inputs of 32 × 32 pixels. Notably, the LR student network guided by RKD achieves higher recognition accuracies than the HR teacher network. When using inputs of 16 × 16 pixels, RKD gains an improvement of 6.43%/4.07% in rank-1/mAP for the student network. Therefore, RKD is more effective for distilling knowledge to the student networks when coping with lower resolution inputs. Finally, we compare our approach with the state-of-the-art methods under cross-resolution knowledge distillation framework. The experiments demonstrate that CRKD achieves the best recognition performance for resolutions of both 32 × 32 pixels and 16 × 16 pixels.

TABLE IV
EXPERIMENTAL RESULTS ON CIFAR-100

Method	Resolution	Rank-1 (%)	Rank-5 (%)
ResNet50$_t$	64 × 64	79.19	94.80
ResNet50	32 × 32	75.21	92.63
ResNet50$_s$+Hinton	32 × 32	77.74	94.83
ResNet50$_s$+Fitnet	32 × 32	78.38	94.91
ResNet50$_s$+PKT	32 × 32	79.66	95.13
ResNet50$_s$+AT	32 × 32	79.59	94.93
ResNet50$_s$+SP	32 × 32	79.68	95.14
ResNet50$_s$+AB	32 × 32	79.55	95.14
ResNet50$_s$+Relation KD	32 × 32	79.38	95.29
ResNet50$_s$+CC	32 × 32	79.39	95.08
ResNet50$_s$+CRKD	32 × 32	**79.95**	**95.42**
ResNet50	16 × 16	63.89	86.55
ResNet50$_s$+Hinton	16 × 16	68.30	89.49
ResNet50$_s$+Fienet	16 × 16	68.60	89.85
ResNet50$_s$+PKT	16 × 16	69.35	90.22
ResNet50$_s$+AT	16 × 16	69.93	90.28
ResNet50$_s$+SP	16 × 16	70.12	90.18
ResNet50$_s$+AB	16 × 16	70.06	90.20
ResNet50$_s$+Relation KD	16 × 16	70.31	90.18
ResNet50$_s$+CC	16 × 16	69.99	90.45
ResNet50$_s$+CRKD	16 × 16	**70.32**	**90.62**

C. Face Recognition

Datasets: We conduct experiments on face recognition task to verify that RKD also works for large-scale benchmarks. Specially, we use the CASIA-WebFace [56] and part of VGG-Face [57] to train the face recognition models. The training set contains 680,933 image samples from 11,303 identities. The involved images are semi-automatically detected from the Internet. The effects of the RKD approach is evaluated on the LFW dataset [58]. The LFW dataset is one of the most popular face recognition benchmark consisting of 13,233 images from 5,749 people captured in the unconstrained environments. We will adopt the standard verification protocol [59] to ensure fair comparisons between RKD and the other models.

Implementation Details: We adopt ResNet50 as the baseline network for both the teacher network and the student network for RKD framework. The teacher network extracts features from inputs with 256 × 256 pixels and the student network is implemented using inputs with a wide range of resolutions including 128 × 128, 64 × 64, and 32 × 32. Besides, we adopt ResNet50 as the teacher network and use ResNet18 as the student network for evaluating the KD framework. The distillation process is optimized in a SGD optimizer with an initial

TABLE V
EXPERIMENTAL RESULTS ON LFW

Method	Resolution	ACC (%)	FLOPS
ResNet50$_t$	256 × 256	98.21	5.4×10^9
ResNet18	256 × 256	97.38	2.3×10^9
ResNet18$_s$+KD		98.25	
ResNet50	128 × 128	97.60	1.4×10^9
ResNet50$_s$+RKD		**98.33**	
ResNet50	64 × 64	95.60	3.5×10^8
ResNet50$_s$+RKD		**98.03**	
ResNet50	32 × 32	93.42	9.7×10^7
ResNet50$_s$+RKD		**96.87**	

learning rate of 1×10^{-1}, which decreases by 90% at epoches 10, 15, and 20, and the batch size is set to be 128. The training of the student network ends when the epoch number reaches 40. The hyperparameters are set to be $\alpha = 0.1$, $\beta = 1$, $T = 8$, and $\gamma = 1$.

Table V demonstrates the experimental results on LFW. The teacher network achieves higher verification accuracy (98.2%) than both the network using ResNet18 structure (97.4%) and the network using inputs of 128 × 128 pixels (97.6%). Obviously, both the RKD and the KD techniques improve the performance of the student network in Table V. Note that RKD achieves higher verification accuracy than the KD approach (98.33% V.S. 98.25%) with a reduction of almost 40% computation complexities (1.4×10^9 FLOPs V.S. 2.3×10^9 FLOPs), proving that RKD is more efficient in advancing the overall performance of a deep network considering the recognition accuracy and computation cost. Interestingly, the student models guided by both KD and RKD surpass the teacher model by a little margin. Furthermore, RKD achieves impressive results when dealing with huge resolution disparities between the teacher network and the student network. RKD improves the recognition accuracies of the student network based on inputs with 64 × 64 / 32 × 32 pixels by a margin of 2.43%/3.45%. Compared with the teacher network, RKD achieves similar performance (98.03% V.S. 98.21%) while reducing more than 90% computation cost (3.5×10^8 FLOPs V.S. 5.4×10^9 FLOPs). Note that RKD can control the computation complexities by simply changing the resolution of inputs for the student network, which is flexible in realistic environments.

Figure 3 shows the ROC curves of the compared models. RKD dramatically improves the performance of the student network and achieves superior results compared with other methods. Figure 4 shows the relationship between the performance of ResNet50 and the resolution of input on LFW. The verification accuracy of ResNet50 reduces significantly with the decrease in the resolution of the input. RKD is effective for suppressing the performance degradation and making LR student model competitive against the HR teacher model.

D. Industrial Surface Defect Segmentation

Datasets: For industrial applications, the input images may be very large, which results in high computation consumptions for deep networks. We conduct experiments on industrial surface defect segmentation task to show that RKD also works when coping with high-resolution images. The experiments are implemented on Kolektor[1] dataset [60]. The Kolektor dataset contains defective electronic commutator images including 50 physical objects and deformed electronic steering devices. Each item is composed of 8 surfaces and 399 sample images, among which 52 images are with visible defects and the other 347 images are without defects. The resolution of the defect images is resized to 1408 × 512.

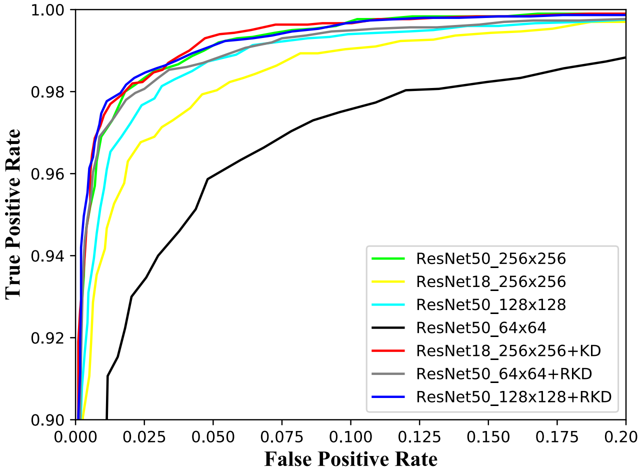


Fig. 3. The ROC curves on LFW. RKD remarkable improves the performance of student models and achieves similar results to the teacher method using low-resolution inputs. Best viewed in color.

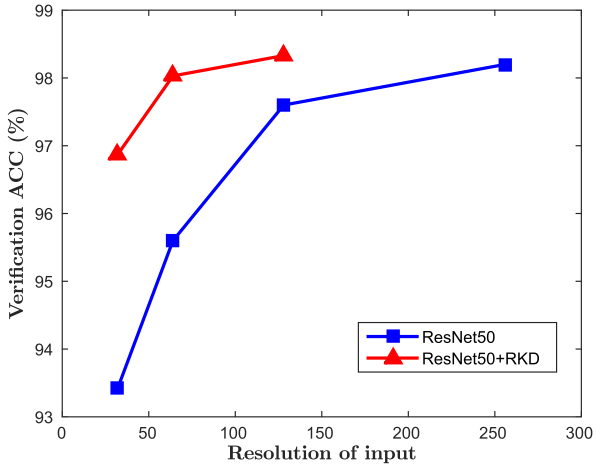


Fig. 4. Relationship between the performance of ResNet50 and the resolution of input on LFW.

Implementation Details: We adopt UNet [5] as the baseline network to extract deep features. The magnification factor is 4. Therefore, the resolution of the student inputs is downsampled to 352 × 128. The distillation process is optimized in an Adam optimizer with an initial learning rate of 5×10^{-4}, $\beta 1 = 0.9$ and $\beta 2 = 0.999$, and the batch size is set to be 30. The training of the student network ends when the epoch number reaches 200. Notably, we adopt mIOU (mean IOU) and DICE to evaluate the performance of involved methods.

[1] The Kolektor surface-defect dataset is publicly available at http://www.vicos.si/Downloads/KolektorSDD

TABLE VI

EXPERIMENTAL RESULTS ON KOLEKTORSDD

Method	Resolution	mIOU	DICE
UNet$_t$	1408 × 512	77.37	85.38
UNet	352 × 128	67.48	75.93
UNet$_s$+Hinton	352 × 128	70.98	79.58
UNet$_s$+Fitnet	352 × 128	71.58	80.17
UNet$_s$+PKT	352 × 128	71.83	80.41
UNet$_s$+AT	352 × 128	71.37	79.97
UNet$_s$+SP	352 × 128	72.51	81.05
UNet$_s$+AB	352 × 128	72.72	81.26
UNet$_s$+Relation KD	352 × 128	72.6	81.15
UNet$_s$+CC	352 × 128	72.65	81.19
UNet$_s$+CRKD	352 × 128	**73.48**	**81.97**

The IOU and DICE can be calculated as follows.

$$IOU = \frac{TP}{TP + FN + FP}, \tag{12}$$

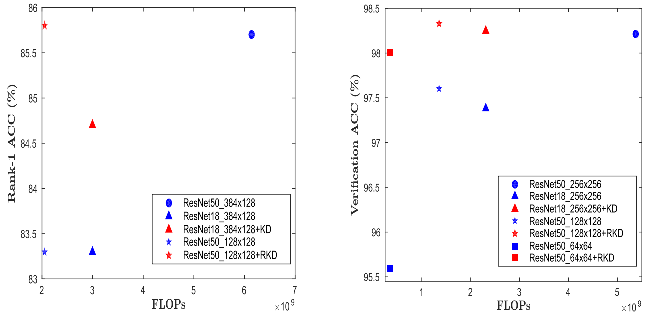
$$DICE = \frac{2TP}{(TP + FN) + (TP + FP)}, \tag{13}$$

where TP denotes true positive, FN denotes false negative, and FP denotes false positive.

Table VI illustrates the experimental results. Although the resolution of the LR images (352 × 128) is enough for some visual recognition tasks such as face recognition and person re-identification, experimental results show that the performance of the HR network outperforms the LR network by a margin of 9.89% and 9.45% in terms of mIOU and DICE. Obviously, the RKD significantly improves the performance of the student network. RKD results in a promotion of 6%/6.04% in mIOU/DICE for the student network, indicating that RKD is also beneficial for high-resolution applications. Because very high-resolution applications have proliferated over the last few years, e.g. remote sensing image classification and 4K games execution, the RKD technique may become crucial for future applications. Furthermore, the proposed objective function achieves the best segmentation results compared with the state-of-the-art methods. RKD beats the competitors by 0.76%/0.71% in mIOU/DICE. Note that the teacher network takes 1.48 seconds to extract features from the high-resolution image (1408 × 512) using a GTX 1080Ti GPU. Meanwhile, the execution time for the student network with the low-resolution image (352×128) is 0.13 seconds, which is 10 times faster than the teacher network without losing too much segmentation accuracy. Therefore, RKD is proven effective for enhancing the overall performance of deep networks against high-resolution task.

E. Overall Comparison

In this section, we compare the overall performance of different models to verify the efficiency of the proposed approach. Figure 5 shows the comparison results of different models by evaluating the accuracy and computation complexities (FLOPs) on DukeMTMC and LFW benchmarks. Figure 5(a) shows the experimental results on DukeMTMC and Figure 5(b) shows the evaluation results on LFW. The teacher network achieves high recognition accuracy with heavy computation burdens caused by cumbersome parameters and HR inputs. On the other hand, when using lightweight network or LR inputs, although the computation cost is reduced, the performance of the model is seriously reduced, which is unfavorable for applications. Obviously, we can see that both KD and RKD are efficient in improving the performance of deep networks without adding any computation burden, indicating that distillation techniques promote the overall performance of the student network. Compared with the KD approach, the RKD approach achieves better recognition accuracy with lower computation complexities. Because RKD focuses on different perspective to exploit the knowledge of the teacher network, the proposed approach should be complementary for the existing researches. Finally, RKD achieves similar performance to the teacher network with much lower computation complexities. The student network outperforms the teacher network using only one quarter computation cost.



(a) DukeMTMC (b) LFW

Fig. 5. Overall evaluations on DukeMTMC and LFW. Compared with the teacher model, RKD achieves competitive accuracy with a much lower computation cost. Best viewed in color.

TABLE VII

RESULTS OF THE ABLATION EXPERIMENTS

Methods	Resolution	DukeMTMC		Market-1501	
		Rank-1	mAP	Rank-1	mAP
ResNet50	128 × 128	83.3	69.2	90.8	78.4
ResNet50$_s$+PC	128 × 128	83.8	70.7	92.8	81.3
ResNet50$_s$+PC+RD	128 × 128	85.1	72.5	93.0	83.2
ResNet50$_s$+PC+RD+MMD	128 × 128	85.3	72.6	93.2	83.4
Proposal	128 × 128	85.8	73.0	93.4	83.7
ResNet50$_t$	384 × 128	85.7	73.6	93.9	84.7

F. Ablation Experiments

Eventually, we conduct ablation experiments on DukeMTMC and Market-1501 datasets to illustrate the effects of different components in the RKD framework. Table VI illustrates the results of the ablation experiments. The proposed pair-wise constraint (PC) improves the student network from 83.3%/69.2% to 83.8%/70.7% on DukeMTMC and from 90.8%/78.4% to 92.8%/81.3% on Market-1501 in terms of rank-1/mAP accuracies. Moreover, the resolution-aware distillation loss (RD) manages to further promote the student network and achieves a gain of 1.3%/1.8% on

DukeMTMC and 0.2%/1.9% on Market-1501 considering rank-1/mAP statistics. Finally, the MMD loss and the adversarial loss are also proven effective in improving deep networks using LR images. The MMD loss achieves an improvement of 0.2%/0.1% on DukeMTMC and 0.2%/0.2% on Market-1501. Meanwhile, the adversarial loss improves the proposal by a margin of 0.5%/0.4% on DukeMTMC and 0.2%/0.3% on Market-1501. Consequently, each component in the RKD framework is valuable for transferring knowledge from the teacher network, and the proposed framework is effective for improving the performance of a student model.

V. Conclusion

In this paper, we focus on addressing the contradiction that using HR inputs leads to powerful features but high computation complexities, whereas employing LR inputs results in fast execution process but unreliable features. We propose a novel resolution-aware knowledge distillation (RKD) framework to conduct knowledge distillation across different resolutions. The proposed framework consists of a HR teacher network and a LR student network, and the distillation process is implemented in a two-step manner. First, we train the teacher network to learn the structural knowledge from the HR inputs. Then we employ the cross-resolution knowledge distillation (CRKD) loss between the teacher output and the student output to transfer knowledge from HR domain to LR domain by forcing the student network to mimic the behaviors of the teacher network. The CRKD loss is combined by a resolution-aware distillation loss, a pair-wise constraint, and a maximum mean discrepancy loss. Finally, we introduce the adversarial learning to shrink the cross-resolution gap between HR and LR features. Because RKD reduces the computation cost of a deep network from a different aspect to traditional KD approach, the proposed approach should be complementary to current literatures. Furthermore, RKD is flexible in controlling the computation efficiency of the student network by changing the resolution of inputs, making it adaptive to realistic environments with different computation requirements. We conduct extensive experiments on person re-identification, image classification, face recognition, and defect segmentation to verify the effectiveness of the RKD framework. Experimental results demonstrate that RKD is beneficial to accelerating deep networks while maintaining the discriminative power, which is essential for practical applications that are lack of computation resources. Furthermore, RKD outperforms the other knowledge distillation methods for transferring knowledge across different resolutions.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 2012, pp. 1097–1105.

[2] F. Wang *et al.*, "Residual attention network for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3156–3164.

[3] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.

[4] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 447–456.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, 2015, pp. 234–241.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.

[7] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.

[8] Z. Feng, J. Lai, and X. Xie, "Learning view-specific deep networks for person re-identification," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3472–3483, Jul. 2018.

[9] Y. Lin *et al.*, "Improving person re-identification by attribute and identity learning," *Pattern Recognit.*, vol. 95, pp. 151–161, Nov. 2019.

[10] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8933–8940.

[11] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 286–301.

[12] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2377–2385.

[13] Y. Sun, D. Liang, X. Wang, and X. Tang, "DeepID3: Face recognition with very deep neural networks," 2015, *arXiv:1502.00873*. [Online]. Available: http://arxiv.org/abs/1502.00873

[14] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*. [Online]. Available: http://arxiv.org/abs/1503.02531

[15] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: http://arxiv.org/abs/1704.04861

[16] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[17] J. Wu, C. Leng, Y. Wang, Q. Hu, and J. Cheng, "Quantized convolutional neural networks for mobile devices," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4820–4828.

[18] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1135–1143.

[19] J. Lin, Y. Rao, J. Lu, and J. Zhou, "Runtime neural pruning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2181–2191.

[20] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 391–407.

[21] Z. Feng, J. Lai, X. Xie, and J. Zhu, "Image super-resolution via a densely connected recursive network," *Neurocomputing*, vol. 316, pp. 270–276, Nov. 2018.

[22] T. Chen, L. Lin, W. Zuo, X. Luo, and L. Zhang, "Learning a wavelet-like auto-encoder to accelerate deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1–9.

[23] M. Gao, R. Yu, A. Li, V. I. Morariu, and L. S. Davis, "Dynamic zoom-in network for fast object detection in large images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6926–6935.

[24] S. Gu, Y. Li, L. Van Gool, and R. Timofte, "Self-guided network for fast image denoising," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2511–2520.

[25] K. Zhang, W. Zuo, and L. Zhang, "FFDNet: Toward a fast and flexible solution for CNN-based image denoising," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4608–4622, Sep. 2018.

[26] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4320–4328.

[27] B. Peng *et al.*, "Correlation congruence for knowledge distillation," 2019, *arXiv:1904.01802*. [Online]. Available: http://arxiv.org/abs/1904.01802

[28] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4133–4141.

[29] A. Wu, W.-S. Zheng, X. Guo, and J.-H. Lai, "Distilled person re-identification: Towards a more scalable system," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1187–1196.

[30] B. Heo, M. Lee, S. Yun, and J. Young Choi, "Knowledge distillation with adversarial samples supporting decision boundary," 2018, *arXiv:1805.05532*. [Online]. Available: http://arxiv.org/abs/1805.05532

[31] N. Passalis and A. Tefas, "Learning deep representations with probabilistic knowledge transfer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 268–284.

[32] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "FitNets: Hints for thin deep nets," 2014, *arXiv:1412.6550*. [Online]. Available: http://arxiv.org/abs/1412.6550

[33] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," 2016, *arXiv:1612.03928*. [Online]. Available: http://arxiv.org/abs/1612.03928

[34] B. Heo, M. Lee, S. Yun, and J. Y. Choi, "Knowledge distillation with adversarial samples supporting decision boundary," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 3771–3778.

[35] Y. Liu, K. Chen, C. Liu, Z. Qin, Z. Luo, and J. Wang, "Structured knowledge distillation for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2604–2613.

[36] Y. Tian, D. Krishnan, and P. Isola, "Contrastive representation distillation," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–19.

[37] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3967–3976.

[38] D. Wang, Y. Li, L. Wang, and B. Gong, "Neural networks are more productive teachers than human raters: Active mixup for data-efficient knowledge distillation from a blackbox model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1498–1507.

[39] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 3903–3911.

[40] Y. Wen, K. Zhang, and Z. Li, "A discriminative feature learning approach for deep face recognition," in *Proc. Comput. Vis. (ECCV)*, Oct. 2016, pp. 499–515.

[41] X. Wang, R. Zhang, Y. Sun, and J. Qi, "Kdgan: Knowledge distillation with generative adversarial networks," in *Proc. NeurIPS*, 2018, pp. 783–794.

[42] Z. Xu, Y.-C. Hsu, and J. Huang, "Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks," 2017, *arXiv:1709.00513*. [Online]. Available: http://arxiv.org/abs/1709.00513

[43] C. Bucilu, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 535–541.

[44] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[45] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[46] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3754–3762.

[47] D. T. Nguyen, H. G. Hong, K. W. Kim, and K. R. Park, "Person recognition system based on a combination of body images from visible light and thermal cameras," *Sensors*, vol. 17, no. 3, p. 605, 2017.

[48] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2353–2367, May 2016.

[49] T. Chen *et al.*, "ABD-Net: Attentive but diverse person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8351–8361.

[50] Z. Feng, J. Lai, and X. Xie, "Learning modality-specific representations for visible-infrared person re-identification," *IEEE Trans. Image Process.*, vol. 29, pp. 579–590, 2020.

[51] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2119–2128.

[52] N. Passalis, M. Tzelepi, and A. Tefas, "Probabilistic knowledge transfer for lightweight deep representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2030–2039, May 2021.

[53] F. Tung and G. Mori, "Similarity-preserving knowledge distillation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1365–1374.

[54] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Tech. Rep. 7, 2009.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[56] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*. [Online]. Available: http://arxiv.org/abs/1411.7923

[57] O. M. Parkhi *et al.*, "Deep face recognition," in *Proc. BMVC*, vol. 1, no. 3, 2015, p. 6.

[58] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. ECCV Workshop Faces Real-Life Images*, 2008, vol. 1, no. 6.

[59] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4690–4699.

[60] D. Tabernik, S. Sela, J. Skvarc, and D. Skocaj, "Segmentation-based deep-learning approach for surface-defect detection," *J. Intell. Manuf.*, vol. 31, pp. 759–776, May 2010.

Zhanxiang Feng (Member, IEEE) received the Ph.D. degree in information and communication engineering from Sun Yat-sen University, China, in 2018. In 2019, he joined Sun Yat-sen University as a Postdoctoral Fellow. He has authored more than ten articles, including IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), *Neurocomputing*, and *ICPR*. His research interests include person re-identification, face recognition, face hallucination, image super-resolution, and visual surveillance.

Jianhuang Lai (Senior Member, IEEE) received the Ph.D. degree in mathematics from Sun Yat-sen University, China, in 1999. In 1989, he joined Sun Yat-sen University as an Assistant Professor, where he is currently a Professor of the School of Computer Science and Engineering. He has published over 100 scientific papers in the international journals and conferences on image processing and pattern recognition, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI), IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS (TNNLS), IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING (TKDE), IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART B (TSMC), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), *Pattern Recognition*, ICCV, CVPR, and ICDM. His current research interests are computer vision, digital image processing, pattern recognition, multimedia communication, and multiple target tracking. He is a fellow of the Image and Graphics Society of China. He serves as the Deputy Director of the Image and Graphics Association of China.

Xiaohua Xie received the B.S. degree in mathematics and applied mathematics from Shantou University in 2005, and the M.S. degree in information and computing science and the Ph.D. degree in applied mathematics from Sun Yat-sen University, China, in 2007 and 2010, respectively. He was an Associate Professor with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. He is currently an Associate Professor with Sun Yat-sen University. He has authored or coauthored over 50 papers in prestigious international journals and conferences. His current research fields cover image processing, computer vision, pattern recognition, and computer graphics.