IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 11, 2020, accepted March 29, 2020, date of publication April 1, 2020, date of current version April 15, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2984915

MSBA: Multiple Scales, Branches and Attention Network With Bag of Tricks for Person Re-Identification

HANLIN TAN, HUAXIN XIAO, XIAOYU ZHANG, BIN DAI, SHIMING LAI, YU LIU, AND MAOJUN ZHANG

College of Systems Engineering, National University of Defense Technology, Changsha 410073, China

Corresponding author: Shiming Lai (shiming413@nudt.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61703415 and Grant 61906206.

ABSTRACT Person re-identification (Re-ID) has become a hot topic in both research and industry. We joined in a person Re-ID challenge of the First National Artificial Intelligence Challenge (China, 2019) and found some model designs and training tricks work great or not on a super big private dataset. In this paper, we propose a model that combines the most effective designs, including multi-scale, multi-branch and attention mechanism, and report training tricks that are no less or even more important in improving person Re-ID performance. We analyze four commonly used public datasets: Market1501, DukeMTMC-ReID, CUHK03, and MSMT17, and achieve the state-of-the-art performance. Besides, we analyze and confirm the effectiveness of the designs by ablation studies. We also share strategies that play a key role in the challenge and experience of model designs that do not generalize well on large datasets.

INDEX TERMS Person re-identification, convolutional neural network.

I. INTRODUCTION

Person re-identification (Re-ID) is to search in a gallery (database) for images that contain the same person given a probe image (query) [1]. Person Re-ID has attracted attention in computer vision due to its applications in surveillance and security. Similar to other computer vision tasks, person Re-ID suffers from different viewpoints, illumination changes, occlusions, and various image quality. Apart from those, person Re-ID faces some unique challenges such as unconstrained poses and different scenes [2].

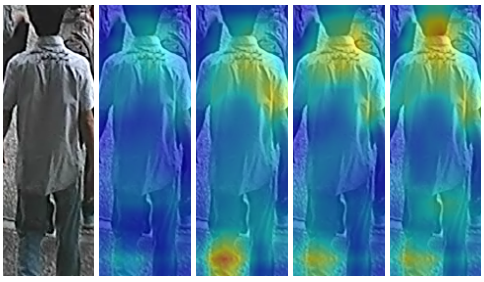
We have participated in the Person Identification Channel of the First National Artificial Intelligence Challenge (China, 2019) and ranked 11 out of 1900 teams in the first round and 30 out of 100 teams in the second round. In this challenge, we can evaluate Re-ID methods on a small private dataset with good quality and a large private test set of 160,000+ images from real scenes with a training set of 70,000+ images. During the challenge, we implement the state-of-the-art methods such as [3]–[5], MGN [6] and find it is easy to achieve good performance on small datasets. However, some model designs induce poor performance on large datasets. Besides, we also find many tricks that contribute to Re-ID performance. Some of them are even more important than model designs.

This work has three motivations. First, we survey recent Convolutional Neural Network (CNN) based works mainly in 2019, which has dominated person Re-ID [3] and summarizes effective model designs and tricks in training and testing. By combining those effective designs and tricks in training, we can provide impressive results on both private datasets of the challenge and public datasets for research.

Second, we find some of the existing works that reported good performance on small public datasets generalize poor on large datasets. There are several reasons. Small public datasets such as Market1501 [1] have relatively good image quality, on which it is easy for CNN models to get high metrics. Some methods use designs to over-fit those properties from the good image quality. Thus, those designs may become useless or even harmful on large datasets. Apart from that, we discover some works use inference tricks to improve performance in their codes without mentioning them in their papers. Therefore, those model designs are not as good as reported.

The associate editor coordinating the review of this manuscript and approving it for publication was Shaojun Wang.

(a) Person 1



(b) Person 2

FIGURE 1. Visualization of intermediate feature maps. Left to right: input image, no multi-scale, no attention, no local branch, proposed MSBA.

Third, we try to design industry-friendly Re-ID models. We keep in mind industry requirements that models are expected to be effective without introducing heavy computations, especially at inference time. It is known in deep learning that using more parameters is a possible direction to improve performance. However, some Re-ID algorithms tend to use a large number of model parameters to improve performance. Those methods such as [5], [6] introduce many branches for different purposes with many parameters. The result is that some model file size reaches 200MB to 300 MB despite that the backbone network weight file is less than 60 MB. Those model designs are industry-unfriendly.

In this work, we propose a highly effective CNN model for person Re-ID with little extra computation introduced at inference time. We introduce in detail the effective designs as briefly depicted in Figure 1 and introduce effective tricks in training a Re-ID network. We also analyze properties of four major public datasets, including Market1501 [1], DukeMTMC-ReID [7], CUHK03 [8] and MSMT17 [9]. By using the practical designs, training tricks and analyzed results, the proposed method achieves new state-of-the-art performance on all the four public datasets.

II. RELATED WORK

In our experience in participating in the challenge, effective designs consist of multi-scale, multi-branch and attention mechanism. We will introduce related work of them in this Section.

A. MULTI-SCALE MODEL

Multi-scale structures are widely used in Re-ID and proved to be effective. Cai *et al.* [10] proposed a three-scale feature of the whole body, upper body, and lower body of the person obtained using the attention mechanism as a method of multi-scale features. This method can reduce the negative impact of changes in the posture of the person and the background confusion. Liu *et al.* [11] proposed a multi-scale feature enhancement (MFE) person re-identification model. The multi-scale information of this model is derived from the shallow, middle and deep feature maps generated by the backbone network. The final feature combines the rich color or texture features of the shallow features, the structured features contained in the middle feature map and the semantic information contained in the deep feature map. Therefore, this model can effectively improve the performance of supervised person re-identification tasks.

Wang *et al.* [12] proposed to extract feature maps with different scales from different stages of the backbone network, and added these multi-scale features as a new feature for further processing, which obtained an advanced result. Wu *et al.* [13] proposed an attention deep architecture with multi-scale deep supervision for person re-identification. This model adds attention modules at different stages of the backbone network to achieve more efficient multi-scale feature extraction. Moreover, the anti-attention module is used to supplement the information lost by the attention module during the training process, achieving higher performance. Zhou *et al.* [14] proposed a light-weighted Omni-Scale Network (OSNet). The network implemented multi-scale feature extraction through the Omni-Scale Residual Block, and fuse these multi-scale features through the unified aggregation gate.

B. MULTI-BRANCH MODEL

Multi-branch is used to regularize the model by different but consistent goals. Dai *et al.* [15] proposed a novel Batch-Drop-Block (BDB) Network, which consists of a global feature branch and a focused feature learning branch of Batch-Drop-Block that enhances local features. The network connects the features of these two branches to provide a more comprehensive and spatially distributed feature representation. Guo *et al.* [16] proposed another novel two-branch network, which consists of a human body branch using a modern human body analysis model and a potential part branch using a self-focus mechanism. This network solves the problem of misalignment of human body parts and misalignment of non-human body parts.

Quan *et al.* [17] proposed a novel multi-branch part-aware module that incorporates information from various parts of the body, which is used to enhance the body structure information of a given input feature tensor. Yang *et al.* [18] proposed a three-branch enhanced Class Activation Maps (CAM) model. Its backbone network consists of a series of ordered partial branches that share the same input. This model uses a loss function called OPA to guide the network training to find distinguishing features from areas where the previous branches have less activation.

C. ATTENTION MECHANISM

Attention mechanisms are known to be useful in Re-ID. Liu *et al.* [19] proposed an end-to-end Comparative Attention Network (CAN). The network simulates the human perception process, that is, learning to integrate the information related to the person's identification in the image to determine whether the two images are from the same person. Rahimpour *et al.* [20] proposed a novel approach based on using a gradient-based attention mechanism in a deep convolution neural network for solving the person re-identification problem. This approach is capable of extracting information from an image by adaptively selecting the most informative image regions and only processing the selected regions at high resolution. Zheng *et al.* [21] proposed a new attention-driven Siamese learning framework, called the Consistent Attentive Siamese Network (CASN). The framework uses the original ID signal to guide the model to find consistent attention areas for images of the same identity, and also learns the identity-aware constant representation for cross-view matching.

Li *et al.* [22] proposed a novel Harmonious Attention CNN (HA-CNN) model for joint learning of soft pixel attention and hard regional attention along with simultaneous optimization of feature representations, dedicated to optimizing person Re-ID in uncontrolled (misaligned) images. Xu *et al.* [23] proposed a novel framework called Attention-Aware Compositional Network (AACN) for person re-identification. Pose-Guided Part Attention (PPA) in this framework is used to estimate finer part attention to extract features more accurately and can extract effective features for occlusion problems.

Xia *et al.* [24] proposed a novel attention mechanism. It can directly model remote relationships through second-order feature statistics to find the correspondence between local features in two images. Therefore, it achieves more precise position matching for similarity measures. Chen *et al.* [25] proposed a novel high-order attention (HOA) module. This module can use sophisticated high-level statistics and attention mechanisms to capture nuances between pedestrians and generate distinctive attention suggestions to achieve better discrimination.

III. METHOD

A. MODEL ARCHITECTURE

1) BASELINE

First, we introduce our baseline architecture: Strong Baseline [3]. Strong Baseline, as illustrated by Figure 2 (a), adopts Resnet-50 as the backbone. The stride parameter of Stage 4 is set to 1 instead of 2 to remain more details. Output feature maps of Stage 4 then go through a global average pooling layer and are converted into feature vectors, which are fed to triplet loss for metric learning at the training stage. The feature vectors are linearly scaled to final feature vectors, which are fed to ID loss (usually cross-entropy loss) at the training stage. The final feature vectors are directly used to compute the distance matrix at the inference stage. In the network, the linear scale operation is implemented by a batch normalization (BN) layer with a randomly selected but fixed scale parameter and zero bias.

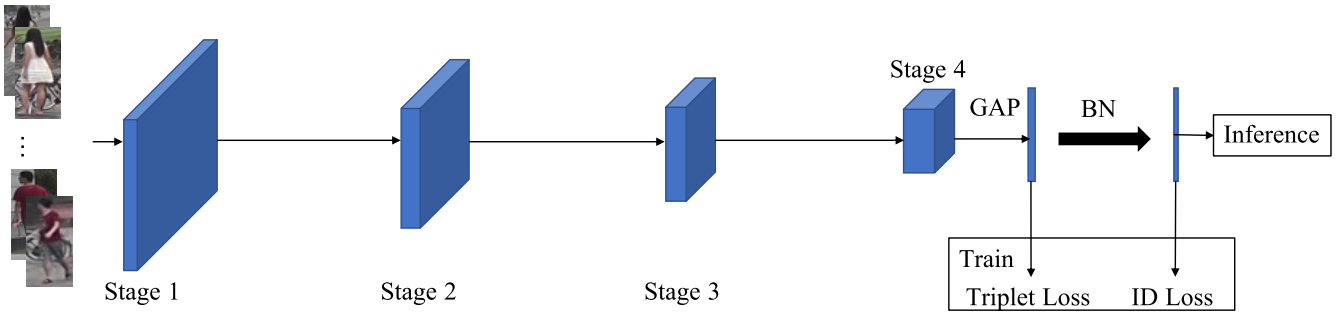
Strong Baseline [3] finds that the descent directions of triplet loss and ID loss are not the same and thus the two losses should not be used on the same feature vector. It alleviates the contradiction by introducing a BN layer to separate the original feature vector into two feature vectors and separately apply triplet loss and ID loss on them. Thus, the model can better fit to data. The modification is both light-weighted and effective in improving Re-ID performance.

An extra finding of Strong Baseline is that using instance batch normalization (IBN) [26] network as a backbone will bring performance gain in Re-ID. IBN nets replace part of BN layers of residual blocks with Instance Normalization (IN) layers, which improve the generalization power of the model. Moreover, this change improves both classification and Re-ID performance with little extra computation.

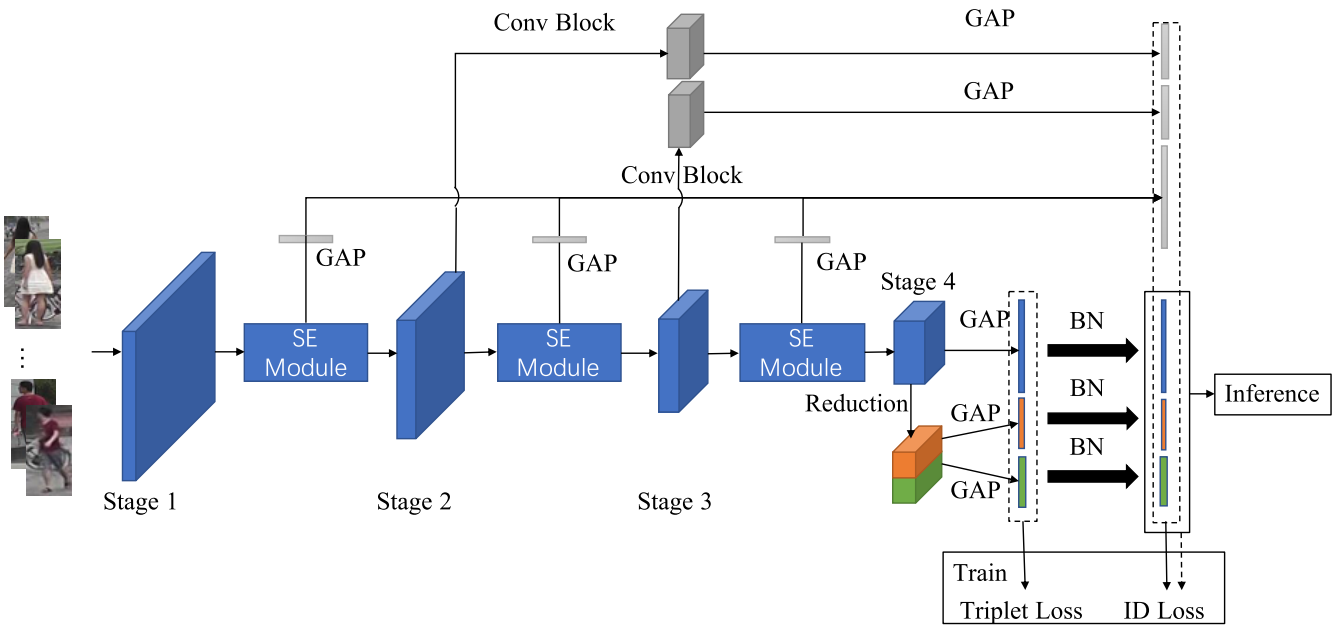
2) KEY DESIGNS

Our model architecture is depicted in Figure 2 (b), which is modified from the baseline depicted in (a). The input part, four stages of the backbone and separation of triplet loss and ID loss are inherited. Major updates of the architecture include: (a) multi-scale branch that extracts features from early Stages 2 and 3. Multi-scale features improve performance in both traditional computer vision methods and CNNs. (b) Addition of local branch with sliced feature at Stage 4, which helps the final feature vector to focus on the upper body and lower body separately. (c) Insertion of attention modules between the stages, along with an inverse attention branch that takes advantage of features abandoned by the attention mechanism. Here we use the squeeze-and-excitation (SE) modules [27] as the attention modules. (d) We use Resnet-50 with instance-batch-normalization to replace the original Resnet-50 backbone. The following part of this subsection will introduce those designs in detail.

Our model architecture improves the baseline in several aspects. First, we introduce the idea of a multi-scale feature at the training stage. Multi-scale feature is known to be effective in both hand-crafted features and CNN features in computer vision since it helps solve the problem of various scales. We introduce two branches at Stages 2 and 3 of the backbone. Each branch is a convolutional block followed by a global max-pooling (GAP) layer. Output feature vectors are fed to ID loss. In summary, we use output features of Stages 2, 3, and 4 as multi-scale features to improve model performance. The two extra branches, as depicted by gray color in Figure 2 (b), are removed at the inference stage. Thus, they will not increase computation at the inference time and are somehow a type of regularization. Figure 1 visualizes the output feature maps of the ablation study of our network, comparison of the second column to the last column clearly shows the design greatly strengthens the whole feature by introducing a multi-scale glance.



(a) Strong Baseline



(b) Our Model Architecture

FIGURE 2. **Model Architecture. (a) Illustrates the model architecture of strong baseline [3]. The backbone of the network is Resnet-50 with four stages. The convolution stride of Stage 4 (last stage) is changed from 2 to 1 to retain more image details. Then, the features of Stage 4 go through a global average pooling (GAP) and become a vector feature, which is fed to triplet loss at the training stage. (b) Illustrates our model architecture, which is modified from (a). First, the backbone is Resnet50-IBN [26] with four stages, which is proven to generalize significantly better than Resnet-50 in both classification and Re-ID tasks. Then, we introduce three attention modules among the four stages. The attention modules are squeeze-and-excitation (SE) modules which strengthen channel attention. Besides, feature maps abandoned by SE modules are also constrained by ID loss. Next, we introduce two branches from both Stage 2 and Stage 3 to generate multi-scale features to improve model generalization. Last, we introduce an additional local branch at Stage 4 to strengthen the representation power of features for the upper body and lower body.**

Second, we add sliced features. The 2048-channel feature maps of Stage 4 are first reduced to 512 channels by a basic convolutional block. Then, it is horizontally sliced into upper and lower parts as depicted by orange and green in Figure 2 (b). The two parts go through a routing similar to the original global feature and output local feature vectors. Global feature vector and local feature vectors are concatenated as the final feature vector. The final feature vector is under an ID loss constraint at the training stage and directly used to compute the distance matrix at inference stage. A comparison of the fourth column to the last one of Figure 1 indicates the local branch strengthens features at the upper and lower parts separately.

Third, we introduce attention modules between stages of backbone as depicted in Figure 2 (b). We directly use the squeeze-and-excitation (SE) module [27] as the attention modules,as depicted in Figure 3. The input feature map $X \in R^{H\times W\times C}$ are first processed by a global average pooling layer and only $1 \times 1 \times C$ elements remain. Then, the C elements are *squeezed* by a fully connected layer to only C/r elements. Next, the squeezed elements are activated by ReLU and *excited* back to C elements. At last, the C elements are activated by a sigmoid layer to form channels weights. Those channel weights are scale factors that applied to the original input feature X channel-wise.

SE module gives scalar weights to channels of output feature maps and is a type of self-attention module. SE module forces the attention of the network to fewer channels and thus is also a channel attention module. It has been successfully used in image classification [27] and we find it also useful

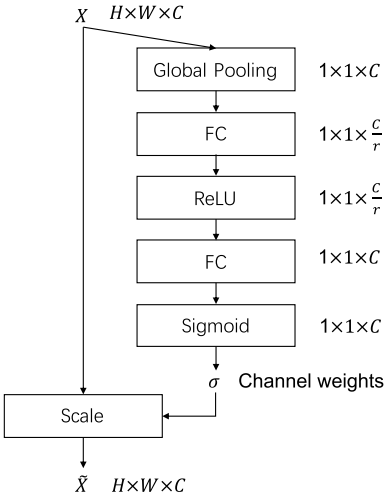
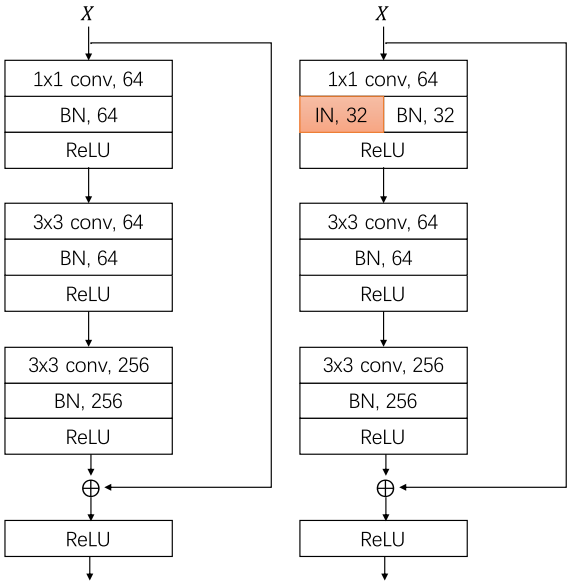


FIGURE 3. Squeeze-and-excitation (SE) module.



(a) Original Bottleneck (b) Bottleneck with IBN

FIGURE 4. Bottleneck with or without IBN.

in Re-ID. Apart from that, feature maps abandoned by SE modules are also constrained by ID loss, named inverse attention [13]. It is a type of regularization that improves Re-ID performance. A comparison of the third column to the last one of Figure 1 indicates the attention design strengthens features near the head and feet, which makes the feature more distinguishing.

At last, we use Resnet-50 with instance batch normalization (IBN) as backbone. The original bottleneck module used in Resnet is depicted in 4 (a). The IBN bottleneck we used in our backbone is depicted in 4 (b). The only difference is that we changed half the number of channels of the first BN layers into instance normalization [28].

Instance normalization executes similar computation as batch normalization. Unlike batch normalization, instance normalization normalizes across each channel in each training example instead of normalizing across input features in a training mini-batch. Instance normalization is a type of contrast normalization [28]. It is first used in image style transfer [28] but also helps CNN models generalize better to process images with different light conditions and contrast.

A combination of instance normalization with batch normalization in Resnet has been proven to improve performance in person Re-ID [3] and image classification [26] with almost no extra computation. Therefore, we recommend using IBN as a new standard backbone for person Re-ID researches.

B. LOSS FUNCTIONS

Apart from model designs, loss functions play key roles in training a Re-ID network. We use only two types of loss functions that are commonly used in person Re-ID: triplet loss and cross-entropy loss as ID loss.

Triplet loss requires triplet samples. Given a batch of person images with labels, we first get a batch of image features f by our network. Then, we use each image as an anchor and search the batch to find the hardest positive image (with the biggest distance) and negative image (with the smallest distance), and thus we get the triplet sample consists of the hardest positive pair p and negative pair n. Triplet loss L_{tri} is defined as

$$L_{tri}(f) = L_{tri}(d_p, d_n) = \max(0, d_p - d_n + \alpha) \tag{1}$$

where d_p and d_n are feature distances of the hardest positive pair and negative pair of batch features f; α is a constant margin that if the distance difference is less than it, the triplet sample is ignored in this optimization.

Cross-entropy loss here is for multi-label. Given an image i with feature f_i and label l_i, cross-entropy loss L_{ce} is defined as

$$L_{ce}(f_i) = \begin{cases} \log(\textbf{softmax}(Wf_i + b)), & l_i = j \\ 0, & l_i \neq j \end{cases} \tag{2}$$

where W and b is a parameter matrix and vector, separately; $\textbf{softmax}(x)$ is a math operation that normalize a vector x as

$$\textbf{softmax}(x) = \exp(x)/\|\exp(x)\|_1 \tag{3}$$

Cross-entropy loss $L_{ce}(f)$ for batch features f is defined as the average of cross-entropy loss of every image feature f_i.

Figure 2 (b) intuitively demonstrate how we use the two types of loss functions. Now we introduce it formally. First, denote multi-scale features from Stages 2 and 3 as f_2, f_3; denote global branch feature before BN as g and after BN as g_{bn}; denote local branch features of the upper and lower parts before BN as u, l and after BN as u_{bn} and l_{bn}; denote concatenated inverse attention feature as f_{inv}. Then, triplet loss is applied to global and local branch features BEFORE BN layers, namely g, u, l. ID loss is applied to global and local branch features AFTER BN layers, namely g_{bn}, u_{bn}, l_{bn}; In addition, ID loss is also applied

TABLE 1. Statistics of used datasets.

Dataset	Time	#total ID	#training ID	#gallery ID	#image	#gallery image	#camera	Label	Resolution
Market1501 [1]	2015	1,501	751	752	32,668	19,732	6	hand & auto	fixed (128x64)
DukeMTMC-ReID [7]	2017	1,404	702	1,110	36,411	17,661	8	hand & auto	vary
CUHK03(Detected) [8]	2014	1,467	767	700	13,161	5,332	2	auto	vary
MSMT17 [9]	2018	4,101	1,041	3,060	126,441	82,161	15	auto	vary

to multi-scale features f_2, f_3. In summary, the total loss function is

$$L_a = L_{tri}(g) + L_{tri}(u) + L_{tri}(l) + L_{ce}(g_{bn}) + L_{ce}(u_{bn}) + L_{ce}(l_{bn}) + \lambda_s(L_{ce}(f_2) + L_{ce}(f_3)) + \lambda_i L_{ce}(f_{inv}) \quad (4)$$

where λ_s and λ_i are scalar weights of multi-scale features and inverse attention feature. We call Eq. 4 as Loss-a and thus model trained with Loss-a as MSBA-a.

The final feature for inference f_f is concatenation of the global and local branch features, such that

$$f_f = [g_{bn}, u_{bn}, l_{bn}]^T \quad (5)$$

Now the final feature f_f is indirectly restricted by parts. If we add an ID loss as a direct restriction to the final feature, there is another total loss

$$L_b = L_a + \lambda_f L_{ce}(f_f) \quad (6)$$

where λ_f is a scalar weight. We name Eq. 6 Loss-b and model trained with Loss-b is MSBA-b.

We find Loss-a 4 and Loss-b 6 achieve good performance on different datasets, which will be analyzed in Section V.

IV. DATASETS AND TRAINING TRICKS

Datasets and training tricks are critical to person Re-ID model design and performance. However, they are often underestimated without sufficient proportion in research papers. Performance gains of some papers come from training tricks rather than models themselves [3]. Therefore, we use a separate section to analyze datasets we use and introduce training tricks, which are supposed to be beneficial for both research and industry.

A. DATASETS

Understanding datasets is the most important step in machine learning, including person Re-ID. A Re-ID dataset consists of training set, gallery set, and query set. We select four commonly used datasets to design and evaluate our method: Market-1501 [1], DukeMTMC-ReID [7], CUHK03(Detected) [8] and MSMT17 [9]. Statistics of the four datasets are listed in Table 1.

Market1501 [1]. Market1501 contains 32,668 images with 1,501 person IDs, which are almost equally divided into a training set and a test set (gallery + query). Images are captured by 6 cameras with fixed resolution (128 × 64). Labels are marked by algorithms with manual correction. The dataset is of high quality and Re-ID methods tend to get high performance on it.

DukeMTMC-ReID [7]. DukeMTMC-ReID contains 36,411 images with 1,404 person IDs. Train set contains half of the person IDs while the gallery set contains 1,110 person IDs, which means about one-third of person IDs in the test set are known. This generally leads to better performance metrics. Images are captured by 8 cameras with various resolutions. Therefore, the choice of input image resolution will be critical for this dataset. Labels of the dataset are also marked by algorithms with manual correction.

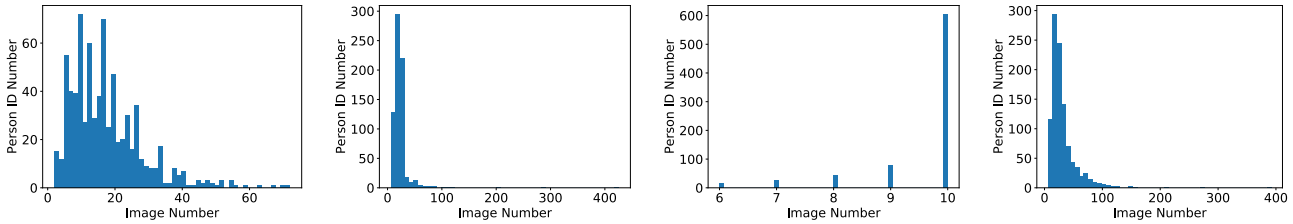
CUHK03(Detected) [8]. CUHK03 consists of two separate datasets: Detected and Labeled. The difference is how labels are generated. Labels of CUHK03 (Detected) are without manual correction, which is easy to obtain, close to the industry but more difficult to cope with. The dataset contains 13,161 images with 1,467 person IDs split into training and test sets without overlap. Images are captured by only 2 cameras with various resolutions.

MSMT17 [9]. By Jan. 2020, MSMT17 is the largest public dataset for Re-ID to the best of our knowledge. It contains 126,441 images with 4,101 person IDs, which are also split into training and test sets without overlap. Note the gallery set contains 82,161 images of 3,060 person IDs. The numbers are much greater than those in the training set. Moreover, the labels are automatically generated with various resolutions. Therefore, the dataset is the most challenging public dataset at the writing time.

In addition to statistics of the datasets, the distribution of image numbers with person ID is also important in training, as plotted in Figure 5. CUHK03(Detected) has a unique distribution which indicates the training images are manually selected. In contrast, Market1501, DukeMTMC-ReID, and MSMT17 share a long tail distribution of different widths and heights. **It means the person ID distributions of the datasets are unbalanced.** Design targeted strategies in data sampling might help improve performance.

B. TRAINING TRICKS

Tricks from Strong Baseline [3]. Strong baseline introduced tricks including warm-up learning rate, adjust the last stride of Stage 4 as 1, data argumentation with random erasing, which we confirm to be effective. Label smoothing is not harmful, which we keep. Center loss brings a slight performance drop in some datasets, thus we do not use it.



(a) Market1501 (b) DukeMTMC-ReID (c) CUHK03(Detected) (d) MSMT17

FIGURE 5. Person ID number with image number of training images.

Learning Rate schedule. The learning rate $lr(t)$ of the epoch t is as follows:

$$lr(t) = \begin{cases} 4 \times 10^{-4} \times \frac{t}{10}, & t \leq 10 \\ 4 \times 10^{-4}, & 10 < t \leq 40 \\ 4 \times 10^{-5}, & 40 < t \leq 70 \\ 4 \times 10^{-6}, & 70 < t \leq 100 \end{cases} \tag{7}$$

Compared with [3], we increase the basic learning from 3.5×10^{-4} to 4×10^{-4} and reduce total training epochs from 120 to 100. We find the changes slightly improve performance while saving training time.

Sampling Strategy. The sampling strategy dramatically affects person Re-ID performance but seldom discussed in depth. We have noticed three out of four public Re-ID datasets share a characteristic of unbalanced person ID distribution, as depicted in Figure 5. We also find the characteristic in private datasets in the Re-ID competition we participated in. To cope with that, we first sample the training set but restrict the max instances from a single person ID with threshold σ. That is to say, **if a person ID has more images than σ, we randomly sample σ instances at the beginning of each epoch.** The strategy is effective to prevent the model from over-fitting person IDs with big numbers of instances. Besides, sampling images within an epoch is also crucial to performance and we follow [3]. In summary, the strategy is as the following steps:

1) At the beginning of each epoch, randomly sample the training set that images from a single person ID must be less than threshold σ.
2) Generate a person ID set from the training set where we first randomly select P person IDs.
3) Sample K images for every selected person ID. Note if a person ID has fewer images than K, it is sampled with the return; otherwise, it is sampled without the return. Considering we use random erasing data argumentation, sampling with the return will also result in different images.
4) Use the select $P \times K$ images to train the model and remove those images from the training set. If all images of a person ID is removed, the person ID is also removed from the person ID set.
5) Go back to Step 2) unless there is no image left in the training set, which means the epoch is finished.

TABLE 2. Parameters in implementation.

Parameter	Value	Description
λ_s	0.03	Weight of multi-scale losses.
λ_i	0.1	Weight of inverse attention loss.
λ_f	1	Weight of the final feature in MSBA-b.
σ	50	Maximum image number of a person ID in an epoch.
P	16	Number of person IDs sampled in every batch.
K	4	Number of images sampled for every person ID.

V. EXPERIMENTS

We evaluate our method in the names of MSBA-a, MSBA-b, which are trained by Loss-a and Loss-b, separately. We do not use any enhanced inference tricks in our evaluation.

A. EVALUATION METRICS

There are two commonly used metrics to evaluate person Re-ID result sequences. First, Rank-1 (R1) of cumulative match characteristics (CMC) to assess Top-1 accuracy. Second, mean average precision (mAP) to assess the quality of the whole result sequences.

Since the two metrics may show different preferences, we also add a score as the final metric, which is average of the two metrics and also used by the competition we joined in.

B. IMPLEMENTATION DETAILS

We train our model with a single Nvidia Titan V GPU (12 GB GPU memory) on the four datasets at different image resolutions. The training procedure takes one to three hours for each dataset. Parameter values we used in losses, data sampling are listed in Table 2. One exception is that when we use input resolution 384×192, we set $P = 14$ to avoid GPU (CUDA) out of memory error.

C. COMPARISON WITH THE STATE-OF-THE-ART

We mainly compare our method with the state-of-the-art methods published in 2019 on the four datasets. Different from previous works, we also list input image size as a factor and find it is crucial for some datasets. Our results are without flipping inference, re-ranking or any other post-processing steps.

We replicate experiments of ABD [5] and DBD+Cut [15] using code and pre-trained weights that their

TABLE 3. Comparison on Market1501.

Method	Image Size	mAP	R1	Score
PNGAN [29] (ECCV18)	256x128	72.6	89.4	81.0
IANet [30] (CVPR19)		83.1	94.4	88.8
CAMA [18] (CVPR19)		84.5	94.7	89.6
Auto-ReID [17] (ICCV19)		85.1	94.5	89.8
OSNet [14] (ICCV19)		84.9	94.8	89.9
Strong Baseline [3] (CVPRW19)		85.9	94.5	90.2
DGNet [31] (CVPR19)		86.0	94.8	90.4
MSBA-a (Ours)		88.4	95.5	92.0
MHN-6 [25] (ICCV19)	288x144	85.0	95.1	90.0
PCB+RPP [32] (ECCV18)	384x128	81.6	93.8	87.7
CASN [21] (CVPR19)		82.8	94.4	88.6
*ABD [5] (CVPR19)		84.1	94.6	89.4
*DBD+Cut [15] (ICCV19)		85.5	94.6	89.6
P^2-Net [16] (ICCV19)		85.6	95.2	90.4
MMGA [10] (CVPRW19)		87.2	95.0	91.1
MGN [6] (CVPR18)		86.9	95.7	91.3
MSBA-b (Ours)		88.2	94.7	91.5
MSBA-a (Ours)		89.0	95.8	92.4

TABLE 4. Comparison on DukeMTMC-ReID.

Method	Image Size	mAP	R1	Score
CAMA [18] (CVPR19)	256x128	72.9	85.8	79.4
IANet [30] (CVPR19)		73.4	87.7	80.6
OSNet [14] (ICCV19)		74.8	86.6	80.7
DGNet [31] (CVPR19)		74.8	86.6	80.7
Strong Baseline [3] (CVPRW19)		76.4	86.4	81.4
Auto-ReID [17] (ICCV19)		75.1	88.5	81.8
MSBA-a (Ours)		80.1	90.1	85.1
MHN-6 [25] (ICCV19)	288x144	77.2	89.1	83.2
PCB+RPP [32] (ECCV18)	384x128	69.2	83.3	76.3
P^2-Net [16] (ICCV19)		73.1	86.5	79.8
CASN [21] (CVPR19)		73.7	87.7	80.7
*ABD [5] (CVPR19)		76.3	87.4	81.9
*DBD+Cut [15] (ICCV19)		76.3	87.4	81.9
MGN [6] (CVPR18)		78.4	88.7	83.6
MMGA [10] (CVPRW19)		78.1	89.5	83.8
MSBA-a (Ours)		79.8	89.7	84.8
MSBA-b (Ours)		79.7	90.3	84.9

provided. DBD+Cut produces the same results as reported in the paper on CUHK03(Detected). Note on MSMT17, replicate results of ABD is much worse than that reported in the paper. Except for those, all other replicate results are a little worse than those in their papers. We put in Tables 3, 4, 5 and 6 the results we got by our replications and mark results that are worse than reported with a star (*).

Table 3 lists results on Market1501 with the best metrics underlined. Our MSBA-a in resolution 256×128 outperforms compared methods in the same resolution by 1.6 in the score. Besides, it outperforms previous the state-of-the-art method MGN [6] in resolution 384 × 128 by 0.5 in the score. Our MSBA-a in resolution 384 × 128 outperforms MGN by 0.9 in the score. Note MSBA-b in resolution 384 × 128 performs slightly worse than MSBA-a on this dataset, but still better than all other methods.

Table 4 lists results on DukeMTMC-ReID with the best metrics underlined. In resolution 256 × 128, our MSBA-a outperforms the second-best method Auto-ReID [17] by 4.3 in the score. In fact, it achieves the best score across the three resolutions listed. In resolution 384 × 128, our MSBA-b outperforms the second-best method MMGA [10] by 1.1 in the score. We can see that resolution is not crucial on this dataset.

TABLE 5. Comparison on CUHK03(DETECTED).

Method	Image Size	mAP	R1	Score
MSBA-a (Ours)	256x128	67.9	71.2	69.6
OSNet [14] (ICCV19)		67.8	72.3	70.0
MSBA-b (Ours)		72.9	76.2	74.6
MHN-6 [25] (ICCV19)	288x144	65.4	71.8	68.6
MGN [6] (CVPR18)	384x128	66.0	66.8	66.4
CASN [21] (CVPR19)		64.4	71.5	68.0
Auto-ReID [17] (ICCV19)		69.3	73.3	71.3
P^2-Net [16] (ICCV19)		68.9	74.9	71.9
DBD+Cut [15] (ICCV19)		73.5	76.4	75.0
MSBA-a (Ours)		74.3	77.1	75.7
MSBA-b (Ours)		75.9	78.5	77.2

TABLE 6. Comparison on MSMT17.

Method	Image Size	mAP	R1	Score
IANet [30] (CVPR19)	256x128	46.8	75.5	61.1
DGNet [31] (CVPR19)		52.3	77.2	64.8
MSBA-a (Ours)		56.4	73.7	65.1
OSNet [14] (ICCV19)		52.9	78.7	65.8
PCB+RPP [32] (ECCV18)	384x128	40.4	68.2	54.3
*ABD [5] (CVPR19)		48.4	73.8	61.1
Auto-ReID [17] (ICCV19)		52.5	78.2	65.2
MSBA-b (Ours)		58.4	74.5	66.5
MSBA-a (Ours)		59.0	75.3	67.2
MSBA-a (Ours)	384x192	60.2	76.1	68.2

Table 5 lists results on CUHK03(Detected) with the best metrics underlined. Our MSBA-b in resolution 384 × 128 outperforms the second-best method DBD+Cut [15] by 2.2 in the score. Besides, it surpasses other methods by a large margin. We also find that our results in resolution 384 × 128 is much better than that in 256 × 128. It suggests a large resolution on the dataset can significantly improve person Re-ID performance.

Table 6 lists results on MSMT17 with the best metrics underlined. We mentioned the gallery set of the dataset is much larger than all other datasets and metrics on the dataset are relatively smaller. In resolution 384 × 128, our MSBA-a outperforms the second-best method Auto-ReID by 2.0 in the score. In resolution 256 × 128, however, OSNet gets a better score. We find the mAP of our method is significantly better than all compared methods while the R1 is worse than some other methods. We trained an extra MSBA-a in resolution 384 × 192, which suggests that an increase in resolution can further improve performance.

In summary, our method achieves new state-of-the-art performance on all four datasets. Big resolutions are cru-

TABLE 7. Ablation study on CUHK03(Detected).

Method & Change	mAP	R1	Score
No attention	73.8	76.3	75.1
No inverse Attention ($\lambda_i = 0$)	74.3	77.1	75.7
No multi-scale ($\lambda_s = 0$)	74.4	77.7	76.1
Strengthen multi-scale ($\lambda_s = 1$)	72.1	74.7	73.4
No local branch	75.4	78.4	76.9
MSBA-b	75.9	78.5	77.2

cial on CUHK03(Detected) and MSMT17 but not important on Market1501 and DukeMTMC-ReID. MSBA-a performs better than MSBA-b on three out of the four datasets. Except for CUHK03(Detected), MSBA-b is significantly better. We already see CUHK03 uses fewer cameras (only two cameras) and has a unique image number with person ID distribution, which may be the cause.

D. ABLATION STUDY

We study the contributions of attention, the multi-scale branches and the local branch on CHUK03(Detected). Table 7 lists MSBA-b and compared cases with a single change. Score drops by 1.5 when we remove inverse attention and 2.1 when we remove all attention modules, which indicates all parts of attention contributes to performance. We also set the weight of multi-scale loss $\lambda_s = 0$ and $\lambda_s = 1$, which all result in performance drops. We conclude that multi-scale is beneficial but requiring a proper weight. Too big of weight of multi-branch will lead to performance drop since it interferes with other more important losses. At last, we test the effects of the local branch and find it also beneficial for performance. If the computation resource is limited, cutting off the local branch is also an option without a big performance drop. The results can also be confirmed by Figure 1, which we analyzed in model designs.

VI. SOME OTHER TRICKS IN THE RE-ID CHALLENGE

We participated in the First National Artificial Intelligence Challenge (2019, China) where we tested the model on a much bigger private dataset. In this section, we share some findings and experience that is hard to be noticed in other research papers.

Be careful of Stripped Parts. Some research works design CNN models with many stripped parts. In those works, output feature maps of the backbone are horizontally sliced into three or more parts and go through separate branches. We find this design improves performance on small datasets with high-quality labels. However, it is harmful to big datasets with automatic labels. There is little chance that the stripped features can align across images even with aligning steps.

Flipped inference and post-processing. We can flip images and compute a flipped distance matrix. By averaging the flipped distance matrix with the original one, performance can be improved by about 1 in mAP and $R1$ on big or difficult datasets. If we can obtain the whole test set, posting processing steps such as Re-ranking [33] will dramatically increase mAP and R1. We do not use those tricks for a fair comparison in this work but they are helpful in other cases.

Take advantage of the gallery set. When the gallery set is big, fixed and accessible, we can use unsupervised learning strategies such as self-similarity grouping [34] to improve Re-ID performance by a large margin (about 5 in the score in the challenge). However, unsupervised learning will take a large amount of time and make the model over-fitted to the specific gallery set only.

VII. CONCLUSION

In this work, we propose a novel and effective CNN model, analyze four commonly used datasets and introduce important training tricks for person Re-ID. In the proposed model, we combine the three most effective designs, including multi-scale, multi-branch and attention mechanism, with carefully designed architecture and training losses. We analyze the four commonly used datasets and propose an improved sampling strategy accordingly. We find that training tricks, including but not limited to learning rate schedule, sampling strategy and image resolutions, are important to person Re-ID performance and propose our solutions. With the effective model designs and training tricks, we achieve new state-of-the-art performance on all of the four public datasets. We also share findings in participating in the First National Artificial Intelligence Challenge (2019, China) where we tested the model on a much bigger private dataset.

REFERENCES

[1] 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.
[2] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," 2020, *arXiv:2001.04193*. [Online]. Available: http://arxiv.org/abs/2001.04193
[3] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Trans. Multimedia*, early access, Dec. 9, 2019, doi: 10.1109/TMM.2019.2958756.
[4] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "AlignedReID: Surpassing human-level performance in person re-identification," 2017, *arXiv:1711.08184*. [Online]. Available: http://arxiv.org/abs/1711.08184
[5] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "ABD-net: Attentive but diverse person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8351–8361.
[6] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple granularities for person re-identification," in *Proc. ACM Multimedia Conf. Multimedia Conf. MM*, 2018, pp. 274–282.
[7] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 17–35.
[8] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep filter pairing neural network for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 152–159.
[9] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.
[10] H. Cai, Z. Wang, and J. Cheng, "Multi-scale body-part mask guided attention for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019.

[11] X. Liu, H. Tan, X. Tong, J. Cao, and J. Zhou, "Feature preserving GAN and multi-scale feature enhancement for domain adaptation person re-identification," *Neurocomputing*, vol. 364, pp. 108–118, Oct. 2019.

[12] C. Wang, L. Song, G. Wang, Q. Zhang, and X. Wang, "Multi-scale multi-patch person re-identification with exclusivity regularized softmax," *Neurocomputing*, vol. 382, pp. 64–70, Mar. 2020.

[13] D. Wu, C. Wang, Y. Wu, and D.-S. Huang, "Attention deep model with multi-scale deep supervision for person re-identification," 2019, *arXiv:1911.10335*. [Online]. Available: http://arxiv.org/abs/1911.10335

[14] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3702–3712.

[15] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch DropBlock network for person re-identification and beyond," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3691–3701.

[16] J. Guo, Y. Yuan, L. Huang, C. Zhang, J.-G. Yao, and K. Han, "Beyond human parts: Dual part-aligned representations for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3642–3651.

[17] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-ReID: Searching for a part-aware ConvNet for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3750–3759.

[18] W. Yang, H. Huang, Z. Zhang, X. Chen, K. Huang, and S. Zhang, "Towards rich feature discovery with class activation maps augmentation for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1389–1398.

[19] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-End comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492–3506, Jul. 2017.

[20] A. Rahimpour, L. Liu, A. Taalimi, Y. Song, and H. Qi, "Person re-identification using visual attention," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 4242–4246.

[21] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5735–5744.

[22] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, p. 2285.

[23] 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.

[24] B. Bryan, Y. Gong, Y. Zhang, and C. Poellabauer, "Second-order non-local attention networks for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3760–3769.

[25] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 371–381.

[26] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via IBN-net," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 464–479.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[28] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: http://arxiv.org/abs/1607.08022

[29] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 650–667.

[30] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-And-Aggregation network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9317–9326.

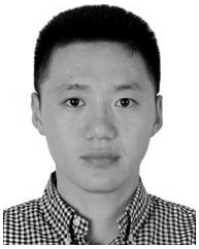
[31] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2138–2147.

[32] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 480–496.

[33] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-Reciprocal encoding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1318–1327.

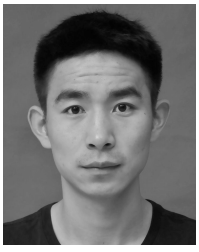
[34] Y. Fu, Y. Wei, G. Wang, Y. Zhou, H. Shi, U. Uiuc, and T. Huang, "Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6112–6121.

HANLIN TAN received the B.S. and M.S. degrees in system engineering from the National University of Defense Technology, Changsha, China, in 2013 and 2016, respectively, where he is currently pursuing the Ph.D. degree in system engineering. His current research interests focus on image denoising and image processing pipeline.



HUAXIN XIAO received the B.E. degree from the University of Electronic Science and Technology of China, China, in 2012, and the Ph.D. degree from the National University of Defense Technology, China, in 2018. He was a Visiting Student with the National University of Singapore, from 2016 to 2018. He is currently a Lecturer at the College of System Engineering, National University of Defense Technology. His current research interest focuses on saliency detection and image/video object segmentation. He received the winner prize in object localization task from ILSVRC 2017.

XIAOYU ZHANG received the M.S. degree from the National University of Defense Technology, Changsha, China, where he is currently pursuing the Ph.D. degree with the Science and Technology on Information System Engineering Laboratory. His research interests include natural language processing, machine learning, and sequence-to-sequence learning.



BIN DAI received the B.S. degree from Chang'an University, Xi'an, China, in 2019. He is currently pursuing the M.S. degree in system engineering with the National University of Defense Technology. His current research interests focus on image object segmentation and image processing pipeline.

SHIMING LAI received the B.S. and Ph.D. degrees in system engineering from the National University of Defense Technology, Changsha, China, in 2008 and 2014, respectively. He is currently a Lecturer at the Department of System Engineering, National University of Defense Technology. His research interests include computer vision, computational photography, and imaging systems.

YU LIU received the B.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2005, and the M.Sc. degree in image processing and the Ph.D. degree in computer graphics from the University of East Anglia, Norwich, U.K., in 2007 and 2011, respectively. He is currently an Associate Professor at the Department of System Engineering, National University of Defense Technology. His research interests include image/video processing, computer graphics, and visual haptic technology.

MAOJUN ZHANG received the B.S. and Ph.D. degrees in system engineering from the National University of Defense Technology, Changsha, China, in 1992 and 1997, respectively. He is currently a Professor at the Department of System Engineering, National University of Defense Technology. His current research interest focuses on computer vision, information system engineering, and system simulation.