Knowledge-Supervised Learning: Knowledge Consensus Constraints for Person Re-Identification

Li Wang[1,2], Baoyu Fan[1,2,*], Zhenhua Guo[1,2], Yaqian Zhao[1,2], Runze Zhang[1,2], Rengang Li[1,2], Weifeng Gong[1,2], Endong Wang[1,2]
[1]Inspur Electronic Information Industry Co.,Ltd.
[2]State Key Laboratory of High-end Server & Storage Technology
wang.lilc@inspur.com,fanbaoyu@foxmail.com,{guozhenhua,zhaoyaqian,zhangrunze,lirg,gongwf,wangend}@inspur.com

ABSTRACT

The consensus of multiple views on the same data will provide extra regularization, thereby improving accuracy. Based on this idea, we proposed a novel Knowledge-Supervised Learning (KSL) method for person re-identification (Re-ID), which can improve the performance without introducing extra inference cost. Firstly, we introduce isomorphic auxiliary training strategy to conduct basic multiple views that simultaneously train multiple classifier heads of the same network on the same training data. The consensus constraints aim to maximize the agreement among multiple views. To introduce this regular constraint, inspired by knowledge distillation that paired branches can be trained collaboratively through mutual imitation learning. Three novel constraints losses are proposed to distill the knowledge that needs to be transferred across different branches: similarity of predicted classification probability for cosine space constraints, distance of embedding features for euclidean space constraints, hard sample mutual mining for hard sample space constraints. From different perspectives, these losses complement each other. Experiments on four mainstream Re-ID datasets show that a standard model with KSL method trained from scratch outperforms its ImageNet pre-training results by a clear margin. With KSL method, a lightweight model without ImageNet pre-training outperforms most large models. We expect that these discoveries can attract some attention from the current de facto paradigm of "pre-training and fine-tuning" in Re-ID task to the knowledge discovery during model training.

CCS CONCEPTS

• **Computing methodologies** → **Image representations**; **Object identification**.

∗ Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475340

KEYWORDS

Person retrieval, knowledge distillation, consensus constraints, isomorphic auxiliary training

ACM Reference Format:
Li Wang[1,2], Baoyu Fan[1,2,*], Zhenhua Guo[1,2], Yaqian Zhao[1,2], Runze Zhang[1,2], Rengang Li[1,2], Weifeng Gong[1,2], Endong Wang[1,2] . 2021. Knowledge-Supervised Learning: Knowledge Consensus Constraints for Person Re-Identification. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475340

1 INTRODUCTION

Person re-identification (Re-ID) is an important technique towards automatic search of a person's presence from a large-scale gallery collected by non-overlapping cameras. In recent years, it has made remarkable progress thanks to the development of deep learning [33]. To further improve person Re-ID performance, scholars tend to build network structures in a more complex way [3, 13, 32, 37]. Generally, it will lead to an obvious growth of parameters and system latency. In practice, person Re-ID requires fast retrieving the true matches across massive gallery data and tracking pedestrians, especially in limited computing power scenarios.

Improving the performance of Re-ID without introducing extra inference cost is a more graceful scheme. For image classification, auxiliary training boosts the accuracy through appending auxiliary supervision branches on top of certain intermediate network layers without affecting the inference graph. It has been widely used in [2, 5, 8, 9, 31]. However, auxiliary training does not consider the consensus constraints between branches.

As for consensus constraints, multi-task learning [14, 15, 20] enhances the model's generalization and performance by adding multiple related tasks to the main task. However, it is not usable for a single task. [28] proposes a method that treats other branches classified outputs as soft labels, which was similar to knowledge distillation in principle. Moreover, knowledge distillation [1, 19, 21] enables smaller networks to achieve higher performance by distilling knowledge from larger networks. However, knowledge distillation cannot be trained end-to-end, and knowledge transfer is a complex process that requires training experience and consumes more training time.

Our motivation is utilizing knowledge to constrain the consensus of multiple views on the same data to improve accuracy without introducing extra inference cost. Besides the problems

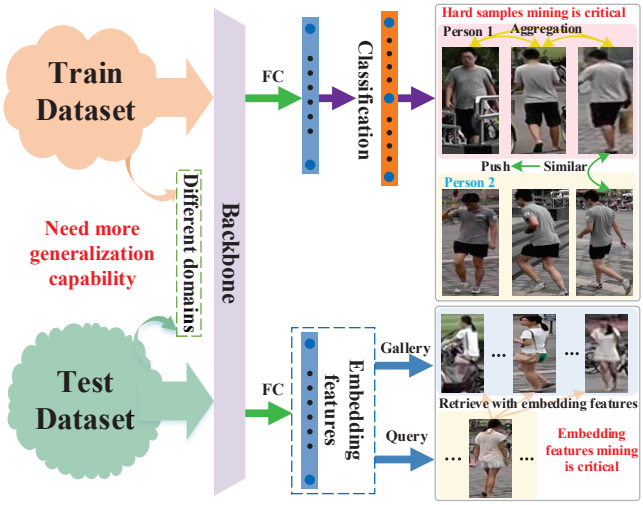

Figure 1: Unique Properties of Person Re-ID. (1) IDs in the test dataset do not overlap with the training dataset; (2) Embedding features are critical, which are actually used to retrieve the same person in gallery at the inference phase; (3) Hard sample mining is significant due to the properties of metric learning that training to pull semantically similar features to nearby while pushing dissimilar image features apart.

mentioned above, person Re-ID has unique properties that differ from image classification, as shown in Figure 1. First, person Re-ID is a retrieval task. Identities (IDs) in the test dataset do not overlap with the training dataset, which requires the model to be more generalizable. Second, embedding features are critical. They are the low-dimensional representations of input images in embedding space and are actually used to retrieve the same person in gallery at the inference phase. Third, person Re-ID is a form of metric learning. It focuses on training to embed the image features that are semantically similar to nearby locations while pushing dissimilar image features apart, which means that hard sample mining is significant.

Based on the above observations, we propose a novel Knowledge-Supervised Learning method (KSL) for person re-identification, which improves the performance with no extra inference cost. First, we introduce an isomorphic auxiliary training strategy to conduct basic multiple views. It appends auxiliary supervision branches on top of certain intermediate network layers, which have the same structure as the classifier heads of the backbone network. The natural similarity of isomorphic branches makes consensus constraints possible. Second, to maximize the agreement among multiple views, we design three novel knowledge consensus constraint losses: similarity of predicted classification probability (KSL-P) for cosine space constraints, distance of embedding features (KSL-E) for euclidean space constraints, hard sample mutual mining (KSL-H) for hard sample space constraints. From different perspectives, these losses complement each other.

The major contributions are summarized as follows. 1) We propose an innovative paradigm that trained a standard model from scratch outperforms its ImageNet pre-training results through knowledge discovery during model training, without introducing extra inference cost. 2) Three novel synergy losses (KSL-P, KSL-E, and KSL-H) are proposed from different perspectives but complement each other for consensus constraints of multiple views. 3) With KSL method, a lightweight model without ImageNet pre-training outperforms most large models.

2 RELATED WORK

2.1 Deep Person Re-ID

The prevailing success of machine learning, especially deep learning, in the field of computer vision [26, 27] in recent years has also made person Re-ID no exception. Most existing deep person Re-ID models adapt an off-the-shelf backbone for pedestrian feature learning. [7, 30, 36] used ResNet to optimize the representation of pedestrian features. Luo *et al.* [16] applied multiple tricks based on ResNeXt to further improve the retrieval accuracy of the Re-ID task. Generally, this kind of method usually has a good performance. However, in terms of computing cost, it could be overburdened for scenarios with limited computing power. Therefore, some works tend to design lightweight models for person Re-ID. [11, 18] apply MobileNet and Inception for person retrieval, respectively. Zhou *et al.* [41] developed a novel lightweight OSNet to learn omni-scale features of pedestrians. Fan *et al.* [4] developed a novel lightweight CMSNet to learn common and contextual multi-scale representations for person Re-ID. Nevertheless, lightweight networks usually have worse performance in person Re-ID, and the gap even larger when training from scratch. In this paper, we proposed the KSL method that outperforms most large models even adopted to lightweight models without pre-training on ImageNet (as shown in Table 2).

2.2 Collaborative Training Methods

We mention collaborative training to summarize a class of methods that add extra supervision in the training phase without affecting the inference graph. Wu *et al.* [35] introduced knowledge distillation to person Re-ID through injecting the teacher model's prior knowledge to assist training. However, it can not be used for end-to-end training. GoogleNet [31] is the first work adopted auxiliary training strategy by adding auxiliary classifiers to two intermediate layers. Lee *et al.* [9] proposed to add auxiliary classifiers to all hidden layers of the network. Based on auxiliary training, [28, 29] further implemented the information interaction among different branches by adding soft labels, in which [28] employed the mean prediction probability as a bridge for information interaction and [29] created mutual learning between paired branches via being soft labels for each other. These methods inspire the core idea of our proposed method. However, much pertinence design has been done for person Re-ID, and it is the first time that knowledge consensus constraints have been systematically studied.

3 KNOWLEDGE-SUPERVISED LEARNING

The framework of KSL is shown in Figure 2. Essential steps to implement Knowledge-Supervised Learning (**KSL**) can be divided into building basic multiple views and designing consensus constraints. We introduce isomorphic auxiliary training for

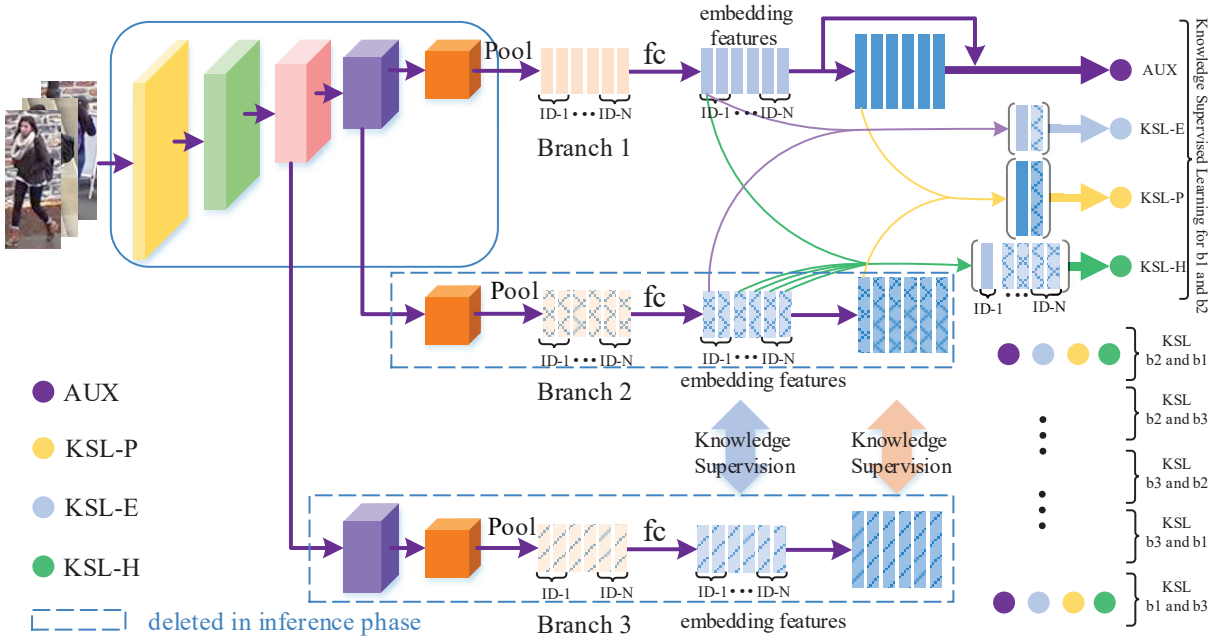

Figure 2: Knowledge-Supervised Learning. AUX represents isomorphic auxiliary training loss consists of cross-entropy and triplet loss. KSL-P, KSL-E, KSL-H stands for knowledge consensus constraints of dynamic classified probability, distance of embedding distance, and hard sample mutual mining. Each batch contains N groups of pedestrian images $ID-1 \cdot \cdot ID-N$, $b\{i\}$ is short for $Branch\{i\}$. Each vertical bar represents the feature of one pedestrian image.

basic multiple views on the same input data. In theory, auxiliary branches with the same structure should have the same properties at all perspectives. The consensus constraints aim to maximize the agreement among multiple views. Three novel constraints losses are proposed to distill the knowledge that needs to be transferred across different branches: similarity of predicted classification probability (KSL-P) for cosine space constraints, distance of embedding features (KSL-E) for euclidean space constraints, hard sample mutual mining (KSL-H) for hard sample space constraints. From different perspectives, these losses complement each other.

To illustrate the detail implementation of knowledge-supervised learning, let $\mathbb{D}_t = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)|n \in [1, N]\}$ be training dataset collected from K image classes which has N samples in total. Here, $\boldsymbol{x}_n$ is the training sample and $\boldsymbol{y}_n$ is the corresponding ground truth label which represented by a one-hot vector with K-dimensions. A CNN model can be expressed as $f(\boldsymbol{x}_n, \boldsymbol{\theta})$, where f is determined by the graph architecture, and $\boldsymbol{\theta}$ represents the network parameters. The optimization objective of the neural network with knowledge consensus constraints can be defined as:

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \mathcal{L}_{aux}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) + \mathcal{L}_{ksl\_p}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) \\ & + \mathcal{L}_{ksl\_e}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) + \mathcal{L}_{ksl\_h}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) \end{aligned} \tag{1}$$

where $\mathcal{L}_{aux}$ is the loss of isomorphic auxiliary training, $\mathcal{L}_{ksl\_p}$, $\mathcal{L}_{ksl\_e}$ and $\mathcal{L}_{ksl\_h}$ are the losses of KSL-P, KSL-E and KSL-H, respectively.

3.1 Isomorphic Auxiliary Training

Isomorphic auxiliary training (**AUX**) boosts the accuracy by appending auxiliary branches with the same structure as the main branch to intermediate layers. Each branch adopts a cross-entropy loss and a triplet loss [12]. Triplet loss is an option that is primarily used to support KSL-H method. The genetic similarity of isomorphic branches laid the foundation for our subsequent optimization. The loss function of isomorphic auxiliary training can be defined as:

$$\begin{aligned} \mathcal{L}_{aux}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) &= \mathcal{L}_{aux\_c}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) + \mathcal{L}_{aux\_t}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) \\ &= \sum_{b=1}^{B} \alpha^b \mathcal{L}_c^b\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}^b\right) + \sum_{b=1}^{B} \beta^b \mathcal{L}_t^b\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}^b\right) \end{aligned} \tag{2}$$

where B stands for the number of branches, $\mathcal{L}_t^b$ represents the triplet loss of b^{th} branch. $\mathcal{L}_c^b$ represents the cross-entropy loss of b^{th} branch. α^b, β^b in $(0, 1]$ are hyper-parameters used to balance $\mathcal{L}_t^b$ and $\mathcal{L}_c^b$.

It is important to note that adding auxiliary triplet loss directly at the beginning of training will cause the process to diverge when training from scratch. Because too many false positive difficult samples will confuse the optimization process unless the classifiers reach a certain degree of accuracy. We solve this problem by adding auxiliary triplet loss in the middle of the training process, and the sampler is switched synchronously. The detailed setting is discussed in the experimental section.

Moreover, the similarity of isomorphic branches makes it the best choice for each branch to allocate the same confidence weight. Therefore, in all the experiments, we set $\alpha^b = 1$, $\beta^b = 1$, which means that our proposed method does not introduce any extra hyper-parameters.

3.2 Knowledge Constraints of Dynamic Probability

Due to the similarity, each paired branch can be treated as soft labels for imitation learning. The basic idea is inspired by Hinton *et al.* [6], when a model has the ability to maximize the average log probability of the correct answer, it can also assign probabilities to all of the incorrect answers. The relative probabilities of incorrect answers tell us a lot about how the model tends to generalize. In other words, more generalized information is distilled with the supervision of dynamic classified probability.

As claimed by Luo [16], cross-entropy between predicted probability and ground truth implicitly expressed the distance under cosine space. Therefore, we term similarity of predicted classification probability (**KSL-P**) as knowledge consensus constraints in cosine space.

We adopted dynamic probability constraints in two manners, as shown in Figure 3: densely pairwise knowledge distillation (KSL-PP) for keeping diversity and distilling from virtual average branch calculating by other branches (KSL-PA) for filtering noise. The loss function is described as:

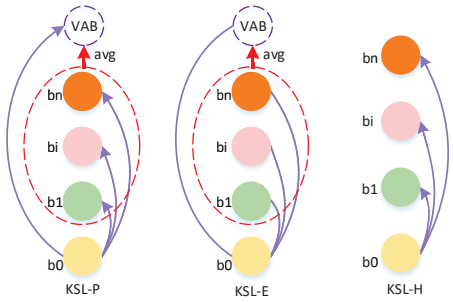

Figure 3: Manners of Knowledge Supervision (An example for branch0). b_i is the i_{th} branch, avg stands for average operation, VAB is virtual average branch, the arrow represents direction of imitation learning.

$$\begin{aligned}\mathcal{L}_{ksl\_p}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) &= \mathcal{L}_{ksl\_pp}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) \\ &+\mathcal{L}_{ksl\_pa}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})\end{aligned} \tag{3}$$

Here

$$\begin{aligned}\mathcal{L}_{ksl\_pp}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{\substack{(u,v)\in\mathbb{B}\\ u\neq v}}\sum_{k=1}^{K} \\ f_c^k(\boldsymbol{x}_n,\boldsymbol{\theta}^u)\log\left(\frac{f_c^k(\boldsymbol{x}_n,\boldsymbol{\theta}^u)}{f_c^k(\boldsymbol{x}_n,\boldsymbol{\theta}^v)}\right)\end{aligned} \tag{4}$$

where $\mathbb{B}$ represents all possible combinations of paired branches, u and v stand for the number of branch, $f_c(\boldsymbol{x}_n,\boldsymbol{\theta})$ represents the predicted classification probability.

$$\begin{aligned}\mathcal{L}_{ksl\_pa}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{b=1}^{B}\sum_{k=1}^{K} \\ f_c^k\left(\boldsymbol{x}_n,\boldsymbol{\theta}^{avg\_b}\right)\log\left(\frac{f_c^k\left(\boldsymbol{x}_n,\boldsymbol{\theta}^{avg\_b}\right)}{f_c^k\left(\boldsymbol{x}_n,\boldsymbol{\theta}^b\right)}\right)\end{aligned} \tag{5}$$

where

$$f_c^k\left(\boldsymbol{x}_n,\boldsymbol{\theta}^{avg\_b}\right) = \frac{1}{B-1}\sum_{j\neq b}^{B} f_c^k\left(\boldsymbol{x}_n,\boldsymbol{\theta}^j\right) \tag{6}$$

3.3 Knowledge Constraints of Embedding Distance

Embedding features are the low-dimensional representations of input images in embedding space (as shown in Figure 2). Different from image classification, the classifier in Re-ID only provides supervised signals in the training phase. In inference phase, embedding features are actually used to retrieve the same person in gallery. Therefore, we term the distance of corresponding embedding features in different branches (**KSL-E**) as knowledge consensus constraints in euclidean space.

We also adopted embedding distance constraints in two manners (KSL-EP and KSL-EA) as same as KSL-P, but with no direction. Because euclidean distance can be calculated symmetrically. The loss function is described as:

$$\begin{aligned}\mathcal{L}_{ksl\_e}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) &= \mathcal{L}_{ksl\_ep}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) \\ &+\mathcal{L}_{ksl\_ea}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta})\end{aligned} \tag{7}$$

Here

$$\begin{aligned}\mathcal{L}_{ksl\_ep}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{u=1}^{B-1}\sum_{v=u+1}^{B} \\ \left(\sum_{l=1}^{L}\left|f_e^l(\boldsymbol{x}_n,\boldsymbol{\theta}^u) - f_e^l(\boldsymbol{x}_n,\boldsymbol{\theta}^v)\right|^2\right)^{\frac{1}{2}}\end{aligned} \tag{8}$$

where $f_e(\boldsymbol{x}_n,\boldsymbol{\theta})$ represents the embedding features. L represents the dimension of the embedding features.

$$\begin{aligned}\mathcal{L}_{ksl\_ea}(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{b=1}^{B} \\ \left(\sum_{l=1}^{L}\left|f_e^l\left(\boldsymbol{x}_n,\boldsymbol{\theta}^{avg\_b}\right) - f_e^l\left(\boldsymbol{x}_n,\boldsymbol{\theta}^b\right)\right|^2\right)^{\frac{1}{2}}\end{aligned} \tag{9}$$

where

$$f_e^l\left(\boldsymbol{x}_n,\boldsymbol{\theta}^{avg\_b}\right) = \frac{1}{B-1}\sum_{j\neq b}^{B} f_e^l\left(\boldsymbol{x}_n,\boldsymbol{\theta}^j\right) \tag{10}$$

3.4 Knowledge Constraints of Hard Sample

Hard sample mining is significant due to the properties of metric learning. Meanwhile, KSL based on dynamic probability and embedding distance only consider the knowledge transfer between features pointing to the same identity (ID) in paired branches(as shown in Figure 2) but do not consider the interaction of different IDs in different branches. Inspired by triplet loss carrying out mining within single branch, we extend it to a multi-branch version to conduct mutual mining of hard samples between different branches. Therefore, we term hard sample mutual mining (**KSL-H**) as knowledge consensus constraints in hard sample space. The loss function is described as:

$$\mathcal{L}_{ksl\_h}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \frac{1}{N_t} \sum_{\substack{(m,n)\in\mathbb{N}_t \\ m\neq n}} \sum_{\substack{(u,v)\in\mathbb{B} \\ u\neq v}} \left[\max\left(d_{pos}\left(f_{em}^{u}, f_{en}^{v}\right)\right) - \min\left(d_{neg}\left(f_{em}^{u}, f_{en}^{v}\right)\right) + \gamma\right]_{+} \tag{11}$$

where $[\cdot]_+$ represents $\max(\cdot, 0)$, N_t represents the number of triplets in each batch, $\mathbb{N}_t$ represents all possible combinations of triplets in paired branches, m and n stands for the number of triplets, $\max(d_{pos}(\cdot))$ represents the maximum intra-class distance of anchor samples, $\min(d_{neg}(\cdot))$ represents the minimum inter-class distance of anchor samples, γ is the margin of triplet loss.

4 EXPERIMENTS

4.1 Evaluation on Person Re-Identification

We selected four mainstream challenging Re-ID datasets, Market-1501 [38], DukeMTMC-ReID [23], MSMT17 [34] and CUHK03 [10] to verify the proposed model. Market-1501 is the most commonly used person Re-ID dataset, which contains 32,668 images from 1,501 pedestrians. DukeMTMC-ReID dataset comes from video-based person tracking and re-identification tasks, containing 36411 images of 1,812 pedestrians. MSMT17 is a larger and more realistic Re-ID dataset published in 2018, containing 126411 pedestrian images with 4101 identities. CUHK03 consists of 14097 pedestrian images of 1467 identities, and the dataset contains two subsets that provide hand-labeled and DPM detected [4] bounding boxes, respectively. We evaluate our proposed model on DPM detected subset. For all the experiments, we use single query evaluation and simultaneously adopt both Rank-1 (R1) accuracy and the mean average precision (mAP) to evaluate KS performance.

4.2 Implementation Details

We trained proposed models on the Re-ID datasets by following the training strategy. For training from scratch, there is a trick that needs to be noticed. Because adding triple loss directly at the beginning of the training phase will hinder the convergence. In our experiment, the triplet loss would be added to the final loss at the 150^{th} epoch and switched to RandomIdentity Sampler at the same time. We used the stochastic gradient descent algorithm to optimize the model, and the epoch is set to 350. The learning rate is decayed using the cosine annealing strategy with the initialization value of 0.0015. In all experiments, the batch size is set to 64, and the weight decay is set to 5e-4. Images are resized to 256×128, and the corresponding data enhancement methods are adopted. In the verification stage, we extracted the 512-D embedding features from the fully-connected layer(as shown in Figure 2) and use the cosine distance for measurement. All experiments are conducted based on PyTorch, and we use NVIDIA V100 GPU to train the model.

4.3 Validity of Knowledge-Supervised Learning

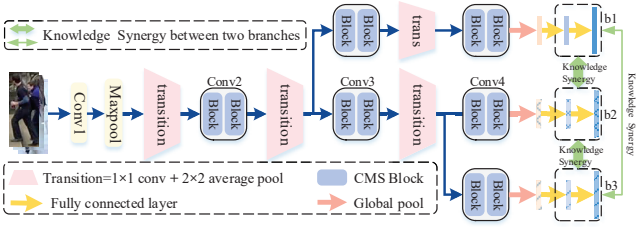

Figure 4: Network architecture of CMSNet (KSL-scratch). Isomorphic auxiliary branches are inserted at the input of Conv3 and Conv4, and the knowledge between branches is regularized to each other.

We applied the Knowledge-Supervised Learning (KSL) method in three advanced lightweight networks (CMSNet, ShuffleNet, and MobileNet). For CMSNet, two auxiliary branches are attached before the block Conv3 and Conv4. The auxiliary branches are constructed with the same building block as in the backbone network. The insertion positions are shown in Figure 4. Similarly, we introduced isomorphic branches before the 3rd and 4th stages of ShuffleNet and the 4th and 6th bottleneck block of MobileNet. Table 1 shows that all models with KSL method trained from scratch outperform their ImageNet pre-training results and significantly improve the accuracy compared to original models.

4.4 Comparison to State-of-the-art Methods

We compared the CMSNet (KSL-scratch) with advanced works that pre-training from the ImageNet. Table 2 show that KSL outperforms most large models even adopted to an extremely lightweight model (CMSNet) without ImageNet pre-training, both with or without RK optimization.

4.5 Ablation Study

As shown in Table 4, the Baseline model is CMSNet, which containing one branch and being trained from scratch. Firstly, each component is effective and boost accuracy. Secondly, AUX can improve performance. Moreover, with knowledge consensus constraints, the performance is further improved. Thirdly, from different perspectives, KSL-P, KSL-E, and KSL-H complement each other. Because combining them bring an extra gain.

As shown in Table 5 and Table 6, $\mathcal{L}_{pp}$, $\mathcal{L}_{pa}$, $\mathcal{L}_{ep}$, $\mathcal{L}_{ea}$ are the short of $\mathcal{L}_{ksl\_pp}$, $\mathcal{L}_{ksl\_pa}$, $\mathcal{L}_{ksl\_ep}$, $\mathcal{L}_{ksl\_ea}$, respectively. Each component is effective and boost accuracy. This indicate that KSL-PP/KSL-EP and KSL-PA/KSL-EA complement each other. Because KSL-PP/KSL-EP keeps the diversity and KSL-PA/KSL-EA filter noise.

Table 1: Validity of Knowledge-Supervised Learning. †: Cited from [41]. All models with KSL method trained from scratch outperform their ImageNet pre-training results and significantly improve the accuracy compared to original models.

Method	Venue	Params (M)	Duke		Market1501		CUHK03		MSMT17	
			R1	mAP	R1	mAP	R1	mAP	R1	mAP
CMSNet (scratch) [4]	ACM MM'20	2.1	86.6	74.3	94.3	84.7	61.2	58.8	70.1	47.6
CMSNet (pre-training) [4]	ACM MM'20	2.1	**90.0**	77.9	95.3	87.1	71.2	68.4	80.0	57.1
CMSNet(KSL-scratch)	***ours***	2.1	89.1	**79.2**	**96.0**	**89.1**	**76.8**	**73.7**	**81.6**	**62.1**
MobileNet(scratch)† [24]	ICCV'19	2.2	75.2	55.8	87.0	69.5	46.5	46.0	50.9	27.0
MobileNet(pre-training) [24]	ICCV'19	2.2	**80.9**	65.3	90.0	76.4	64.3	60.0	57.1	36.2
MobileNet(KSL-scratch)	***ours***	2.2	80.5	**66.4**	**91.8**	**80.3**	**64.9**	**62.4**	**57.4**	**36.5**
ShuffleNet(scratch)† [17]	ICCV'19	1.2	71.6	49.9	84.8	65.0	38.4	37.2	41.5	19.9
ShuffleNet(pre-training) [17]	ICCV'19	1.2	79.4	61.3	88.5	73.2	54.4	51.2	56.2	34.3
ShuffleNet(KSL-scratch)	***ours***	1.2	**80.6**	**67.3**	**91.5**	**80.1**	**62.3**	**59.8**	**56.8**	**38.1**

Table 2: Comparison to SOTA Methods. RK: kreciprocal re-ranking method [39]. KSL outperforms most large models even adopted to an extremely lightweight model (CMSNet) without ImageNet pre-training, both with or without RK optimization).

Methods Pre-training from ImageNet										
Method	Venue	Params (M)	Duke		Market1501		CUHK03		MSMT17	
			R1	mAP	R1	mAP	R1	mAP	R1	mAP
LUO [16]	TMM19	46.9	90.1	79.1	95.0	88.2	-	-	-	-
LEAP-AF [13]	CVPR'20	> 23.5	86.9	74.2	93.5	83.2	-	-	76.3	51.3
HOReID [32]	CVPR'20	> 23.5	86.9	75.5	94.2	84.9	-	-	-	-
RGA-SC [37]	CVPR'20	> 23.5	-	-	**96.1**	88.4	81.1	77.4	80.3	57.5
Dense121+M^3 [40]	CVPR'20	> 7.9	84.9	68.0	95.3	81.2	-	-	61.6	54.4
CMSNet [4]	ACM MM'20	2.1	90.0	77.9	95.3	87.1	71.2	68.4	80.0	57.1
CMSNet(KSL-scratch)	***ours***	**2.1**	89.1	**79.2**	96.0	89.1	76.8	73.7	81.6	**62.1**
CMSNet (KSL-scratch + RK)	***ours***	**2.1**	**91.4**	**89.4**	95.9	**94.4**	**82.8**	**84.1**	**83.9**	**75.7**
Trained from scratch										
HAN [11]	CVPR'18	4.5	80.5	63.8	91.2	75.7	41.7	38.6	-	-
Auto-ReID [22]	ICCV'19	13.1	-	-	90.7	74.6	-	-	-	-
OSNet [41]	ICCV'19	2.2	84.7	68.6	93.6	81.0	57.1	54.2	71.0	43.3
CMSNet [4]	ACM MM'20	2.1	86.6	74.3	94.3	84.7	61.2	58.8	70.1	47.6
CMSNet(KSL-scratch)	***ours***	**2.1**	**89.1**	**79.2**	**96.0**	**89.1**	**76.8**	**73.7**	**81.6**	**62.1**

4.6 Mechanisms of KSL method

To illustrate the mechanisms of Knowledge-Supervised Learning, we visualized the statistical weight distribution of each block in the trained CMSNet, as shown in Figure 5. The weight distribution of baseline model occurred a very large spike near zero in the bottom blocks due to the vanishing gradients. Small gradients make the weight decay part taking the major impact and cause near-zero "dead" weights. Figure 5 shows that each component of KSL method effectively helps reduce the number of "dead" weights.

Decreasing the "dead" weights means more neurons being activated, which easing the dependence on individual neurons, thus improve the generalization capability. Inherently, it benefits from the network structure of isomorphic auxiliary training: the weights of bottom layers are shared by more branches. Training multiple

Table 3: Fine-tuning from ImageNet pre-trained models. †: From OSNet [41] experimental results. ImageNet pre-training will further improve model performance compared to the KSL-scratch method.

Method	Params (M)	FLOPs (M)	Duke		Market1501		CUHK03		MSMT17	
			R1	mAP	R1	mAP	R1	mAP	R1	mAP
CMSNet(scratch) [4]	2.1	911.0	86.6	74.3	94.3	84.7	61.2	58.8	70.1	47.6
CMSNet(finetune) [4]	2.1	911.0	**90.0**	77.9	95.3	87.1	71.2	68.4	80.0	57.1
CMSNet(KSL-scratch)	2.1	911.0	89.1	**79.2**	**96.0**	**89.1**	**76.8**	**73.7**	**81.6**	**62.1**
CMSNet(KSL-pre-training)	2.1	911.0	89.8	**79.3**	**96.1**	**88.4**	**79.7**	**75.8**	**82.3**	**61.1**
MobileNet(scratch)† [24]	2.2	203.9	75.2	55.8	87.0	69.5	46.5	46.0	50.9	27.0
MobileNet(finetune) [24]	2.2	203.9	80.9	65.3	90.0	76.4	64.3	60.0	57.1	36.2
MobileNet(KSL-scratch)	2.2	203.9	**80.5**	**66.4**	**91.8**	**80.3**	**64.9**	**62.4**	**57.4**	**36.5**
MobileNet(KSL-pre-training)	2.2	203.9	**84.2**	**66.8**	**91.9**	**80.5**	**66.6**	**62.2**	**60.3**	**38.2**
ShuffleNet(scratch)† [17]	1.2	93.9	71.6	49.9	84.8	65.0	38.4	37.2	41.5	19.9
ShuffleNet(finetune) [17]	1.2	93.9	79.4	61.3	88.5	73.2	54.4	51.2	56.2	34.3
ShuffleNet(KSL-scratch)	1.2	93.9	**80.6**	**67.3**	**91.5**	**80.1**	**62.3**	**59.8**	**56.8**	**38.1**
ShuffleNet(KSL-pre-training)	1.2	93.9	**83.4**	**68.5**	**92.0**	**81.1**	**72.3**	**68.2**	**61.6**	**39.6**

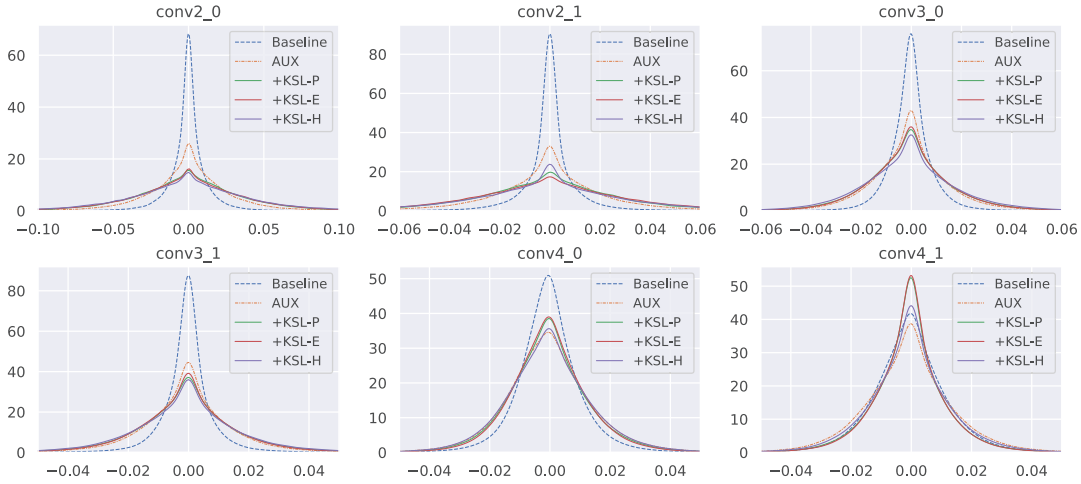

Figure 5: Per-block weight distribution in trained CMSNet (KSL-scratch). *conv* x_y **represents the** y^{th} **block in** x^{th} **stage. The two auxiliary branches are appended on top of** *conv*2_1 **and** *conv*3_1**. Models share the notations in Table 4.**

branches synchronously make the learned low-level features more generalized.

The weight distribution centralizes more to zero in the top blocks due to the more guaranteed low-level features (lines, edges, Etc.). Sparser high-level semantic features make the model behaving more similar to the process of object recognition in the human cerebral cortex [25].

4.7 Noisy Label Robustness

For Re-ID task, noisy labels are hard to completely avoid during model training either due to incorrect labels or data augmentation. For example, the most discriminative regions are cropped or erased. KSL method is, by nature, more robust to label noise. Because multiple views on the same sample have diversity of predictions. Consensus constraints will make multiple branches close to common features, especially the case of classification errors, which has a filtering effect. As shown in Figure 6, each component of KSL method effectively enhances the robustness to label noise.

Table 4: Ablation of KSL

Model	Loss				CUHK03	
	$\mathcal{L}_{aux}$	$\mathcal{L}_{ksl\_p}$	$\mathcal{L}_{ksl\_e}$	$\mathcal{L}_{ksl\_h}$	R1	mAP
Baseline					57.1	54.2
AUX	✓				63.8	61.1
+KSL-P	✓	✓			73.6	71.4
+KSL-E	✓	✓	✓		75.8	72.3
+KSL-H	✓	✓	✓	✓	**76.8**	**73.7**
KSL-E	✓		✓		73.9	71.6
KSL-H	✓			✓	66.8	63.6

Table 5: Ablation of KSL-P

Model	Loss		CUHK03	
	$\mathcal{L}_{pp}$	$\mathcal{L}_{pa}$	R1	mAP
AUX			63.8	61.1
KSL-PP	✓		68.0	64.7
KSL-PA		✓	72.7	70.2
KSL-P	✓	✓	**73.6**	**71.4**

Table 6: Ablation of KSL-E

Model	Loss		CUHK03	
	$\mathcal{L}_{ep}$	$\mathcal{L}_{ea}$	R1	mAP
AUX			63.8	61.1
KSL-EP	✓		72.6	70.8
KSL-EA		✓	72.4	70.4
KSL-E	✓	✓	**73.9**	**71.6**

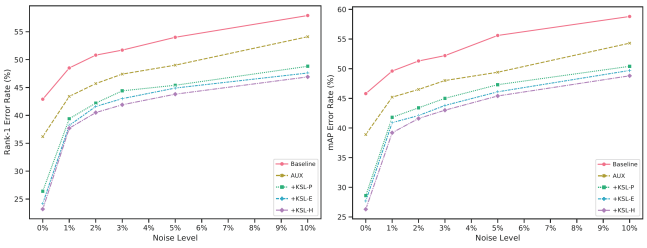

Figure 6: Test error on CUHK03 with label noise. Noisy labels are randomly generated in each epoch. Noise Level is the ratio of noisy labels over the all training set.

4.8 Qualitative Analysis of Generalization Capability

Generalization means being capable of dealing with complex changes with little or no cost. To qualitatively analyze the generalization capability of KSL method, we simulate the complex changes through data augmentation (random flip, random crop, random patch, random erase, Etc.) at inference phase. As shown in Figure 7, with regard to the feature distribution in the embedding space, KSL

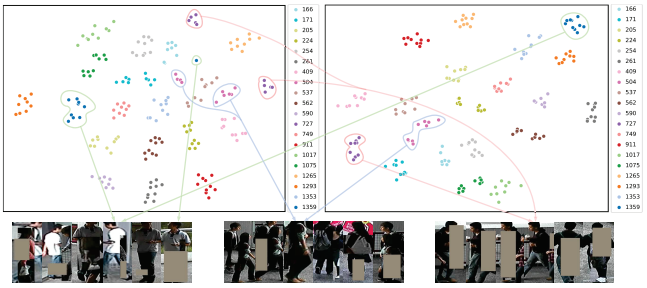

Figure 7: t-SNE visualization of the feature representation on CUHK03 (20 identities randomly sampled from the test set, each image is augmented by a composed of random flip/crop/patch/erase). The left and right figures are the results of Baseline and +KSL-H model in Table 4, respectively. Obviously, KSL method is more generalized to deal with complex image changes simulated by data augmentations.

method is more compact than baseline model for all identities and the gap between identities is larger, especially for difficult samples (id=504,727,1359). It is obvious that the KSL method can improve the generalization capability of the model.

4.9 Further Performance Improvement through ImageNet Pre-training

The results of KSL method fine-tuning from the ImageNet are not mentioned in above section. We intentionally do this to highlight the performance of the KSL method training from scratch. So that people can focus more on "training from scratch + knowledge discovery" instead of the traditional "pre-training and fine-tuning" paradigm. But there is no doubt that pre-training will further improve model accuracy.

We adopted KSL method to CMSNet, MobileNet and ShuffleNet. The experimental results of fine-tuning from ImageNet are as shown in Table 3. Firstly, ImageNet pre-training will further improve model performance compared to the KSL-scratch method, yet the gap is minimal. This implicit indicates that KSL method can near entirely release the potential of a network. Secondly, the performance gap of MobileNet/ShuffleNet between ImageNet pre-training and training from scratch is larger than CMSNet. The core insight of KSL method is the consensus of multiple views on the same data. We conjecture that the view of data in MobileNet/ShuffleNet contains more noise due to the lower accuracy. Providing initialization weights with higher accuracy helps to achieve a more impressive performance.

5 CONCLUSION

In this paper, we proposed a Knowledge-Supervised Learning method that can improve the performance with no extra inference cost. Experiments show that a standard model with KSL method trained from scratch outperforms its ImageNet pre-training results by a clear margin. In the future, we will do further research to investigate the potential of Knowledge-Supervised Learning in other visual recognition tasks.

REFERENCES

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. 2019. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9163–9171.

[2] Leulseged Tesfaye Alemu, Marcello Pelillo, and Mubarak Shah. 2019. Deep constrained dominant sets for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 9855–9864.

[3] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang. 2020. Salience-Guided Cascaded Suppression Network for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3300–3310.

[4] Baoyu Fan, Li Wang, Runze Zhang, Zhenhua Guo, Yaqian Zhao, Rengang Li, and Weifeng Gong. 2020. Contextual Multi-Scale Feature Learning for Person Re-Identification. In *Proceedings of the 28th ACM International Conference on Multimedia*. 655–663.

[5] Hrayr Harutyunyan, Kyle Reing, Greg Ver Steeg, and Aram Galstyan. 2020. Improving Generalization by Controlling Label-Noise Information in Neural Network Weights. *arXiv preprint arXiv:2002.07933* (2020).

[6] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[7] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. 2019. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9317–9326.

[8] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q Weinberger. 2017. Multi-scale dense networks for resource efficient image classification. *arXiv preprint arXiv:1703.09844* (2017).

[9] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. 2015. Deeply-supervised nets. In *Artificial intelligence and statistics*. 562–570.

[10] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. 2014. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 152–159.

[11] Wei Li, Xiatian Zhu, and Shaogang Gong. 2018. Harmonious attention network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2285–2294.

[12] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. 2017. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing* 26, 7 (2017), 3492–3506.

[13] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. 2020. Deep Representation Learning on Long-tailed Data: A Learnable Embedding Augmentation Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2970–2979.

[14] Shikun Liu, Edward Johns, and Andrew J Davison. 2019. End-to-end multi-task learning with attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1871–1880.

[15] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. 2020. Multi-task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10034–10043.

[16] Hao Luo, Wei Jiang, Youzhi Gu, Fuxu Liu, Xingyu Liao, Shenqi Lai, and Jianyang Gu. 2019. A Strong Baseline and Batch Normalization Neck for Deep Person Re-identification. *IEEE Transactions on Multimedia* (2019).

[17] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. 2018. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 116–131.

[18] Paul Marchwica, Michael Jamieson, and Parthipan Siva. 2018. An evaluation of deep cnn baselines for scene-independent person re-identification. In *2018 15th Conference on Computer and Robot Vision (CRV)*. IEEE, 297–304.

[19] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 5191–5198.

[20] Duy-Kien Nguyen and Takayuki Okatani. 2019. Multi-task learning of hierarchical vision-language representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10492–10501.

[21] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3967–3976.

[22] Ruijie Quan, Xuanyi Dong, Yu Wu, Linchao Zhu, and Yi Yang. 2019. Auto-ReID: Searching for a part-aware ConvNet for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*. 3750–3759.

[23] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. 2016. Performance measures and a data set for multi-target, multi-camera tracking. In *European Conference on Computer Vision*. Springer, 17–35.

[24] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4510–4520.

[25] Thomas Serre, Aude Oliva, and Tomaso Poggio. 2007. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences* 104, 15 (2007), 6424–6429.

[26] Shuai Shao, Lei Xing, Wei Yu, Rui Xu, Yan-Jiang Wang, and Bao-Di Liu. 2021. SSDL: Self-Supervised Dictionary Learning. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[27] Shuai Shao, Rui Xu, Weifeng Liu, Bao-Di Liu, and Yan-Jiang Wang. 2020. Label embedded dictionary learning for image classification. *Neurocomputing* 385 (2020), 122–131.

[28] Guocong Song and Wei Chai. 2018. Collaborative learning for deep neural networks. In *Advances in Neural Information Processing Systems*. 1832–1841.

[29] Dawei Sun, Anbang Yao, Aojun Zhou, and Hao Zhao. 2019. Deeply-supervised Knowledge Synergy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6997–7006.

[30] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*. 480–496.

[31] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[32] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. 2020. High-Order Information Matters: Learning Relation and Topology for Occluded Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6449–6458.

[33] Li Wang, Baoyu Fan, Zhenhua Guo, Yaqian Zhao, Runze Zhang, Rengang Li, and Weifeng Gong. 2020. Dense-Scale Feature Learning in Person Re-Identification. In *Proceedings of the Asian Conference on Computer Vision*.

[34] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 79–88.

[35] Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, and Jian-Huang Lai. 2019. Distilled person re-identification: Towards a more scalable system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1187–1196.

[36] Wenjie Yang, Houjing Huang, Zhang Zhang, Xiaotang Chen, Kaiqi Huang, and Shu Zhang. 2019. Towards rich feature discovery with class activation maps augmentation for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1389–1398.

[37] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. 2020. Relation-Aware Global Attention for Person Re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3186–3195.

[38] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*. 1116–1124.

[39] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li. 2017. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1318–1327.

[40] Jiahuan Zhou, Bing Su, and Ying Wu. 2020. Online Joint Multi-Metric Adaptation from Frequent Sharing-Subset Mining for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2909–2918.

[41] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. 2019. Omni-Scale Feature Learning for Person Re-Identification. In *The IEEE International Conference on Computer Vision (ICCV)*.