Horizontal Flipping Assisted Disentangled Feature Learning for Semi-Supervised Person Re-Identification

Gehan Hao[1], Yang Yang[2], Xue Zhou[1,3,*], Guanan Wang[2], and Zhen Lei[2]

[1] School of Automation Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China
[2] National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[3] Shenzhen Institute of Advanced Study, UESTC, Shenzhen, China
*Corresponding author:zhouxue@uestc.edu.cn

Abstract. In this paper, we propose to learn a powerful Re-ID model by using less labeled data together with lots of unlabeled data, *i.e.* semi-supervised Re-ID. Such kind of learning enables Re-ID model to be more generalizable and scalable to real-world scenes. Specifically, we design a two-stream encoder-decoder-based structure with shared modules and parameters. For the encoder module, we take the original person image with its horizontal mirror image as a pair of inputs and encode deep features with identity and structural information properly disentangled. Then different combinations of disentangling features are used to reconstruct images in the decoder module. In addition to the commonly used constraints from identity consistency and image reconstruction consistency for loss function definition, we design a novel loss function of enforcing consistent transformation constraints on disentangled features. It is free of labels, and can be applied to both supervised and unsupervised learning branches in our model. Extensive results on four Re-ID datasets demonstrate that by reducing 5/6 labeled data, Our method achieves the best performance on Market-1501 and CUHK03, and comparable accuracy on DukeMTMC-reID and MSMT17.

1 Introduction

Person Re-Identification (Re-ID) aims to automatically match the underlying identities of person images from non-overlapping camera views [1]. As an essential task in video surveillance of distributed multi-cameras, Re-ID is very important for individual-specific long-term behavior analysis. Due to variations of view angles, poses and illuminations in different cameras, it's very challenging to tackle this task.

In person Re-ID community, many models have been proposed, which mainly focus on three parts: hand-crafted descriptor design, metric learning and deep Re-ID models. Hand-crafted person descriptors [2–4] try to design features that are robust to different view angles, poses, and illuminations. Metric learning

[5, 6] aims to learn a feature projected space wherein the similarity of same person is higher than that of different person. With successful application of deep Convolution Neural Network (CNN), deep Re-ID [7–13], models are able to straightly learn robust and discriminative features in a compact end-to-end manner, which have gained more and more attention.

Based on whether identity labels are used, deep Re-ID models can be roughly divided into supervised and unsupervised ones. The former trains models with a supervised loss such as classification loss [10] and triplet loss [14]. The latter solves unsupervised Re-ID with cross-dataset domain-adaptation learning [15–17], tracklet information [18], or clustering-based pseudo-labels learning [19]. Although the above two kinds of Re-ID methods have achieved promising progress, they still suffer from their inherent weaknesses. Supervised Re-ID methods require massive cross-camera identity-labels to avoid over-fitting a training set. Obtaining such labels can be very time-consuming and expensive. In unsupervised Re-ID methods, domain-adaptation learning also needs source labeled data, tracklet-based methods rely on accurate tracking results, and pseudo-labels are sensitive to initial parameters. Besides, existing unsupervised Re-ID methods are still far from supervised ones in terms of accuracy.

In this paper, we formulate Re-ID problem in a semi-supervised way by leveraging a few labeled and lots of unlabeled data. Our semi-supervised Re-ID method enjoys the following two merits. 1) Compared with pure supervised Re-ID methods, our method requires less labeled data. Besides, auxiliary by mega unlabeled data, it avoids over-fitting on the training set. 2) Compared with pure unsupervised Re-ID methods, ours can achieve better accuracy by exhaustively exploring the limited labeled data.

In order to learn more robust and discriminative deep global features for Re-ID task, disentangled feature learning (DFL) is introduced in our work. Usually, DFL requires multi-inputs to disentangle features with different semantics. In our work, we found that a pair of horizontally flipped images changed the person structural information while maintaining the identity and attribute characteristics unchanged, and showed a symmetrical distribution. Then, we simply take the original person image with its horizontal mirror image as two inputs of DFL module, which does not need extra complicated operations and costs like other methods [20, 21, 11, 22, 23]. DFL module is designed in an encoder-decoder way to disentangle identity-aware features and structure-aware features, which has been applied on the original image and its horizontal mirror image respectively. With respect to the four disentangled features, two identity-aware features from the original image and its mirror image should be the same, meanwhile, two structure-aware features should satisfy mirror symmetry. However, the above invariance and equivariance constraints are often missing in the normal Re-ID training process, which only considers image-level identity labels. After that, four different combinations of disentangling features are used to reconstruct images in the decoder module.

Our main contributions can be summarised as below:

(1) We propose a novel semi-supervised Re-ID framework, which consists of two branches with shared feature disentanglement models, one for supervised task and the other for unsupervised task. It alleviates limitation of labeled data by exploiting lots of unlabeled data.

(2) We exploit unsupervised data by disentangling images and its horizontal flipping images into structure-aware and identity-aware features in a self-supervised way. A consistent transformation constrained loss function including identity invariance and structure equivariance is defined on disentangled features, which is free of labels.

(3) Extensive results on four Re-ID datasets demonstrate that by reducing 5/6 labeled data, Our method achieves the best performance on Market-1501 and CUHK03, and comparable accuracy on DukeMTMC-reID and MSMT17.

2 Related Work

2.1 Semi-supervised Person Re-identification

There are a few early semi-supervised work on person Re-ID. Figueira et al. [24] propose a method that combines multiple semi-supervised feature learning frameworks to deal jointly with the appearance-based and learning-based Re-ID problem. Liu et al. [25] propose a semi-supervised coupled dictionary learning method, which jointly learns two coupled dictionaries in the training phase from both labeled and unlabeled images. However, these non-deep-learning methods can only achieve good results on small-scale datasets.

In recent years, with the development of deep CNN, some deep semi-supervised person Re-ID methods have been presented. The first semi-supervised approach [26] that performs pseudo-labeling by considering complex relationships between unlabeled and labeled training samples in the feature space. They adopt a generative adversarial network to generate additional artificial sample data as unlabeled data. Huang et al. [27] introduce multi-pseudo regularized labels and distribute them to the generated data to supplement the real training data in a semi-supervised manner. Liu et al. [22] design a simple but effective learning mechanism that merely substitutes the last fully-connected layer with the proposed Transductive Centroid Projection (TCP) module. Fan et al. [28] propose a simple and progressive unsupervised deep learning framework, whose purpose is to use k-means clustering to estimate the labels of unlabeled training samples, and extend it to semi-supervised Re-ID. Xin et al. [29] propose a semi-supervised method that combines multi-view clustering and deep metric learning to repeatedly update the pseudo-labels of unlabeled training samples.

Different from the above methods, our method does not follow the idea of pseudo labelling and clustering, but construct an encoder-decoder feature disentanglement framework which can be learned not relying on the labels.

2.2 DFL-based Person Re-identification

In recent years, disentangled feature learning(DFL)-based person Re-ID has gained more and more attention [22, 20, 11, 23, 30–34]. DFL is expected to pro-

vide gains by separating the underlying structure of data into uncorrelated meaningful variables, which helps determine what types of hidden features are actually learned [35]. Current DFL-based Re-ID methods usually adopt GAN or auto-encoder model to separate different attributes (i.e., appearance or pose, etc.) from multi-inputs of a person. Among them, posture is the most considered attribute. Ma et al. [22] use a complex multi-branch model to decompose the input person image into the foreground, background, and pose to generate a specific image, but the model cannot be trained end-to-end. Qian et al. [20] generate a normalized pose image for each person, but there are only 8 predefined poses. Ge et al. [11] also guide the network to learn pose-invariant features, but utilizing human key points to describe pose features, which is time-consuming. Based on previous work, Li et al. [23] not only extract pose features, but also use additional key features of human body.

Therefore, most of the current work can be summarized as follows: 1) Additional annotations are used, such as human keypoint features. They define characteristics of human posture information as constraints to guide the network to learn identity-invariant features. 2) Requiring person samples with different postures for learning identity-invariant features. However, both methods have their shortcomings. The first requires the introduction of additional annotations, which increases the complexity of the network. Samples that meet the conditions of second method are difficult to find. Either you need to select samples with different poses or using GAN to generate these multi-pose samples. Even if you find these kinds of samples, different posture images caused by different perspectives will bring confusion in attributes, resulting ambiguity in identity. For example, the chaos of carrying school bag due to changes in camera view, or the chaos of long hair because the person turns around.

In order to avoid the above disadvantages and make full use of existing data, we simply horizontally flip the original image without introducing extra annotations or complicated GAN model. The horizontal mirror image implicitly reflects structural information and enjoys several merits: identity and attributes invariance, and structural symmetrical equivariance.

3 Our Approach

In this section, we firstly describe the overall architecture of our network. Then, we introduce Disentangled Feature Learning for semi-supervised Re-ID task, followed by loss functions explanation.

3.1 Overall Framework

The overall architecture of the proposed framework is shown in Fig.1. Our semi-supervised framework consists of two branches: a supervised branch and an unsupervised branch. For each branch, we design an Encoder-Decoder network to realize feature disentanglement and reconstruction. We take a pair of original image I_O and its horizontal mirror image I_T along with the label Y as three

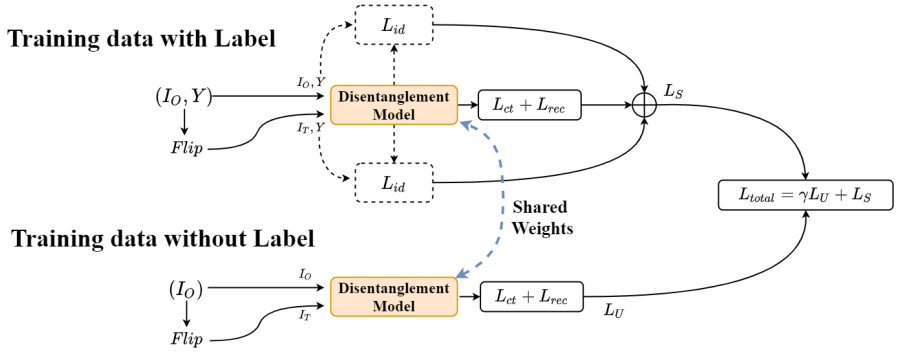

Fig. 1. Overview of our framework. Our Semi-supervised framework consists of two branches with shared feature disentanglement models, one for supervised task and the other for unsupervised task. The labeled and unlabeled data are simultaneously adopted to train the whole framework.

inputs for supervised branch, and omit the label for unsupervised branch. The final loss L_{Total} equals to the weighted summation of supervised branch loss L_S and unsupervised branch loss L_U. Constraints on label consistency L_{id}, image reconstruction L_{rec}, and consistent transformations on disentangled features L_{ct} are considered for designing supervised loss L_S. While only L_{ct} and L_{rec} are considered in unsupervised loss L_U. The detailed description about feature disentanglement model and loss functions design are referred to next following subsections. Due to sharing parameters and training as a whole, under the strong label guidance learning in the supervised branch, the unsupervised branch can effectively make full use of a large amount of unlabeled data.

3.2 Disentangled Feature Learning

For person Re-ID tasks, it is very important to mine person identity information with different structural information under different views. We hope to guide the network to learn how to disentangle the mixed global features into independent structure-aware features and identity-aware features. Previously, some methods build pairs of images which have same identity but different structures, effectively disentangling features through Siam network. However, for unlabeled data, we cannot find samples that have same identity but different structures. Here we are inspired by data augmentation, and can obtain new mirror structural samples through flipping and horizontal displacement operations. Mirror samples meet our requirements for a pair of person samples: 1) the same identity 2) different structure.

Thus, we design an encoder-decoder-based feature disentanglement network which requires a pair of inputs, shown in Fig.2(a). DenseNet-121 [36] pretrained on ImageNet [37] is chosen as our auto-encoder backbone by removing the final

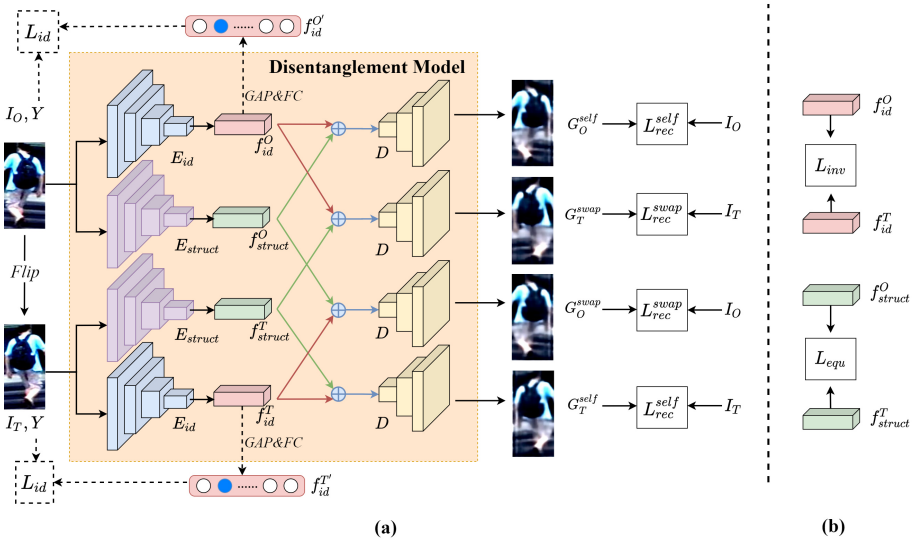

Fig. 2. Disentangled feature learning network for Re-ID task. (a) Encoder-decoder-based backbone to disentangle identity-aware features from structure-aware features. Both self-reconstruction and swap-reconstruction are considered in the reconstruction loss function. $G_{\{O,T\}}^{\{self,swap\}}$ denotes the generated reconstruction image after decoder module in each path, corresponding to operation $D(\cdot,\cdot)$ in Eqn.(2) and Eqn.(3). The dotted lines indicate that these parts only exist in supervised branch. (b) Illustration of consistent transformation constraints on disentangled features.

pooling and fully-connected layers. Please note that, although most existing Re-ID methods take ResNet-50 as CNN backbone, we choose a smaller DenseNet-121 compared with ResNet-50 ($8M$ params *vs.* $25.5M$ params). Taking the original image input I_O as an example, two auto-encoders (E_{id} and E_{struct}) with the same structures but different parameters are respectively applied to encode identity-aware features f_{id}^O and structure-aware features f_{struct}^O. We define a Horizontal Flipping Transformation $T(\cdot)$, which is used to generate the horizontal flipped image $I_T = T(I_O)$. By analogy, f_{id}^T and f_{struct}^T can be also decomposed from the horizontal flipped image I_T. The superscript O and T denote the original image and its horizontal flipping image, respectively. Then, we concatenate two disentangled features with different semantics, resulting four different combinations followed by a decoder network D to reconstruct images. Decoder D consists of 5 transposed convolutional layers followed by batch normalization [38], leaky ReLU [39] and dropout [40].

In order to guarantee that the disentangled features encoding semantic information, the reconstructed images should satisfy the following criteria: 1) Self-reconstruction. if both identity-aware features and structure-aware features are decomposed from the same image, i.e., (f_{id}^O and f_{struct}^O) or (f_{id}^T and f_{struct}^T), the

reconstructions are certainly similar with themselves corresponding input images, i.e., I_O or I_T; 2) Swap-reconstruction. if the decomposed identity-aware features and structure-aware features are from different input images, i.e., (f_{id}^O and f_{struct}^T) or (f_{id}^T and f_{struct}^O), the reconstructions are consistent with the image from which the structure-aware features are disentangled, i.e., I_T or I_O.

Therefore, by obeying the above criteria, we define the final reconstruction loss function L_{rec} composed of two kinds of reconstructions:

$$L_{rec} = L_{rec}^{self} + L_{rec}^{swap} \tag{1}$$

The first item L_{rec}^{self} follows the first criteria, each reconstruction is similar to itself, i,e.,

$$L_{rec}^{self} = ||I_O - D(f_{id}^O, f_{struct}^O)||_2 + ||I_T - D(f_{id}^T, f_{struct}^T)||_2 \tag{2}$$

where $D(\cdot,\cdot)$ denotes the reconstructed image by concatenating two decomposed features, $||\cdot||_2$ is the pixel-wise L_2 loss. The second item L_{rec}^{swap} follows the second criteria. Disentangled identity-aware features and structure-aware features are expected to be independent of each other. Obviously, identity features do not change after flipping the image, and the reconstructed image is determined by the structure-aware features. Thus, the second reconstruction loss can be defined as follows:

$$L_{rec}^{swap} = ||I_O - D(f_{id}^T, f_{struct}^O)||_2 + ||I_T - D(f_{id}^O, f_{struct}^T)||_2 \tag{3}$$

3.3 Consistent Transformation Constraints

Traditional supervised Re-ID frameworks [22, 23, 21] are trained under the identity label guidance to encode the global person features. Different from them, in our work we adopt a disentangled feature learning framework to decompose semantic mixed features into independent features with different characteristic.

As described in subsection 3.2, with respect to a pair of image and its horizontal mirror inputs, four disentangled features are obtained. Among them, two are identity-aware features, the other two are structure-aware features. Because horizontal flipping will not change the person identity, these two disentangled identity-aware features should satisfy invariant properties. At the same time, the two structure-aware features accordingly presents equivariant transformation as two images, i.e.,the output feature maps of flipped images are also flipped to ensure the consistency of structure features. Fig.3 is an illustration of these constraints. The left part in Fig.3 displays identity invariance constraint, therein the generated two identity-aware features f_{id}^O and f_{id}^T should maintain invariant. The right part in Fig.3 displays structure equivariance constraint, therein two structure-aware features f_{struct}^O and f_{struct}^T should maintain horizontal symmetry.

Therefore, following the above ideas we respectively design the identity invariance transformation loss function L_{inv} and structure equivariance transformation loss function L_{equ} as:

$$L_{inv} = D_{KL}(f_{id}^O || f_{id}^T) \tag{4}$$

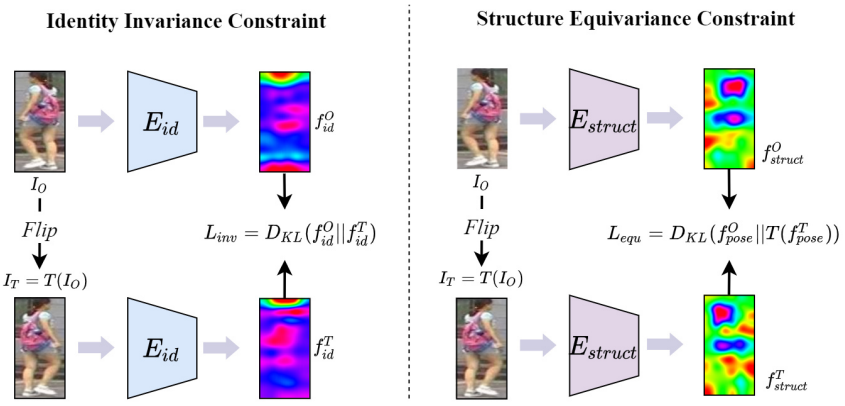

Fig. 3. Illustration of consistent transformation constraints on disentangled features. **Identity Invariance Constraint (left)** The identity information does not change as the image is horizontally flipped. **Structure Equivariance Constraint (right)** The structure information is flipped equivariantly as the image is horizontally flipped. Grad-CAM [41] is adopted for feature maps visualization.

$$L_{equ} = D_{KL}(f^O_{struct}||T(f^T_{struct})) \tag{5}$$

where $D_{KL}(\cdot)$ is the Kullback–Leibler divergence distance. With respect to KL divergence applied on feature maps $(f^O_{id}, f^T_{id}, f^O_{struct}, f^T_{struct})$, we firstly apply Softmax operation along channel dimension to enforce a discrete distribution at each element location, then element-wise KL divergence loss is applied between pre-transform and after-transform feature maps.

The overall loss function under these two consistent transformation constraints could be defined as:

$$L_{ct} = L_{inv} + L_{equ} \tag{6}$$

These two constraints reflect the intrinsic correlation among disentangled features, which guarantees the disentangled feature learning well conducted. This loss function is free of labels, can be applied both supervised learning and unsupervised learning.

3.4 Semi-Supervised Training and Testing

Annotating person Re-ID datasets is a very time-consuming task. We introduce a semi-supervised method to train Re-ID model using less labeled data by making full use of unlabeled data. Our model consists of a supervised branch and an unsupervised branch, where two branches share the same parameters.

For the unsupervised branch, the disentangled features get the consistent transformation constrained loss L_{ct}, and then different combinations of features are concatenated into the decoder to reconstruct images under the reconstruction

loss L_{rec}. In this case, the unsupervised training loss L_U is defined as:

$$L_U = \alpha L_{ct} + \beta L_{rec} \tag{7}$$

where α and β control the relative importance of the corresponding objectives. We empirically set $\alpha = 5$, β=0.3 in our experiments.

For the supervised branch, in addition to the unsupervised loss mentioned above, we use the identity label as a strong supervised signal to guide our model to disentangle identity-aware and structure-aware features more effectively. Here we use cross-entropy loss function L_{id} applied on the two identity feature vectors ($f_{id}^{O'}$ and $f_{id}^{T'}$), which are generated by GAP&FC operation based on the disentangled feature maps (f_{id}^{O} and f_{id}^{T}). As shown in Fig.1 and Fig.2, L_{id} denoted by dashed line is only valid for supervised branch. In this case, the supervised training loss L_S is defined as:

$$L_S = \alpha L_{ct} + \beta L_{rec} + L_{id} \tag{8}$$

In our training process, supervised branch and unsupervised branch are trained as a whole. We define the overall loss function L_{total} as follows:

$$L_{total} = \gamma L_U + L_S \tag{9}$$

where γ is the weighting parameter. The training loss L_{total} is used to optimize the whole network. The unsupervised branch is trained under the guidance of supervised branch to make the feature disentangling be more successful and effective.

During testing, we input each test image in conjunction with its horizontal flipping image into ID encoder model (E_{id} in Fig.2) respectively, and take the mean of two disentangled id-aware feature vectors ($f_{id}^{O'}$ and $f_{id}^{T'}$) as the final global feature vector. Cosine similarity is used for matching with gallery images.

4 Experiments

4.1 Experimental Configurations

We evaluate our proposed method on 4 datasets (Market-1501 [3], DukeMTMC-reID [13], CUHK03 [42] and MSMT17 [43]) under both semi-supervised and fully-supervised settings. When using the semi-supervised setting, we split training set into the labeled and unlabeled data according to identities. Under the fully-supervised setting, we view all images in training set as labeled ones. Cumulative match curve (CMC) and mean average precision (mAP) are used as evaluation protocols. The detailed description about 4 datasets are given as follows. **Market-1501** [3] consists of 32,668 labeled images of 1,501 identities captured by 6 cameras; wherein 12,936 images are for training and 19,732 images are for testing. **DukeMTMC-reID** [13] collects from 8 cameras and is comprised of 36,411 labeled images of 1,404 identities. Especially, 702 identities are for training and the others are for testing. **CUHK03** [42] contains 14,096

Table 1. Comparison with five state-of-the-art Re-ID methods when labeled data *ratio* is set to 1/6. Our method achieves the best performance on Market-1501 and CUHK03, and comparable accuracy on DukeMTMC-reID and MSMT17.

Methods	Market-1501		DukeMTMC-reID		CUHK03		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
IDE (arXiv) [10]	30.4	18.5	40.1	21.5	11.7	10.5	13.7	6.4
MGN (MM'2018) [45]	75.4	52.0	**69.1**	50.1	19.5	20.4	55.3	21.6
PCB (ECCV'2018) [7]	74.1	48.2	68.4	45.8	23.2	21.4	23.2	12.4
BoT (CVPRW'2019) [9]	65.6	42.3	60.5	41.0	16.5	16.4	34.6	14.5
ABD-Net (ICCV'2019) [8]	68.0	48.1	68.0	48.2	26.0	25.2	**45.4**	21.0
Ours	**77.8**	**54.5**	69.0	**50.5**	**32.9**	**29.8**	44.5	**29.3**

images of 1,467 identities, captured by 6 camera views. Among these identities, 767 identities with a total of 7368 images are used for training and 700 identities with a total of 6728 images are for testing. **MSMT17** [43] has 126,441 images of 4,101 identities captured by a 15-camera network (12 outdoor, 3 indoor). This large dataset is closer to the real scene in terms of environment diversity and amount of identities. 1041 identities with 32621 samples are for training and 3060 identities with 93820 samples are for testing.

During training, the input images are resized to 256×128 augmented by random erasing operation [44]. The batch size is set 96. Considering that the encoders are pre-trained on ImageNet, its initial learning rate is 0.01, which is smaller than the two fully connected layers and decoder, whose learning rates are 0.1. The learning rates are decayed to its $0.1\times$ and $0.01\times$ at 70th and 80th epochs, and end at 90 epochs.

4.2 Comparison with State-of-the-Art under Semi-Supervised Setting

We denote the proportion of labeled data in the training set as *ratio* and the rest are used as unlabeled data. We evaluate the effectiveness of our approach with different settings of *ratio*. For example, in the Market-1501 dataset, there are 751 pedestrians in the training set. If we define *ratio* is 1/3, we only select 250 identities as the labeled data, and the remaining images of 501 identities as the unlabeled data.

We compare ours with 5 state-of-the-art Re-ID methods, including IDE [10], MGN [45], PCB [7], BoT [9] and ABD-Net [8]. Multiple experiments are conducted on different datasets by setting *ratio* to 1/3, 1/6 and 1/12. Quantitative comparison results are shown in Fig.4. We found that the lower the percentage of labeled data, the better our method worked. When the proportion of labeled data is 1/12, our method has got Rank-1 scores increased by 51.9%, 16.7%, 6.0%, 3.7% and 1.8%, and has got mAP increased by 31.9%, 12.9%, 5.2%, 4.8% and 5.1%, compared with IDE, BoT, MGN, PCB and ABD-Net methods. Among them, MGN and PCB methods extract stripe-level features of the target, IDE

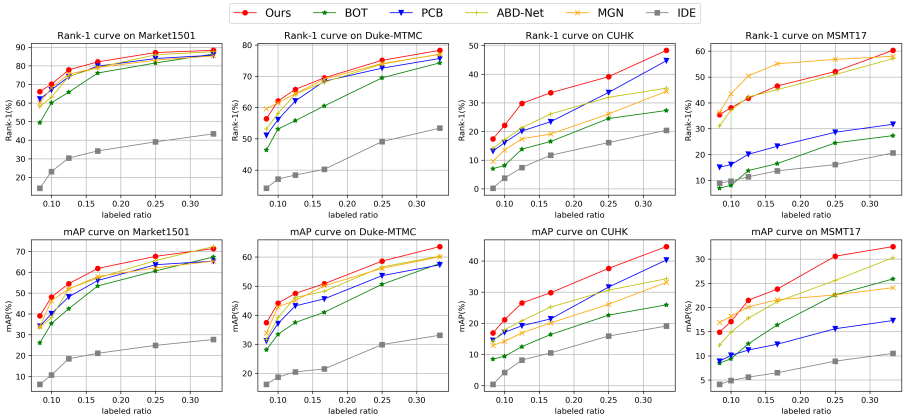

Fig. 4. Semi-supervised quantitative comparison results of six methods on Market-1501, DukeMTMC-reID, CUHK03, MSMT17 under different *ratio* values.

and BoT directly use global features, and ABD-Net extracts features through attention mechanism based on channel dimension and spatial dimension.

We especially show the experimental results by setting *ratio* to 1/6 in Table 1. For example, on Market-1501, the global feature-based methods IDE and BoT have poor results. Their Rank-1 are 30.4% and 65.6%, respectively. Global features learned in Re-ID task have limited capability of extracting effective and discriminative features. Stripe-based methods perform well on Market-1501 and DukeMTMC-reID. For example, MGN achieves 75.4%, 69.1% Rank-1 scores and 52.0%, 50.1% mAP on Market-1501 and DukeMTMC-reID datasets. This shows that on small datasets, local features help improve performance. Attention-based ABD-Net perform well in CUHK03, with 26.0% Rank-1 and 25.2% mAP scores, respectively. This also shows that the attention-based method can effectively mine deeper features. Through feature disentangling and consistent transformation constraints, our method only considering global features achieves the best accuracy on Market-1501 and CUHK03, and comparable accuracy on DukeMTMC-reID and MSMT17, which demonstrates that our proposed method is capable of effectively extracting more robust features.

4.3 Comparison with State-of-the-Art under Supervised Setting

In this section, fully supervised setting is applied. The unsupervised branch is invalid in our method. We report the performance comparisons of ours and 11 state-of-the-art Re-ID models including hand-crafted methods [3], attention-based methods [8], stripe-based methods [45, 7], global feature methods [10, 9, 47], and GAN-based methods [46, 20, 21, 11]. The quantitative comparison results are shown in Table 2.

As we can see, the hand-crafted features has got the worst accuracy on all four datasets. For example, on Market-1501, its Rank-1 is 44.4%, much lower

Table 2. Comparison with state-of-the art Re-ID methods under supervised setting, *i.e.* using all training images as labeled ones. Unsupervised branch is invalid in our method.

Methods	Market-1501		DukeMTMC-reID		CUHK03		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
BoW (ICCV'2015) [3]	44.4	20.8	25.1	12.2	6.4	6.4	-	-
IDE (arXiv) [10]	72.5	46.0	65.2	44.9	-	-	-	-
Cam-GAN (CVPR'2018) [46]	89.5	71.5	78.3	57.6	-	-	-	-
Pose-Normalized (CVPR'2018) [20]	89.4	72.6	73.6	53.2	-	-	-	-
MGN (MM'2018) [45]	**95.7**	86.9	88.7	78.4	66.8	**66.0**	-	-
PCB (ECCV'2018) [7]	92.3	73.3	81.7	66.1	63.7	57.5	68.2	40.4
DG-Net (CVPR'2019) [21]	94.8	86.0	86.6	74.8	65.6	61.1	77.2	52.3
BoT (CVPRW'2019) [9]	94.5	85.9	86.4	76.4	-	-	-	-
FD-GAN (NIPS'2019) [11]	90.5	77.7	80.0	64.5	-	-	-	-
VCFL (ICCV'2019) [47]	89.3	74.5	-	-	61.4	55.6	-	-
ABD-Net (ICCV'2019) [8]	95.6	**88.2**	**89.0**	**78.6**	-	-	-	-
Ours	95.0	86.7	88.9	78.0	**68.8**	64.9	**78.8**	**55.9**

than deep Re-ID methods, which achieve around 90%. Among deep Re-ID methods, GAN-based methods [46, 20, 21, 11] are not so satisfying. The reasons may be that GAN import some noise to the generated images, and attention mechanism is not so useful for well-cropped images. For example, Cam-GAN and Pose-Normalized perform less than 90% on Market-1501. Compared with the above two kinds of methods, global feature-based methods [10, 9, 47] achieve very good performance. For example, BoT achieves 94.5% and 86.4% Rank-1 scores, and 85.9% and 76.4% mAP scores on Market and DukeMTMC-reID, respectively. Unsurprisingly, stripe-based methods achieve better accuracy than the global feature-based methods. Take MGN as an example, it gets 95.7% and 88.7% Rank-1 scores, and 86.9% and 78.4% mAP scores on Market-1501 and DukeMTMC-reID, respectively. Finally, our method uses only global features and achieves comparable performance with stripe-based methods on Market-1501 and DukeMTMC-reID, even better accuracy on CUHK03 and MSMT17. The analysis above demonstrates the effectiveness of our proposed methods under supervised setting.

4.4 Ablation Analysis and Effect of Hyper-parameters γ, α and β

Our method introduces two main parts, namely, disentangling feature learning (DFL) and the Consistent Transformation loss (CT loss), which are systematically analyzed through experiments. The performance of each component in fully-supervised task and semi-supervised task are shown in Fig.5.

Baseline If CT loss and DFL are disabled, our model degenerates into a classification model containing only one DenseNet-121 branch. Since our model only uses the cross entropy loss function, our baseline model also uses the cross entropy loss function.

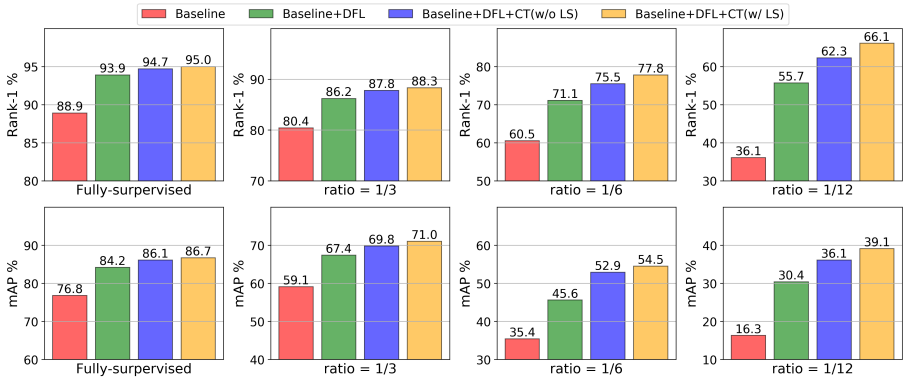

Fig. 5. Ablation analysis of three components on Market-1501. **DFL** refers to disentangled feature learning. **CT** means Consistent Transformation loss. **LS** represents label smooth adopted as a trick in our method. The first column is for fully-supervised case, and the last three columns are for semi-supervised cases with different labeled *ratio* values.

Baseline+DFL We simultaneously input a pair of structure mirror images and add a DFL module.

Baseline+DFL+CT loss Consistent transformation constraints are introduced for disentangled identity-aware features and structure-aware features.

As each component is applied one by one, we can observe significant performance improvements from Fig.5. For the fully-supervised case (the first column in Fig.5), Rank-1 is 88.9% when no strategies are used. When incrementally applying DFL, CT, SL, Rank-1 is increased to 93.9%, 94.7% and 95.0%, respectively. When reducing the proportion of labeled data in semi-supervised case, the effect of adding these strategies to the model is becoming more and more obvious. For example, when the *ratio* is 1/12, after adding DFL and CT, Rank-1 increases by 26.2%, and mAP increases by 19.8% compared with baseline. These three modules also have different impacts on model performance. As the results showing, DFL has the greatest impact on model performance. It also verifies that the combination of these components is complementary and conducive to achieve better performance.

The effect of hyper-parameters γ, α and β in the loss function In the total loss function, we set a parameter γ for unsupervised loss. The performance of our model is also related to this parameter. As can be seen from Fig. 6, when the parameter γ is 0.01, the performance of the model is the best. Particularly, when $\gamma = 0$ means only the supervised branch of the framework is valid. Therefore, it has verified that by utilizing the labeled data together with unlabeled data can bring performance improvement. Fig.7 shows the performance analysis on α and β. We choose $\alpha = 5$ and $\beta = 0.3$ because of their better performance.

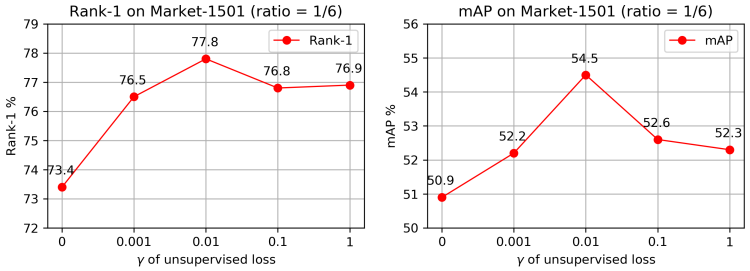

Fig. 6. Analysis on hyper-parameter γ of unsupervised loss

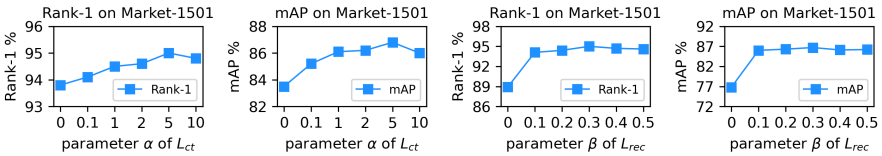

Fig. 7. Analysis on hyper-parameters α and β.

5 Conclusion

In this paper, we proposed a novel semi-supervised Re-ID framework, which consists of two branches with shared feature disentanglement models, one for supervised task and the other for unsupervised task. It alleviates limitation of labeled data by exploiting lots of unlabeled data. Furthermore, we design a free-of-label loss function to enforce consistent transformation constraints on disentangled features, which can be applied to both supervised and unsupervised learning branches. We have shown in ablation analysis experiments, the combination of above components play a very important role in performance improvement. A series of comparison results with stat-of-the-art methods have shown the good performance of ours in both semi-supervised and supervised tasks, and also demonstrated that our method can make full use of labeled data and unlabeled data. In the future, we plan to establish deeper connections between supervised and unsupervised branches and design a better training strategies.

6 Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 61972071), the National Key Research & Development Program (No. 2020YFC2003901), the 2019 Fundamental Research Funds for the Central Universities, the Research Program of Zhejiang lab (No.2019KD0AB02), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR No.201900014) and Sichuan Science and Technology Program (No.2020YJ0036).

References

1. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person re-identification. Springer (2014).
2. Satta, R.: Appearance descriptors for person re-identification: a comprehensive review. arXiv preprint arXiv:1307.5748 (2013).
3. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: 2015 IEEE International Conference on Computer Vision (ICCV). (2015) 1116–11244.
4. Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., Li, S.Z.: Salient color names for per-son re-identification. In: ECCV 2014: European Conference on Computer Vision.(2014) 536–551
5. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. (2010) 2360–2367
6. Yang, Y., Liao, S., Lei, Z., Li, S.Z.: Large scale similarity learning using simi-larpairs for person verification. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. (2016) 3655–3661
7. Sun, Y., Zheng, L., Yang, Y., Tian, Q., Wang, S.: Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 501–518
8. Chen, T., Ding, S., Xie, J., Yuan, Y., Chen, W., Yang, Y., Ren, Z., Wang, Z.:Abd-net: Attentive but diverse person re-identification. (2019) 8351–8361
9. Luo, H., Gu, Y., Liao, X., Lai, S., Jiang, W., Bao, L., Ma, B., Chang, H., Chen, X.,Li, P., et al.: Bag of tricks and a strong baseline for deep person re-identification.(In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)Workshops)
10. Zheng, L., Yang, Y., Hauptmann, A.G.: Person re-identification: Past, present and future. arXiv preprint arXiv:1610.02984 (2016)
11. Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., Li, H.: Fd-gan: Pose-guided feature distilling gan for robust person re-identification. (2018) 1222–1233
12. Chen, W., Chen, X., Zhang, J., Huang, K.: Beyond triplet loss: a deep quadruplet network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 403–412
13. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision, Springer (2016) 17–35
14. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
15. Liu, J., Zha, Z., Chen, D., Hong, R., Wang, M.: Adaptive transfer network for cross-domain person re-identification. (2019) 7202–7211
16. Zhong, Z., Zheng, L., Luo, Z., Li, S., Yang, Y.: Invariance matters: Exemplar memory for domain adaptive person re-identification. (2019) 598–607
17. Tang, H., Zhao, Y., Lu, H.: Unsupervised person re-identification with iterative self-supervised domain adaptation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE (2019) 1536–1543
18. Li, M., Zhu, X., Gong, S.: Unsupervised person re-identification by deep learning tracklet association. (2018) 772–788
19. Yu, H., Zheng, W., Wu, A., Guo, X., Gong, S., Lai, J.: Unsupervised person re-identification by soft multilabel learning. (2019) 2148–2157

20. Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y., Xue, X.: Pose-normalized image generation for person re-identification. (2018) 661–678
21. Zheng, Z., Yang, X., Yu, Z., Zheng, L., Yang, Y., Kautz, J.: Joint discriminative and generative learning for person re-identification. (2019) 2138–2147
22. Liu, Y., Song, G., Shao, J., Jin, X., Wang, X.: Transductive centroid projection for semi-supervised large-scale recognition. (2018) 72–89
23. Li, Y.J., Lin, C.S., Lin, Y.B., Wang, Y.C.F.: Cross-dataset person re-identification via unsupervised pose disentanglement and adaptation. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 7919–7929
24. Figueira, D., Bazzani, L., Minh, Q.H., Cristani, M., Bernardino, A., Murino, V.:Semi-supervised multi-feature learning for person re-identification. AVSS (2013)111–116
25. Liu, X., Song, M., Tao, D., Zhou, X., Chen, C., Bu, J.: Semi-supervised coupled dictionary learning for person re-identification. (2014) 3550–3557
26. Ding, G., Zhang, S., Khan, S., Tang, Z., Zhang, J., Porikli, F.: Feature affinity-based pseudo labeling for semi-supervised person re-identification. IEEE Transactions on Multimedia 21(2019) 2891–2902
27. Huang, Y., Xu, J., Wu, Q., Zheng, Z., Zhang, Z., Zhang, J.: Multi-pseudo regularized label for generated data in person re-identification. IEEE Transactions on Image Processing 28(2019) 1391–1403
28. Fan, H., Zheng, L., Yan, C., Yang, Y.: Unsupervised person re-identification:Clustering and fine-tuning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)14(2018) 1–18
29. Xin, X., Wang, J., Xie, R., Zhou, S., Huang, W., Zheng, N.: Semi-supervised person re-identification using multi-view clustering. Pattern Recognition 88(2019) 285–297
30. Wang, G., Zhang, T., Cheng, J., Liu, S., Yang, Y., Hou, Z.: Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In: 2019IEEE/CVF International Conference on Computer Vision (ICCV). (2019) 3622–3631
31. Wang, G., Zhang, T., Yang, Y., Cheng, J., Chang, J., Hou, Z.: Cross-modality paired-images generation for rgb-infrared person re-identification. In: AAAI 2020: The Thirty-Fourth AAAI Conference on Artificial Intelligence. (2020)
32. Wang, G., Yang, Y., Cheng, J., Wang, J., Hou, Z.: Color-sensitive person re-identification. In: IJCAI'19 Proceedings of the 28th International Joint Conference on Artificial Intelligence. (2019) 933–939
33. Wang, G., Yang, S., Liu, H., Wang, Z., Yang, Y., Wang, S., Yu, G., Zhou, E., Sun,J.: High-order information matters: Learning relation and topology for occluded person re-identification. 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
34. Wang, G., Gong, S., Cheng, J., Hou, Z.: Faster person re-identification. Proceedings of the European Conference on Computer Vision (ECCV) (2020)
35. Li, X., Makihara, Y., Xu, C., Yagi, Y., Ren, M.: Gait recognition via semi-supervised disentangled representation learning to identity and covariate features.In: CVPR 2020: Computer Vision and Pattern Recognition. (2020) 13309–13319
36. Huang, G., Liu, Z., Der Maaten, L.V., Weinberger, K.Q.: Densely connected convolutional networks. (2017) 2261–2269
37. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009.CVPR 2009. IEEE Conference on, Ieee (2009) 248–255

38. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. international conference on machine learning(2015) 448–456
39. Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
40. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.:Improving neural networks by preventing co-adaptation of feature detectors. arXivpreprint arXiv:1207.0580 (2012)
41. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? arXiv preprint arXiv:1611.07450 (2016)
42. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural networkfor person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 152–159
43. Wei, L., Zhang, S., Gao, W., Tian, Q.: Person transfer gan to bridge domain gapfor person re-identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 79–88
44. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. arXiv preprint arXiv:1708.04896 (2017)
45. Wang, G., Yuan, Y., Chen, X., Li, J., Zhou, X.: Learning discriminative features with multiple granularities for person re-identification. (2018) 274–282
46. Zhong, Z., Zheng, L., Zheng, Z., Li, S., Yang, Y.: Camera style adaptation for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5157–5166
47. Liu, F., Zhang, L.: View confusion feature learning for person re-identification.(2019) 6639–6648