# Camera Style and Identity Disentangling Network for Person Re-identification

Ruochen Zheng
m201772447@hust.edu.cn
Lerenhan Li
lrhli@hust.edu.cn
Chuchu Han
hcc@hust.edu.cn
Changxin Gao*
cago@hust.edu.cn
Nong Sang
nsang@hust.edu.cn

National Key Laboratory of Science and Technology on Multispectral Information Processing
School of Artificial Intelligence and Automation
Huazhong University of Science and Technology
Wuhan, China

**Abstract**

Camera style (camstyle) is a main factor that affects the performance of person re-identification (ReID). In the past years, existing works mainly exploit implicit solutions from the inputs by designing some strong constraints. However, these methods cannot consistently work as the camstyle still exists in the inputs as well as in the intermediate features. To address this problem, we propose a Camstyle-Identity Disentangling (CID) network for person ReID. More specifically, we disentangle the ID feature and camstyle feature in the latent space. In order to disentangle the features successfully, we present a Camstyle Shuffling and Retraining (CSR) scheme to generate more ID-preserved and camstyle variation samples for training. The proposed scheme ensures the success of disentangling and is able to eliminate the camstyle features in the backbone during the training process. Numerous experimental results on the Market-1501 and DukeMTMC-reID datasets demonstrate that our network can effectively disentangle the features and facilitate the person ReID networks.

## 1 Introduction

Person re-identification (ReID) aims at matching pedestrian images from non-overlapping cameras, which is challenging due to the camera style (camstyle) variations. It is difficult to match the changing appearance of the same person due to various camera imaging conditions, *e.g.*, viewpoints, image resolutions, illuminations, and background biases.

With the advances of deep CNNs, state-of-the-art methods focus on learning a feature representation that is robust to camstyle variations, which are mainly divided into explicit and implicit approaches. An explicit approach [27] presents a camstyle-insensitive network to deal with various camstyles. They address the problem by generating fake images in each

* indicates the corresponding author.

Table 1: **Person ReID vs. Camera classification.** We apply extracted features to evaluate the accuracy (%) of person ReID and camera classification tasks on the Market-1501 [21] dataset, respectively. Features from the IDE and ID+Tri can be directly used for both ReID and camera classification tasks. The proposed method can effectively disentangle the ID features thus improve the performance of person ReID.

| Task | IDE [22] | ID+Tri [4] | ID feature (multi-task) | Camstyle faeture (multi-task) | ID feature (ours) | Camstyle feature (ours) |
|---|---|---|---|---|---|---|
| Person ReID (rank-1) | 89.40 | 92.15 | 85.17 | 0.43 | 93.62 | 0.21 |
| Camera Classificaiton | 72.67 | 70.41 | 36.43 | 75.91 | 33.57 | 77.20 |

camera condition based on the CycleGAN [28]. However, this kind of method relies heavily on the generated images, which are not always in good conditions. On the other hand, a number of implicit methods [4, 14, 15, 19, 20, 23] propose to learn camstyle invariant features. They add specific constraints (*e.g.*, triplet loss [4], verification loss [23]) on the networks and project the features of the same person on a unified feature space. However, these methods still suffer from the remaining camstyle information in the projected features. We note that the useful features for ReID (i.e., ID features) are tangled with ambiguous camstyles (i.e., camstyle features) in the extracted features. As shown in the first two columns of Table 1, the extracted features from two state-of-the-art methods [4, 22] can be simultaneously applied for both person ReID and camera classification tasks. The results in first two columns of Table 1 also demonstrate that higher ReID performance leads to lower camera classification performance.

A natural way to address this problem is to directly extract the ID features and camstyle features using the multi-task learning framework, as shown in Figure 1(a). The third and fourth columns of Table 1 show the results of the multi-task method. The person ReID task performance even decreases, further proving the conflict between the camstyle classification and person ReID tasks. The tangled features cannot be effectively used for person ReID as the camstyle features play a negative role in the ReID task. Thus, it is of great interest to disentangle the ID features and camstyle features in a proper way.

In this work, our goal is to extract useful ID features from the input images. More specifically, we propose a Camstyle-Identity Disentangling (CID) network to divide the original features from a backbone into ID features and camstyle features. The proposed network consists of two encoders and one decoder. Two encoders, including ID encoder and camstyle encoder, aim at extracting high-level features from the backbone network. However, it is challenging to separate the features as ID features and camstyle features are highly tangled. Directly adding specific constraints (*e.g.* ID loss, Triplet loss, and Camera classification loss) may not work. To address the problem, we present a Camstyle Shuffling and Retraining (CSR) strategy to ensure the success of disentangling. We randomly re-tangle the ID and camstyle features back to the original features by the proposed decoder. With the CSR strategy, the proposed network can effectively disentangle the ID features and camstyle features from a backbone network. We then apply the disentangled ID features as the final representations for the person ReID task. We conduct extensive experiments to discuss and analyze the effectiveness of the CID network and CSR strategy. Furthermore, numerous experimental results show that the proposed CID network performs favorably against the state-of-the-art person ReID algorithms.

The contributions of this work are summarized as follows:

- We propose a Camstyle-Identity Disentangling Network to extract robust ID features for person ReID.

- We present a Camstyle Shuffling and Retraining strategy to ensure the success of disentangling by randomly re-tangling the ID and camstyle features in a cycle way.
- We conduct extensive experiments to demonstrate the effectiveness of the proposed CID network and CSR strategy and also show that the proposed method performs favorably against the state-of-the-art person ReID approaches.

## 2 Related Work

In this section, we focus our discussion on the explicit and implicit deep learning person ReID methods and the closest disentangling approaches.

**Implicit strategies.** Zheng *et al.* [23] propose a joint training on the classification loss and verification loss to obtain a more robust feature representation for each pedestrian image. Herman *et al.* [4] propose hard mining triplet losses to constrain the intra-class distance while enlarging the inter-class distance, thus the camstyle variations problem is relieved to some extent. Attention model is also an optional method to improve feature robustness toward camstyle implicitly. Zhao *et al.* [20] adopt an attention based method to learn a model focusing on more discriminative parts. Xu *et al.* [18] introduce pose information into attention model to obtain pose aligned features. These implicit solutions remove camera interference by designing more robust models and obtain some improvements.

However, as ID features and camstyle features are heavily tangled with each other, it is hard to extract the ID features from the tangled representations.

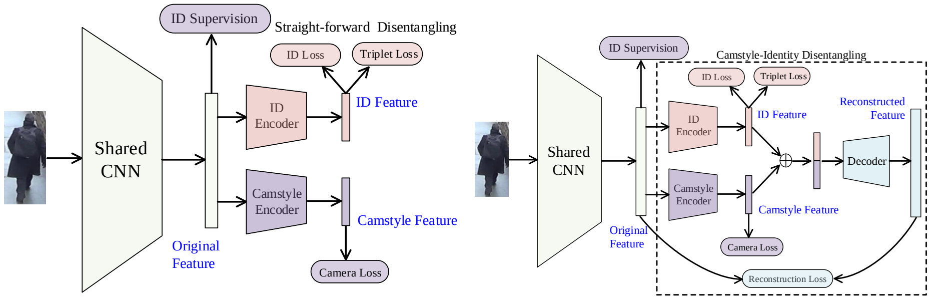
**Explicit strategy.** Zhong *et al.* [28] propose a camstyle adaptation method based on CycleGAN for ReID. For every two camera sources, they build generation models, which will generate cross-camstyle fake images. Fake images preserve the original ID and participate in training stage as ancillary data. However, it is complicated to build so many generative models for each two camera pairs, and the fake images introduce undesirable noise. On the contrary, our method generates data on feature level, which is more robust than generating samples on image level.

**Disentangling approaches.** Disentangling approaches have been used in some other computer vision tasks. Ma *et al.* [10] propose a multi-branched reconstruction network to disentangle the foreground, background and pose of the input image. The three disentangling factors can be manipulated to generate designated images. In face image generation field, Zheng *et al.* [25] present a method for disentangling the latent space into the label relevant and irrelevant dimensions, which is assumed to follow a Gaussian mixture distribution.

Different from these disentangling approaches, we present a novel model for person ReID by disentangling ID features and camstyle features from a backbone network.

## 3 Proposed Method

In this section, we first present our Camstyle-Identity Disentangling (CID) network. Then, we introduce the proposed Camstyle Shuffling and Retraining (CSR) strategy, which further ensures the success of disentangling camstyle and ID feature. Finally, we present some detailed training strategies.



(a) Baseline multi-task network

(b) Our disentangling model with CSR

Figure 1: **Camstyle-identity disentangling for person ReID.** (a) A multi-task model learns both the ID information and camstyle information from a network, (b) We train the CID network and propose a CSR strategy to ensure the success of the disentangling.

## 3.1 Camstyle-identity disentangling

**Backbone network.** Our CID network can be plugged into any network that is designed for the classification task and outputs the vector representation, *e.g.*, ResNet [3] and DenseNet [5]. In this paper, we adopt the ResNet50 as our backbone for its succinct structure and competitive performance. More specifically, we use the ResNet50 to extract features from an input pedestrian image and output a 2048-dimensional representation.

**Multi-task network.** We first design a multi-task network and take the camstyle as a kind of attribute to help the training of the backbone, as shown in Figure 1(a). Two encoders are designed independently to extract the low dimension ID feature and camstyle feature from the original feature. However, the addition of camstyle recognition task reduces the performance of ReID, which results from the direct conflict between the ReID task and camstyle recognition task. This phenomenon indicates the challenge of the straight disentangling method. To address this problem, we propose a novel CID network which involves a Camstyle Shuffling and Retraining strategy.

**CID architecture.** Based on the challenge of multi-task network, we propose the CID network to achieve the camstyle-identity disentangling in the latent space. The proposed CID network is shown in Figure 1(b). Particularly, we design two independent encoders with the same structure. It includes two fully connected layers and a hyperbolic tangent layer to disentangle the camstyle feature and ID feature in the latent space. Both of the features are 128-dimensional. The details of our network can be found in our supplementary materials.

We first formulate our CID as follows. We record a pair of camstyle feature and pedestrian ID feature in the latent space as $\{x_i, y_i\}_{i=1}^{m}$, with camera label $\{c_i\}_{i=1}^{m}$ and ID label $\{p_i\}_{i=1}^{m}$, where $m$ is the number of images in a batch.

To force the network to learn camstyle feature and ID feature, classification loss is taken to supervise them independently. To supervise the camstyle recognition, we use the camera IDs as the labels. In the commonly used ReID datasets, the camera ID labels of all the samples are available. In the practical scenes, camera label is also easy to obtain. Thus, the classification loss for camstyle can be formulated as $L_{cam}$:

$$L_{cam} = -\frac{1}{m}\sum_{i=1}^{m} \log \frac{e^{W_{c_i}^T X_i + b_{c_i}}}{\sum_{j=1}^{S} e^{W_j^T X_i + b_j}} \tag{1}$$

where $S$ is the number of cameras in the dataset.

For the identity decoder branch, we supervise the ID feature with the classification loss $L_{ID}$ and triplet hard [4] loss $L_{triplet}$ jointly. $L_{ID}$ is similar with $L_{cam}$ in form, not to be described in detail here.

As for the triplet hard loss, it defines triplet in vector representation space. It first chooses a sample as the anchor, and then selects the hardest positive sample and the hardest negative samples for the anchor. Then, distance between the anchor and positive sample should be less than the distance between the anchor and negative sample by a margin $\alpha$. Thus the loss can be formulated as:

$$L_{triplet} = \frac{1}{m}\sum_{i=1}^{m} [\max_{P_a = P_p} d(y_a, y_p) - \min_{P_a \neq P_n} d(y_a, y_n) + \alpha]_+ \tag{2}$$

Based on the aforementioned encoder and loss functions, we disentangle the camstyle feature and ID feature from the original 2048 dimension feature. Compared to the original dimension, the ID feature in the latent space contains more refined ID information by using the constraints mentioned above. However, the refined ID feature still not thoroughly solve the camstyle problem, due to two main factors: 1) uneven distribution of data under each camera, some cameras own obviously fewer images; 2) camstyle is a hybrid factor including changes of viewpoints, image resolutions, illuminations, and background biases, which is quite difficult to describe using the camera ID.

To tackle the problem mentioned above, we propose to further refine the ID feature using a feature-level sample generation method. That means, we try to reconstruct to the original 2048 dimension vector and overcome camera variations interference with controllable camstyle representation. We connect the ID feature and camera style feature to get a combined feature. Thus, the combined feature owns a clear distribution of camstyle and ID information. Next, we hope to use this known and determined distribution to remove the interference of camera factors on the ID features. We record the combined ID feature and corresponding camstyle representation as $f_{cam+ID}$. Then, we design a decoder $D$ to reconstruct the 2048 dimension feature $f_{ori}$ by $f_{cam+ID}$. The decoder can be learned automatically by the the reconstruction loss:

$$L_{rec} = ||f_{ori} - D(f_{cam+ID})||_2^2. \tag{3}$$

The overall loss for the CID part can be formulated as:

$$L_{all} = L_{ID} + L_{triplet} + L_{cam} + L_{rec}. \tag{4}$$

**Training strategy.** In the training stage, we train the backbone network and CID separately. Specifically, the gradient caused by encoder and decoder will not flow to the backbone

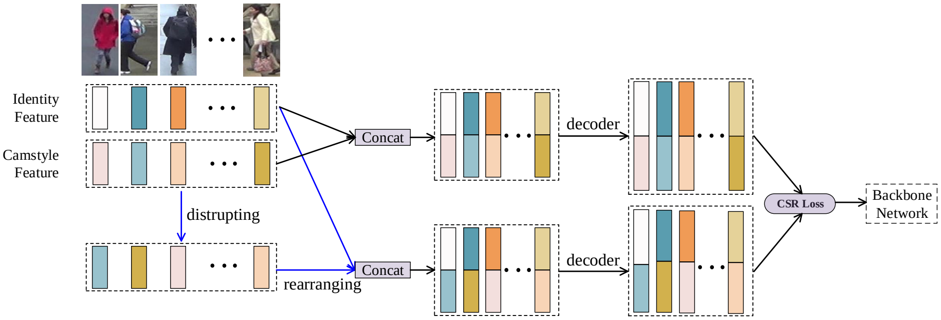


Figure 2: **Illustration of Camstyle Shuffling and Retraining.**

network in the back propagation. That is to say, the backbone can only be influenced by the loss of ID surprised original feature, while our CID will not influence the backbone at all in this stage. This strategy allows our approach to be applied to many popular ReID methods. The experiments in Section 4.4 will show the performance of our method in several ReID baseline.

## 3.2 Camstyle shuffling and retraining

As presented in Section 3.1, we use two encoders to extract refined ID information and camstyle information respectively, a decoder to reconstruct the original feature. Aiming at reducing the camstyle factor to the ReID branch, we propose Camstyle Shuffling and Retraining (CSR) to randomly generate new samples in the features space. In each batch, the new sample comes from a random combination of ID feature and camstyle feature at the latent space. The newly generated samples still keep the ID of the ID feature corresponding, which means we obtain more samples owning rich camstyle variations for each ID. The detailed CSR progress is shown in Figure 2.

We adopt the generated samples to help us retrain the network and reduce the camstyle variation for ID feature. There are two obvious advantages for taking the new samples into consideration: 1) more ID-preserving but camstyle variations samples are put into the training process, making up for the lack of multiple camera samples under some IDs. Besides, such data augmentation in feature level is easier to obtain; 2) restricting the distance between the new combined features and the features $con(cam, ID)$ can explicitly make the network pay more attention to the distinguishing pedestrian ID features, because this stage the reconstructed feature is mainly affected by the ID feature. We record the randomly combining strategy as $fuscon$. Thus the $L_{CSR}$ loss can be formulated as:

$$L_{CSR} = ||D(con(f_{ID}, f_{cam})) - D(fuscon(f_{ID}, f_{cam}))||_2^2. \tag{5}$$

In the CSR stage, we will concentrate on the reconstructed feature, including features reconstructed in order and features from the random reconstruction. The ID loss and triplet hard loss will be used to restrict the reconstructed features. Besides, the parameter of the encoder and decoder will be fixed to keep the encoding and decoding methods remaining unchanged.

### 3.3 Training details

To prove the universality of our method, we conduct the experiments on three representative person ReID baseline: the ID loss baseline(IDE) [22], ID+verification baseline [23] and ID+triplet hard baseline [4]. For the IDE baseline and ID+triplet hard baseline, the backbone is the same as Resnet50 without the final fully 1000 way connected layers. For the ID+verification baseline, extra two fully connected layers are added as verification branch to answer whether two images are the same person or not. All possible pairs in a batch are selected for the verification part.

All the baseline adopt $P \times K$ sampling method, there are $P$ persons and each person owns $K$ images in a batch. $P$ is set as 8 while $K$ is 4 for all experiments.

We adopt data augmentation strategy for increasing training data. The training images are resized as 384×128, then the random parsing strategy [26] is adopted. Finally, all the images are flipped horizontally with the probability of 0.5.

The all training epoch is set as 200. The former 20 epoch adopts a warm up strategy[2]. The learning rate increases linearly from 0 to 0.03. Then the learning rate is fixed until 180th epoch. At last, the learning rate drops to 0.003 until the end of the training. The training of CSR stage includes 60 training epoch, the learning rate is set as 0.003. As for testing stage, the refined ID feature after the CSR method is taken as the final ID representation.

## 4 Experimental Results

### 4.1 Evaluation datasets

We conduct experiments on two large scale person ReID datasets, i.e., , Market-1501 [21] and DukeMTMC-reID [11, 24].

The Market-1501 dataset is captured from 6 non-overlapping cameras, containing 1,501 pedestrians and 32,688 bounding boxes. For the partitioning of the dataset, 12,936 images of 751 pedestrians are chosen for training, while 3,368 query images and 19,732 gallery images are selected for testing. The DukeMTMC-reID dataset consists of 1,812 identities from 8 spatially disjoint cameras. The dataset is split into two parts: 16,522 images of 702 persons for training, 17,661 gallery images and 2,228 query images of another 702 persons for testing.

### 4.2 Protocols

In our experiments, we adopt the cumulative matching cure (CMC) and the mean Average Precision (mAP) as evaluation standard. The results of single query setting are reported in our experiments.

### 4.3 Comparisons with state-of-the-arts

As shown in Table 2 and Table 3, the proposed CID+CSR method obtains both state of the art results in the Market-1501 and DukeMTMC-reID datasets. For Market-1501, our CID+CSR obtains 93.8% rank-1 accuracy, which is comparable with the existing state of the art method PCB+RPP [15], and higher mAP(+0.8%). For DukeMTMC-reID, we also get the best performance in this benchmark, showing the effectiveness of our method.

Table 2: **Quantitative evaluations (%) on the Market-1501 dataset.** The proposed method performs favorably against the state-of-the-art person ReID approaches.

| Methods | Rank-1 | Rank-5 | Rank-10 | mAP |
|---|---|---|---|---|
| Bow+Kissme[24] | 44.4 | 63.9 | 72.2 | 20.8 |
| KLFDA[6] | 46.5 | 71.1 | 79.9 | - |
| MultiRegion[16] | 66.4 | 85.0 | 90.2 | 41.2 |
| PAR[20] | 81.0 | 92.0 | 94.7 | 63.4 |
| MultiLoss[7] | 83.9 | - | - | 64.4 |
| MultiScale[1] | 88.9 | - | - | 73.1 |
| PDC[13] | 84.4 | 92.7 | 94.9 | 63.4 |
| HA-CNN[8] | 91.2 | - | - | 75.7 |
| GAN_Camstyle[27] | 89.5 | - | - | |
| PCB[15] | 92.3 | 97.2 | 98.2 | 77.4 |
| PCB+RPP[15] | 93.8 | 97.5 | 98.5 | 81.6 |
| Mancs[17] | 93.1 | - | - | 82.3 |
| Ours | **93.8** | **97.7** | **98.7** | **82.4** |

Table 3: **Quantitative evaluations (%) on the DukeMTMC-reID dataset.** Our CID network performs favorably against the state-of-the-art person ReID algorithms.

| Methods | Rank-1 | Rank-5 | Rank-10 | mAP |
|---|---|---|---|---|
| Bow+KISSME[24] | 25.1 | - | - | 12.2 |
| LOMO+XQDA[9] | 30.8 | - | - | 17.0 |
| GAN[24] | 67.7 | - | - | 47.1 |
| GAN_camstyle[27] | 78.3 | - | - | 57.6 |
| SVDNet[14] | 76.7 | 86.4 | 89.9 | 56.8 |
| AACN[18] | 76.8 | - | - | 59.3 |
| PSE[12] | 79.8 | 89.7 | 92.2 | 62.0 |
| PCB[15] | 81.8 | - | - | 66.1 |
| PCB+RPP[15] | 83.3 | - | - | 69.2 |
| HA-CNN[8] | 80.5 | - | - | 63.8 |
| Mancs[17] | 84.9 | - | - | 71.8 |
| Ours | **85.6** | **93.8** | **95.4** | **72.4** |

## 4.4 Effectiveness of the CID network

To evaluate the effect of the proposed CID network and CSR strategy, we conduct experiments on three commonly used baseline, the results are show in Table 4 and Table 5. It can be seen that our method obtains breakthrough in all three baselines, with both of the CID network and CSR strategy. Overall, as for Market-1501, mAP on three baselines increases from 73.57%, 75.05% and 79.33% to 82.45% (+8.88%), 82.38% (+7.33%) and 82.28% (+2.95%), respectively. This indicates our method works well for these different kinds of baselines. In DukeMTMC-reID, our methods also shows stable improvement. The improvements of rank-1 accuracy is +8.22%, +5.97%, 2.47%.

## 4.5 Sensitivity to key parameters

**The number of generated data for CSR stage.** In the CSR stages, the number of newly generated data is a determined parameter to influence the performance. We record the number of raw samples as M, the generated samples in CSR as N. We evaluate the performance with different ratio of fake data and real data (N : M) in Figure 4. We can find when N:M is set as 1:1, we can obtain the highest performance. This result also shows that the number of generated data should stay in a apposite level. Too many generated data will damage the

Table 4: **Effectiveness of the proposed CID network.** We compare the baseline models and the proposed CID and CSR on the Market-1501 dataset.

| Methods | Rank-1 | Rank-5 | Rank-10 | mAP |
|---|---|---|---|---|
| IDE | 89.40 | 96.38 | 98.16 | 73.57 |
| CID | 90.67 | 97.11 | 98.43 | 79.35 |
| CID + CSR | **93.44** | **97.62** | **98.72** | **82.45** |
| ID+Ver | 90.41 | 96.47 | 98.13 | 75.05 |
| ID+Ver CID | 91.06 | 97.11 | 98.43 | 80.65 |
| ID+Ver CID+CSR | **93.50** | **97.71** | **98.57** | **82.38** |
| ID+Tri | 91.76 | 97.30 | 98.34 | 79.33 |
| ID+Tri CID | 92.23 | 97.18 | 98.54 | 79.64 |
| ID+Tri CID+CSR | **93.82** | **97.62** | **98.66** | **82.28** |

Table 5: **Effectiveness of the proposed CID network.** We compare the baseline models and the proposed CID and CSR on the DukeMTMC-reID dataset.

| Methods | Rank-1 | Rank-5 | Rank-10 | mAP |
|---|---|---|---|---|
| IDE | 77.51 | 88.87 | 92.06 | 60.44 |
| CID | 83.29 | 90.37 | 93.53 | 67.62 |
| CID+CSR | **85.73** | **93.49** | **95.38** | **72.18** |
| ID+Ver | 79.44 | 90.08 | 93.04 | 63.66 |
| ID+Ver CID | 84.47 | 91.43 | 94.52 | 68.15 |
| ID+Ver CID+CSR | **85.41** | **93.45** | **95.69** | **72.26** |
| ID+Tri | 83.08 | 92.01 | 94.61 | 68.39 |
| ID+Tri CID | 83.98 | 92.73 | 94.93 | 71.34 |
| ID+Tri CID+CSR | **85.55** | **93.76** | **95.38** | **72.41** |

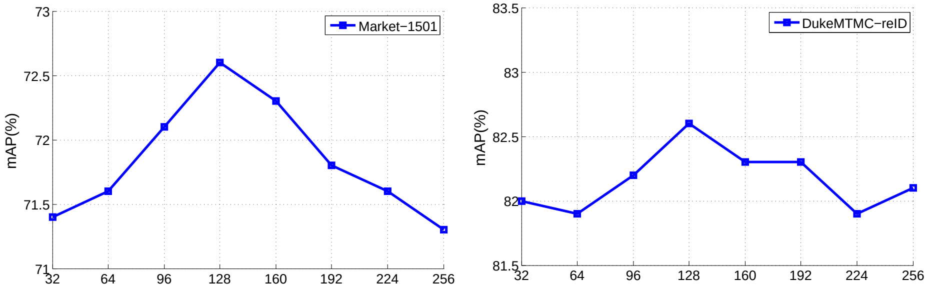
method's performance. In the other parts of the paper, if not specified, we set N : M = 1 : 1.

**Dimension of camstyle represebtation.** The dimension of the camstyle determines the compositional relationship between discriminative ID features and camstyle representations in the original feature. Higher dimension makes the camstyle information take larger proportion. We evaluate the sensitivity by fixing the dimension of ID feature in latent space as 128, and setting dimension of camstyle from 32 to 256.

Figure 3 shows that the performance of CSR improves as camstyle dimension increases. However, when the camstyle dimension becomes larger than the ID feature in latent space, the final performance drops at this scene. Therefore, the dimension of camstyle is set as 128, which is the same as the dimension of ID in latent space.

## 5 Conclusions

In this paper, we propose a novel disentangling network to disentangle the camstyle and refined ID feature. Different from the multi-task network, we take an encoder-decoder architecture to achieve such disentangling in the latent space. We also present a CSR strategy to ensure the success of the disentangling. The CSR strategy is also a novel data augmentation method in the feature level, which is light-weight compared with existing data augmentation in the image level. The competitive performance on Market-1501 and DukeMTMC-reID shows the effectiveness of our method.



(a) Market-1501

(b) DukeMTMC-reID

Figure 3: **Sensitivity to the dimension of feature representation.** We evaluate the mAP(%) on the Market-1501 and DukeMTMC-reID datasets.

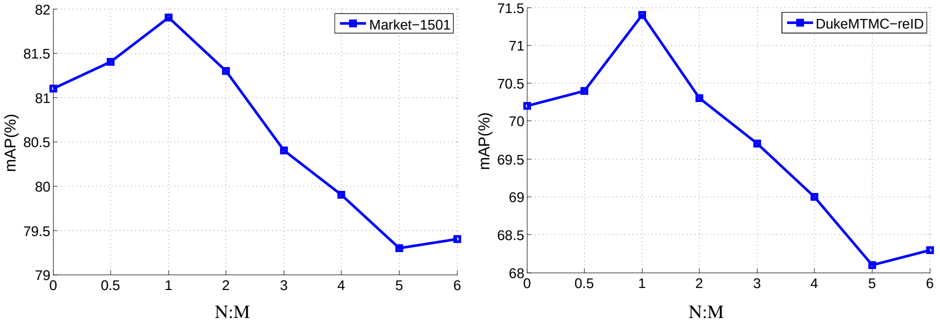


Figure 4: **Different ratio of fake data and real data (N:M).** We evaluate map(%) on the Market-1501 and DukeMTMC-reID datasets.

# Acknowledgements

This work was supported by the Project of the National Natural Science Foundation of China (No. 61876210), National Key R&D Program of China (No. 2018YFB1004600), and Natural Science Foundation of Hubei Province (No. 2018CFB426).

# References

[1] Yanbei Chen, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep learning multi-scale representations. In *IEEE International Conference on Computer Vision*, 2017.

[2] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision*, 2016.

[4] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[6] Srikrishna Karanam, Mengran Gou, Ziyan Wu, Angels Rates-Borras, Octavia Camps, and Richard J Radke. A comprehensive evaluation and benchmark for person re-identification: Features, metrics, and datasets. *arXiv preprint arXiv:1605.09653*, 2 (3):5, 2016.

[7] Wei Li, Xiatian Zhu, and Shaogang Gong. Person re-identification by deep joint learning of multi-loss classification. *arXiv preprint arXiv:1705.04724*, 2017.

[8] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[9] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[10] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

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

[12] M Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 420–429, 2018.

[13] Chi Su, Jianing Li, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Pose-driven deep convolutional model for person re-identification. In *IEEE International Conference on Computer Vision*, 2017.

[14] Yifan Sun, Liang Zheng, Weijian Deng, and Shengjin Wang. Svdnet for pedestrian retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

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

[16] Evgeniya Ustinova, Yaroslav Ganin, and Victor Lempitsky. Multi-region bilinear convolutional neural networks for person re-identification. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[17] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *European Conference on Computer Vision*, 2018.

[18] Jing Xu, Rui Zhao, Feng Zhu, Huaming Wang, and Wanli Ouyang. Attention-aware compositional network for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[19] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[20] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *IEEE International Conference on Computer Vision*, 2017.

[21] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *IEEE International Conference on Computer Vision*, pages 1116–1124, 2015.

[22] Liang Zheng, Yi Yang, and Alexander G Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[23] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):13, 2017.

[24] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *IEEE International Conference on Computer Vision*, 2017.

[25] Zhilin Zheng and Li Sun. Disentangling latent space for vae by label relevant/irrelevant dimensions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12192–12201, 2019.

[26] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. *arXiv preprint arXiv:1708.04896*, 2017.

[27] Zhun Zhong, Liang Zheng, Zhedong Zheng, Shaozi Li, and Yi Yang. Camera style adaptation for person re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[28] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017.