DC-Former: Diverse and Compact Transformer for Person Re-identification

Wen Li[1], Cheng Zou[1], Meng Wang[1], Furong Xu[1], Jianan Zhao[1],
Ruobing Zheng[1], Yuan Cheng[2]*, Wei Chu[1]

[1]Ant Group
[2]Artificial Intelligence Innovation and Incubation (AI[3]) Institute, Fudan University
{yinian.lw,wuyou.zc,darren.wm,booyoungxu.xfr,zhaojianan.zjn,zhengruobing.zrb,
weichu.cw}@antgroup.com, cheng_yuan@fudan.edu.cn

Abstract

In person re-identification (ReID) task, it is still challenging to learn discriminative representation by deep learning, due to limited data. Generally speaking, the model will get better performance when increasing the amount of data. The addition of similar classes strengthens the ability of the classifier to identify similar identities, thereby improving the discrimination of representation. In this paper, we propose a Diverse and Compact Transformer (DC-Former) that can achieve a similar effect by splitting embedding space into multiple diverse and compact subspaces. Compact embedding subspace helps model learn more robust and discriminative embedding to identify similar classes. And the fusion of these diverse embeddings containing more fine-grained information can further improve the effect of ReID. Specifically, multiple class tokens are used in vision transformer to represent multiple embedding spaces. Then, a self-diverse constraint (SDC) is applied to these spaces to push them away from each other, which makes each embedding space diverse and compact. Further, a dynamic weight controller (DWC) is further designed for balancing the relative importance among them during training. The experimental results of our method are promising, which surpass previous state-of-the-art methods on several commonly used person ReID benchmarks. Our code is available at *https://github.com/ant-research/Diverse-and-Compact-Transformer*.

Introduction

Person re-identification (ReID) aims at identifying person across different camera views, which is very important in many applications, such as intelligent surveillance, cross camera tracking and smart city. While ReID has attracted great research interest and gained considerable development in recent years, there still exist some challenges (Yan et al. 2021; Yang et al. 2021), such as blur, low resolution, occlusion, illumination and viewpoint variation. These factors cause the intra-class distance of samples to be larger than the inter-class distance, which makes it challenging to retrieve pedestrians of correct identities.

Plenty of efforts (He et al. 2021; Gu et al. 2022; Tan et al. 2022) have been made recently to improve the performance
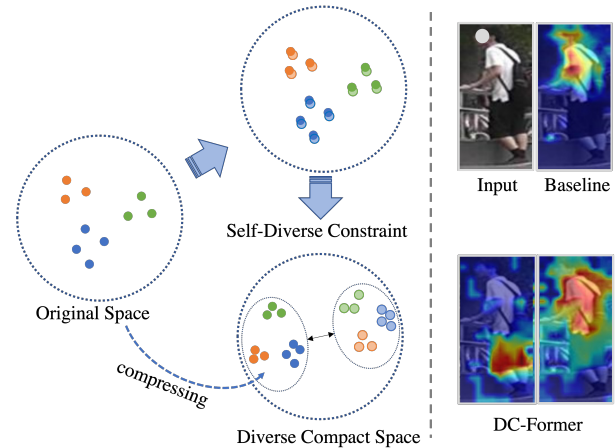

Figure 1: Illustration of DC-Former for representation learning. On the left, each circle with dots denotes an embedding space. DC-Former uses multiple embeddings to represent each sample, and a self-diverse constraint is imposed on these embeddings to push them away. Finally, DC-Former obtains multiple diverse embedding subspaces for representation. And each subspace is more compact than the original space, which increases the identity density of embedding space to help model improve its discrimination for identifying similar classes. Figures on the right visualized by Grad-Cam (Selvaraju et al. 2017) show that diverse embeddings from DC-Former focus on multiple different discriminative regions. And the fusion of them can provide more fine-grained information.

of ReID, and among which increasing the amount of training data may be the most powerful way. On the one hand, increasing more instances for each identity helps to recognize one person under different circumstances, extracting the most common and discriminative features for the same class, thus reducing intra-class distance. On the other hand, increasing more identities means that there is a higher chance to place more similar (easy-to-confuse) classes in the embedding space, which would help the model to extract discriminative features with larger inter-class distance for similar classes. Therefore, more identities and more instances

for each identity, resulting in larger inter-class distance and smaller intra-class distance, make ReID task easier.

As mentioned above, great potential lies in data. However, due to limited data, existing works concentrate more on other aspects, such as stronger backbone (Liu et al. 2021), metric loss (Sun et al. 2020), pre-trained model (Fu et al. 2021), etc. As far as we know, few studies consider ReID task from the perspective of data except for (Gong et al. 2021; Zhong et al. 2020), all of which are data augmentations, finally increasing the instances of each identity, but no study tries to increase the number of identity. Directly increasing the number of identity is impractical because it equals to adding more labeled data, which is expensive, but if there is a certain way that can simulate to increase the number of identity, it will also improve the performance of ReID. Here we hypothesize that increasing the number of identity is equal to increasing the identity density in a given space. Reducing the size of the embedding space can make all the identities become more compact, so the relative amount of identity (identity density) increases. In a more compact embedding space, the reduction of inter-class distance makes it more difficult for model to distinguish samples from similar classes. Therefore, more discriminative and robust information is extracted to ensure that the classifier correctly identifies similar classes.

In this paper, we propose a Diverse and Compact Transformer (DC-Former) that can achieve a similar effect of increasing identities of training data. As shown in Figure 1, it compresses the embedding space by dividing the original space into multiple subspaces. More compact representations in the compressed embedding subspace helps model extract more discriminative feature to identify similar classes. And the embeddings of different subspaces are diverse, the fusion of them contains more information that can further improve performance. Specifically, multiple class tokens (CLSes) are used in vision transformer, and each CLS is supervised by an identity loss to obtain multiple representations. Then, a self-diverse constraint (SDC) is applied to CLSes to make the distribution of them as far as possible. In this way, the original space is divided into multiple subspaces. Due to the different learning status of different CLSes when dividing the space, some CLSes are pushed away while others are very close. A dynamic weight controller (DWC) is further designed for balancing the relative importance among them during training. Finally, each of compact subspaces learns a more robust representation. The experimental results of our method are promising, which surpass previous state-of-the-art methods on three commonly used person ReID benchmarks.

The contributions of this paper are summarized as follows:

- We propose a DC-Former to get multiple diverse and compact embedding subspaces, each embedding of these compact subspaces is more robust and discriminative to identify similar classes. And the fusion of these diverse embeddings can further improve the effect of ReID.

- We propose a self-diverse constraint (SDC) to make embedding subspaces presented by each class token do not overlap. And a dynamic weight controller (DWC) is devised to balance the relative importance among multiple class tokens during training.

- Our method surpasses previous methods and sets state-of-the-art on three person ReID benchmarks including *MSMT17*, *Market-1501* and *CUHK03*.

Related Work

Image-Based ReID

ReID can be conducted on either images (He et al. 2021) or videos (Zhao et al. 2021) . Recent image-based tasks mainly focus on person ReID and vehicle ReID. The studies of person ReID have paid attention to feature representation learning (Chen et al. 2019; Suh et al. 2018) and deep metric learning (Zheng et al. 2017; Deng et al. 2018). For the feature learning strategy, Typically, there are three main categories for the feature learning strategy, including global feature (Chen et al. 2019), local feature (Suh et al. 2018) and auxiliary feature (e.g., domain information (Lin et al. 2017), semantic attributes (Lin et al. 2019), viewpoint information (Zhu et al. 2020)). As for the deep metric learning, many existing works have designed loss functions to guide the feature representation learning, which can be divided into three groups, i.e., identity loss (Zheng et al. 2017), verification loss (Deng et al. 2018) and triplet loss (Hermans, Beyer, and Leibe 2017).

Representation Learning In ReID

ReID is one of fine-grained task which has far intra-class distance and close inter-class distance. Extracting discriminative features to distinguish similar classes is challenging. To minimize intra-class distance and maximize inter-class distance, proxy-based loss (Zheng, Zheng, and Yang 2017; Krizhevsky, Sutskever, and Hinton 2012) and pair-based loss (Liu et al. 2017; Hermans, Beyer, and Leibe 2017) are commonly used to push the different class away and pull the same class close. Some works use attention-base method (Huynh and Elhamifar 2020) to discover discriminative features or enhance low-discriminative features, which helps disciminative feature representation. For example, DAM (Xu et al. 2021) iteratively identifies insufficiently trained elements and improves them. And some works (Zhong et al. 2020; Srivastava et al. 2014) impose some regulation operations (e.g., random erasing and dropout) to prevent overfitting. Self-supervised representation learning (He et al. 2019; Chen et al. 2020; Caron et al. 2021; Isobe et al. 2021) use contrastive loss to maximize the agreement of features from same image with different augmentations.

Transformer-Based ReID

Before vision transformer, CNN-based methods have achieved absolute advantages on ReID task. Methods like PCB (Sun et al. 2018), MGN (Wang et al. 2018), etc., partition an image into several stripes to obtain local feature representations for each stripe with multiple granularities. With the success of transformer in the field of computer version, it's also widely used in person ReID. Compared with
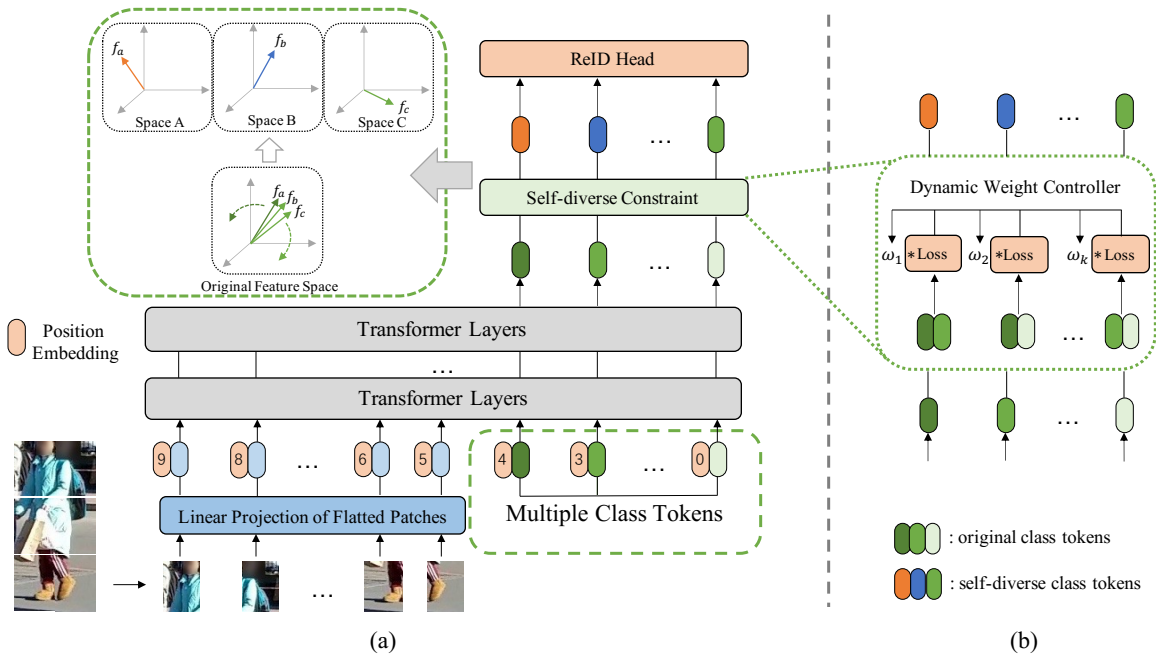
Figure 2: The framework of DC-Former. Multiple class tokens concatenated with patch embeddings, adding positional embeddings, are fed into transformer encoder. Self-diverse constraint is employed on these class tokens in the last transformer layer to push them far way from each other, leading to diverse representation spaces. Then they are each supervised by a ReID head, which contains a triplet loss and a classification loss. During training, a dynamic weight controller is used to dynamically adjust the constraint loss for easier optimization. f_a, f_b and f_c are different representations of the same object.

CNN-based method, ViT can keep more detailed information because these is no convolution and dowmsampling operators, which is more suited for fine-grained task like ReID. TransReID (He et al. 2021) is the first pure transformer-based method on ReID, it proposes a jigsaw patches module (JPM) which shuffles patch embeddings and re-groups them for further feature learning to extract several local features and aggregates them to get robust feature with global context. TransReID-SSL (Luo et al. 2021) uses a massive person ReID dataset $LUPerson$ (Fu et al. 2021) to train a stronger pre-trained model by DINO (Caron et al. 2021).

Many works (Zhu et al. 2021; Li et al. 2021; Chen et al. 2021; Lai, Chai, and Wei 2021) devote to extract partial region representation by transformer. For example, AAformer (Zhu et al. 2021) uses the additional learnable vectors of 'part tokens' to learn the part representations by clustering the patch embeddings into several groups and integrates the part into the self-attention for alignment. Some methods (Li et al. 2021; Lai, Chai, and Wei 2021) use transformer to learn several different part-aware masks to get partial feature. And some works aim to fuse features at different granularities. For example, HAT (Zhang et al. 2021) put hierarchical features at different granularities from CNN backbone into transformer to aggregate them. These methods have one thing in common that multiple different features are extracted and integrated to obtain more robust representation. However, they impose constraints on specific goals like focusing only on local receptive fields or different gran-

ularity which are limited to extract representation that the model really needs. It is insightful to learn multiple feature space by model itself through certain implicit constraint.

Methodology

Overview

For enhancing the identity density to increase the ability of discriminating similar classes, the original embedding space is divided into multiple diverse and compact subspaces. The proposed method consists of three parts: a ViT-based network with multiple class tokens to produce multiple embedding spaces, a self-diverse constraint loss (SDC) to push these embedding spaces far away from each other, and a dynamic weight controller (DWC) to balance the relative importance among class tokens during training as the token number increases.

Network Architecture

To construct multiple different embedding spaces for one single input image, an architecture with multiple parallel output features is required. Most multi-feature representation methods usually extract different features from different granularities or different partial regions by multiple branches (Wang et al. 2018; Sun et al. 2018). The multiple features generated by these branches are often individual and have no interactions with each other during train, which may
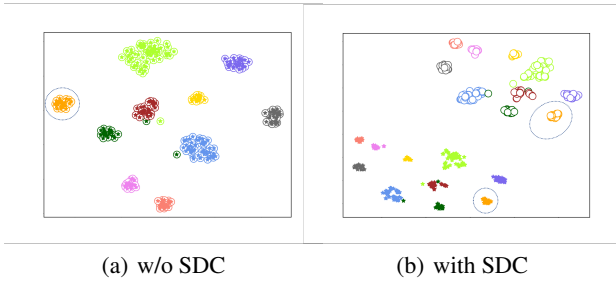
(a) w/o SDC   (b) with SDC

Figure 3: Feature distributions visualized by t-SNE of two class tokens. Different colors represent different classes, marker ⋆ stands for feature points from one class token and ◯ from the other. The data points are sampled from *MSMT17* test set. (a) Without SDC, the feature points from two class tokens are heavily overlapped, which implies that the embedding spaces of these two class tokens are very close. (b) With SDC, the feature points from two class tokens can be easily separated into two parts without any overlap, and each embedding space become more compact.

cause multiple embedding spaces overlapping and homogenous.

In vision transformer, thanks to stacked self-attention modules, information flows from patch embeddings to one class token, layer by layer gradually and autonomously. The class token acts as an information collector here, it receives information from each patch and outputs the summary according to its prior knowledge (learnable parameters). It's natural to come up with an idea that if there exist multiple class tokens acting as multiple information collectors, there will be a chance to gather different information from patch embeddings. Then, multiple different embedding spaces can be supported by multiple class tokens. Therefore, ViT with additional class tokens is chosen to be the main achitecture.

Figure 2 illustrates the proposed framework. The input image is divided into $H \times W$ patches, and then each patch is mapped to a vector of dimension D through a trainable linear projection. The output of this projection is denoted as patch embeddings $P^{C \times D}$, where C is the number of patches. Then multiple learnable embeddings of dimension D are concatenated as a sequence, denoted as class tokens $f^{N \times D}$, where N is the number of class tokens. After that, class tokens and patch embeddings are concatenated to get a vector sequence $X^{(N+C) \times D} = [f^{N \times D}; P^{C \times D}]$. Next, this sequence added with a learnable positional embedding sequence is fed into transformer layers to get multiple representations.

Notably in our design, multiple class tokens are mutually visible in self-attention layers, which helps the model to converge because of sharing intermediate features. For one certain class token, it receives information not only from all the patch embeddings but also from other class tokens, which improves the information acquisition efficiency.

Self-Diverse Constraint Loss

If there are multiple different embeddings to represent each samples, the embedding space will become more compact, which helps model improve embedding's discrimination of similar classes. With multiple class tokens, the proposed framework can output multiple embeddings in parallel. But in practice, the learned embedding spaces overlap with each other because these class tokens are homogeneous in structure, their learning goals are the same, and there is no external energy to push them to be different.

One way to learn different embedding spaces is to change the learning objectives, DeiT (Touvron et al. 2021) proposed a distillation token to learn the output logits of another CNN model, which requires training two models and much manual efforts, making the training procedure complex. Considering that class token is computed as a weighted sum of V, although multiple class tokens share the same K-V pairs from patch embeddings $P^{C \times D}$, they can still get different attentions from $P^{C \times D}$ if the query Q is different. In other words, class tokens can learn different information as long as they are different.

Here we humbly hypothesize that the more difference between class tokens, the farther the distance between their embedding space, which pushes overlapping embedding spaces away from each other. Not only do we want the class tokens to be different from each other, but we want to maximize the difference between them. In the metric of cosine similarity, if two vectors are orthogonal, i.e., $cos(*, *) = 0$, they are irrelevant, and the distance between them are extremely large. If the class tokens are orthogonal to each other, each embedding space of them is compressed more tightly in the finite space to ensure their distance is maximized.

We propose a self-diverse constraint loss (SDC) to constrain the relationship between class tokens. It forces the class tokens to be orthogonal to each other and can be expressed as:

$$L_{sdc} = \frac{1}{C_N^2} \sum_i \sum_j \nu_{ij}, \ i < j, \ \ i,j = 1,...,N \quad (1)$$

where $\nu_{ij} = |cos(f_i, f_j)|$, and f_i, f_j indicates any two class tokens of $f^{N \times D}$. Self-diverse constraint loss is employed on these class tokens to make sure that these embedding subspaces is far apart from each other. It makes each embedding subspace compact and helps model extract more discriminative feature to identify similar classes.

The representations from different embedding subspaces contain not only identity information, but also features from different perspectives, i.e., coarse/fine-grained granularity, global/local region, and other unrecognized complementary aspects. The fusion of them can further facilitate robust and perturbation-invariant feature representation for ReID tasks.

Dynamic Weight Controller

To obtain more compact and robust features, more class tokens are needed to get more different embedding subspaces. However, as the number of class tokens increases, it becomes harder to optimize because there are more pairs of class tokens, each of which is required to be orthogonal.

Methods	Publications	MSMT17		Market-1501		CUHK03-l		CUHK03-d	
		mAP	R1	mAP	R1	mAP	R1	mAP	R1
PFD (Wang et al. 2022)	AAAI	64.4	83.8	89.7	95.5	-	-	-	-
TransReID (He et al. 2021)	ICCV	67.4	85.3	88.9	95.2	-	-	-	-
DPM (Tan et al. 2022)	ACM'MM	-	-	89.7	95.5	-	-	-	-
GASM$^{384\uparrow}$ (He and Liu 2020)	ECCV	52.5	79.5	84.7	95.3	-	-	-	-
CDNet$^{384\uparrow}$ (Li, Wu, and Zheng 2021)	CVPR	54.7	78.9	86.0	95.1	-	-	-	-
AutoLoss$^{384\uparrow}$ (Gu et al. 2022)	CVPR	63.0	83.7	<u>90.1</u>	<u>96.2</u>	74.3	75.6	-	-
AAformer$^{384\uparrow}$ (Zhu et al. 2021)	-	63.2	83.6	87.7	95.4	<u>77.8</u>	<u>79.9</u>	<u>74.8</u>	<u>77.6</u>
OH-former$^{368\uparrow}$ (Chen et al. 2021)	-	69.2	<u>86.6</u>	88.7	95.0	-	-	-	-
TransReID$^{384\uparrow}$ (He et al. 2021)	ICCV	<u>69.4</u>	86.2	89.5	95.2	-	-	-	-
DC-Former	-	69.8	86.2	90.4	96.0	79.4	81.6	77.5	**80.1**
DC-Former$^{384\uparrow}$	-	**70.7**	**86.9**	**90.6**	96.0	**83.3**	**84.4**	**77.5**	79.6

Table 1: Comparisons with state-of-the-art methods on person ReID benchmarks. $368\uparrow$ and $384\uparrow$ denote the input images are resized to 368×128 and 384×128, otherwise 256×128. Best results for previous methods are underlined and best of our methods are labeled in bold.

In experiments, we find that when the number of class tokens increases to a certain number, the loss in Eq. 1 cannot be minimized as expected. Although some of the pairs have low cosine similarity, others are still very similar (cosine similarity close to 1). Actually it only learns less embedding space than we expect, and those class token pairs with higher self-diverse constraint loss are not optimized well in training. The reason for this is that randomness in the training process makes it easier for some pairs to be pushed farther apart while others harder. Therefore, it's necessary to change the relative importance among class token pairs for self-diverse constraint loss while training.

We propose a dynamic weight controller (DWC) to dynamically adjust the loss weight of each pair during training on the fly. Instead of simply averaging these pair losses as in Eq. 1, the loss of each pair is re-weighted by its own softmax-normalized loss. The weight of each loss can be defined as:

$$\omega_{ij} = \frac{exp(\nu_{ij})}{\sum_m \sum_n exp(\nu_{mn})}, m < n, \ \ m, n = 1, ..., N \quad (2)$$

So, the balanced self-diverse constraint loss is defined as:

$$L_{SDC} = \sum_i \sum_j \omega_{ij} \nu_{ij}, \ i < j, \ \ i, j = 1, ..., N \quad (3)$$

Those pairs with smaller cosine similarities are given smaller weights, while larger ones are given larger weights to make the model focus more on similar pairs. In this way, pairs can be learned more evenly so that they can be all orthogonal to each other.

Objective Function

In training, in order to ensure that each class token has the ability to distinguish identities, they are each supervised by cross-entropy loss for classification (ID loss) after normalized by BNNeck (Luo et al. 2019). To pull the samples of the same class closer and push the samples of different classes far away, the triplet loss with soft-margin is used to mine

hard example in each embedding subspace and can be calculated as:

$$L_{triplet} = log[1 + exp(||f_a - f_p||_2^2 - ||f_a - f_n||_2^2)] \quad (4)$$

The overall objective function is:

$$L_{total} = \frac{1}{N} \sum_{i=1}^N (L_{ID}^i + L_{triplet}^i) + \lambda L_{SDC} \quad (5)$$

During inference, all the class tokens are concatenated to represent an image.

Experiments

Experimental Settings

Implementation Details. We apply VIT-B/16 (Dosovitskiy et al. 2020) as our backbone, it contains 12 transformer layers with hidden size of 768 dimensions. Overlapping patch embedding (step size = 12) and SIE (He et al. 2021) are also used in our experiments. All the images are resized to 256×128 unless other specified. The training images are augmented with random horizontal flipping, padding, random cropping, random erasing (Zhong et al. 2020), and random grayscale (Gong et al. 2021). The initial weights of the models are pre-trained on ImageNet.

The batch size is set to 64 with 4 images per ID. SGD optimizer is employed with a momentum of 0.9 and a weight decay of 1e-4. The learning rate is initialized as 0.032 with cosine learning rate decay. All the experiments are performed on 4 Nvidia Tesla V100 GPUs.

Datasets and Evaluations. The proposed method is evaluated on four widely used person ReID benchmarks, i.e., MSMT17 (Wei et al. 2018), Market-1501 (Zheng et al. 2015) and CUHK03 (Li et al. 2014). Mean Average Precision (mAP) and Cumulative Matching Curve (CMC) are used to evaluate the performance of ReID tasks.

Comparisons with State-of-the-Arts

To verify the effectiveness of the proposed method, experiments are conducted on three commonly used person ReID benchmarks in Table 1. On MSMT17, our method outperforms previous SOTA methods (e.g., TransReID) by a large
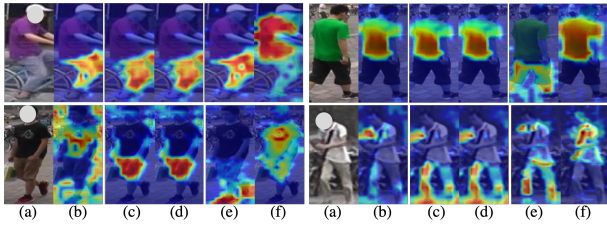
Figure 4: Grad-Cam visualization of attention map on *Market-1501*. (a) Input image. (b) Baseline. (c)-(d) Two class tokens without SDC. (e)-(f) Two class tokens with SDC. As can be seen, self-diverse constraint loss makes multiple class tokens to focus on different discriminative regions.

(a) Baseline   (b) $Cat(T_1, T_2)$

(c) T_1 (d) T_2

Figure 5: The distance of positive and negative pairs in *MSMT17* test set. The positive/negative pair denotes that two sample from $Query$ and $Gallery$ are the same/different class(es). The ordinate is the number of pairs. For better visualization, only partial negative pairs sampled randomly are shown in this figure.

	mAP			Rank-1		
	T_1	T_2	$Cat.$	T_1	T_2	$Cat.$
Baseline	-	-	66.1	-	-	84.6
w/o SDC	66.9	66.9	66.9	84.8	84.8	84.8
with SDC	67.9	68.1	**68.3**	85.3	85.4	**85.5**

Table 2: The ablation study of SDC on *MSMT17*. T_1 and T_2 denote two class tokens, $Cat.$ denotes the concatenation of the two class tokens.

	$Confusion \downarrow$	mAP/R1 (%) $\uparrow$
Baseline	26,469	66.1/84.6
T_1	25,738	67.9/85.3
T_2	25,064	68.1/85.4
$Cat.$	24,390	68.3/85.5

Table 3: The confusion of positive and negative pairs on *MSMT17*. $Confusion$ means the number of overlapped pairs in Figure 5.

margin especially on mAP (+2.4%) with 256×128 resolution, and also achieves the best performance with higher resolution 384×128. On *Market-1501*, our method achieves the best performance on mAP and comparable performance on Rank-1. On *CUHK03*, our method achieves absolute superiority which outperforms previous SOTA method (AAformer) by a large margin both on mAP (+4.2%) and Rank-1 (+2.5%).

Ablation Study

Experiments are conducted to study the effectiveness of self-diverse constraint loss (SDC), intra-class and inter-class distance of compressed embedding space, the hyper-parameters λ and token number N, the effectiveness of dynamic weight controller (DWC), and the effectiveness of training with a smaller amount of identity.

Impact of Multiple Class Tokens and SDC. The effectiveness of multiple class tokens is validated on *MSMT17* in Table 2. The baseline has only one class token. Adding one additional class token to the baseline (totally two class tokens) provides +0.8% mAP improvement on *MSMT17*. During training, each class token aggregates information not only from patch embeddings but also from other class tokens, which improves the information acquisition efficiency. But in further study, we find that the similarity between these two class tokens is very high, i.e., 0.999, which implies that there is no difference between them, so there is no improvement after fusion. And continuing to increase the number of tokens cannot continue to improve the performance.

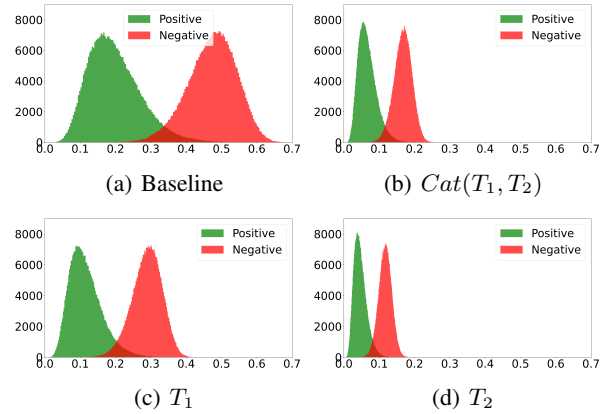With SDC imposed on these two class tokens, the cosine similarity of them becomes 0.007, which is very close to 0.0, implying that these two class tokens have learned two non-overlapping representation subspaces. Figure 3 illustrates the feature distributions of two class tokens, and they are separated to each other. Also, the performance of these two class tokens are both improved by a large margin especially on mAP, more than +1.0% improvement, which means both tokens have learned more robust features. After concatenation, feature fusion further improves the performance, reaching 68.3% mAP and 85.5% Rank-1, which is +2.2% mAP and +0.9% Rank-1 higher than the baseline. The attention map visualized in Figure 4 shows that both two tokens correctly represent the foreground part of the object. Moreover, the two tokens represent different embedding spaces, and the information they represent is also different. Compared with the baseline, two class tokens with SDC has captured more fine-grained features. The fusion of multiple class tokens helps the model to learn more discriminative representation.

Intra-class and inter-class distance. DC-Former divides the original embedding space into multiple compact subspaces, reducing intra-class distance but also reducing inter-class distance. To verify the effectiveness of the compact space proposed by DC-Former, the intra-class and inter-class distance of DC-Former's embeddings are calculated and visualized in Figure 5. And the confusion of positive and negative pairs in Figure 5 is further calculated in Table 3.
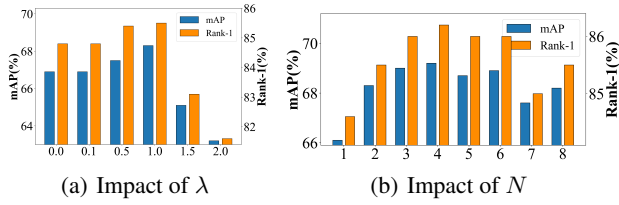
(a) Impact of λ (b) Impact of N

Figure 6: Visualization of ablation studies on *MSMT17*. Impact of two hyper-parameters of SDC.

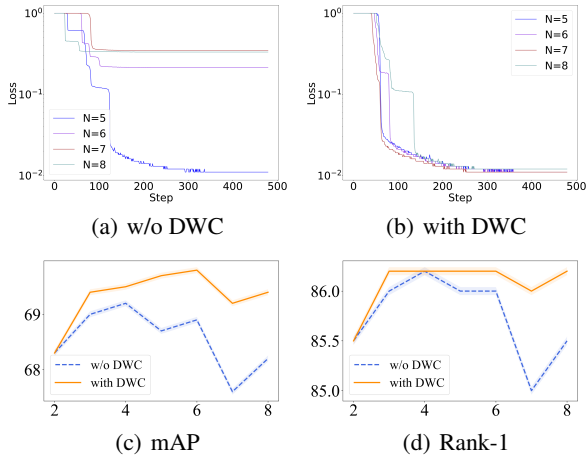

(a) w/o DWC (b) with DWC

(c) mAP (d) Rank-1

Figure 7: The evaluation of DWC on *MSMT17*. N denotes the number of class tokens. (a-b) DC-Former's SDC loss when training with/without DWC. (c-d) The mAP and Rank-1 of DC-Former with/without DWC.

Compared to baseline, the embedding space of each token in DC-Former is smaller, as is their concatenation. Moreover, the embedding space of T_2 is more compact than that of T_1, so T_2 achieves higher performance. And the confusion of DC-Former is less than baseline, which means that compact embedding space pushes the embeddings of the same class more tightly than the embeddings of different classes.

Hyper-parameters of SDC. SDC has two hyper-parameters which are the weight of loss λ and the number of class tokens N. We analyze the influence of λ on the performance in Figure 6(a). When $\lambda = 0$, baseline achieves 66.9% mAP and 84.8% Rank-1 on *MSMT17*. When $\lambda = 0.1$, SDC doesn't work because it's too small. As λ increases, the performance increases. When $\lambda = 1$, the mAP and Rank-1 are improved to 68.3% and 85.5%, respectively. When continuing to increase λ, the performance is degraded because excessive weights make the model pull apart two features at the beginning, making it difficult to optimize the classification loss. Therefore, $\lambda = 1$ is the best beneficial for learning multiple diverse features.

The experiments on the number of class tokens N is in Figure 6(b). Increasing the number of class tokens improves the performance of model. When N is between 3 and 6, mAP and Rank-1 are higher. And $N = 4$ reaches the best performance at 69.2% mAP and 86.2% Rank-1. Continuing
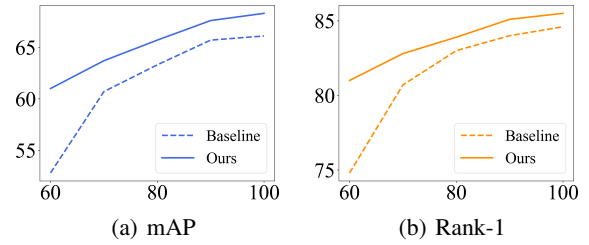

(a) mAP (b) Rank-1

Figure 8: Performance under training with a smaller amount of identity on *MSMT17*. Randomly select partial categories as training set, and keep the test set unchanged.

to increase N, performance is degraded because finding too much embedding space makes training difficult. As we can see SDC loss in Figure 7(a), too much tokens cannot optimize model well because the subspaces cannot be separated due to competition between tokens. N could be a little different in different datasets. In the SOTA experimental configuration (Table 1), N is [6,5,4] for *MSMT17*, *Market-1501* and *CUHK03*, respectively.

Dynamic Weight Controller. The effectiveness of the proposed DWC module is validated in Figure 7. The SDC loss in training phase illustrated in Figure 7(a-b) shows that DWC can make each pairs learn evenly so that they are all orthogonal to each other, which shows effectiveness on balancing the relative importance among class tokens during training as the token number increases. The performance in Figure 7(c-d) shows that DWC has limited effect when N is small (less than 5). While it provides about +1.0% improvement when N is large. When $N \geq 7$, the performance decreases because the number of subspaces in a finite embedding space has reached its limit. Too small space makes the embeddings lose their discrimination.

Effectiveness with a smaller amount of identity. We evaluate the effectiveness of DC-Former with a smaller amount of identity in Figure 8. DC-Former achieves comparable results by using less than 20% identities of baseline. And the smaller the amount of identity, the more obvious the advantages of DC-Former. Increasing the amount of identity strengthens the ability of the model to identify similar identities. And more compact embedding space of DC-Former also enhances the discrimination of similar classes' representations, which is similar to the effect of increasing the amount of identity.

Conclusions

In this paper, we propose a transformer-based network for ReID to learn multiple diverse and compact embedding subspaces, which improves the robustness of representation by increasing identity density of embedding space. And the fusion of these representations from different subspaces further improves performance. Our method outperforms previous state-of-the-arts on three person ReID benchmarks. Based on this promising results, we believe our method has great potential to be further explored in other area, and we hope it can bring new insights to the community.

References

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

Chen, G.; Lin, C.; Ren, L.; Lu, J.; and Zhou, J. 2019. Self-critical attention learning for person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9637–9646.

Chen, X.; Fan, H.; Girshick, R.; and He, K. 2020. Improved Baselines with Momentum Contrastive Learning. *arXiv preprint arXiv:2003.04297*.

Chen, X.; Xu, J.; Xu, J.; and Gao, S. 2021. OH-Former: Omni-Relational High-Order Transformer for Person Re-Identification. *arXiv preprint arXiv:2109.11159*.

Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; and Jiao, J. 2018. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 994–1003.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fu, D.; Chen, D.; Bao, J.; Yang, H.; Yuan, L.; Zhang, L.; Li, H.; and Chen, D. 2021. Unsupervised pre-training for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14750–14759.

Gong, Y.; Zeng, Z.; Chen, L.; Luo, Y.; Weng, B.; and Ye, F. 2021. A Person Re-identification Data Augmentation Method with Adversarial Defense Effect. *CoRR*, abs/2101.08783.

Gu, H.; Li, J.; Fu, G.; Wong, C.; Chen, X.; and Zhu, J. 2022. AutoLoss-GMS: Searching Generalized Margin-Based Softmax Loss Function for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4744–4753.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2019. Momentum Contrast for Unsupervised Visual Representation Learning. *arXiv preprint arXiv:1911.05722*.

He, L.; and Liu, W. 2020. Guided saliency feature learning for person re-identification in crowded scenes. In *European Conference on Computer Vision*, 357–373. Springer.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. TransReID: Transformer-Based Object Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 15013–15022.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Huynh, D.; and Elhamifar, E. 2020. Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Isobe, T.; Li, D.; Tian, L.; Chen, W.; Shan, Y.; and Wang, S. 2021. Towards Discriminative Representation Learning for Unsupervised Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 8526–8536.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lai, S.; Chai, Z.; and Wei, X. 2021. Transformer Meets Part Model: Adaptive Part Division for Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 4150–4157.

Li, H.; Wu, G.; and Zheng, W.-S. 2021. Combined Depth Space Based Architecture Search for Person Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6729–6738.

Li, W.; Zhao, R.; Xiao, T.; and Wang, X. 2014. DeepReID: Deep Filter Pairing Neural Network for Person Re-identification. In *CVPR*.

Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; and Wu, F. 2021. Diverse Part Discovery: Occluded Person Re-Identification With Part-Aware Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2898–2907.

Lin, J.; Ren, L.; Lu, J.; Feng, J.; and Zhou, J. 2017. Consistent-aware deep learning for person re-identification in a camera network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5771–5780.

Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; and Yang, Y. 2019. Improving person re-identification by attribute and identity learning. *Pattern Recognition*, 95: 151–161.

Liu, H.; Feng, J.; Qi, M.; Jiang, J.; and Yan, S. 2017. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7): 3492–3506.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. Bag of Tricks and a Strong Baseline for Deep Person Re-Identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Luo, H.; Wang, P.; Xu, Y.; Ding, F.; Zhou, Y.; Wang, F.; Li, H.; and Jin, R. 2021. Self-Supervised Pre-Training for Transformer-Based Person Re-Identification. *arXiv preprint arXiv:2111.12084*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Suh, Y.; Wang, J.; Tang, S.; Mei, T.; and Lee, K. M. 2018. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, 402–419.

Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; and Wei, Y. 2020. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6398–6407.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 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.

Tan, L.; Dai, P.; Ji, R.; and Wu, Y. 2022. Dynamic Prototype Mask for Occluded Person Re-Identification. *arXiv preprint arXiv:2207.09046*.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357. PMLR.

Wang, G.; Yuan, Y.; Chen, X.; Li, J.; and Zhou, X. 2018. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM international conference on Multimedia*, 274–282.

Wang, T.; Liu, H.; Song, P.; Guo, T.; and Shi, W. 2022. Pose-guided Feature Disentangling for Occluded Person Re-identification Based on Transformer. *AAAI*.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 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.

Xu, F.; Wang, M.; Zhang, W.; Cheng, Y.; and Chu, W. 2021. Discrimination-Aware Mechanism for Fine-Grained Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 813–822.

Yan, C.; Pang, G.; Jiao, J.; Bai, X.; Feng, X.; and Shen, C. 2021. Occluded person re-identification with single-scale global representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11875–11884.

Yang, J.; Zhang, J.; Yu, F.; Jiang, X.; Zhang, M.; Sun, X.; Chen, Y.-C.; and Zheng, W.-S. 2021. Learning To Know Where To See: A Visibility-Aware Approach for Occluded Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11885–11894.

Zhang, G.; Zhang, P.; Qi, J.; and Lu, H. 2021. *HAT: Hierarchical Aggregation Transformers for Person Re-Identification*, 516–525. New York, NY, USA: Association for Computing Machinery. ISBN 9781450386517.

Zhao, J.; Qi, F.; Ren, G.; and Xu, L. 2021. PhD Learning: Learning with Pompeiu-hausdorff Distances for Video-based Vehicle Re-Identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2225–2235.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable Person Re-identification: A Benchmark. In *Computer Vision, IEEE International Conference on*.

Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; and Tian, Q. 2017. Person re-identification in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1367–1376.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 14(1): 1–20.

Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; and Yang, Y. 2020. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13001–13008.

Zhu, K.; Guo, H.; Zhang, S.; Wang, Y.; Huang, G.; Qiao, H.; Liu, J.; Wang, J.; and Tang, M. 2021. Aaformer: Auto-aligned transformer for person re-identification. *arXiv preprint arXiv:2104.00921*.

Zhu, Z.; Jiang, X.; Zheng, F.; Guo, X.; Huang, F.; Sun, X.; and Zheng, W. 2020. Aware loss with angular regularization for person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13114–13121.