Strip-Cutmix for Person Re-Identification

Yuxiang Sun[1], Ke Qi[1*], Yicong Zhou[2], Yutao Qi[3]
[1]School of Computer Science and Cyber Engineer, Guangzhou University, Guangzhou 510006, China
[2]School of Computer and Information Science, University of Macau, Macau 999078, China
[3]School of Computer and Information Engineering, Guangzhou Huali College, Guangzhou 510000, China
sunyuxiang99@163.com, qike@gzhu.edu.cn, yicongzhou@um.edu.mo, 1826870760@qq.com

***Abstract*—Person re-identification is a very challenging image retrieval task that aims to match the specific person images from different camera views. Person re-identification model requires a large amount of training data to improve its generalization ability, however the current datasets of person re-identification are not enough that tend to make the model overfit. Therefore, some data augmentation methods are used to increase the amount of training data to improve the generalization ability of the model. Cutmix is a common data augmentation method in the field of deep learning, but it is rarely used in person re-identification task because the triple loss cannot handle the decimal similarity label generated by cutmix. In order to put the cutmix method for data augmentation in person re-identification, we extend the triplet loss that is commonly used in person re-identification to a form which can handle decimal similarity label from the perspective of optimizing image similarity. In addition, we propose Strip-Cutmix data augmentation method, which is more suitable for person re-identification, and discuss the strategies about using Strip-Cutmix in the field of person re-identification. Extensive experiments show that our approach can prevent model overfit and achieve impressive performance on DukeMTMC-ReID, Market-1501 and MSMT17 benchmark datasets.**

***Index Terms*—Person Re-Identification, Strip-Cutmix, Data Augmentation**

I. Introduction

Person re-identification is a very challenging image retrieval task which the goal is to match images of the same person in different camera views. Person re-identification models are often implemented using deep neural networks, which require a large amount of training data to ensure the generalization ability of the model. Due to existing person re- identification datasets are not large enough and are labor intensive to collect and label the data, so data augmentation methods are often used to augment the amount of training data. However, currently there are only few available data augmentation methods for person re-identification task. For this reason, we adapt triplet-loss [1] and cutmix [2] to apply the cutmix to the person re-identification task. In addition, we propose a strip-cutmix data augmentation method that is more suitable for person re-identification task.

Cutmix [2] crops a part of an image and pastes it onto another image to create a new combined image, which is a common method of data augmentation in the field of deep learning. However, cutmix is rarely used in person re-identification task because the triplet loss commonly used in person re-identification task cannot handle the decimal similarity label generated by cutmix. The triplet loss play an important role in the metric learning process for the person re-identification task. The positive and negative sample pairs of triple loss are determined according to the ground truth. The current metric learning loss function cannot handle the decimal similarity label generated by cutmix. We modify the triplet loss so that it can handle decimal similarity label, allowing the use of cutmix and triple loss together in person re-identification task. Based on the target similarity of the model output, we make two modifications to the triplet loss. First, the optimization direction is determined dynamically. If the output similarity of the network is higher than the target similarity, the output similarity of the network needs to be reduced. Otherwise, the output similarity needs to be increased. Second, the decision-making conditions of the triplet loss are rewritten from the original conditions related to the $\{0,1\}$ label to the conditions related to the target similarity label, and the modified triplet loss is also compatible with the original conditions.

One feature of person re-identification is that a whole pedestrian is contained in a single image, which allows the same class of features of the pedestrian to appear at the approximately same location in the image. Based on this feature, we propose Strip-Cutmix for person re-identification task. Strip-Cutmix cuts the image into image blocks in a horizontal manner, and we can assume that images of the same person have similar features at the same image block locations. Based on this assumption, we can treat cutmix in terms of image block combinations, where the similarity of any two images can be obtained by the corresponding image block label. Strip-Cutmix has three advantages over the original cutmix. First, it is possible to obtain a similarity between the two mixed images, but cutmix cannot. Second, strip-cutmix can be more efficient to exploit pedestrian image features to generate better quality generated images. Third, strip-cutmix can obtain better boundary conditions for triplet loss based on the combination of image blocks. Our proposed method can be used together with other proposed methods as a data enhancement method to improve the performance of the model.

Our contributes as follows:

- We propose a new method to extend loss function of deep metric learning from which is only being able to handle $\{0,1\}$ label to being able to handle decimal similarity label. This allows us to use cutmix and triplet loss together in person re-identification task.

2023 International Joint Conference on Neural Networks (IJCNN) | 978-1-6654-8867-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IJCNN54540.2023.10191865

- We propose Strip-Cutmux, which is more suitable for person re-identification task. Strip-cutmix performs a cutmix from the perspective of image block combination, so that similarity label can be obtained from any combined image.
- We investigate which types of image pairs should be included in the mini-batch for training under the person re-identification task in order to form an optimal model, and so attain the best strategy of strip-cutmix.
- Extensive experiments show that our method is significantly superior to other competitive methods and can alleviate model overfitting

II. Related Works

A. Data augmentation for person re-identification

Most person re-identification models merely refers random erasing [3] and random horizontal flip [4] as their data augmentation methods. Huang et al. [5] augmented the training data by generating occlusion sample images. Bak et al. [6] generated virtual person images under different lighting conditions. There are also approaches [7]–[9] to improve model generalization by generating images using generative adversarial networks (GAN). However, the quality of sample images generated by these methods is not good enough. Cutmix and mixup can generate high-quality images. However, cutmix [2] and mixup [10] data augmentation methods are rarely used in person re-identification task. This is because the loss of triplet used in person re-identification task cannot handle the decimal similarity label generated by cutmix. To exploit cutmix in the person re-identification task, we extend the commonly used metric learning function triplet loss to a form that can handle decimal similarity labels. In addition, we design Strip-Cutmix, which generates higher quality combine images, based on the feature of person re-identification images.

B. Deep metric learning

Deep metric learning aims to make the similarity between positive sample pairs higher and the similarity between negative sample pairs lower. Metric learning uses a neural network to learn a non-linear mapping of the input image into a one-dimensional vector, and then measures the similarity between sample pairs in terms of cosine similarity or euclidean distance. The two types of metric learning loss are proxy loss and pairwise loss. Proxy loss optimizes the similarity between samples indirectly by optimizing the similarity between each sample and its central identity. Its representative works include Proxy NCA [11]], ProxyNCA++ [12], softmax loss, cosine softmax loss [13] and so on. Pairwise loss optimises the similarity between samples directly, and it is representative work refer to triplet loss [1], quadruplet loss [14] and contrastive loss [15]. There also remain a few loss functions that keep both forms of proxy loss and pairwise loss, such as unified pairwise loss [16] and circle loss [16]. However, these loss functions are designed for 0,1 similarity labels and cannot handle decimal similarity label. They cannot be used with the data enrichment method like cutmix. Proxy loss is similar to the classification function in that it produces multiple classes of similarity, so it can be extended to a form that can handle decimal similarity labels using a weighted approach to the multi-label problem. However, pairwise loss cannot be handled in this simple way. Both proxy loss and pairwise loss are generally used in person re-identification task. Although the cutmix data augmentation method can be used in conjunction with proxy loss, the model only using proxy loss does not perform well comparing with using both proxy loss and pairwise loss. As a result, cutmix is rarely used in person re-identification task. To enable pairwise loss to fit in conjunction together with cutmix for person re-identification task and from the point of view of generating target similarities, we modifiy the triplet loss to handle decimal similarity labels and it also compatible with the original triplet loss.

C. Partial models

The images for the person re-identification task are pedestrian kinds, thus some pedestrian features (such as clothing, trousers or more) are highly correlated in terms of where they appear in the image. Some person re-identification methods exploit this property at the feature level by dividing the feature map into several parts and then optimising the similarity of the corresponding parts between images separately, such as PCB [17], MGN [18] and Pyramid [19]. However, there are two problems with the local model-based approach. Due to feature misalignment and the fact that different people have the same local features, the local feature-based methods is equivalent to train model using labels with noise. To address this problem, [17]–[19]proposed different methods of local feature segmentation and the application of loss functions to local features. PCB [17] and Aligned-ReID [20] also propose algorithms for local feature alignment. Besides, the local feature-based approach requires more computational effort comparing with the global feature-based model during the inference phase. Our proposed strip-cutmix also uses the characteristics of the pedestrian image, but we exploit it not at the feature level but at the pixel level. Comparing with models based on local feature methods, our method does not have serious problems about feature alignment as most sample pairs are clean-clean and mixed-clean type. The final features of the network output are global features and so our approach requires no additional computational effort in the inference phase.

III. Our Method

A. Extended Triplet Loss

Pairwise loss optimizes the similarity between sample pairs of images. Given an anchor image $x_a \in \mathbb{R}^{H\times W\times C}$, where H, W, C denote its height, width and number of channels respectively. The set of positive samples with the same identity as which the anchor x_a has is $P(a)$, and the set of negative samples with different identities from the anchor is $N(a)$. Pairwise loss requires that the positive sample pair similarity s_p should close to 1 and the negative sample pair similarity s_n should close to 0. A typical pairwise loss is triplet loss:

$$L_{triplet} = [s_n - s_p + m]_+ \tag{1}$$

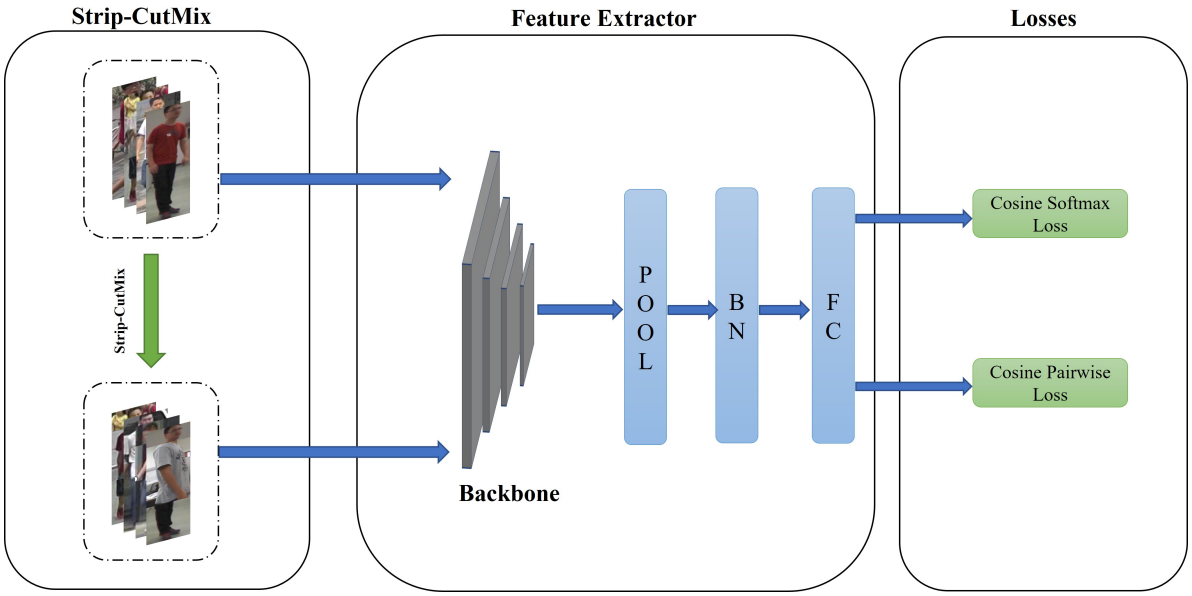

Fig. 1. Overall framework of the proposed approach. It consists of Stride-Cutmix, Feature Extractor Module and Losses. Stride-Cutmix is used for data augmentation to increase the amount of training data. Feature Extractor is used to extract features from an image, mapping the extracted features to a public embedding space. The losses component is used to bring features of the same identity closer together in the embedding space and to push features of different identities apart.

Where $[x]_+ = max(x, 0)$ denotes the standard hinge loss, m is a margin. s_p and s_n denote the similarity of positive and negative sample pairs respectively. Triplet loss contains a relaxation condition that as long as $s_p - s_n > m$, i.e. the positive sample similarity is higher than the negative sample similarity by a margin m, it is acceptable and the set of s_n and s_p will not be optimized further. The meaning of the relaxation condition is that s_n and s_p do not need to both be 0 and 1 strictly as long as the positive and negative sample pairs are all sufficiently distinguishable.

However, triplet loss cannot handle decimal similarity labels, making it impossible to use together with cutmix. First is that the optimizing direction of sample pair similarity is determined by the positive and negative sample pairs, which in turn depend on the 0,1 labels. The division of positive and negative sample pairs is represented as follows:

$$\begin{aligned} y(a,i) &\in \{0,1\}, \quad i = 1,2,\ldots,M \\ P(a) &= \{x_i, i = 1,2,...,M | y(a,i) = 1\} \\ N(a) &= \{x_i, i = 1,2,...,M | y(a,i) = 0\} \end{aligned} \tag{2}$$

where M denotes the total number of sample images and $y(a,i)$ denotes the similarity label of the anchor image and the i-th image sample. If they have the same identity, $y(a,i) = 1$ means they are positive sample pair, otherwise $y(a,i) = 0$ means they are negative sample pair. The original triplet loss does not give an idea for dividing positive and negative sample pairs for decimal similarity labels. The second reason is that the relaxation condition of triplet loss is designed on the basis of that there only exist 0,1 labels. It causes positive sample pairs to be distributed around 1 and for negative sample pairs is around 0. If the distribution aggregation around 1 is more compact, the distribution aggregation around 0 can be slightly looser, and vice versa. However, if the target similarity label of a sample is a decimal, such as 0.5, the corresponding similarity we want to get is distributed around 0.5. This is impossible for triplet loss.

In order to allow triplet loss to handle decimal similarity labels in the interval [0,1], we make two improvements to triplet loss. The first is to determine the direction of optimization of the sample pair similarity dynamically in each mini-batch. If the similarity of the network output is less than the similarity label, the similarity needs to be increased to optimize in the positive direction. Otherwise, it will be optimized in the negative direction. If the output of the network yields a sample pair that the similarity $s(a,i)$ is greater than the sample pair similarity label $y(a,i)$, the sample pair is a negative one and the similarity $s(a,i)$ has to be optimised towards 0 to reducing the similarity. If $s(a,i)$ is less than or equal to the sample pair similarity label $y(a,i)$, then the sample pair should be a positive one and the similarity $s(a,i)$ has to be optimised towards 1 to improve the similarity. So through this method for decimal similarity labels, it is also possible to divide positive and negative sample pairs for samples. Noted that our triplet loss uses cosine similarity, so the resulting similarity $s(a,i) \in [-1,1]$. The two vectors are linearly uncorrelated when the cosine similarity is 0, and negatively correlated when the cosine similarity is -1. Therefore for a sample pair with a similarity label of 0, the obtained similarity $s(a,i) \leq 0$ is acceptable. We set which meet the condition $y(a,i) = 0$ as negative samples. While $s(a,i) \leq 0$, there is no requirement to improve its similarity, which is also in terms of the compatibility with the original triplet loss. The dynamic division of positive and negative sample pairs is as follows:

$$\begin{aligned} P(a) &= \{x_i, i = 1,2,...,M | s(a,i) \leq y(a,i) \; and \; y(a,i) \neq 0\} \\ N(a) &= \{x_i, i = 1,2,...,M | s(a,i) > y(a,i) \; or \; y(a,i) = 0\} \end{aligned} \tag{3}$$

Where $P(a)$ is the set of positive sample pairs of anchor image x_a and $N(a)$ is the set of negative sample pairs. M is the total number of samples, and $y(a,i)$ refers to the similarity label of anchor image and the i-th image. $s(a,i)$ denotes the similarity between the anchor image a and the i-th image, and we use the cosine similarity to calculate it.

We have also rewritten the relaxation condition for triplet loss. Instead of only allowing the output similarity to be distributed around 0 and 1, it has been modified to allow the distribution being around the target similarity label. The new form is as follows:

$$\begin{aligned} \hat{L}_{triplet} &= [(s_n - \Delta_n) - (s_p - \Delta_p)]_+ \\ \Delta_p &= y_p(1-\hat{m}) \\ \Delta_n &= y_n(1-\hat{m}) + \hat{m} \\ \hat{m} &= 0.5 - \frac{m}{2} \end{aligned} \tag{4}$$

Where y_p denotes positive sample pair similarity label and y_n denotes negative sample pair similarity label. When $s_n < \Delta_n$ and $s_p > \Delta_p$, it's unnecessary to optimize this triplet. Since we are dividing positive and negative sample pairs dynamically, a sample pair may be either a positive or negative sample pair in a iteration. We can obtain $\Delta_p < s < \Delta_n$, while Δ_p and Δ_n are the lower and upper bounds respectively of the receptive domain of s, which allows similarity s to be distributed around the similarity label y. We set $\hat{m} = 0.5 - \frac{2}{m}$ so that when $y_n = 0$ and $y_p = 1$, the relaxation condition is compatible with the original triplet loss. The exact derivation process is given in the next section. Our extended method can also be applied with other triplet style losses, such as unified pairwise loss [16], contrastive loss [15] and multi- similarity loss [33]. Unified pairwise loss can be called cosine pairwise loss and we use it to train the model. It can be expressed as follows:

$$\begin{aligned} \mathcal{L}_{uni} = \frac{1}{M}\sum_{a=1}^{M} log[1+ \\ \sum_{p\in P(a)}\sum_{n\in N(a)} exp(\gamma((s_n-\Delta_n)-(s_p-\Delta_p)))] \end{aligned} \tag{5}$$

where γ is a scale factor. In addition, we also use cosine softmax loss [13] to train our model. It can be expressed as follows:

$$\mathcal{L}_{cos} = \frac{1}{M}\sum_{i=1}^{M} -log\frac{e^{\gamma(S(W_{y_i}^T,x_i)-m)}}{e^{\gamma(S(W_{y_i}^T,x_i)-m)} + \sum_{j\neq y_i} e^{\gamma S(W_j^T,x_i)}} \tag{6}$$

where M is the number of sample,γ is a scale factor and m is a margin term. $S(x,y)$ computes the cosine similarity between x and y. W_{y_i} is the weight vector of the y_i-th class. The total loss is expressed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{uni} + \mathcal{L}_{cos} \tag{7}$$

Our model is jointly trained in an end-to-end manner. The overall framework of our model is shown in figure 1.

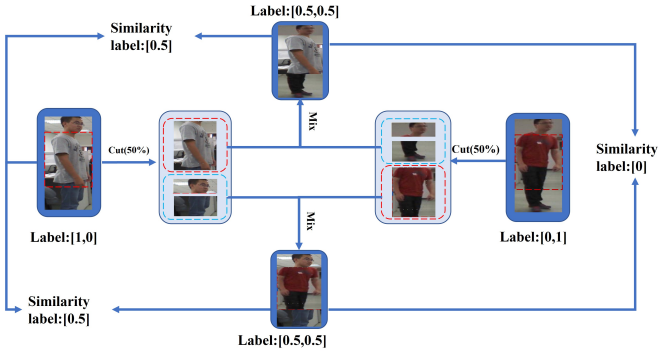

Fig. 2. The overview of Strip-Cutmix and similarity label generation process. We cut image block at the same location and generate similarity label of sample pair according to image block at the same location in each mini-batch.

B. Strip-Cutmix

The improved triplet loss can handle the decimal similarity labels generated by cutmix, however the direct use of cutmix does not work well enough. Cutmix does not make good use of the pedestrian image's characteristics that contain a complete pedestrian shape and does not obtain the similarity label between mixed images. The pedestrian image contains the whole pedestrian structure information. Therefore, pedestrian features are very much correlated with their position in the image. Some methods [17]–[20] propose local feature based methods according to this characteristic. We propose Strip-Cutmix from the point of view of combining image blocks.

Strip-Cutmix starts by determining the position and the shape of the cut in the image where we use random short-strip. Short-strip means that we select 20% - 50% of the image height for cutting. In each mini-batch, all the images are cut horizontally at the same position and the cut image blocks are shuffled and pasted back to the image. In this way each image involved consists of two blocks respectively from the original image and the pasted image. We record the identity label to which each image block belongs, so that each image can obtain two types of labels. One type of similarity label is image itself, which is assigned according to the area shares of each identity's image block, the other is the inter-image similarity label, which compares the identity labels of image blocks at the same location of two images, and the similarity label between them stands for the percentage of the both image blocks with the same identity label. It is worth mentioning that the mixed-mixed image pair remains the problem of feature misalignment while comparing with the clean-mixed image pair, so we have added an overlap factor term to alleviate this issue. There are three differences between our strip-cutmix and the original cutmix. The first is that strip-cutmix emphasises that all images are cut and mixed on the same location. The reason for doing so is to make use of the characteristics of the pedestrian image. The second is that the shape of our cut is a short-strip and the cutmix is a rectangle. Because pedestrian features are vertically distributed, using

TABLE I
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON DUKEMTMC-REID, MARKET1501 AND MSMT17

Backbone	Method	DukeMTMC-ReID		Market1501		MSMT17	
		mAP	R1	mAP	R1	mAP	R1
ResNet50 [31]	PCB [17]	66.1	81.8	77.4	92.3	-	-
	Auto-ReID [21]	-	-	85.1	94.5	52.5	78.2
	BoT [22]	76.4	86.4	85.9	94.5	50.2	74.1
	SFT [23]	79.6	90.0	87.5	94.1	58.3	79.0
	AGW [24]	79.6	89.0	87.8	95.1	55.6	78.3
	ABD-Net [25]	78.6	89.0	88.3	95.6	**60.8**	**82.3**
	FastReID [26]	79.8	89.6	88.2	95.4	58.4	81.8
	3D-SF [27]	76.1	88.2	87.3	95.0	-	-
	CDNet [28]	76.8	88.6	86.0	95.1	54.7	78.9
	AD-SO [29]	74.9	87.4	87.7	94.8	-	-
	FA-Net [30]	77.0	88.7	84.6	95.0	51.0	76.8
	ours	**81.4**	**90.6**	**89.4**	**95.5**	60.1	81.7
RegNetY-1.6GF [32]	ours	**83.6**	**92.0**	**90.8**	**96.0**	**67.6**	**85.8**

strip generates a higher probability of cutting to the same type of feature. Short-strip is used for that we consider that cutting a large chunk of the image from the middle of the origin might cut out the whole person's features, and such cutting will not achieve the purpose of mixing features. The third is that by using strip-cutmix we can give a similarity label between any two images, which is the vanilla cutmix cannot do. In figure 2, we take two pedestrian images belonging to different identity labels as an example to show our strip-cutmix and similarity tag generation process.

In addition, we derive a better upper and lower bound for triplet loss from an approach based on the combination of image blocks. For two images of the same identity label, we set its upper bound to 1 and lower bound to $(1-\hat{m})$. For an image pair with a similarity label of 0,we set the upper bound to $\hat{m}$ and the lower bound to 0. If the similarity label is y, the two images that have similar area share of y belong to the identical person and the corresponding similar part is considered as similar block. The image blocks with (1-y) area proportion belong to different people, and they are recorded as different blocks. We calculate the upper and lower bounds value of each image block separately. Similar blocks have an upper bound as y and a lower bound as $y(1-\hat{m})$. The upper bound of the different block sets as $(1-y)\hat{m}$ and the lower bound sets as 0. We can form the upper and lower bounds of these two images in the way of combining the upper and lower bounds of these two blocks and the lower bound of the image pair is $y(1-\hat{m})$ and the upper bound is $y(1-\hat{m})+\hat{m}$. At this stage, we complete the derivation of the new upper and lower bounds, which can be shown in Eq 4s.

C. Scheme for the use of Strip-Cutmix

Different schemes could make significant impacts on Strip-Cutmix. For a original image, there would form six sample pairs when using strip-cutmix, which include clean-clean (similarit label is 1), clean-clean (similarit label is 0), clean-mixed (similarit label between 0 and 1), clean-mixed (similarit label is 0), mixed-mixed(similarit label between 0 and 1) and mixed-mixed(similarit label is 0). Clean is the original image, mixed is the image obtained using strip-cutmix. In each mini-batch, we use PK sampling strategy. P denotes the quantity of pedestrian identities in each mini-batch and k denotes how many images each person has. We developed 3 schemes based on the proportion of the sample participation in strip-cutmix. In the first scheme, all p and all k take part in strip-cutmix, so that the sample pairs in each iteration are all of mixed-mixed type. In the second scheme, half p and all k take part in strip-cutmix and each iteration probably contains 25% clean-clean, 25% mixed-mixed and 50% clean-mixed sample pairs. In the third scheme, all p and half k take part in strip-cutmix so that each iteration probably contains a combination of 25% clean-clean, 25% mixed-mixed, and 50% clean-mixed (similarit label is 0) and clean-mixed (similarit label between 0 and 1) sample pairs. After experimentation we choose the third scheme and carry out strip-cutmix with a probability of P=0.8.

IV. EXPERIMENTS

A. Datasets and evaluation protocols

We evaluate the effectiveness of our proposed method on three datasets:Market-1501 [34], DukeMTMC-ReID [35] and MSMT17 [36]. Market-1501 contains 32668 images of a total of 1501 persons. DukeMTMC-ReID contains a total of 36411 images of 1812 persons from 8 cameras. MSMT17 consists of 1041 persons with a total of 32621 training set images, 11659 query set images and 82161 gallery set images.

We use mean average precision (mAP) accuracy and standard cumulated matching characteristics (CMC) curve as evaluation protocols to assess the performance of our proposed method.

B. Implementation details

We use ResNet-50 [31] pretrained on ImageNet as a backbone in our experiments. Each mini-batch is 64 in size, where both P and K are set to 8. The training data are augmented by means of random horizontal flipping, random cropping and random erasing. We use Adam optimiser with a weight decay of $1e^{-4}$. For ResNet-50 backbone, we set its initial learning rate to 0.0006. We the cosine annealing lr scheduler for learning rate tuning.

TABLE II
ABLATION STUDY OF EACH COMPONENT ON DUKEMTMC AND MARKET1501

Component	Market1501		DukeMTMC	
	R-1	mAP	R-1	mAP
Baseline	85.3	72.1	82.6	68.1
+Ex-Tri	91.3	81.6	86.0	74.2
+Strip-Cutmix	93.0	83.9	87.4	76.6
+Cosine Softmax	95.2	88.7	89.7	80.3
+Erasing	95.5	89.4	90.6	81.4

C. Comparison with state-of-the-art methods

We compare our proposed method with state-of-the-art methods on the DukeMTMC-ReID, Market1501 and MSMT17 datasets and the results are shown in Table 1. Our method is the best one when using ResNet-50 as our backbone. The mAP of our method is higher than what of ABD-Net for over 1%. After using the RegNetY-1.6GF as the backbone network, the performance of our method can be further improved. On the Market dataset, the mAP can reach 90.8% while the Rank-1 can reach 96.0%. Our method is state-of-the-art on convolutional neural networks. These results greatly outperform comparing with all the other competing methods, which further verifies the validity of our proposed approach.

D. Ablation Studies

1) The effectiveness of each component: We carried incremental validation of each component on DukeMTMC-ReID and Market-1501 datasets, while the experimental results are shown in Table 2. Baseline refers to only using cosine pairwise loss. After adding our proposed extended triplet loss method and using it together with cutmix, the mAP on the market-1501 dataset reached 81.6%. Performance can be further improved by replacing the cutmix with strip-cutmix. Cosine Softmax refers to the use of cosine softmax loss. After adding cosine softmax loss, the performance can be greatly improved. Erasing refers to random erasing. Our strip-cutmix can be used with random erasing to further improve the performance of the model. These experimental results prove that each component is effective.

TABLE III
THE RESULTS OF ABLATION EXPERIMENTS WITH DIFFERENT SCHEME

Scheme	P	Market1501	
		mAP	R1
No	-	72.1	85.3
A	0.2	**79.2**	**90.4**
	0.5	77.5	88.8
	0.8	78.0	90.4
B	0.2	75.4	87.0
	0.5	75.7	87.2
	0.8	**75.8**	**87.7**
C	0.2	79.5	89.3
	0.5	83.0	92.2
	0.8	**83.9**	**93.0**

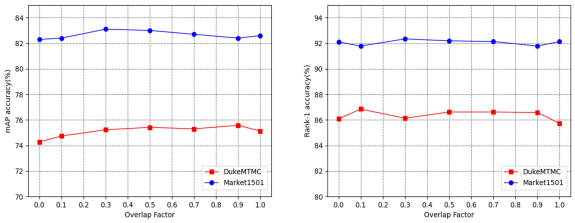

Fig. 3. The influence of the number of overlap factor on market1501 and dukemtmc dataset

2) Analysis of the strip-cutmix usage scheme: The proposed scheme for the use of strip-cutmix is analysed and experimented are shown in Table 3. No indicates strip-cutmix is not used. Scheme A indicates all p and all k are involved in strip-cutmix and scheme B indicates half p and all k are involved in strip-cutmix, while scheme C indicates all p and half k are involved in strip-cutmix. The performance obtained by using the scheme is obviously better than that obtained by not using the scheme. In addition, the probability P of participating in strip-cutmix will also affect performance. Among the three schemes, scheme C can achieve the best performance by performing strip-cutmix with a probability of P = 0.8. So we use scheme C to train the model.

3) Analysis of overlap factor parameters for mixed-mixed sample pair: We have carried out ablation experiments on the overlap factor parameter both in our proposed short-strip cut manner settings. The experimental results are shown in Figure 3. The best results can be obtained when the overlap factor sets to 0.3. In addition, the inclusion of mixed-mixed samples significantly improves the generalisability of the model, which demonstrates the effectiveness of the mixed-mixed sample pair.

TABLE IV
THE RESULTS OF ABLATION EXPERIMENTS WITH ALLEVIATE OVERFITTING

Backbone	Schedule	DukeMTMC		Market1501	
		mAP	R1	mAP	R1
Regnety-1.6GF	1x	82.8	91.2	90.7	95.9
	3x	**83.6**	**92.0**	**90.8**	**96.0**
Resnet-50	1x	80.5	90.2	88.8	94.9
	3x	**81.4**	**90.6**	**89.4**	**95.5**

4) The results of ablation experiments with alleviate overfitting: Our strip-cutmix can train the model for a long time without overfitting. We conducted experiments on regnety-1.6gf and resnet-50 backbone networks. The experimental results are shown in Table 4. 1x refers 6000 iterations and 3x refers 18000 iterations. The experimental results show that when strip-cutmix is not used, long training will lead to overfitting and performance decline. When using strip-cutmix, long time training can improve the performance of the model. The effect of using strip-cutmix on both datasets is better than that of not using strip-cutmix. When regNetY-1.6gf is used as the backbone, it can reach the state of the art performance on pure

convolutional network after a long time of training.

V. Conclusion

In this paper, in order to put the cutmix data augmentation method in use of the field of person re-identification, we have extended the triplet loss commonly used in person re-identification to a form that is able to handle decimal similarity labels. The extend triplet loss is not only compatible with the original form but also can handle decimal similarity labels. In addition, we propose strip-cutmix that is better suited to the task of person re-identification. Finally, we study the use schemeof strip-cutmix and obtain the optimal scheme. Compared with other convolutional neural network-based person re-identification models, the proposed method of our work makes it possible to achieve the best performance on pure convolutional network.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (u1936116), the innovation training program for college students of Guangzhou University (202111078028, s202011078043) and Guangzhou University Graduate Innovation Ability Cultivation Funding Program (2022GDJC-M31).

References

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

[2] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[3] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 001–13 008.

[4] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normneuralization neck for deep person reidentification," *arXiv preprint arXiv:1906.08332*, 2019.

[5] H. Huang, D. Li, Z. Zhang, X. Chen, and K. Huang, "Adversarially occluded samples for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5098–5107.

[6] S. Bak, P. Carr, and J.-F. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 189–205.

[7] J. Liu, B. Ni, Y. Yan, P. Zhou, S. Cheng, and J. Hu, "Pose transferrable person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4099–4108.

[8] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, and X. Xue, "Pose-normalized image generation for person re-identification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 650–667.

[9] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5157–5166.

[10] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[11] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 360–368.

[12] E. W. Teh, T. DeVries, and G. W. Taylor, "Proxynca++: Revisiting and revitalizing proxy neighborhood component analysis," in *European Conference on Computer Vision*. Springer, 2020, pp. 448–464.

[13] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[14] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 403–412.

[15] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

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

[17] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 480–496.

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

[19] F. Zheng, C. Deng, X. Sun, X. Jiang, X. Guo, Z. Yu, F. Huang, and R. Ji, "Pyramidal person re-identification via multi-loss dynamic training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8514–8522.

[20] X. Zhang, H. Luo, X. Fan, W. Xiang, Y. Sun, Q. Xiao, W. Jiang, C. Zhang, and J. Sun, "Alignedreid: Surpassing human-level performance in person re-identification," *arXiv preprint arXiv:1711.08184*, 2017.

[21] R. Quan, X. Dong, Y. Wu, L. Zhu, and Y. Yang, "Auto-reid: Searching for a part-aware convnet for person re-identification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3750–3759.

[22] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[23] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 1151–1157.

[24] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 6, pp. 2872–2893, 2021.

[25] T. Chen, S. Ding, J. Xie, Y. Yuan, W. Chen, Y. Yang, Z. Ren, and Z. Wang, "Abd-net: Attentive but diverse person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8351–8361.

[26] L. He, X. Liao, W. Liu, X. Liu, P. Cheng, and T. Mei, "Fastreid: A pytorch toolbox for general instance re-identification," *arXiv preprint arXiv:2006.02631*, 2020.

[27] J. Chen, X. Jiang, F. Wang, J. Zhang, F. Zheng, X. Sun, and W.-S. Zheng, "Learning 3d shape feature for texture-insensitive person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8146–8155.

[28] H. Li, G. Wu, and W.-S. Zheng, "Combined depth space based architecture search for person re-identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6729–6738.

[29] A. Zhang, Y. Gao, Y. Niu, W. Liu, and Y. Zhou, "Coarse-to-fine person re-identification with auxiliary-domain classification and second-order information bottleneck," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 598–607.

[30] Y. Liu, W. Zhou, J. Liu, G.-J. Qi, Q. Tian, and H. Li, "An end-to-end foreground-aware network for person re-identification," *IEEE Transactions on Image Processing*, vol. 30, pp. 2060–2071, 2021.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," in *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, 2020, pp. 10 428–10 436.

[33] X. Wang, X. Han, W. Huang, D. Dong, and M. R. Scott, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5022–5030.

[34] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1116–1124.

[35] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[36] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer gan to bridge domain gap for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 79–88.