PERSON RE-IDENTIFICATION WITH GRADUAL BACKGROUND SUPPRESSION

Yingzhi Tang[1], *Xi Yang*[1*], *Nannan Wang*[1], *Xinrui Jiang*[1], *Bin Song*[1], *Xinbo Gao*[2]

[1]State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, China

[2]State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, China

ABSTRACT

Person re-identification plays an important role in public security. However, owing to the interference of background clutters, its performance still needs to be improved. Several mask-based methods aim to solve this problem by totally removing the background clutters, but the promotion is limited because of the mask sharpening effect. In this paper, we propose a novel person re-identification method with **G**radual **B**ackground **S**uppression (GBS). The GBS adopts several CNN branches to extract deep features of images with different weight distributions between background and human body. Thus, it can not only reduce the background clutters but also keep the smoothness of target pedestrians. Afterwards, deep features from different CNN branches are integrated with a fusion scheme, and the fused feature is capable of balancing the influence of background clutter and mask sharpening. Extensive experiments have been conducted and the results prove the superiority of the proposed GBS over the background removal approach. Comparing with the state-of-the-art methods, our method achieves remarkable performance with 6.6%, 7.58% and 8.26% improvement of mAP on dataset Market-1501, CUHK03-labeled and CUHK03-detected, respectively.

Index Terms— person re-identification, background clutters suppression, mask sharpening

1. INTRODUCTION

Person Re-Identification (Re-ID) is a task aiming to recognize target pedestrians cross different surveillance cameras. Because of its significant application in public security, person Re-ID has attracted much attention from both academia and industry. Though it has been researched for years, person Re-ID still faces many challenges and is far from real-world application, due to low resolution images, variation of human poses and background clutters, *etc.*

Conventional methods [1, 2, 3] extract features from the entire images, which contain many background clutters. There are two general solutions to suppress the impact of background clutters. One solution is learning the identity features from the body partitions [4, 5] which are generated by the pose or keypoint estimation methods [6] . Another solution is getting the human body by mask segmentation. With the deep-learning based instance segmentation methods like FCN [7] and Mask RCNN [8], we can obtain much better body mask which is able to remove background clutters in pixel level.

One straightforward way to use mask is cutting the human body region with mask segmentation in image-level, as shown in Fig. 1. It is supposed that masked images should perform better than the entire images, however, in our experiments, we found that the masked images even have worse performance than the entire images (rank 1 accuracy on Market-1501 are 82.15% on masked images and 86.91% on entire images, which drops 4.76%). Since the RGB values will be dramatically changed at the mask line, the image's structure and smoothness are destroyed. Hence, mask sharpening happens when acquiring human body by this hard manner, which seriously depresses the person Re-ID model's performance. Song *et al.* proposed a MGCAM [9] to remove the background clutters in the feature level, however, the triplet-net based MGCAM [9] framework costs much more time and has limited promotion.

To solve this problem, we propose a novel person Re-ID method with **G**radual **B**ackground **S**uppression (GBS). Firstly, in the image preparation process, we distribute different weights between background and human body region, *e.g.* 0.6 weight value for background, 1 weight value for human body region. Then we utilize several CNN branches to train and extract person features with different weight distributions between human body region and background. Afterwards, a feature fusion scheme is proposed to generate integrated feature. In practice, the distances between features extracted from different weight distributions in query image and gallery images are first computed and then fused as one

[*]Corresponding author: Xi Yang (yangx@xidian.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (61602355, 61432014, 61772402, U1605252, 61671339, 61876142), the National Key Research and Development Program of China (2016QY01W0200), the National High-Level Talents Special Support Program of China (CS31117200001), the China Post-Doctoral Science Foundation (2018T111020, 2016M590926), the Shaanxi Province Natural Science Foundation (2017JQ6007), the Young Talent fund of University Association for Science and Technology in Shaanxi, China (20180104), the Shaanxi Province Post- Doctoral Science Foundation, the CCF-Tencent Open Fund, and the Fundamental Research Funds for the Central Universities (XJS17086).

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00127

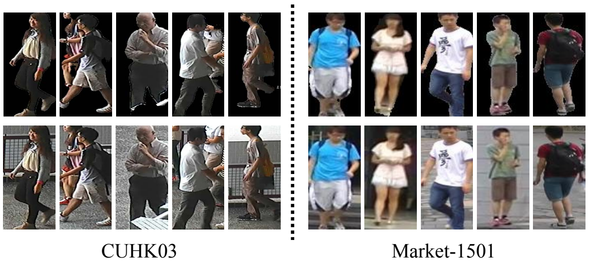

Fig. 1. The masked images (top) and the entire images (bottom) of CUHK03 (left row) and Market-1501 (right row), respectively.

distance. With this manner, the human body region can be reinforced and the background clutters can be suppressed. Additionally, because the entire images features are covered in our GBS, it also achieves favorable robustness with the condition when mask segmentation is incorrect.

2. RELATED WORK

In this section, we will first review several related works on deep-learning based person Re-ID, and then introduce some instance segmentation methods.

Deep-learning based Person Re-ID. Person Re-ID has attracted much attention in recent years. Because traditional methods exhibit poor feature representation ability, numerous deep-learning person Re-ID methods have been proposed and perform well on person Re-ID tasks. According to the understanding of image components, person Re-ID tasks can be divided into two categories, *i.e.* **Global** feature learning and **Local** feature learning. In global feature learning, numerous methods [1, 2, 3] extracted the entire image features containing too much background clutters. However, background clutters impact the person Re-ID performance seriously. The local feature learning methods are capable to solve this problem. There are two manners to do local feature learning, one is separating images to background and human body, another is separating person into several body partitions. In the former one, Tian *et al.* [10] proposed a method to eliminate the background-bias to do robust person Re-ID. Song *et al.* [9] utilized a triplet net to remove background clutters in feature level. As for the latter one, Wei *et al.* proposed GLAD [4] to separate human body into three parts to get body partition features. Zhao *et al.* [5] proposed Spindle Net which assigned human body to seven sets based on fourteen body keypoints.

Instance Segmentation. Instance segmentation plays an important role in separating images into background and human body region. Until now, only several works [11, 10] introduce mask segmentation into person Re-ID tasks because the traditional segmentation methods have poor performance and are computational consuming. Recently, some deep-learning based segmentation methods like Fully Convolutional Networks (FCN) [7], Mask RCNN [8] and some human parsing methods have been proposed [12, 11], with which we can get much better body masks.

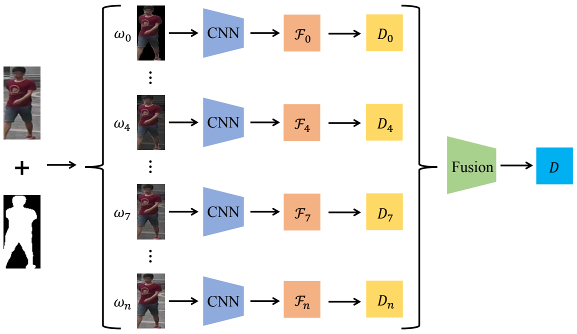

Fig. 2. The framework of proposed GBS, including n CNN branches and the $Fusion$ component . $\omega = (\lambda_1, \lambda_2)$, λ_1 and λ_2 represent the weight of human body and background, respectively. $\mathcal{F}$ represents the image feature, D in yellow is the distance tensor computed from the $\mathcal{F}$ and the D in blue is the final distance tensor fused by yellow Ds.

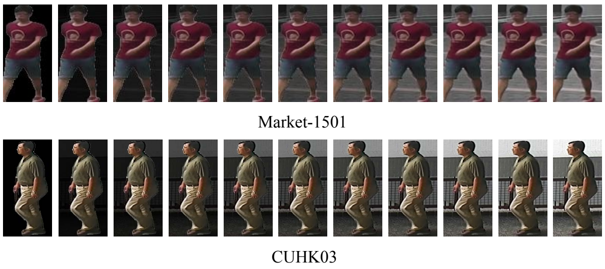

Fig. 3. The samples of fusion pictures in Market-1501 and CUHK03, both identities are selected from the training set. In addition, the λ_2 values is 0 to 1 from left to right, increase step is 0.1.

3. PROPOSED METHOD

3.1. Method Overview

As mentioned before, the smoothnesses and structures of images will be destroyed when only uses the masked images, which seriously impact the performance. We consider that in masked images, the RGB values will suddenly change from 0 to other non-zero values in the human body mask line. The dramatic change makes CNNs pay more attention on the mask and less on the human body region. To smooth the dramatic change, we propose a novel person Re-ID method with gradual background suppression with two main components. Firstly, learning and extracting the feature of distributions with different weights between background and human body. Secondly, the different features will be fused to reduce computation consumption and information redundancy for metric.

The structure of GBS is shown in Fig. 2. The image with its mask is used to generate several images with different weights ω, and then the images are inputted into CNN branches to extract features and compute distances. At the last, the distances are fused as one distance.

3.2. Impact of Background

With full background. The image's structure and smoothness can be preserved when trained with full background. The performance of these person Re-ID models is fair, meanwhile, lots background clutters are also retained which seriously impact the performance.

Without background. The RGB values nearby the mask are changed dramatically, image's structure and smoothness will be destroyed, and then results in the decline on person Re-ID accuracy.

With part background. As shown in Table. 1 2 3, we found that the performance of person Re-ID model is improved sometimes, due to the background clutters are suppressed in some degree. We attempt to combine the images distances of masked images and entire images and achieve the rank 1 and mAP accuracy of 88.77% and 70.73% respectively, which means the combine manner works. The results shows that if the background clutters can be removed and the image's structure and smoothness can be preserved in the meanwhile, the performance of person Re-ID models have great potential to be improved. To solve this problem, the background clutters should be suppressed gradually and the human body region needs reinforcement.

3.3. Gradual Background Suppression

The first step of GBS is distributing different weights between background and human body region. With the person mask, the image will be separated into two partitions, the background and the human body region. We distribute weights between background and the human body region respectively and then combine the two partitions as one image.

$$FPicture = \lambda_1 HumanBody + \lambda_2 Background \tag{1}$$

Where $FPicture$ represents the image fused by human body region and background, $HumanBody$ and $Background$ represent the tensor of human body region and background, respectively. λ_1 and λ_2 are the weights of human body region and background, respectively. Follow Eq. 1, we distribute k weights between background and human body region. These weights are not certain and it should be decided by different conditions, which needs further research. The samples of generated images are shown in Fig. 3.

λ_2	Market-1501			
	rank-1	rank-5	rank-10	mAP
a	86.91	94.80	96.64	68.17
b	87.50	94.69	96.82	69.22
c	86.28	94.53	96.55	67.78
d	86.34	94.41	96.08	67.15
e	85.99	93.85	95.90	66.15
f	83.99	93.46	95.66	64.85

Table 1. The rank-1, rank-5, rank-10 and mAP (%) results of different λ_2 values on Market-1501, the λ_1 is set as 1. Where a to f means 1 to 0.5, step is 0.1, and this is the setting for all tables.

For feature learning and extraction, we design a framework with k CNN branches, where k is the counts of the distribution weights. There are varieties of CNN structures introduced to learn and extract features for person Re-ID tasks, $e.g.$ VGGNet, Resnet-50, and Densenet-121, we select Resnet-50 [13] as the CNN backbone. The different weights datasets will be trained separately, and there will be k extracted features which are 2048 dim. The features for metric will be too long which is computation consuming and information redundance if just simply concat the k features, and will seriously impact the person Re-ID performance. To solve this problem, we propose a new scheme to fusion the features.

3.4. Distance Fusion

There are two main processes to fuse the features, $i.e.$ the distance computing and distance fusion. There are many proposed metric methods for distance computing like KISSME [14], XQDA [1], Re-ranking [2], to prove our proposed GBS's effectiveness, we utilize the simplest Euclidean for distance computing.

$$G = \{F_1, F_2, F_3, \cdots, F_N\} \tag{2}$$

$$q = \{f_1, f_2, f_3, \cdots, f_n\} \tag{3}$$

Suppose G as the Gallery and q as the query, as shown in Eq. 2 3, and F_N represents the feature of the N-th image in Gallery, f_n is the feature of the n-th image in query, there are N and n images in Gallery and query, respectively.

For each image in query, we compute the distances between it and each image in gallery by Eq. 4 to obtain its distance tensor.

$$d = \|f - F\|_2 \tag{4}$$

$$D = \{d_1, d_2, d_3, \cdots, d_N\} \tag{5}$$

To reduce computational consumption and information redundancy, we fuse the k features in distance level by summing the distance tensors together. To prove the effectiveness

λ_2	CUHK03-detected			
	rank-1	rank-5	rank-10	mAP
a	42.28	63.35	73.57	39.04
b	45.14	66.35	74.78	41.37
c	45.85	66.42	74.71	41.35
d	43.64	64.28	75.00	39.82
e	42.79	65.14	73.71	39.32
f	42.42	62.21	71.42	38.56

Table 2. The rank-1, rank-5, rank-10 and mAP (%) results of different λ_2 values on CUHK03-detected, the λ_1 is set as 1.

λ_2	CUHK03-labeled			
	rank-1	rank-5	rank-10	mAP
a	46.50	67.35	76.28	43.49
b	46.35	67.35	75.85	42.46
c	46.35	65.64	75.64	42.33
d	47.57	67.28	76.85	44.02
e	47.07	68.50	76.28	43.13
f	47.71	67.64	76.00	43.10

Table 3. The rank-1, rank-5, rank-10 and mAP (%) results of different λ_2 values on CUHK03-labeled, the λ_1 is set as 1.

of GBS, we utilize this simple scheme, there are still much space for research in this domain.

$$D = \sum_{n=1}^{N} D_n \tag{6}$$

Where D is the fused distance we use for person Re-ID metric.

4. EXPERIMENTS

4.1. Datasets and Evaluation Protocol

We select two widely used person Re-ID datasets to evaluate the proposed GBS, including CUHK03 [15] and Market-1501 [16].We adopt the new train and test protocol CUHK03-NP [2], which contains 767 identities for training and 700 identities for testing. In DPM-detected manner, there are 7368 images in training set, 5328 images in testing set and 1400 images as query. In Manually manner, there are 7365 images in training set, 5332 images in testing set and 1400 images as query. In Market-1501, there are 12936 images of 751 pedestrians in training set, 19732 images of 750 pedestrians in testing set and 3368 images as query. We adopt the cumulative matching characteristic (CMC) and mean Average Precison (mAP) as evaluate protocol. And we use the masks proposed in [9].

GBS	Market-1501			
	rank-1	rank-5	rank-10	mAP
a	86.91	94.80	96.64	68.17
a+b	89.69	95.60	97.41	72.89
a+b+c	90.38	96.05	97.53	74.23
a+b+c+d	90.49	96.02	97.59	74.73
a+b+c+d+e	90.49	96.02	97.65	74.74
a+b+c+d+e+f	90.83	96.11	97.44	74.77

Table 4. The rank-1, rank-5, rank-10 and mAP (%) results of different GBS assembles on Market-1501.

GBS	CUHK03-labeled			
	rank-1	rank-5	rank-10	mAP
a	46.50	67.35	76.28	43.49
a+b	50.85	70.07	78.92	47.31
a+b+c	51.92	70.57	80.14	48.85
a+b+c+d	53.64	71.50	79.85	50.19
a+b+c+d+e	53.64	72.50	80.85	50.77
a+b+c+d+e+f	54.28	72.64	81.00	51.07

Table 5. The rank-1, rank-5, rank-10 and mAP (%) results of different GBS assembles on CUHK03-labeled.

4.2. Implementation Details and Feature Learning

For each CNN branch, we implement an ID-discriminative Embedding [17] framework with Resnet-50 [13] as backbone. The batchsize, epoch and learning rate of pre-trained Resnet-50 are 64, 60 and 0.01. The learning rate of modified fully-connected layer is set as 0.1. SGD is used as optimizer during training process.

4.3. Evaluation

4.3.1. Evaluate the effect of background

As shown in Table. 1 2 3, when $\lambda_2 < 1$ there are always datasets have better performance than the origin datasets.

GBS	CUHK03-detected			
	rank-1	rank-5	rank-10	mAP
a	42.28	63.35	73.57	39.04
a+b	48.14	68.50	78.00	44.51
a+b+c	49.35	70.64	79.21	46.29
a+b+c+d	49.57	70.78	79.57	46.98
a+b+c+d+e	50.35	71.14	80.35	47.17
a+b+c+d+e+f	50.71	70.50	79.85	47.30

Table 6. The rank-1, rank-5, rank-10 and mAP (%) results of different GBS assembles on CUHK-detected.

Methods	Market-1501	
	rank-1	mAP
BOW [16]	34.40	14.09
LOMO [1]	43.8	22.20
Re-rank [2]	77.11	63.63
SVDNet [18]	82.30	62.10
DaF [3]	82.30	62.10
MGCAM [9]	83.79	74.33
GBS	**90.83**	**74.77**

Table 7. Comparison with the state-of-the-art methods on Market-1501. Evaluated by Rank-1 and mAP (%).

Methods	CUHK03-labeled	
	rank-1	mAP
BOW [16]	7.93	9.29
LOMO [1]	14.80	13.60
Re-rank [2]	38.10	40.30
SVDNet [18]	43.00	40.50
DaF [3]	27.50	31.50
MGCAM [9]	50.14	50.21
GBS	**54.28**	**51.07**

Table 8. Comparison with the state-of-the-art methods on CUHK03-labeled. Evaluated by Rank-1 and mAP (%).

On Market-1501, when $\lambda_2 = 0.9$, it gets rank 1 and mAP accuracy 87.50% and 69.22% respectively. While the origin dataset achieves rank 1 and mAP accuracy 86.91% and 68.17% respectively. On CUHK03-detected, only when $\lambda_2 = 0.5$, it has worse rank 1 and mAP performance than origin dataset. On CUHK03-labeled, it has better rank 1 performance when $\lambda_2 = 0, 7, 0.6, 0.5$, and it has better mAP performance when $\lambda_2 = 0.7$.

We can conclude that background clutters really impact the person Re-ID performance and distributing less weight for background can improve the performance of person Re-ID models.

4.3.2. *Evaluate the effect of GBS*

In our experiments, we found that the performance will drop when $\lambda_2 \leq 0.4$ on both Market-1501 and CUHK03 because the mask sharpening is serious. There may be different results on other datasets which need further research and experiments. We test the GBS with five different weight assembles, start with $\lambda_2 = 1$, reduce step is 0.1, *e.g.* models consist of $\lambda_2 = 1$ and $\lambda_2 = 0.9$, models consist of $\lambda_2 = 1$, $\lambda_2 = 0.9$ and $\lambda_2 = 0.8$.

As shown in Table. 4 5 6, the performance of person Re-ID increase with the gradually adding models of less weight of background. On Market-1501, we finally get rank 1 and mAP accuracy 90.83% and 64.85% respectively, which gains rank 1 accuracy 3.92% and mAP accuracy 6.6% improvement. On CUHK03-labeled, we finally get rank 1 and mAP accuracy 54.28% and 51.07% respectively, which gains rank 1 accuracy 7.78% and mAP accuracy 7.58% improvement. On CUHK03-detected, we finally get rank 1 and mAP accuracy 50.71% and 47.30% respectively, which gains rank 1 accuracy 8.43% and mAP accuracy 8.26% improvement.

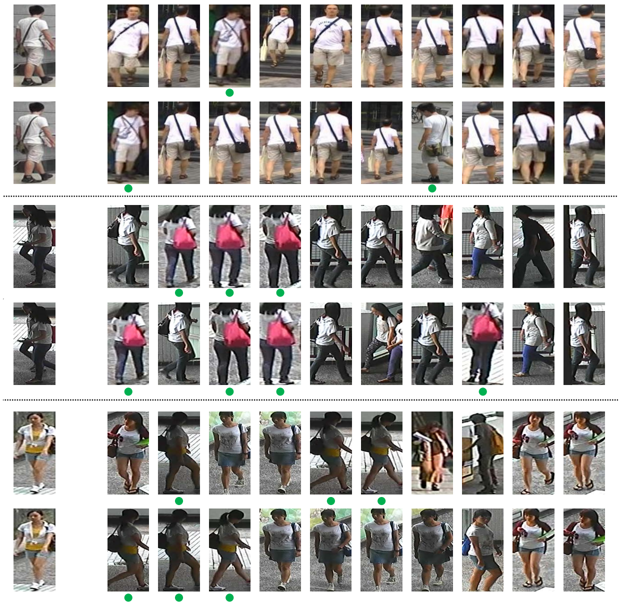

Fig. 4. The results comparison between baseline (entire images) and GBS on Market-1501 (first group), CUHK03-detected (second group) and CUHK-labeled (third group). Image with green dot below means it is correct.

Some results comparison between the entire images and the masked images are shown in Fig. 4, with GBS the model can retrieval correct results at Rank 1 and more correct results in Rank 10. For example, in the second group, GBS can retrieval the correct result at Rank 1 and get four correct results in Rank 10 while the baseline gets correct result at Rank 2 and only obtain three correct results in Rank 10. We can conclude that the person Re-ID performance can be improved by gradually suppress the background.

4.3.3. *Compare with the State-of-the-art Methods*

As shown in Table. 7 8 9, we compare the GBS with several state-of-the-art methods on Market-1501 and CUHK03, including the BOW [16], LOMO [1], Re-rank [2], SVDNet [18], DaF [3] and MGCAM [9]. On Market-1501, the GBS achieves rank 1 and mAP accuracy 90.83% and 74.77% respectively. It is higher rank 1 accuracy 7.04% and mAP accuracy 0.44% than the second best results. On CUHK03-

Methods	CUHK03-detected	
	rank-1	mAP
BOW [16]	6.36	6.39
LOMO [1]	12.80	11.50
Re-rank [2]	34.70	37.40
SVDNet [18]	40.70	37.00
DaF [3]	26.40	30.00
MGCAM [9]	46.71	46.87
GBS	**50.71**	**47.30**

Table 9. Comparison with the state-of-the-art methods on CUHK03-detected. Evaluated by Rank-1 and mAP (%).

labeled, the GBS achieves rank 1 and mAP accuracy 54.28% and 51.07% respectively. The result is higher rank 1 accuracy 4.14% and mAP accuracy 0.86% than the second best results. On CUHK03-detected, the GBS achieves rank 1 and mAP accuracy 50.71% and 47.30% respectively. The rank 1 and mAP accuracy are higher 4% and 0.43% respectively than the second best results. The results demonstrate that GBS can greatly improve the person Re-ID performance and outperform the state-of-the-art methods.

5. CONCLUSION

In this paper, we propose a novel person Re-ID method which can greatly improve the performance by gradually suppressing background clutters. We distribute different weights between background and human body region, and then use several Resnet-50 branches to learn and extract the deep features. Extensive experiments have shown that the proposed GBS is effective and achieves the state-of-the-art results. Meanwhile, GBS has much potential to improve person Re-ID performance furthermore. For instance, how to make weight distributions be adaptive to background and human body region among different datasets, and design a better manner to fuse features.

6. REFERENCES

[1] S. C. Liao, Y. Hu, and et al., "Person re-identification by local maximal occurrence representation and metric learning," in *CVPR*, 2015, pp. 2197–2206.

[2] Z. Zhong, L. Zheng, and et al., "Re-ranking person re-identification with k-reciprocal encoding," in *CVPR*, 2017, pp. 1318–1327.

[3] R. Yu, Z. C. Zhou, and et al., "Divide and fuse: A re-ranking approach for person re-identification," in *BMVC*, 2017.

[4] L. H. Wei, S. L. Zhang, and et al., "GLAD: global-local-alignment descriptor for pedestrian retrieval," in *ACM MM*, 2017, pp. 420–428.

[5] H. Zhao, M. Tian, and et al., "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *CVPR*, 2017, pp. 907–915.

[6] T. Simon, H. Joo, and et al., "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017, pp. 1145–1153.

[7] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.

[8] K. M. He, G. Gkioxari, and et al., "Mask r-cnn," in *ICCV*, 2017, pp. 2980–2988.

[9] C. F. Song, Y. Huang, and et al., "Mask-guided contrastive attention model for person re-identification," in *CVPR*, 2018, pp. 1179–1188.

[10] M. Q. Tian, S. Yi, and et al., "Eliminating background-bias for robust person re-identification," in *CVPR*, 2018, pp. 5794–5803.

[11] C. F. Song, Y. Z. Huang, and et al., "1000fps human segmentation with deep convolutional neural networks," in *ACPR*, 2015, pp. 474–478.

[12] Z. F. Wu, Y. Z. Huang, and et al., "Early hierarchical contexts learned by convolutional networks for image segmentation," in *ICPR*, 2014, pp. 1538–1543.

[13] K. M. He, X. Y. Zhang, and et al., "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[14] M. Koestinger, M. Hirzer, and et al., "Large scale metric learning from equivalence constraints," in *CVPR*, 2012, pp. 2288–2295.

[15] W. Li, R. Zhao, and et al., "Deepreid: Deep filter pairing neural network for person re-identification," in *CVPR*, 2014, pp. 152–159.

[16] L. Zheng, L. Y. Shen, and et al., "Scalable person re-identification: A benchmark," in *ICCV*, 2015, pp. 1116–1124.

[17] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *CoRR*, vol. abs/1610.02984, 2016.

[18] Y. F. Sun, L. Zheng, and et al., "Svdnet for pedestrian retrieval," in *ICCV*, 2017, pp. 3800–3808.