FLIPREID: CLOSING THE GAP BETWEEN TRAINING AND INFERENCE IN PERSON RE-IDENTIFICATION

Xingyang Ni *Esa Rahtu*

Tampere University, Finland

ABSTRACT

Since neural networks are data-hungry, incorporating data augmentation in training is a widely adopted technique that enlarges datasets and improves generalization. On the other hand, aggregating predictions of multiple augmented samples (*i.e*., test-time augmentation) could boost performance even further. In the context of person re-identification models, it is common practice to extract embeddings for both the original images and their horizontally flipped variants. The final representation is the mean of the aforementioned feature vectors. However, such scheme results in a gap between training and inference, *i.e*., the mean feature vectors calculated in inference are not part of the training pipeline. In this study, we devise the FlipReID structure with the flipping loss to address this issue. More specifically, models using the FlipReID structure are trained on the original images and the flipped images simultaneously, and incorporating the flipping loss minimizes the mean squared error between feature vectors of corresponding image pairs. Extensive experiments show that our method brings consistent improvements. In particular, we set a new record for MSMT17 which is the largest person re-identification dataset. The source code is available at `https://github.com/nixingyang/FlipReID`.

Index Terms— Person re-identification, test-time augmentation

1. INTRODUCTION

The purpose of person re-identification is retrieving the person of interest across multiple cameras, based on either images or videos [1]. While large-scale datasets [2, 3, 4] have been collected, academic research is progressing in three major directions: feature representation learning, deep metric learning, and ranking optimization.

Feature representation learning refers to developing strategies for feature construction [1]. Early works only extract a global representation for each person image [5, 6]. Later on, incorporating local features from body parts/regions has been proven beneficial [7, 8]. In addition, Lin *et al*. [9] propose leveraging the auxiliary attributes, while Wei *et al*. [4] generate synthetic images to reduce the performance drop when training and testing on different datasets.

Deep metric learning studies objective functions which are utilized to optimize neural networks [1]. With the identity loss, a model classifies the identities, and each identity is a distinct class [5, 6]. By comparison, the verification loss exploits the pairwise relationship between samples. Varior *et al*. [7] adopt the contrastive loss function for learning the embeddings, and Zheng *et al*. [10] use the binary verification loss by feeding positive and negative pairs. Lastly, the triplet loss is based on the assumption that the distance between the positive pair should be smaller than the negative pair [11].

Ranking optimization is a post-processing step that refines the retrieved ranking list [1]. Liu *et al*. [12] devise a method that requires only one negative feedback. Wang *et al*. [13] propose a hybrid model that learns cumulatively from users' feedback, and it is scalable to large-scale gallery sets. Since human interaction is time-consuming, Zhong *et al*. [14] opt for a fully automated solution instead. The pairwise distance between query and gallery samples is updated by comparing their k-reciprocal nearest neighbors.

Apart from person re-identification, data augmentation is the other subject of this study. Transformations are applied on the samples while preserving the labels, thus avoiding the need for re-annotating [15]. It is widely adopted in the training procedure to diversify samples, reduce overfitting and improve model robustness. In the mixup [16] work, models are trained on convex combinations of pairs of samples and their labels. Cubuk *et al*. [17] propose a search algorithm that finds the best data augmentation policies for the task at hand. The learned policies perform better than manually designed policies and are transferable between datasets. Optionally, data augmentation can be utilized in the inference procedure as well, *i.e*., one could generate predictions of multiple augmented samples and aggregate them into a more accurate prediction. Employing test-time augmentation improves performance at the cost of extra computations.

For image classification models such as AlexNet [18] and ResNet [19], a 10-crop testing method is applied. Five patches are extracted from the original image, *i.e*., four corner patches and one center patch. Another five patches are obtained from the horizontal reflection. Since the output of the last dense layer with softmax activation represents the probabilities of classes, it is trivial to use the mean of these ten patches' predictions.

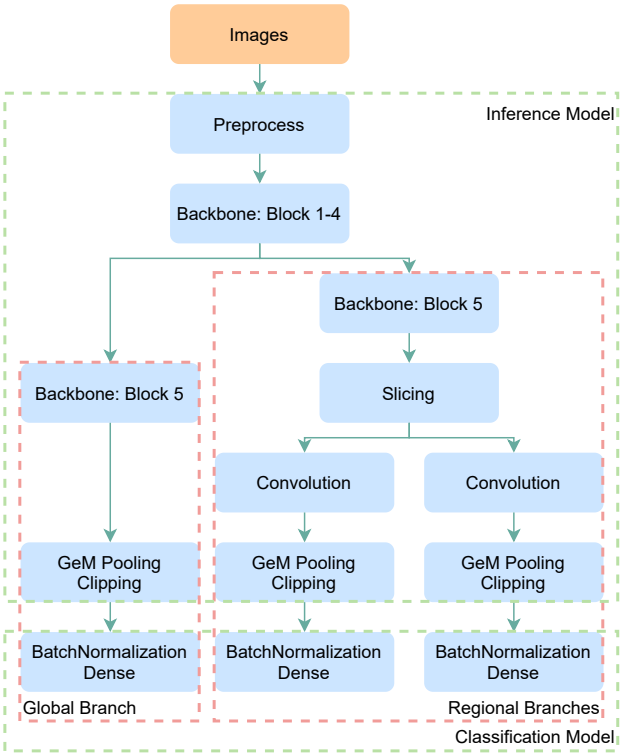

Fig. 1. Overall structure of the baseline method.

For person re-identification models such as MGN [20] and FastReID [21], features are extracted from both the original images and their horizontally flipped variants. Afterward, the mean of these feature vectors is used for evaluation. Even though this practice brings performance improvements, it lacks a thorough explanation for aggregating feature vectors by calculating the mean. More specifically, such mean feature vectors are not involved in optimizing the model, and calculating the mean in inference deviates from the training objectives. As a consequence, there exists a gap between training and inference.

In this study, we examine test-time augmentation in person re-identification. Our contribution is twofold:

- We identify a largely neglected issue, *i.e.*, utilizing the mean feature vectors of multiple augmented samples for evaluation results in suboptimal performance due to the gap between training and inference.
- We devise the FlipReID structure with the flipping loss. For models using the FlipReID structure, the original images and the flipped images are both used for training. Additionally, incorporating the flipping loss minimizes the mean squared error between feature vectors of corresponding image pairs. The resulting method achieves state-of-the-art performance on popular person re-identification datasets.

2. PROPOSED METHOD

2.1. Baseline

We leverage the method described in [22] as the baseline since it achieves high performance among recent works in person re-identification. Figure 1 illustrates the overall structure of the baseline method, and its essential components are explained as follows.

Backbone. Given images with pixel values in $\{0, \dots, 255\}$, the pre-processing layer scales pixels to $[0, 1]$ and normalizes each channel. The backbone model is an image classification model pre-trained on ImageNet [23], *e.g.*, ResNet [19], IBN-ResNet [24] and ResNeSt [25]. Due to its pyramidal architecture, the backbone model can be separated into consecutive blocks, and higher blocks refine feature maps extracted by lower blocks.

Global Branch. Following the last block of backbone, the Generalized-Mean (GeM) pooling [26] layer learns a trainable power parameter, and it generalizes the max and average operations. The subsequent clipping layer performs element-wise value clipping, which introduces constraints on feature vectors. Lastly, the batch normalization [27] and dense layers produce the probabilities of classes.

Regional Branches. The regional branches differ from the global branch in two aspects. On the one hand, the slicing layer [8] divides the feature maps along the height axis. For example, it outputs the upper and lower stripes in the case of using two partitions. On the other hand, the following convolution layer reduces the number of channels so that the resulting feature vector is not excessively long.

Objective Function. In the training procedure, the objective function is a weighted sum of batch hard triplet loss [11] and categorical cross-entropy loss. The batch hard triplet loss utilizes the hardest positive and negative samples within the batch when forming the triplets, and it is applied to the clipping layer's output. By contrast, the categorical cross-entropy loss measures the difference between the ground truth and predicted probability distributions, and it is applied to the dense layer's output.

Sub-Models. In the inference procedure, the outputs of the clipping layers from both global branch and regional branches are concatenated to get the embedded feature vector. From another perspective, the pipeline starting from the pre-processing layer to the clipping layers can be viewed as the inference model, while the remaining batch normalization and dense layers constitute the classification model.

Mini-Batch. Due to the nature of the batch hard triplet loss [11], each mini-batch comprises samples from both the same and different identities so that the positive and negative exemplars can be selected. To alleviate the overfitting issue, random horizontal flipping, random grayscale patch replacement [28] and random erasing [29] are used as the data augmentation policies.

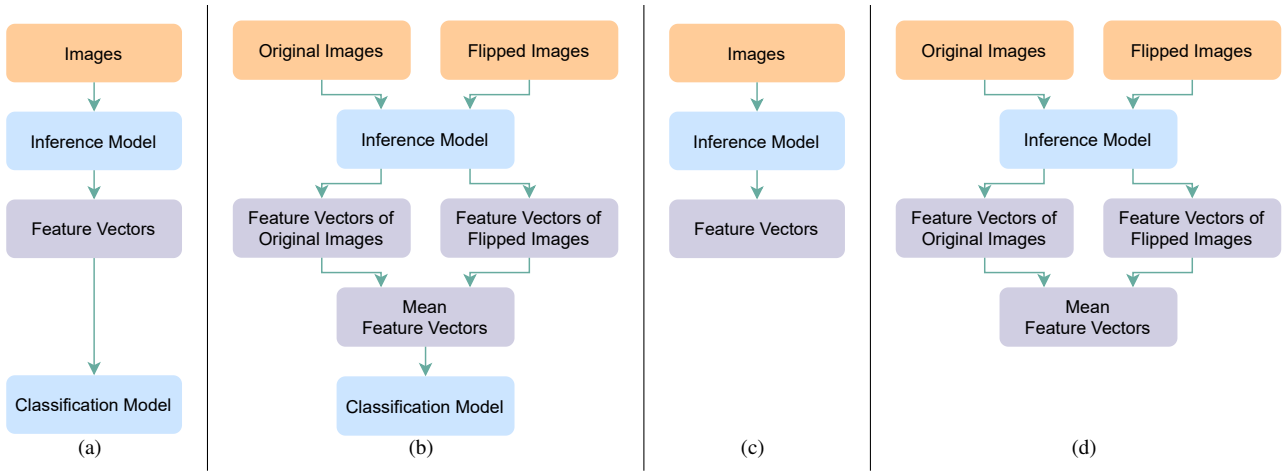

Fig. 2. Overview of the training and inference procedures in different settings: (a) training in the baseline method; (b) training in the FlipReID method; (c) inference using single image; (d) inference using double images.

2.2. FlipReID

Training. Figure 2(a) visualizes a higher level of abstraction of Figure 1. In the baseline method, random horizontal flipping is adopted for data augmentation. However, only one image will be used for one sample at a time, *i.e.*, either the original image or the flipped image. By contrast, Figure 2(b) shows the diagram of the FlipReID method. Both the original images and the flipped images are feed into the inference model, and the classification model takes in the mean feature vectors. As an optional loss term, we introduce the flipping loss, which calculates the mean squared error between the feature vectors of the original images and the flipped images.

Inference. Figure 2(c) and 2(d) show the inference pipeline using single image and double images, respectively. In Figure 2(c), data augmentation is disabled, and feature vectors are extracted from the original images. By comparison, horizontally flipped images are used in Figure 2(d), and the final representation is the mean feature vectors of the original images and horizontal reflections.

Options for Training and Inference. Figure 2(a) and 2(b) provide two choices for training, while Figure 2(c) and 2(d) offer two inference strategies. This gives the following four options for training and inference:

- Figure 2(a), 2(c): Inference is consistent with training, and the classification model is discarded in inference.
- Figure 2(a), 2(d): There exists a gap between training and inference, *i.e.*, the classification model is only trained on feature vectors of single image.
- Figure 2(b), 2(c): There exists a gap between training and inference, *i.e.*, the classification model is only trained on feature vectors of double images.
- Figure 2(b), 2(d): Inference is consistent with training, and the classification model is discarded in inference.

3. EXPERIMENTS

3.1. Datasets

Experiments are carried out on three person re-identification datasets, *i.e.*, Market-1501 [2], DukeMTMC-reID [3] and MSMT17 [4].

Market-1501. It consists of 32,217 images taken from 1,501 pedestrians. Each pedestrian is captured by at least two cameras, and there are six cameras in total.

DukeMTMC-reID. Eight cameras were deployed to collect the dataset. The training set contains 702 pedestrians with 16,522 images, and the test set contains 1,110 pedestrians with 19,889 images. Note that 408 pedestrians appear only in one camera, and those samples serve as distractors.

MSMT17. Compared with the previous two datasets, the MSMT17 dataset is closer to real scenarios due to its diversity. Collected by three indoor cameras and twelve outdoor cameras, it comprises 126,441 images from 4,101 pedestrians. The ratio of training to test samples is set to 1:3 with the intention of limiting available training samples and therefore emphasizing effective training methods.

3.2. Evaluation metrics

Models are evaluated using mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) rank-k accuracy [2]. The mAP score is preferable to the CMC rank-k accuracy because the former metric considers both precision and recall while the latter metric does not report the performance on hard positive samples. Furthermore, gallery samples taken from the same camera as the query sample are discarded during evaluation so that the metrics would emphasize the performance in the cross-camera setting.

Table 1. Performance comparisons among existing studies, our baseline method, and our FlipReID method. †: inference using single image (see Figure 2(c)). ‡: inference using double images (see Figure 2(d)). §: re-ranking [14] is applied.

Method	Backbone	Market-1501		DukeMTMC-reID		MSMT17	
		mAP	R1	mAP	R1	mAP	R1
PCB, ECCV 2018 [8]	ResNet50	81.6	93.8	69.2	83.3	-	-
BoT, CVPRW 2019 [6, 21]	ResNet50	86.1	94.4	77.0	87.2	50.2	74.1
SCSN, CVPR 2020 [30]	ResNet50	88.5	95.7	79.0	91.0	58.5	83.8
GASM, ECCV 2020 [31]	ResNet50	84.7	95.3	74.4	88.3	52.5	79.5
AGW, TPAMI 2020 [1, 21]	ResNet50	88.2	95.3	79.9	89.0	55.6	78.3
FastReID, arXiv 2020 [21]	ResNet50	88.2	95.4	79.8	89.6	59.9	83.3
1.1 Baseline†	ResNet50	88.1	95.0	78.9	89.4	61.7	81.5
1.2 Baseline‡		88.6	95.0	79.5	89.4	62.9	82.1
2.1 FlipReID (without flipping loss)†	ResNet50	86.2	94.7	77.2	88.9	57.1	79.5
2.2 FlipReID (without flipping loss)‡		88.5	95.5	79.8	90.2	64.3	83.6
3.1 FlipReID (with flipping loss)†	ResNet50	87.6	95.2	78.9	89.1	61.4	81.9
3.2 FlipReID (with flipping loss)‡		88.5	95.3	79.8	89.4	64.3	83.3
3.3 FlipReID (with flipping loss)‡§		94.6	96.0	90.9	92.5	79.5	86.3
1.1 Baseline†	IBN-ResNet50	88.4	94.8	79.0	88.7	64.6	83.4
1.2 Baseline‡		88.9	95.5	79.6	88.8	65.7	84.2
2.1 FlipReID (without flipping loss)†	IBN-ResNet50	86.9	94.3	77.5	88.7	60.4	81.3
2.2 FlipReID (without flipping loss)‡		88.7	94.8	79.7	89.4	66.2	84.4
3.1 FlipReID (with flipping loss)†	IBN-ResNet50	87.9	94.7	78.8	88.9	63.2	83.1
3.2 FlipReID (with flipping loss)‡		88.6	95.0	79.8	89.6	65.9	84.5
3.3 FlipReID (with flipping loss)‡§		94.1	95.4	89.8	91.8	80.2	87.3
1.1 Baseline†	ResNeSt50	89.3	95.8	80.0	89.6	66.0	84.2
1.2 Baseline‡		89.7	96.2	80.5	89.9	66.9	84.5
2.1 FlipReID (without flipping loss)†	ResNeSt50	88.6	95.2	79.9	90.0	64.6	83.9
2.2 FlipReID (without flipping loss)‡		89.6	95.7	81.2	90.7	67.6	85.3
3.1 FlipReID (with flipping loss)†	ResNeSt50	88.9	95.2	80.7	90.5	66.0	84.6
3.2 FlipReID (with flipping loss)‡		89.6	95.5	81.5	90.9	68.0	85.6
3.3 FlipReID (with flipping loss)‡§		94.7	95.8	90.7	93.0	81.3	87.5

3.3. Analysis of results

Table 1 compares the mAP scores and the rank-1 accuracies of existing studies and our methods.

Datasets. Limited by the number and quality of samples, the Market-1501 [2] and DukeMTMC-reID [3] datasets are saturated, and scores reported on these two datasets may not be indicative [22]. For example, the FastReID [21] method surpasses the AGW [1] method by a large margin on MSMT17 [4], while the scores on Market-1501 and DukeMTMC-reID are close. In the remainder of this study, we mainly compare the mAP scores on MSMT17.

Existing Studies versus Baseline. Among methods built on the ResNet50 [19] backbone, it is evident that the baseline method performs better than existing studies on MSMT17. This makes a good starting point since the FlipReID method is an extension of the baseline method.

Single Image versus Double Images. Independent of how the training is performed, utilizing data augmentation in inference always boosts performance. The notable downside of test-time augmentation is the extra computations which may pose a constraint for real-time applications, in which execution speed is crucial.

Inconsistency between Figure 2(a) and 2(d), *i.e.*, entries starting with 1.2. If data augmentation is enabled in inference, choosing the baseline scheme during training leads to suboptimal performance because the classification model is only trained on feature vectors of single image. Such gap can be solved by switching to the FlipReID method, and noticeable improvement can be observed.

Inconsistency between Figure 2(b) and 2(c), *i.e.*, entries starting with 2.1 or 3.1. If data augmentation is disabled in inference, selecting the FlipReID mechanism during training results in inferior performance since the classification model is only trained on feature vectors of double images. Adding the flipping loss is an effective approach to suppress this problem. On the one hand, the resulting method is on par with the baseline method when using single image in inference. On the other hand, the flipping loss does not degrade performance when using double images in inference.

Backbone and Re-Ranking. In addition to ResNet [19], experiments have been conducted with IBN-ResNet [24] and ResNeSt [25]. The latter two backbone models improve performance. Moreover, adding re-ranking [14] as a post-processing step introduces significant improvements.

4. CONCLUSION

In this study, we recognize and investigate the gap between training and inference in person re-identification. Prior works typically use the mean of feature vectors extracted from the original images and their horizontally flipped variants in inference. However, such mean feature vectors are not present when optimizing the model in training. In order to close the gap, we propose to utilize the FlipReID structure with the flipping loss. On the one hand, both the original images and the flipped images are feed into models with the FlipReID structure. On the other hand, incorporating the flipping loss minimizes the mean squared error between feature vectors of corresponding image pairs. Using the proposed method, models work as expected regardless of whether test-time augmentation is enabled or not, and the inconsistency issue is solved. An extension of this study is to design a module that learns to aggregate feature vectors from multiple sources, rather than calculating the mean.

5. REFERENCES

[1] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C H Hoi, "Deep learning for person re-identification: A survey and outlook," *arXiv preprint arXiv:2001.04193*, 2020.

[2] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[3] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision*. Springer, 2016, pp. 17–35.

[4] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi 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.

[5] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian, "Person re-identification in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1367–1376.

[6] Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, p. 0.

[7] Rahul Rama Varior, Bing Shuai, Jiwen Lu, Dong Xu, and Gang Wang, "A siamese long short-term memory architecture for human re-identification," in *European conference on computer vision*. Springer, 2016, pp. 135–153.

[8] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 480–496.

[9] Yutian Lin, Liang Zheng, Zhedong Zheng, Yu Wu, Zhilan Hu, Chenggang Yan, and Yi Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, 2019.

[10] Zhedong Zheng, Liang Zheng, and Yi Yang, "A discriminatively learned cnn embedding for person reidentification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 14, no. 1, pp. 1–20, 2017.

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

[12] Chunxiao Liu, Chen Change Loy, Shaogang Gong, and Guijin Wang, "Pop: Person re-identification post-rank optimisation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 441–448.

[13] Hanxiao Wang, Shaogang Gong, Xiatian Zhu, and Tao Xiang, "Human-in-the-loop person re-identification," in *European conference on computer vision*. Springer, 2016, pp. 405–422.

[14] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1318–1327.

[15] Connor Shorten and Taghi M Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[16] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[17] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le, "Autoaugment: Learning augmentation policies from data," *arXiv preprint arXiv:1805.09501*, 2018.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "ImageNet classification with deep convolutional

neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[20] Guanshuo Wang, Yufeng Yuan, Xiong Chen, Jiwei Li, and Xi 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.

[21] Lingxiao He, Xingyu Liao, Wu Liu, Xinchen Liu, Peng Cheng, and Tao Mei, "Fastreid: A pytorch toolbox for general instance re-identification," *arXiv preprint arXiv:2006.02631*, vol. 6, no. 7, pp. 8, 2020.

[22] Xingyang Ni, Liang Fang, and Heikki Huttunen, "Adaptive L2 Regularization in Person Re-Identification," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 9601–9607.

[23] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[24] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 464–479.

[25] Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Zhi Zhang, Haibin Lin, Yue Sun, Tong He, Jonas Mueller, R Manmatha, and others, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[26] Filip Radenovic, Giorgos Tolias, and Ondrej Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 7, pp. 1655–1668, 2018.

[27] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[28] Yunpeng Gong and Zhiyong Zeng, "An Effective Data Augmentation for Person Re-identification," *arXiv preprint arXiv:2101.08533*, 2021.

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

[30] Xuesong Chen, Canmiao Fu, Yong Zhao, Feng Zheng, Jingkuan Song, Rongrong Ji, and Yi Yang, "Salience-Guided Cascaded Suppression Network for Person Re-Identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3300–3310.

[31] Lingxiao He and Wu Liu, "Guided Saliency Feature Learning for Person Re-identification in Crowded Scenes," in *European Conference on Computer Vision*. Springer, 2020, pp. 357–373.