COL780 Project: Person Re-Identification

Mustafa E. Chasmai
Indian Institute of Technology Delhi
cs1190341@iitd.ac.in

Tamajit Banerjee
Indian Institute of Technology Delhi
cs1190408@iitd.ac.in

Abstract

Person Re-Identification (Re-ID) is an important problem in computer vision-based surveillance applications, in which one aims to identify a person across different surveillance photographs taken from different cameras having varying orientations and field of views. Due to the increasing demand for intelligent video surveillance, Re-ID has gained significant interest in the computer vision community. In this work, we experiment on some existing Re-ID methods that obtain state of the art performance in some open benchmarks. We qualitatively and quantitaively analyse their performance on a provided dataset, and then propose methods to improve the results.

1. Introduction

Intelligent video surveillance has been gaining a lot of attention recently. Along with detection and tracking, re-identification is one of the major tasks needed to get a complete surveillance system, and has its own set of challenges. Person Re-ID aims to find the person of a query image in a different set of gallery images that may contain the same person. This becomes particularly challenging when the gallery has images from a wide variety of orientations, occlusions, lighting conditions, and backgrounds.

Many past works [7,13,15,23,26] belong to the distance metric learning paradigm, where the problem is reduced to learning a suitable metric that can provide a partitioning of images having the same and different persons. Equivalently, a suitable feature representation is to be learnt such that the latent vectors corresponding to the same person should have a large similarity, while those corresponding to different persons should have lower similarity. The similarity is determined by selecting an appropriate similarity metric. Various metrics of similarity like the norm and the cosine similarity have been used in literature. The Re-ID problem is different from person classification because the persons are not restricted to a fixed set of classes, and the model is expected to be robust to queries for new persons it had not encountered during training.

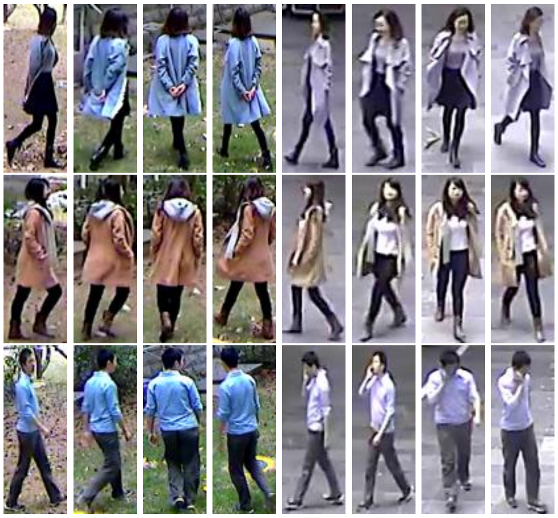

Figure 1. Few sample images from the ReID780 training dataset. Each row corresponds to different images for a single person. The images have varying orientation and poses of the concerned person, along with different lighting conditions and backgrounds.

There have been many different methods proposed to solve Re-ID. From hand crafted features to deep CNN based models to the recent transformer based models, the state of the art has improved considerably. These methods are benchmarked on various openly available datasets like Market-1501 [27], CUHK03 [9] and MSMT17 [21]. We used the dataset provided by the instructors of COL780 , IIT Delhi for all the experimentation and analysis in this work. We qualitatively analyse the failure cases of the baseline and propose novel design and methodological changes to improve the performance further.

2. Related Work

There have been many methods proposed for Re-ID. Most early research in the field involved exploring new hand crafted features that could better represent a person's identity across varying viewpoints, lighting conditions and backgrounds. Some works [3, 5, 10] focused on combining intuitive local features via ensembling or accumulation

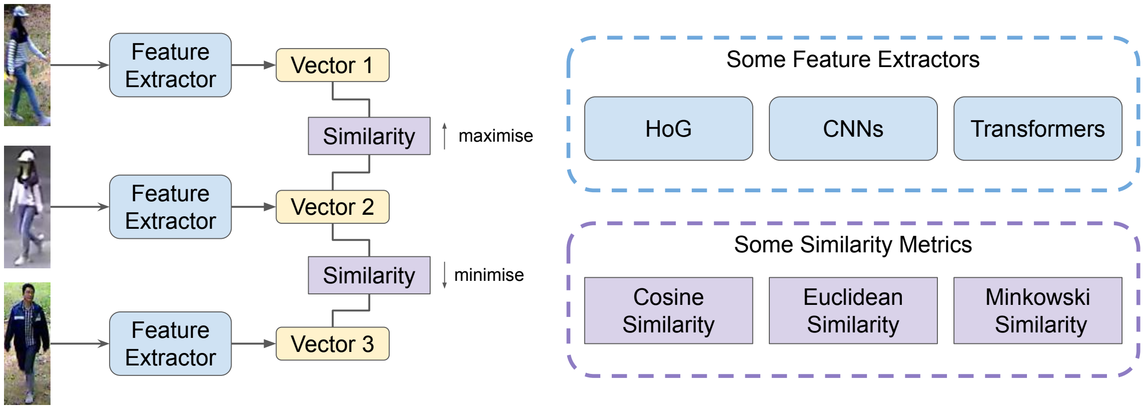

Figure 2. A rough outline of a generic architecture of a metric learning based method. The task is to learn a feature representation that maximises the similarity metric of images of the same person, and minimises the same for images of different persons.

or maximal occurences, while others [14, 24] devised new feature representations.

With the advent of deep learning, most methods shifted to CNNs for richer representations that could be learnt directly. Deep metric learning methods transform raw images into embedding features, then compute the feature similarities. The focus shifted from designing good features to designing suitable loss functions that could train the CNN based feature extractors (e.g. ResNet [6]). The ID loss [28] and triplet loss [11] are most widely used in deep ReID, with some methods [12, 18] using a combination of the two. The triplet loss uses three inputs, two of the same person, while one of a different one. It is designed so as to minimise the distance between the positive pair (two images of the same person), and to maximise the distance between the negative pair (different persons). There has also been some work in improving the strategy of selecting suitable positive and negative pairs. Ideas very similar to the triplet loss can be found in Siamese networks [8], Contrastive Learning [1] and GANs [4].

Many other methods have worked on improving the underlying architecture. AlignedReID [13, 26] extracts a global feature which is jointly learnt with local features by aligning them together. Aligning local features by pose estimation [27] has also been used in the past. Wojke et al. [23] used a conventional softmax classification regime, and used the pre-softmax embeddings with small re-parametrisation to optimise the cosine similarity metric.

The Transformer model [19], proposed originally to handle sequential data in the field of natural language processing (NLP), has been gaining a lot of attention in the computer vision field. The Vision Transformer [2] uses the transformer architecture for image classification. The transformer has since been used for wide ranging tasks like segmentation [20] and pose estimation [17], and some methods have used it for Re-ID as well. TransReID [7] and LA Transformers [15] are two sunch methods. There are many other methods like Centroids ReID [22] that also showed promise on existing datasets.

3. Baselines

We experiment and explore 4 existing methods in person ReID, namely AlignedReID [26], LA Transformers [15], Centroid ReID [22] and TransReID [7]. We also ran Deep Cosine Metric [6] on the dataset, but did not experiment on it further because of low accuracy and tensorflow - pytorch incompatibilities. The results of these baseline methods on the provided dataset can be found in Table 1.

Method	mAP	CMC @R1	CMC @R5
AlignedReID	98.5	100	100
LA-Transformer	95.1	100	100
Centroids-ReID	68.4	82.1	82.1
TransReID	64.3	71.4	82.1
Deep Cosine	35	-	-

Table 1. Performance of some existing methods on the dataset

We have used these codebases provided by the authors of the respective methods: AlignedReID, LA Transformer, Centroids ReID, Trans ReID and Deep Cosine Metric.

4. Analysis of Baselines

We ran multiple experiments on the baselines to identify which is the best performing. We report these experimental observations, along with our interpretations and conclusions from them.

The top 5 predictions based on similarity for a few query images can be seen in Fig 3 above. It can be observed that AlignedReID and LA Transformer methods perform comparatively better than the rest. One interesting observation

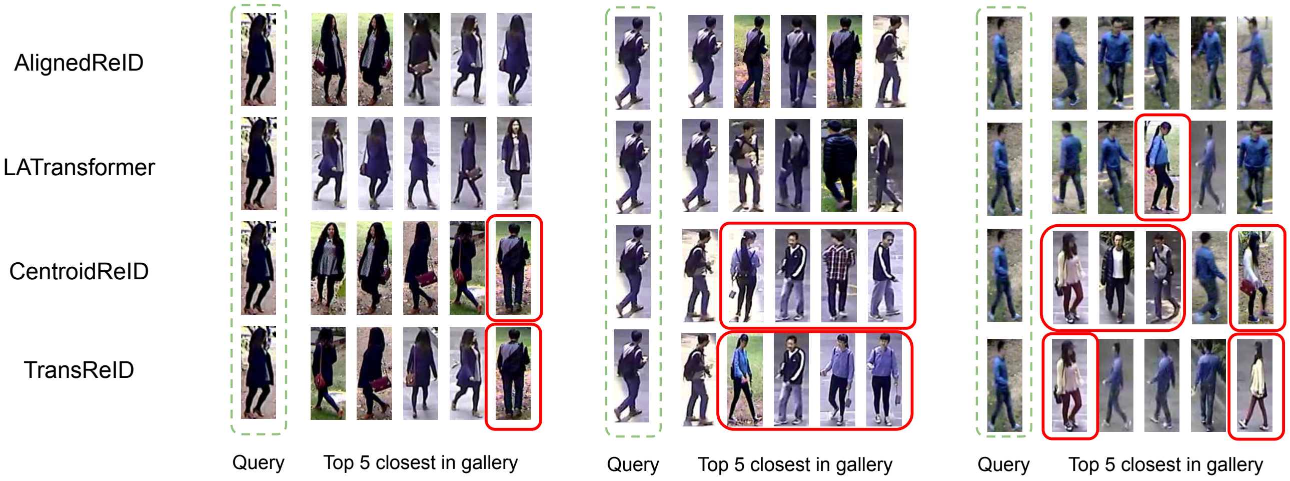

Figure 3. The top 5 closest images in the gallery for a few querry images with the explored baseline methods

is that for the last query image, all of the models predicted the same wrong image, which was even the closest for three of them. We interpret this as an indication that there is some semantic similarity in them even though they look completely different to human eyes. Other than this corner case, all of the models seem to perform quite good, with even the wrong predictions being visually similar to the query image. These examples of visual similarity are especially difficult to deal with.

Method	mAP	CMC @R1	CMC @R5
AlignedReID	69.8	78.6	85.7
LA-Transformer	75.2	89.3	82.1
Centroids-ReID	64.8	75.0	85.7

Table 2. Performance of some existing methods on the dataset, trained on the original data and tested with the background masked

We wanted to explore the sensitivity of the models with the background information. Ideally, we expect a person re-identification model to be background agnostic, with no change in the predictions if the background is removed completely. We run the trained models for the baselines on images with their backgrounds masked out using a semantic segmentation model, and observe that in all of them, the performance drops. We interpret this as meaning that all of the models need some background context to be able to perform well, and in the absence of the background, they are not able to do so. From this, we conclude that the methods are somewhat overfitting on the background information.

Finally, we wanted to explore how the learned features are distributed. We expect that the features of the same persons would be close together, while that of the different persons would be far from each other. Thus, we expect the formation of cluster-like structures in the feature space. We explore this for the case of the gallery images of the validation set, and plot t-SNE graphs for our different models. The plots can be seen in Fig 4. The t-SNE plots for TransReID seem to be slightly more clustered that the same for AlignedReID, and we conclude from this that TransReID learns more robust features.

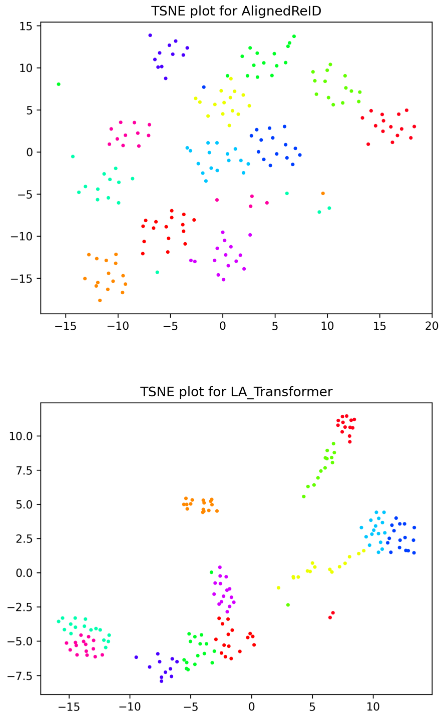

Figure 4. t-SNE plots AlignedReID (top) & LATrans (down)

5. Proposed Methodologies

5.1. Mask Guidance

As we had analysed, removing the background pixels via masking leads to a significant drop in the performance of all the models. We wanted to make a methods that is invariant to the absence of backgrounds and thus, experiment with different ways of including the binary segmentation masks in the training process as well.

Deriving inspiration from [16], we use the binary mask primarily for 2 reasons. It can improve the robustness of ReID models under various background conditions by removing background clutters in the pixel-level. Along with this, the mask contains body shape information which can be regarded as the important feature and also it is robust to illumination, color of clothes and thus is an important feature.

The most straightforward way to utilize the binary body mask is to directly mask the background in the images. With the binary mask, the masked image only contains the body region which is expected to perform better than using the whole image. However, in our experiments, we find the performance of masked images is even slightly worse compared with the one using the original images. This result means that directly removing the background with binary mask in a 'hard' manner is not a good choice, which may affect the structured information and smoothness of an image. In addition, the wrongly segmented masks may contain lots of backgrounds or lose some important body parts which will greatly impact the performance. In this case, removing the backgrounds in the feature-level may be a better solution.

We initially used only binary mask for predictions. Even though the mask did not contain colour information and only had body shapes. It gave an mAP of 29.2 which was better than random selection which showed that the body shape is important.

Our proposed solution is to send the mask along with the RGB image to learn the weights for feature. We add an extra convolutional layer at the beginning of the model. The main purpose of this layer is too merge the RGB image and the binary mask. Thus, with this, we have a model that takes as input the RGB image and binary mask, merges the two into a single feature in the first layer and then passes this through the remaining model same as used earlier.

5.2. Feature position invariance

In Aligned ReId , the dataset on which they have worked does not have accurate bounding boxes and to counter that they have used the shortest distance algorithm through dynamic programming to find the distances between local feature vectors, which finally gets translated to the losses.

On the other hand, our dataset is more organised, that is, each bounding box is nearly perfect. As a result, the local feature vectors should corresponding to each other better. So, the complex DP algorithm can be replaced by a simple one to one correspondence. This both reduced the time complexity and improves the performance on our dataset.

5.3. Self Ensembling

One observation we made during training the methods was that the validation performance metrics would fluctuate a lot as the model is trained. This was particularly prevalant in LA Transformer. To improve upon this short coming, we proposed the use of self ensembling. Self Ensembling is a powerful and easy to implement idea, as used in the Mean Teacher models [25] for contrastive learning.

The principle idea behind the mean teacher model is that of ensembling multiple weak models to obtain a much stronger model. The weak learners for self ensembling are the same model, at different stages of the training, particularly after training different batches. Self ensembling use a non-trainable clone of the actual model, and updates its parameters as the exponentially moving average of the training model. In literature, this exponentially moving averaged model is called the mean teacher model, and the training model as the student model. Finally, they use a consistency loss between the teacher and student such that the student learns the same features as the teacher. The teacher with the average weights is the ensemble of the models in the last few training steps, and is the main driving force for this self ensembling approach.

We observed that upon adding the self ensembling mean teacher approach, and using the consistency loss, the training of LA Transformer was comparatively much more smoother, and also the performance was better. Details can be found in the experimental results later.

5.4. Triplet Loss and Hard Example Mining

LA Transformer uses only the cross entropy loss for training. We added triplet loss along with it, and expected a slightly better performance.

The basic idea is that the distance between the positive pair should be smaller than the negative pair. For calculating the loss, a triplet is formed, containing an anchor sample, a positive sample having the same person, and a negative sample from a different person. The distance between the anchor and positive sample is to be minimised while the same between the anchor and negative sample is to be maximised. This is very similar to the contrastive loss commonly used in self supervised learning.

The performance of the triplet loss depends on the choice of positive and negative samples for given anchors. One widely used approach is Hard example mining. In this approach, the sample of the same person that is farthest and the sample of another class that is closest to the anchor is

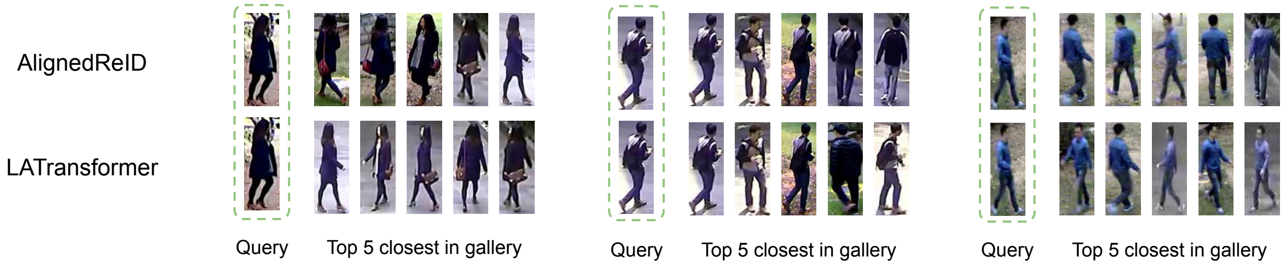

Figure 5. Predictions on the two improved models

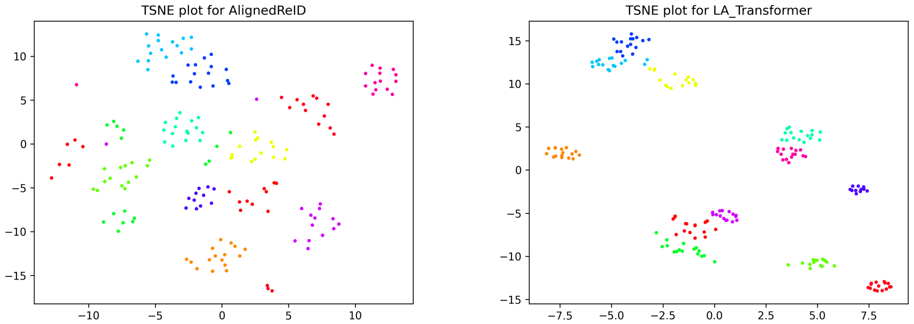

Figure 6. t-SNE plots for improved AlignedReID and LA Transformer

chosen. These samples are expected to be hardest to classify correctly, and thus using these as a triplet should guide the model towards learning better.

5.5. Camera Contextual Learning

The dataset provided to us had a more structured distribution compared to most existing datasets like Market1501. In this dataset, each person has images from 8 different angles, while in most other datasets, the number of camera angles present is often not as rigorous and mostly random. Thus, we though of a novel approach to leverage this structure in the data.

We reason that the change in the feature vectors between camera orientation should be agnostic of the identity of the person, and thus, should be nearly the same for all persons in the gallery. We wanted to explicitly encode this prior in the model, and share weights for the camera orientation dependent features. Thus, keeping the number of cameras as 8 in the dataset, we initialise 8 new residual layers that take as input the complete features and output embeddings of the same dimension. Since the residual layers make it easy to learn a unity function, these extra layers orientation should also be easily able to learn the non-camera contextual features.

The main point is that these 8 layers have the same shared weights for all persons, but are different for different camera angles. Thus, the prior for camera orientation is effectively introduced to the model. In effect, we expected the model to be able to learn features that are clustered around a ring like shape instead of the usual point clusters, where the ring radius roughly corresponds to the extra residuals learned by the added camera layers. Thus, instead of trying to cluster features of the same person to the same point, we effectively try to cluster them in a ring like shape for each person. The fact that we observe a similar ring like pattern for the LA Transformer model is validation of our reasoning. We were not able to experiment this idea thoroughly, and our initial experiments indicated that the performance is actually dropping.

6. Experimental Results and Analysis

We trained and tested all of our proposed methodologies on the provided dataset. The quantitaive results for the different experiments are tabulated in Table 3, while the best performing methods are summarised in Table 4.

In addition to the quantitative results, we report a qualitative analysis of the methods. The visualisation of top 5 predictions for the two improved methods can be seen in Fig 5, while the t-SNE plots for the same can be seen in Fig 6. Comparing these with similar visualisations of the baselines, we can see that the improvements are reasonable, and that we can expect similar performance boosts in the test dataset. The trained weights of our baseline and improved models can be found here. The training code can

Baseline	Experiments	mAP	CMC @Rank 1	CMC @Rank 5
AlignedReID	Feature Position Invariance	99.0	100	100
	Binary Masks	29.2	30.1	52.4
	masked RGB images	76.7	90.8	83.6
	RGB images + masks	88.0	92.9	100
	Camera Contextual Learning	67.3	78.6	85.7
La Transformer	Triplet Loss	94.3	100	100
	Self Ensembling	96.4	100	100
	Triplet Loss + Self Ensembling	98.4	100	100

Table 3. List of experiments performed and the metrics obtained in each

Baseline	Improvement Method	mAP	CMC @Rank 1	CMC @Rank 5
AlignedReID	Feature Position Invariance	99.0	100	100
La Transformer	Triplet Loss + Self Ensembling	98.4	100	100

Table 4. Experiments that led to an improvement over the baaseline

be found in the corresponding colab links: LATransformer, AlignedReID, Centroids-ReID and TransReID.

7. Conclusion

Person Re-identification is a challenging and yet unsolved task in Computer Vision that has seed a rising popularity in recent. We thoroughly analyse few state of the art approaches in this domain on the dataset provided, and try to point out some poossible research gaps. Based on these analyses, we propose novel approaches to bridge these gaps.

First, we experiment with mask guidance. Though the mask guidance method showed promise, we were not able to achieve good results with the method. We suspect that the model was not able to properly incorporate the semantic information contained in the binary mask into its training. We wanted to propagate this mask information into the deeper layers of our model, and expect this to improve the performance of our model by making it more robust to background variation. Next, we observed that while AlignedReID had some mechanisms for images where the bounding boxes are not perfectly aligned, these cases are not present in the provided dataset, and removing them led to some improvements in the method.

We further explore other methods to improve the performance of the model. We use self-ensembling and triplet loss with hard example mining to improve the performance of LA-Transformer. Finally, we make architectural changes in LA-Transformer to introduce the domain specific prior of camera orientations. Although the performance does not improve in some of our proposed approaches, we believe that they are in the right direction, and have scope for improvement.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[2] Alexey Dosovitskiy, Lucas Beyer, and Alexander Kolesnikov et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[3] Michela Farenzena, Loris Bazzani, Alessandro Perina, Vittorio Murino, and Marco Cristani. Person re-identification by symmetry-driven accumulation of local features. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2360–2367. IEEE, 2010. 1

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[5] Douglas Gray and Hai Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008. 1

[6] 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*, pages 770–778, 2016. 2

[7] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object reidentification. *arXiv preprint arXiv:2102.04378*, 2021. 1, 2

[8] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015. 2

[9] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-

identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 152–159, 2014. 1

[10] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015. 1

[11] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017. 2

[12] 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/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2

[13] Hao Luo, Wei Jiang, Xuan Zhang, Xing Fan, Jingjing Qian, and Chi Zhang. Alignedreid++: Dynamically matching local information for person re-identification. *Pattern Recognition*, 94:53–61, 2019. 1, 2

[14] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1363–1372, 2016. 2

[15] Charu Sharma, Siddhant R Kapil, and David Chapman. Person re-identification with a locally aware transformer. *arXiv preprint arXiv:2106.03720*, 2021. 1, 2

[16] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1179–1188, 2018. 4

[17] Lucas Stoffl, Maxime Vidal, and Alexander Mathis. End-to-end trainable multi-instance pose estimation with transformers. *arXiv preprint arXiv:2103.12115*, 2021. 2

[18] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6398–6407, 2020. 2

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[20] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 2

[21] 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*, pages 79–88, 2018. 1

[22] Mikolaj Wieczorek, Barbara Rychalska, and Jacek Dabrowski. On the unreasonable effectiveness of centroids in image retrieval. *arXiv preprint arXiv:2104.13643*, 2021. 2

[23] Nicolai Wojke and Alex Bewley. Deep cosine metric learning for person re-identification. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 748–756. IEEE, 2018. 1, 2

[24] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. Salient color names for person re-identification. In *European conference on computer vision*, pages 536–551. Springer, 2014. 2

[25] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019. 4

[26] Xuan Zhang, Hao Luo, Xing Fan, Weilai Xiang, Yixiao Sun, Qiqi Xiao, Wei Jiang, Chi Zhang, and Jian Sun. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv preprint arXiv:1711.08184*, 2017. 1, 2

[27] 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*, pages 1116–1124, 2015. 1, 2

[28] Zhedong Zheng, Liang Zheng, and Yi Yang. A discriminatively learned cnn embedding for person reidentification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(1):1–20, 2017. 2