NFormer: Robust Person Re-identification with Neighbor Transformer

Wang, H.; Shen, J.; Liu, Y.; Gao, Y.; Gavves, E.

DOI
10.48550/arXiv.2204.09331
10.1109/CVPR52688.2022.00715

Publication date
2022

Document Version
Author accepted manuscript

Published in
2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition

NFormer: Robust Person Re-identification with Neighbor Transformer

Haochen Wang[1], Jiayi Shen[1], Yongtuo Liu[1], Yan Gao[2], Efstratios Gavves[1]
University of Amsterdam[1], Xiaohongshu Inc[2]
{h.wang3, j.shen, y.liu6}@uva.nl, wanjianyi@xiaohongshu.com, egavves@uva.nl

Abstract

Person re-identification aims to retrieve persons in highly varying settings across different cameras and scenarios, in which robust and discriminative representation learning is crucial. Most research considers learning representations from single images, ignoring any potential interactions between them. However, due to the high intra-identity variations, ignoring such interactions typically leads to outlier features. To tackle this issue, we propose a Neighbor Transformer Network, or NFormer, which explicitly models interactions across all input images, thus suppressing outlier features and leading to more robust representations overall. As modelling interactions between enormous amount of images is a massive task with lots of distractors, NFormer introduces two novel modules, the Landmark Agent Attention, and the Reciprocal Neighbor Softmax. Specifically, the Landmark Agent Attention efficiently models the relation map between images by a low-rank factorization with a few landmarks in feature space. Moreover, the Reciprocal Neighbor Softmax achieves sparse attention to relevant -rather than all- neighbors only, which alleviates interference of irrelevant representations and further relieves the computational burden. In experiments on four large-scale datasets, NFormer achieves a new state-of-the-art. The code is released at https://github.com/haochenheheda/NFormer.

1. Introduction

Image-based person re-identification (Re-ID) aims to retrieve a specific person from a large number of images captured by different cameras and scenarios. Most research to date has focused on how to obtain more discriminative feature representations from single images, either by attention modules [17, 26, 32, 34], part representation learning [6, 14, 28, 41], or GAN generation [20, 23, 46]. However, one of the main challenges in Re-ID is that any individual typically undergoes significant variations in their appearance due to extrinsic factors, like different camera settings, lighting, viewpoints, occlusions, or intrinsic factors like dress changing, to name a few examples. As a result, there are high intra-identity variations in the representations corresponding to a specific individual, leading to unstable matching and sensitivity to outliers, see figure 1.

Figure 1. The dots indicate the representation vectors of persons for retrieval in the feature space. The right top figure illustrates person representations distribution obtained by learning from single input images, which typically leads to outliers (hollow dots) caused by occlusions, dress changing, viewpoint changing, etc. The right bottom figure shows the person representations distribution obtained by NFormer, which explicitly models the relations (grey lines) between relevant neighbor persons to alleviate the outlier features caused by the above-mentioned abnormal conditions and maintain the most discriminative features for each identity.

A possible remedy against high intra-identity variations is to exploit the knowledge that exists in the different images from the same identity. Intuitively, one can encourage the model to cluster neighbor representations tightly, as they are likely to correspond to the same individuals. A few works have proposed to model relations between input images in Re-ID, either with conditional random fields [3] or similarity maps from training batches [21]. However, these works focus on modeling relations between a few images *at training time* only, while during the test, they extract representations per image independently due to the computation limitation, which inevitably loses the interactions and leads to a gap between training and test. Moreover, they only build relations between a small group of images within each training batches so that there is limited relevant information that could be learned from each other. To sum up, we argue

that encouraging lower representation variations per identity is crucial during both training and test among all the input images.

Following this train of thought, we propose a Neighbor Transformer Network, or NFormer for short, to efficiently model the relations among all the input images *both at training and test time*. As shown in figure 2, NFormer computes an affinity matrix representing the relations between the individual representations and then conducts the representation aggregation process according to the affinity matrix. The involvement of relation modeling between images suppresses high intra-identity variations and leads to more robust features.

Unfortunately, computing an affinity kernel matrix is typical of quadratic complexity to the number of samples. Such a computational complexity is prohibitively expensive in person Re-ID setting, where the number of input images can easily grow to several thousands during the inference. To this end, we propose a *Landmark Agent Attention* module (LAA) that reduces the computations in the affinity matrix by the introduction of a handful of landmark agents in the representation space. The landmark agents map the representation vectors from a high-dimensional feature space into a low-dimensional encoding space, which factorizes large affinity maps into a multiplication of lower rank matrices. Similarly, the representation aggregation process with a standard softmax attends to all the input representations, which tends to be distracting and computation-consuming caused by a large number of irrelevant representations. We introduce the *Reciprocal Neighbor Softmax* function (RNS) to achieve sparse attention attending to computationally manageable neighbors only. The Reciprocal Neighbor Softmax significantly constrains the noisy interactions between irrelevant individuals, which makes the representation aggregation process more effective and efficient.

Our contributions are summarized as follows:

- We propose to explicitly model relations between person representations with a Neighbor Transformer Network, designed to yield robust and discriminative representations.
- We design a Landmark Agent Attention module to reduce the computational cost of the large affinity matrix by mapping the representations into lower-dimensional space with a handful of landmark agents.
- We propose a Reciprocal Neighbor Softmax function to achieve sparse attention attending to neighbors only, which strengthens the interaction between relevant persons with efficiency.
- We conduct extensive experiments on four person Re-ID datasets to indicate the general improvements which NFormer brings. The results show that NFormer achieves a new state-of-the-art. We further note that NFormer is easy to plug and play with other state-of-the-arts and further boosts the performance.

2. Releated Work

In this section, we first briefly review two main families of Re-ID methods: Feature Representation Learning methods and Ranking Optimization methods. Then we introduce the Transformer and related applications.

2.1. Feature Representation Learning Methods

Learning the discriminative feature representations is crucial for Re-ID. Most of the existing methods [17, 26, 28, 41, 46] focus on how to extract better representation with single images. Some methods introduce local part features with automatic human part detection [28, 41] or horizontal image division [29] to tackle the occlusion and misalignment problems. Some methods design attention modules within single images to enhance representation learning at different levels. For instance, method [17] involves pixel-level attention while methods [32, 34] achieve channel-wise attention for feature re-allocation. Method [27] suppresses the background region to obtain robust foreground person representation. Another kind of method focuses on increasing the richness of training data. [13, 48] generates adversarially occluded samples to augment the variation of training data. [20, 46] utilize GAN to generate images as auxiliary information to help the training. In general, this family of methods makes full use of information from individual images to extract discriminative feature representations.

2.2. Ranking Optimization Methods

Ranking optimization is a strategy to improve the retrieval performance in the test stage. Given an initial ranking list obtained by the distance matrix between query and gallery sets, works [19, 22, 37, 38, 49] optimize the ranking order by the following methods. [38] propose a rank aggregation method by employing similarity and dissimilarity. [19] involves human feedback to optimize the ranking list. Methods [22, 49] propose the query adaptive retrieval strategy to improve the performance. [37, 47] also utilize contextual information from other images. Those methods are directly conducted on each initial ranking list as post-processing, instead of conducted on the representation distribution. Our proposed NFormer is compatible with those re-ranking methods to further boost the performance.

2.3. Transformer

The Transformer [31] is built upon the idea of Multi-Head Self-Attention (MHA), which allows the model to jointly attend to different representation elements. The transformer is proposed to tackle the sequence problem in the beginning. Recently, Tranformer is widely used for

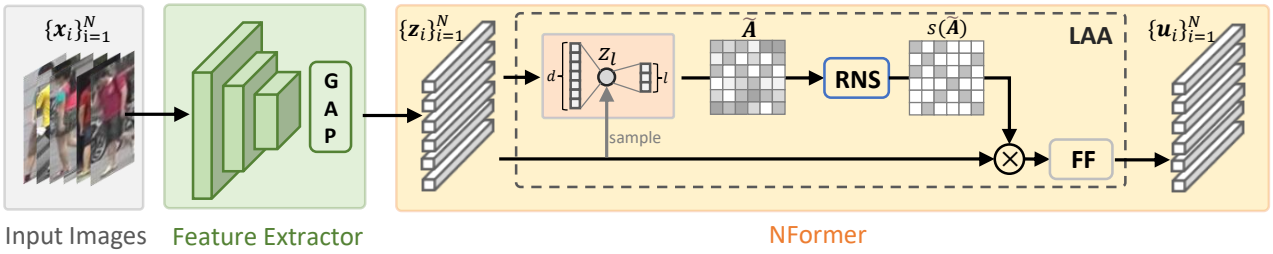

Figure 2. An illustration of NFormer. **GAP**: Global Average Pooling. **LAA**: Landmark Agent Attention. **RNS**: Reciprocal Neighbor Softmax. **FF**: Feed-forward Network. Input with N images $\{\mathbf{x}_i\}_{i=1}^N$, a convolutional network followed by GAP is used to get the representation vectors $\{\mathbf{z}_i\}_{i=1}^N$. $\{\mathbf{z}_i\}_{i=1}^N$ is fed to NFormer, in which LAA is proposed to map the d-dimensional representations into a l-dimensional encoding space with sampled landmark agents $\mathbf{z}_l$ and then obtain the approximate affinity matrix $\widetilde{\mathbf{A}}$ more efficiently. Then the RNS is proposed to get the sparse attention weights $s(\widetilde{\mathbf{A}})$ and the output representations $\{\mathbf{u}_i\}_{i=1}^N$ are obtained by weighted aggregation of $\{\mathbf{z}_i\}_{i=1}^N$. Finally, a ranking algorithm is performed on the representation vectors after NFormer for the retrieval process.

many vision tasks because of its powerful ability to obtain long-distance dependance, such as DETR [1] for object detection, TT [5] for object tracking and ViT [8] for image classification. We first adopt Transformer architecture to learn the relations between input persons in the Re-ID task.

3. Neighbor Transformer Network

We first describe the problem setting and the overview of the proposed method. Then, we describe the Landmark Agent Attention and the Reciprocal Neighbor Softmax.

3.1. Problem Setting

Re-ID is typically cast as a retrieval task. We start with the training set $T = \{\mathbf{x}_i, y_i\}_{i=1}^{N^T}$, where $\mathbf{x}_i$ corresponds to the i-th image with identity $y_i \in S_T$, and S_T contains the identities of all the training images. During training, we learn a model $\mathbf{z}_i = f(\mathbf{x}_i)$ that computes discriminative feature representations $\mathbf{z}_i$ per input image [17, 26, 32]. At test time we have a query set $U = \{\mathbf{x}_i\}_{i=1}^{N^U}$ with persons-of-interest. Then, given a gallery set $G = \{\mathbf{x}_i\}_{i=1}^{N^G}$ for retrieval, we retrieve persons with correct identity when comparing query images in U against the images in the gallery set G. The identities of the persons in the query set S_U are disjoint from the identities available during training, that is $S_U \cup S_T = \emptyset$.

3.2. Learning NFormer

In the described setting, we place no restrictions on the form of the function $f(\cdot)$. Typically, $f(\cdot)$ is computed on single input images, thus ignoring any possible relations that may arise between the representations of the same individual across cameras and scenarios. To explicitly account for such relations, we introduce a function to get the aggregated representation vector $\mathbf{u}_i$:

$$\mathbf{u}_i = g(\mathbf{z}_i, \{\mathbf{z}_j\}_{j=1}^N) = \sum_j \mathbf{w}_{ij}\mathbf{z}_j, \tag{1}$$

where $\{\mathbf{z}_j\}_{j=1}^N$ contains the representation vectors obtained by the feature extraction function $f(\cdot)$ of all the input images $\{\mathbf{x}_i\}_{i=1}^N$. During training, $\{\mathbf{x}_i\}_{i=1}^N \subset T$ is a large batch sampled from training set. While during test, $\{\mathbf{x}_i\}_{i=1}^N = U \cup G$ contains all the images from query set and gallery set. $\mathbf{w}_{ij}$ is learnable weight between $\mathbf{z}_i$ and $\mathbf{z}_j$, where $\sum_j \mathbf{w}_{ij} = 1$. Recently, Transformer [31] has shown to be particularly apt in modeling relations between elements in a set. With a Transformer formulation, we have equation (1) reformed by:

$$\mathbf{u}_i = \sum_j s(\mathbf{A})_{ij}\varphi_v(\mathbf{z}_j), \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is an affinity matrix that contains the similarities between any two pairs of input representation vectors $\mathbf{z}_i, \mathbf{z}_j$, $s(\cdot)$ is a softmax function to turn the affinities into weights, and $\varphi_v(\cdot)$ is a linear projection function. For the affinity matrix $\mathbf{A}$, we have

$$\mathbf{A}_{ij} = K(\varphi_q(\mathbf{z}_i), \varphi_k(\mathbf{z}_j))/\sqrt{d} = \mathbf{q}_i^\top \mathbf{k}_j/\sqrt{d}, \tag{3}$$

where $\varphi_q(\cdot), \varphi_k(\cdot)$ are two linear projections, which map the input representation vectors $\mathbf{z} \in \mathbb{R}^{N\times d}$ to query and key matrices $\mathbf{q}, \mathbf{k} \in \mathbb{R}^{N\times d}$. N is the number of input images and d is the dimension of the representation vectors. The $K(\cdot,\cdot)$ is typically the inner product function.

Unfortunately, considering a conventional transformer network to model relations between the representations of persons in Re-ID is computationally prohibitive. First, computing the affinity matrix $\mathbf{A}$ in equation (3) has quadratic $O(N^2d)$ complexity with respect to the number of images N. Thus, the computation of affinity matrix scales poorly with N, especially when the dimension d of the representation vector is also large. To this end, we introduce the Landmark Agent Attention module to factorize the affinity computation into a multiplication of two lower-dimensional matrices, which relieves the computational burden of the affinity matrix. Second, in equation (2) for the final representation vector $\mathbf{u}_i$ we attend to all the $\mathbf{z}_j, j \in \{1, \ldots, N\}$ to

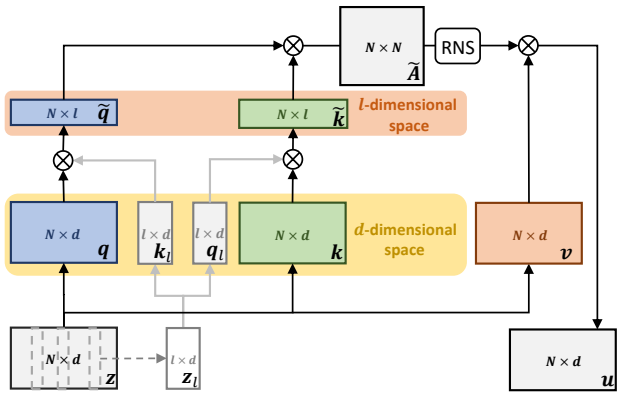

Figure 3. Pipeline of LAA. The horizontal side of the rectangles indicates the first dimension of the according matrices, while the vertical side indicates the second dimension. Input with representation vectors $\mathbf{z} \in \mathbb{R}^{N\times d}$, the query, key and value matrices $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbb{R}^{N\times d}$ are generated by three linear projection functions respectively. The landmark agents $\mathbf{z}_l \in \mathbb{R}^{N\times l}$ are sampled from $\mathbf{z}$ to map the $\mathbf{q}, \mathbf{k}$ of d-dimension to $\widetilde{\mathbf{q}}, \widetilde{\mathbf{k}}$ of l-dimension. Then the approximate affinity matrix $\widetilde{\mathbf{A}}$ is obtained by the multiplication of $\widetilde{\mathbf{q}}$ and $\widetilde{\mathbf{k}}$. In this way, the time complexity of obtaining the affinity matrix reduces from $O(N^2d)$ to $O(N^2l)$, since the l is much smaller than d in practice. Then, the RNS is applied to $\widetilde{\mathbf{A}}$ and turns the affinities into sparse attention weights. The final output $\mathbf{u}$ is obtained by weighted aggregation of value matrix $\mathbf{v}$.

compute the weighted aggregation, which also scales poorly with N. Importantly, the weighted aggregation tends to be noisy and dispersed, caused by a large number of mostly irrelevant images. To tackle those problems, we introduce the Reciprocal Neighbor Softmax function to achieve sparse attention to neighbors, which reduces the noisy interactions with irrelevant individuals and makes the representation aggregation more effective and efficient. We illustrate the full pipeline in figure 2.

3.3. Landmark Agent Attention

Instead of measuring the similarity between high-dimensional representation vectors, we propose a more efficient way to obtain an approximate affinity matrix $\widetilde{\mathbf{A}}$. The key idea is to map the high-dimensional representation vectors z into a lower-dimensional encoding space, making the affinity computations in equation (3) considerably more efficient, as inspired by random Fourier features [24].

As shown in figure 3, following Transformer [31], the query, key and value matrices $\mathbf{q}, \mathbf{k}, \mathbf{v} \in \mathbb{R}^{N\times d}$ are obtained by three separate linear projections $\varphi_q(\cdot), \varphi_k(\cdot), \varphi_v(\cdot)$ using representation vectors $\mathbf{z} \in \mathbb{R}^{N\times d}$ as input. Specifically, we randomly sample l representations $\mathbf{z}_l \in \mathbb{R}^{l\times d}$ from $\mathbf{z}$ as landmark agents, and then obtain the $\mathbf{q}_l$ and $\mathbf{k}_l$ matrices with $\varphi_q(\cdot)$ and $\varphi_k(\cdot)$. Thus we could map the original query and key matrices $\mathbf{q}, \mathbf{k} \in \mathbb{R}^{N\times d}$ to a l-dimensional space by $\widetilde{\mathbf{q}} = \mathbf{q}\mathbf{k}_l^\top, \widetilde{\mathbf{k}} = \mathbf{k}\mathbf{q}_l^\top$, where $\widetilde{\mathbf{q}}, \widetilde{\mathbf{k}} \in R^{N\times l}$. $\widetilde{\mathbf{q}}_{ij}, \widetilde{\mathbf{k}}_{ij}$ indicate the similarity between representation vector $i \in \{1, \ldots, N\}$ and landmark agent $j \in \{1, \ldots, l\}$. Then the equation (3) could be replaced by:

$$\widetilde{\mathbf{A}}_{ij} = (\mathbf{q}\mathbf{k}_l^\top)_i (\mathbf{k}\mathbf{q}_l^\top)_j^\top / \sqrt{d} = \widetilde{\mathbf{q}}_i \widetilde{\mathbf{k}}_j^\top / \sqrt{d}. \tag{4}$$

In this way, we decompose the computation of the large affinity map $\mathbf{A} \in \mathbb{R}^{N\times N}$ into a multiplication of two low-rank matrices $\widetilde{\mathbf{q}}, \widetilde{\mathbf{k}}$. Thus, the multiplication complexity for obtaining the affinity matrix is significantly reduced from $O(N^2d)$ to $O(N^2l)$, since l is typically much smaller than d ($l = 5, d \geq 256$ in our experiments). In the Supplementary Material section A, we further prove that the cosine similarity of $\mathbf{A}$ and $\widetilde{\mathbf{A}}$ is positively correlated with l, with larger l yielding a cosine similarity close to 1,

$$\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}^{l_b})) \geq \cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}^{l_a})), \tag{5}$$

where $l_b > l_a$. In fact, as we show experimentally in figure 6 (a), even with a small number of landmark agents, the NFormer is able to perform stably.

3.4. Reciprocal Neighbor Softmax

After obtaining the approximate affinity matrix $\widetilde{\mathbf{A}}$, a softmax function s is typically used in equation (2) to turn the affinities into attention weights (probabilities). We can rewrite equation (2) as a sum of two parts, $\mathbf{u}_i = \sum_{j:\widetilde{\mathbf{A}}_{ij}\leq\rho} s(\widetilde{\mathbf{A}})_{ij}\varphi_v(\mathbf{z}_j) + \sum_{j:\widetilde{\mathbf{A}}_{ij}>\rho} s(\widetilde{\mathbf{A}})_{ij}\varphi_v(\mathbf{z}_j)$, where ρ is a small threshold. The first part represents the sum of elements with small attention weights and the second part represents the sum of elements with large attention weights. Although each of the attention weights in $\sum_{j:\widetilde{\mathbf{A}}_{ij}\leq\rho} s(\widetilde{\mathbf{A}})_{ij}\varphi_v(\mathbf{z}_j)$ is small, with a growing number of samples N the total summation will still be large and comparable to the second term in the summation, as shown in figure 4 (a). As a result, the final computation of $\mathbf{u}_i$ will be negatively impacted by the significant presence of irrelevant samples. Besides the negative effect in the output representation $\mathbf{u}_i$, the computation complexity of the representation aggregation is $O(N^2d)$, which presents a significant computational burden because of the large input size N.

To mitigate the above problems, we propose the Reciprocal Neighbor Softmax (RNS) to enforce sparsity to few relevant attention weights with a reciprocal neighbor mask. We assume that if two images are reciprocal neighbors with each other in feature space, they are likely to be relevant. To this end, we propose to compute a top-k neighbor mask $\mathbf{M}^k$ from the approximate affinity map $\widetilde{\mathbf{A}}$, which will attend to the top-k value of affinities per row:

$$\mathbf{M}^k_{ij} = \begin{cases} 1, & j \in \mathrm{topk}(\widetilde{\mathbf{A}}_{i,:}) \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

We can then obtain a reciprocal neighbor mask $\mathbf{M}$ by multiplying $\mathbf{M}^k$ with its transposition using Hadamard Product.

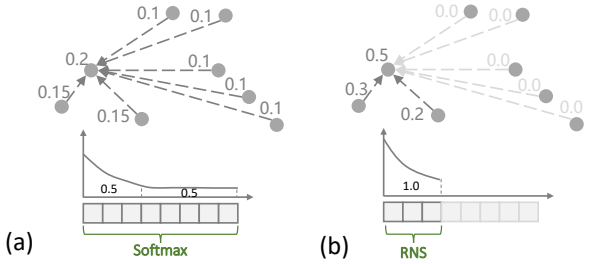

Figure 4. Illustration of Reciprocal Neighbor Softmax. (a) indicates the normal softmax, in which the softmax is performed on all the input representations, thus lots of irrelevant representations will contribute to the feature aggregation and distract the attention. (b) indicates the Reciprocal Neighbor Softmax, in which only the relations between reciprocal neighbors are kept.

$$\begin{aligned}\mathbf{M}_{ij} &= \mathbf{M}^k \circ \mathbf{M}^{k\top} \\ &= \begin{cases} 1, & j \in \text{topk}(\widetilde{\mathbf{A}}_{i,:}), i \in \text{topk}(\widetilde{\mathbf{A}}_{:,j}) \\ 0, & \text{otherwise.} \end{cases}\end{aligned} \tag{7}$$

For each element $\mathbf{M}_{ij}$, the value will be set to 1 if i and j are both top-k neighbors of each other, 0 otherwise. By adding this mask $\mathbf{M}$ to the regular softmax function, we achieve sparse attention only occurring in neighbors, which increases the focus on more relevant images. The formula of RNS is shown as follows:

$$\text{RNS}(\mathbf{A})_{ij} = \frac{\mathbf{M}_{ij}\exp(-\widetilde{\mathbf{A}}_{ij})}{\sum_k \mathbf{M}_{ik}\exp(-\widetilde{\mathbf{A}}_{ik})}, \tag{8}$$

Since most attention values are set to zero, as shown in figure 4 (b), the relations are constrained to the relevant neighbors, making the aggregation in equation (2) more focused and robust. Furthermore, as we do not need to conduct addition operation for the representations with zero weights, the time complexity of feature aggregation significantly decreases from $O(N^2d)$ to $O(Nkd)$.

4. Experiments

We first describe the datasets, the evaluation protocols, and the implementation details of NFormer. Then, we conduct extensive ablation studies to demonstrate the effectiveness and efficiency brought by each of the proposed modules. Finally, we compare NFormer with other state-of-the-art methods on four large-scale Re-ID datasets.

4.1. Datasets and Evaluation Protocols

We conduct experiments on four widely used large-scale person Re-ID datasets: Market1501 [44], DukeMTMC-reID [25], MSMT17 [35] and CUHK03 [16] to validate the effectiveness and efficiency of NFormer. All the above-mentioned datasets contain multiple images for each identity collected from different cameras or scenarios. We follow the standard person Re-ID experimental setups. We use the standard metric in the literature [44] for evaluation: the cumulative matching characteristic (CMC) curve and mean Average Precision (mAP). CMC shows the top K accuracy by counting the true positives among the top K persons in the ranking list. The mAP metric measures the area under the precision-recall curve, which reflects the overall re-identification accuracy among the gallery set rather than only considering the top K persons.

4.2. Implementation

We adopt ResNet-50 [10] pre-trained on ImageNet [7] as the backbone architecture for our feature extractor. To preserve spatial information, we change the stride convolutions at the last stage of ResNet with dilated convolutions, which leads to a total downsampling ratio of 16. We then apply a fully connected layer after the Resnet-50 backbone to reduce the dimension of the embedding vector from 2048 to 256 for efficiency. We stack four LAA modules to build the NFormer. The number of landmark agents l in the LAA module is set to 5 and the number of neighbors k in RNS is set to 20 for a good trade-off between computational cost and performance according to the experimental results. During the inference, the interactions between the different query images are eliminated for fair comparison.

For all experiments, the images are resized to a fixed resolution of 256×128. Random horizontal flipping is utilized as data augmentation during training. We combine the identity loss [45], center loss [36] and triplet loss [11] to form the total loss function. The three loss functions are weighted by 1, 1, 0.0005 respectively. We use Stochastic Gradient Descent (SGD) as the optimizer. The initial learning rate is set to 3e-4 and momentum is set to 5e-4. We train the Resnet-50 and NFormer in turn for 160 epochs. The batch size is set to 128 for training the Resnet-50 feature extractor and is set to 2048 for training NFormer. We freeze the parameters of Resnet-50 during the NFormer training iteration to achieve such a large batch size. All the experiments are conducted with PyTorch on one GeForce RTX 3090.

4.3. Ablation Study

We conduct comprehensive ablation studies on Market-1501 and dukeMTMC-reID datasets to analyze the effectiveness of LAA and RNS with different hyper-parameters. With Res50 we denote the modified ResNet-50 feature extractor without the NFormer, and use it as the baseline.

NFormer vs. Transformer vs. Res50. Table 1 shows the comparison between NFormer, the regular Transformer and Res50 baseline model on Market-1501 and dukeMTMC-reID datasets. The regular Transformer, without any special design, slightly surpasses the baseline model by 0.5%/1.6% and 0.5%/1.3% top-1/mAP on Market-1501 and dukeMTMC-reID. With the LAA module and RNS function, NFormer outperforms the baseline model by a

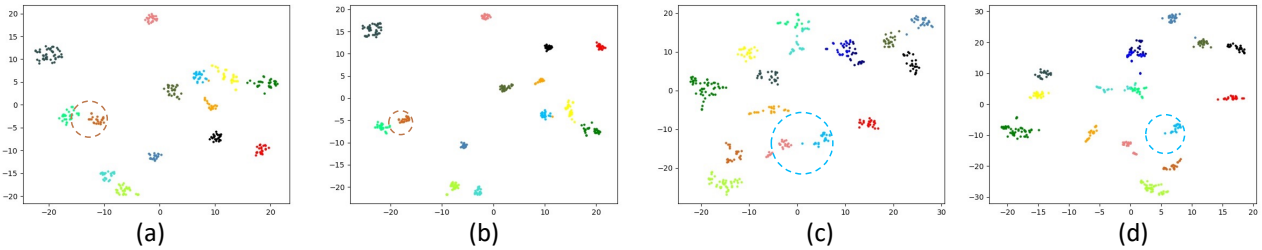

Figure 5. t-SNE visualization of representation vectors. (a)/(b) show several random sampled identities on Market-1501 without/with NFormer. (c)/(d) show several random sampled identities on dukeMTMC-reID without/with NFormer. In this figure, we can see that after NFormer, the representation distribution is more gathered and detached. Specifically, if we choose one of the brown points as query person in figure (a), there will be a lot of cyan points at the top of the ranking list, as shown in the brown circle in (a). On the contrary, the ranking list of the same query person contains fewer negative persons in (b) because of the more gathered and detached distribution. The blue circles in (c) and (d) show the same results.

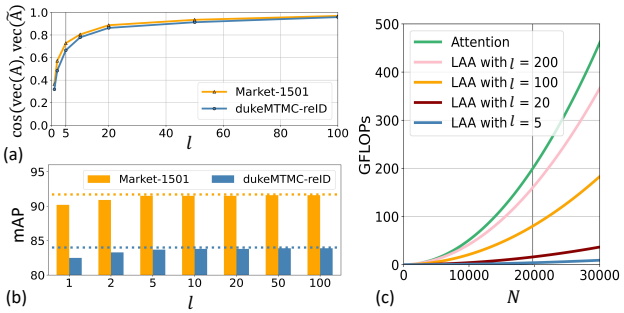

Figure 6. Figure (a) shows how the $\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}))$ changes with the number of landmark agents l. Figure (b) shows how the mAP changes with l, in which the orange and blue dash lines show the mAP performance without LAA (with normal affinity matrix). Figure (c) shows how the total GFLOPs changes with input number N under different l.

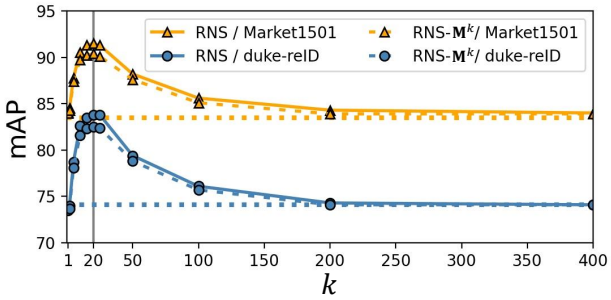

Figure 7. This figure shows how mAP changes with the number of neighbors k in RNS on Market-1501 and dukeMTMC-reID. RNS-$\mathbf{M}^k$ indicates RNS with top-k neighbor mask $\mathbf{M}^k$ instead of reciprocal neighbor mask $\mathbf{M}$. The orange and blue horizontal dash lines show the mAP performance with normal softmax function.

considerably larger margin of 1.5%/7.6% and 3.3%/9.4% top-1/mAP on Market-1501 and dukeMTMC-reID respectively. Notably, NFormer requires 1.5 orders of magnitude fewer computations (0.0025 GFLOPs vs 0.088 GFLOPs per person for the regular Transformer).

We qualitatively demonstrate the effectiveness of NFormer by visualizing the representation distribution before and after NFormer on Market-1501 and dukeMTMC-reID datasets in figure 5. We observe better feature discriminability for the NFormer, while the outliers of each identity are significantly constrained because the relevant and common information of neighbors is integrated into each data point. We conclude that NFormer learns relations between input persons not only effectively but also efficiently.

Method	Market-1501		DukeMTMC		GFLOPs
	T-1	mAP	T-1	mAP	
Res50	93.2	83.5	86.1	74.1	
+Transformer [31]	93.7	85.1	86.6	75.4	0.088
+NFormer	94.7	91.1	89.4	83.5	0.0025

Table 1. mAP and GFLOPs comparison between Res50 baseline model, normal Transformer, and NFormer on Market-1501 and dukeMTMC-reID datasets. GFLOPs mean the average number of floating-point operations for processing each input image.

Influence of Landmark Agent Attention. We first study the effect of the number of landmark agents l for computing the approximate affinity matrix $\widetilde{\mathbf{A}}$. Figure 6 (a) shows that the cosine similarity $\cos(\mathrm{vec}(\mathbf{A}), \mathrm{vec}(\widetilde{\mathbf{A}}))$ is positive relative with the number of landmark agents l, and monotonically increasing, approaching 1 even with a small l. As shown in figure 6 (b), the mAP performance on Market-1501 and dukeMTMC-reID achieves 91.1% and 83.5% when $l = 5$, which causes only 0.3% and 0.3% drops compared with original affinity map without LAA module. When l gets larger, the cosine similarity and mAP performance are saturated while the FLOPs continue to grow, as shown in figure 6 (c). So we choose $l = 5$ as a good balance between effectiveness and efficiency. That is, the LAA module only needs 1.95% of the computations to obtain an approximate affinity map $\widetilde{\mathbf{A}}$, while basically maintaining the performance compared with the original affinity map $\mathbf{A}$.

Influence of Reciprocal Neighbor Softmax. We show the effect of the number of reciprocal neighbors k in figure 7. When k increases, mAP performance of RNS on

Method	Market-1501		DukeMTMC	
	T-1	mAP	T-1	mAP
Res50	93.2	83.5	86.1	76.1
+NFormer	94.7	91.1	89.4	83.5
+NFormer+KR [47]	94.6	93.0	89.5	88.2
*ABDNet [4]	95.4	88.2	88.7	78.6
+NFormer	95.7	93.0	90.6	85.7
+NFormer+KR [47]	95.7	94.1	91.1	89.4

Table 2. Performance of combination of NFormer, ABD-Net and KR re-ranking method on Market-1501 and DukeMTMC datasets. * represents our reproduced performance.

Market-1501 and dukeMTMC-reID firstly reaches the maximum values 91.1% and 83.5% at $k = 20$. This is because more neighbors information benefits the aggregation of the individual representations in the early stage. Then as k continues to increase, the performance gradually decreases because of the introduction of irrelevant interactions. Therefore, we set k to 20 in all the experiments with RNS. As shown, RNS outperforms normal Softmax function (horizontal dash lines in figure 7) by 7.3% and 8.9% on Market-1501 and dukeMTMC-reID in terms of mAP, which indicates that attending to relevant reciprocal neighbors only leads to better feature representations compared with directly incorporating all the images. Besides, RNS outperforms RNS-$\mathbf{M}^k$ under the different number of neighbors k consistently, which shows that the reciprocal neighbor mask $\mathbf{M}$ could provide better prior knowledge of learning relations between input images.

Complementarity to third methods. NFormer is easy to combine with other methods. We showcase this by choosing a SOTA feature extractor ABD-net [4] for representation learning and a re-ranking method RP [47] to combine with NFormer. As shown in table 2, NFormer with ABDNet and RP further boosts the performance by 1.0%/3.0% and 1.7%/5.9% top-1/mAP on Market-1501 and dukeMTMC-reID, which demonstrates the compatibility of NFormer.

Limitation. NFormer learns information from neighbor persons in the feature space. If the number of images for each identity in the testset is small, then the individuals will not be able to obtain a lot of useful information from each other. We conduct an ablation study on Market-1501 and dukeMTMC-reID datasets to analyze the influence of the average number of images per identity. Specifically, we sample 4 sub-testsets from the original testsets of Market-1501 and dukeMTMC-reID respectively. Each sub-testset has a different average number of images per identity. We then evaluate NFormer and Res50 baseline model on each sub-testset. The results are shown in table 3, from which we can see that as the number of images per identity reduces from 20 to 5, the improvements brought by

Methods	Dataset	Market-1501				DukeMTMC-reID			
	Subset	0	1	2	3	0	1	2	3
	n/p	20	15	10	5	20	15	10	5
Res50	mAP	83.7	83.9	84.1	85.5	74.4	74.4	75.2	76.1
+NFormer		91.0	90.6	90.1	87.8	83.6	83.0	81.7	79.9
	ΔmAP	+7.3	+6.7	+6.0	+2.3	+9.2	+8.6	+6.5	+3.8

Table 3. The mAP performance of NFormer and Res50 baseline model on sampled sub-testsets with different n/p of Market-1501 and dukeMTMC-reID datasets. n/p indicates the average number of images per identity.

NFormer ($\Delta\mathrm{mAP}$) drops significantly from 7.3%/9.2% to 2.3%/3.8% on Market-1501 and dukeMTMC-reID datasets. By contrast, the performance of the Res50 baseline model barely changes, and even slightly increases. The reason is that as the number of images decreases, it is easier to search through the new and smaller test sets. The results confirm, therefore, that a limitation of the NFormer is that it expects a large enough number of images of the same person. This makes NFormer particularly interesting in more complex and large-scale settings with many cameras and crowds and less relevant in smaller setups.

4.4. Comparison with SOTA methods

Last, we compare the performance of NFormer with recent state-of-the-art methods on Market1501, DukeMTMC-reID, MSMT17 and CUHK03 in table 4. Overall, our proposed NFormer outperforms other state-of-the-arts or achieves comparable performance.

Results on Market-1501. As shown in table 4, NFormer achieves the best mAP and comparable top-1 accuracy among all the state-of-the-art competitors. Specifically, even with a simple feature extractor Res50, the mAP of NFormer outperforms the second-best method ISP [51] (with HRNet-W30 [33] backbone) by a large margin 2.5%. When combining NFormer with a better feature extractor from ABDNet [4], the mAP/rank-1 accuracy is further boosted by 1.9%/1.0% and outperforms the ISP [51] by 4.4% in terms of mAP. Notably, NFormer outperforms methods STF [21] and GCS [3] which build relations inside each training batch by 8.4% and 9.5% in terms of mAP. This indicates that the relation modeling among all input images both during training and test leads to better representations. The visualization of the ranking lists is shown in the Supplementary Material section B, from which we can see that the NFormer could help to constrain the outliers and improve the robustness of the ranking process.

Results on DukeMTMC-reID. The results are presented in table 4, from which we can see that our method outperforms other state-of-the-arts significantly. Specifically, NFormer with Res50 feature extractor gains 3.5% improvement in terms of mAP over second-best method

Method	Market-1501		duke-reID		MSMT17		CUHK03-L		CUHK03-D	
	T-1	mAP	T-1	mAP	T-1	mAP	T-1	mAP	T-1	mAP
PCB+RPP (ECCV'18) [29]	93.8	81.6	83.3	69.2	68.2	40.4	-	-	63.7	57.5
GCS (CVPR'18) [3]	93.5	81.6	84.9	69.5	-	-	-	-	-	-
MHN (ICCV'19) [2]	95.1	85.0	89.1	77.2	-	-	77.2	72.4	71.7	76.5
OSNet (ICCV'19) [50]	94.8	84.9	88.6	73.5	78.7	52.9	-	-	72.3	67.8
Pyramid (CVPR'19) [43]	95.7	88.2	89.0	79.0	-	-	78.9	76.9	78.9	74.8
IANet (CVPR'19) [12]	94.4	83.1	87.1	73.4	75.5	46.8	-	-	-	-
STF (ICCV'19) [21]	93.4	82.7	86.9	73.2	73.6	47.6	68.2	62.4	-	-
BAT-net (ICCV'19) [9]	94.1	85.5	87.7	77.3	79.5	56.8	78.6	76.1	76.2	73.2
PISNet (ECCV'20) [42]	95.6	87.1	88.8	78.7	-	-	-	-	-	-
CBN (ECCV'20) [52]	94.3	83.6	84.8	70.1	-	-	-	-	-	-
RGA-SC (CVPR'20) [40]	**96.1**	88.4	-	-	80.3	57.5	**81.1**	77.4	79.6	74.5
ISP (ECCV'20) [51]	95.3	88.6	89.6	80.0	-	-	76.5	74.1	75.2	71.4
CBDB-Net (TCSVT'21) [30]	94.4	85.0	87.7	74.3	-	-	77.8	76.6	75.4	72.8
CDNet (CVPR'21) [15]	95.1	86.0	88.6	76.8	78.9	54.7	-	-	-	-
PAT (CVPR'21) [18]	95.4	88.0	88.8	78.2	-	-	-	-	-	-
C2F (CVPR'21) [39]	94.8	87.7	87.4	74.9	-	-	80.6	**79.3**	**81.3**	**84.1**
Res50	93.2	83.5	86.1	76.1	74.9	50.1	74.7	73.8	73.4	71.2
+NFormer	94.7	91.1	89.4	83.5	77.3	59.8	77.2	78.0	77.3	74.7
*ABDNet(ICCV'19) [4]	95.4	88.2	88.7	78.6	78.4	55.5	78.7	75.8	77.3	73.2
+NFormer	95.7	**93.0**	**90.6**	**85.7**	**80.8**	**62.2**	80.6	79.1	79.0	76.4

Table 4. Quantitative results on Market-1501, DukeMTMC-reID, MSMT17 and CUHK03 datasets. T-1 means top-1 accuracy and mAP means mean average precision. The best performance value in each column is marked by bold and the second-best performance value is marked by underline. The symbol "-" indicates that the corresponding value is not provided in the corresponding paper. * represents our reproduced performance.

ISP [51]. NFormer with ABDNet feature extractor outperforms ISP [51] by 1.0%/5.7% in terms of top-1/mAP. We observe that the improvement is more pronounced for the mAP than for the top-1 metric. The reason is that NFormer reforms the representation of all the input persons, and in general impacts positively the overall search, not just the top retrieval.

Results on MSMT17. As shown in table 4, the NFormer with Res50 feature extractor outperforms the second best method RGA-SC [40] (ResNet-50 backbone) by 2.3% in terms of mAP, while NFormer with ABDNet feature extractor outperforms RGA-SC [40] by 0.5%/4.7% in terms of top-1/mAP. The NFormer outperforms the baseline model significantly by 2.4%/9.7% top-1/mAP, which shows that NFormer works even better on larger datasets, as there is rich neighbor information for each person.

Results on CUHK03. We conduct experiments on both the manually labelled version and the detected version of CUHK03 dataset. From table 4, we can see that the NFormer with ABD-net achieves comparable performance on both labelled and detected sets. NFormer with Res50 feature extractor outperforms baseline model by 2.5%/4.2% and 3.9%/3.5% top-1/mAP on Labelled and Detected sets. We further illustrate the reasons for the fewer improvements on CUHK03 dataset. We count that the average number of images per identity in CUHK03 is 9.6, which is much less than 25.7 in Market-1501, 23.4 in DukeMTMC-reID, and 30.7 in MSMT17. So the NFormer can not learn much relevant information from the neighbors. We provide a detailed analysis in the **limitation** part in the ablation study.

5. Conclusion

In this paper, we propose a novel Neighbor Transformer Network for person re-identification, which interacts between input images to yield robust and discriminative representations. In contrast to most existing methods focusing on single images or a few images inside a training batch, our proposed method models the relations between all the input images. Specifically, we propose a Landmark Agent Attention to allow for more efficient modeling of the relations between a large number of inputs, and a Reciprocal Neighbor Softmax to achieve sparse attention to neighbors. As such, NFormer scales well with large input and is robust to outliers. In extensive ablation studies, we show that NFormer learns robust, discriminative representations, which are easy to combine with third methods.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[2] Binghui Chen, Weihong Deng, and Jiani Hu. Mixed high-order attention network for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 371–381, 2019. 8

[3] Dapeng Chen, Dan Xu, Hongsheng Li, Nicu Sebe, and Xiaogang Wang. Group consistent similarity learning via deep crf for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8649–8658, 2018. 1, 7, 8

[4] Tianlong Chen, Shaojin Ding, Jingyi Xie, Ye Yuan, Wuyang Chen, Yang Yang, Zhou Ren, and Zhangyang Wang. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8351–8361, 2019. 7, 8

[5] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8126–8135, 2021. 3

[6] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *Proceedings of the iEEE conference on computer vision and pattern recognition*, pages 1335–1344, 2016. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[9] Pengfei Fang, Jieming Zhou, Soumava Kumar Roy, Lars Petersson, and Mehrtash Harandi. Bilinear attention networks for person retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8030–8039, 2019. 8

[10] 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. 5

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017. 5

[12] Ruibing Hou, Bingpeng Ma, Hong Chang, Xinqian Gu, Shiguang Shan, and Xilin Chen. Interaction-and-aggregation network for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9317–9326, 2019. 8

[13] Houjing Huang, Dangwei Li, Zhang Zhang, Xiaotang Chen, and Kaiqi Huang. Adversarially occluded samples for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5098–5107, 2018. 2

[14] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 384–393, 2017. 1

[15] Hanjun Li, Gaojie Wu, and Wei-Shi Zheng. Combined depth space based architecture search for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6729–6738, 2021. 8

[16] 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. 5

[17] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2018. 1, 2, 3

[18] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907, 2021. 8

[19] 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*, pages 441–448, 2013. 2

[20] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018. 1, 2

[21] Chuanchen Luo, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Spectral feature transformation for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4976–4985, 2019. 1, 7, 8

[22] Andy Jinhua Ma and Ping Li. Query based adaptive re-ranking for person re-identification. In *Asian Conference on Computer Vision*, pages 397–412. Springer, 2014. 2

[23] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–667, 2018. 1

[24] Ali Rahimi, Benjamin Recht, et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, page 5. Citeseer, 2007. 4

[25] 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*, pages 17–35. Springer, 2016. 5

[26] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. End-to-end deep kronecker-product matching for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6886–6895, 2018. 1, 2, 3

[27] 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. 2

[28] Yumin Suh, Jingdong Wang, Siyu Tang, Tao Mei, and Kyoung Mu Lee. Part-aligned bilinear representations for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2018. 1, 2

[29] 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 (ECCV)*, pages 480–496, 2018. 2, 8

[30] Hongchen Tan, Xiuping Liu, Yuhao Bian, Huasheng Wang, and Baocai Yin. Incomplete descriptor mining with elastic loss for person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 8

[31] 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, 3, 4, 6

[32] Cheng Wang, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 365–381, 2018. 1, 2, 3

[33] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 7

[34] Yicheng Wang, Zhenzhong Chen, Feng Wu, and Gang Wang. Person re-identification with cascaded pairwise convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1470–1478, 2018. 1, 2

[35] 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. 5

[36] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016. 5

[37] Mang Ye, Chao Liang, Zheng Wang, Qingming Leng, and Jun Chen. Ranking optimization for person re-identification via similarity and dissimilarity. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1239–1242, 2015. 2

[38] Mang Ye, Chao Liang, Yi Yu, Zheng Wang, Qingming Leng, Chunxia Xiao, Jun Chen, and Ruimin Hu. Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. *IEEE Transactions on Multimedia*, 18(12):2553–2566, 2016. 2

[39] Anguo Zhang, Yueming Gao, Yuzhen Niu, Wenxi Liu, and Yongcheng 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*, pages 598–607, 2021. 8

[40] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 3186–3195, 2020. 8

[41] Liming Zhao, Xi Li, Yueting Zhuang, and Jingdong Wang. Deeply-learned part-aligned representations for person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 3219–3228, 2017. 1, 2

[42] Shizhen Zhao, Changxin Gao, Jun Zhang, Hao Cheng, Chuchu Han, Xinyang Jiang, Xiaowei Guo, Wei-Shi Zheng, Nong Sang, and Xing Sun. Do not disturb me: Person re-identification under the interference of other pedestrians. In *European Conference on Computer Vision*, pages 647–663. Springer, 2020. 8

[43] Feng Zheng, Cheng Deng, Xing Sun, Xinyang Jiang, Xiaowei Guo, Zongqiao Yu, Feiyue Huang, and Rongrong Ji. Pyramidal person re-identification via multi-loss dynamic training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8514–8522, 2019. 8

[44] 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. 5

[45] 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*, pages 1367–1376, 2017. 5

[46] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE international conference on computer vision*, pages 3754–3762, 2017. 1, 2

[47] 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*, pages 1318–1327, 2017. 2, 7

[48] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020. 2

[49] Jiahuan Zhou, Pei Yu, Wei Tang, and Ying Wu. Efficient online local metric adaptation via negative samples for person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2420–2428, 2017. 2

[50] Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-scale feature learning for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3702–3712, 2019. 8

[51] Kuan Zhu, Haiyun Guo, Zhiwei Liu, Ming Tang, and Jinqiao Wang. Identity-guided human semantic parsing for person re-identification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 346–363. Springer, 2020. 7, 8

[52] Zijie Zhuang, Longhui Wei, Lingxi Xie, Tianyu Zhang, Hengheng Zhang, Haozhe Wu, Haizhou Ai, and Qi Tian. Rethinking the distribution gap of person re-identification with camera-based batch normalization. In *European Conference on Computer Vision*, pages 140–157. Springer, 2020. 8