From Poses to Identity: Training-Free Person Re-Identification via Feature Centralization

Supplementary Material

1. Methods Supplementary

1.1. Aggregation role of ReID loss functions

Currently, ReID models commonly use cross-entropy loss to impose ID-level constraints, and contrastive losses (such as triplet loss) to bring features of the same ID closer while pushing apart features of different IDs. Some models also utilize center loss to construct identity centers for dynamically constraining the IDs. These methods lead to one common result: feature aggregation. From the perspective of the gradient of the loss functions, we could prove that the feature vectors of each ID in current ReID tasks naturally aggregate around a center or mean in the followings.

Cross-Entropy Loss is often used in classification tasks, optimizing the model by maximizing the probability of the correct class. Given N samples, each with a feature vector $\mathbf{z}_i \in \mathbb{R}^d$, and its corresponding class label $y_i \in \{1, 2, \ldots, C\}$, the cross-entropy loss is defined as:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\mathbf{w}_{y_i}^\top \mathbf{z}_i + b_{y_i})}{\sum_{j=1}^{C}\exp(\mathbf{w}_j^\top \mathbf{z}_i + b_j)} \tag{1}$$

where $\mathbf{w}_j$ and b_j are the weight vector and bias for class j, respectively.

For simplicity, assume that the final layer is a linear classifier without bias, i.e., $b_j = 0$. When the loss is minimized, the optimization objective is to maximize the score $\mathbf{w}_{y_i}^\top \mathbf{z}_i$ of the correct class while minimizing the scores $\mathbf{w}_j^\top \mathbf{z}_i$ of other classes ($j \neq y_i$).

By gradient descent optimization, we can obtain:

$$\frac{\partial \mathcal{L}_{\text{CE}}}{\partial \mathbf{z}_i} = 1/N\left(p_{y_i} - 1\right)\mathbf{w}_{y_i} + 1/N\sum_{j\neq y_i} p_{ij}\mathbf{w}_j \tag{2}$$

where $p_{ij} = \frac{\exp(\mathbf{w}_j^\top \mathbf{z}_i)}{\sum_{k=1}^{C}\exp(\mathbf{w}_k^\top \mathbf{z}_i)}$.

With the loss function converges, $p_{y_i} \to 1$ and $p_{ij} \to 0 (j \neq y_i)$. The feature $\mathbf{z}_i$ is optimized to be near a linear combination of the class weight vectors $\mathbf{w}_{y_i}$. This indicates that features of the same class will tend toward a common direction, thus achieving feature aggregation.

Contrastive loss (Triplet Loss as example) optimizes the feature space by bringing samples of the same class closer and pushing samples of different classes further apart. A triplet $(\mathbf{z}_a, \mathbf{z}_p, \mathbf{z}_n)$ is defined, where $\mathbf{z}_a$ is the anchor, $\mathbf{z}_p$ is the positive sample (same class), and $\mathbf{z}_n$ is the negative sample (different class). The triplet loss is defined as:

$$\mathcal{L}_{\text{Triplet}} = \max\left(\|\mathbf{z}_a - \mathbf{z}_p\|_2^2 - \|\mathbf{z}_a - \mathbf{z}_n\|_2^2 + \alpha, 0\right) \tag{3}$$

where α is the margin parameter.

To minimize the loss, the optimization objective is:

$$\|\mathbf{z}_a - \mathbf{z}_p\|_2^2 + \alpha < \|\mathbf{z}_a - \mathbf{z}_n\|_2^2 \tag{4}$$

$$\frac{\partial \mathcal{L}_{\text{Triplet}}}{\partial \mathbf{z}_a} = 2(\mathbf{z}_n - \mathbf{z}_p), \tag{5}$$

$$\frac{\partial \mathcal{L}_{\text{Triplet}}}{\partial \mathbf{z}_p} = 2(\mathbf{z}_p - \mathbf{z}_a), \tag{6}$$

$$\frac{\partial \mathcal{L}_{\text{Triplet}}}{\partial \mathbf{z}_n} = 2(\mathbf{z}_a - \mathbf{z}_n). \tag{7}$$

By minimizing triplet loss, the feature $\mathbf{z}_p$ is pulled closer to $\mathbf{z}_a$, while $\mathbf{z}_n$ is pushed away. Through this mechanism, Triplet Loss encourages features of the same class to aggregate together while features of different classes are separated from each other.

Center loss further enhances feature aggregation by introducing a feature center for each class. For each class j, there is a feature center $\mathbf{c}_j$, and the center loss is defined as:

$$\mathcal{L}_{\text{Center}} = \frac{1}{2}\sum_{i=1}^{N}\|\mathbf{z}_i - \mathbf{c}_{y_i}\|_2^2 \tag{8}$$

The goal of minimizing center loss is to make each sample's feature vector $\mathbf{z}_i$ as close as possible to its corresponding class center $\mathbf{c}_{y_i}$. Through gradient descent, we obtain:

$$\frac{\partial \mathcal{L}_{\text{Center}}}{\partial \mathbf{z}_i} = \mathbf{z}_i - \mathbf{c}_{y_i} \tag{9}$$

$$\frac{\partial \mathcal{L}_{\text{Center}}}{\partial \mathbf{c}_j} = \begin{cases} \mathbf{c}_j - \mathbf{z}_i & \text{if } y_i = j \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Thus, the optimization process not only pulls sample features closer to their centers but also dynamically updates each class's center to represent the mean of that class's feature distribution. This directly encourages features of the same class to aggregate together.

1.2. Identity Density (ID^2) Metric

Identity density is one aspect of measuring ReID effectiveness. However, there is currently no quantitative metric for this, and researchers commonly rely on visualization tools like t-SNE to demonstrate model performance. Due to the large number of IDs, this approach is limited to visualizing only a few IDs, making it challenging to assess model performance from a global perspective quantitatively. Some researchers exploit this limitation by selecting the best-performing IDs of their models for visualization. To address this, we propose an Identity Density (ID^2) Metric. This metric evaluates the global ID aggregation performance by taking each ID center across the entire test set (gallery and query) as a benchmark.

$$\text{ID}^2 = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i}\sum_{j=1}^{n_i} d\left(\frac{f_{ij}}{\|f_{ij}\|_2}, c_i\right) \tag{11}$$

where N is the total number of unique IDs in the test set, and n_i is the number of samples for ID i. The feature vector of the j-th sample of ID i is denoted as f_{ij}, and c_i represents the identity center of ID i, computed as follows:

$$c_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\frac{f_{ij}}{\|f_{ij}\|_2} \tag{12}$$

Both the feature vectors f_{ij} and the identity centers c_i are L_2-normalized to ensure consistent feature scaling. The function $d(\cdot,\cdot)$ represents the Euclidean distance.

1.3. Pose Encoder Details

The Pose Encoder module is designed to extract high-dimensional pose embeddings from the input poses.

$$\mathbf{E}_{\text{pose}} = \text{PoseEncoder}(\mathbf{x}^{\text{pose}}) \tag{13}$$

The input is a feature map of size $C_{\text{in}} \times H \times W$, denoted as $\mathbf{x}^{\text{pose}}$, where C_{in} is the number of input channels, and H, W are the height, and width of the input. The first convolution layer is defined as:

$$\mathbf{E}_0 = \text{SiLU}(\text{Conv}_{\text{in}}(\mathbf{x}^{\text{pose}})) \tag{14}$$

where Conv_{in} is a convolution operation with kernel size 3×3, and the number of channels changes from $C_{\text{in}} = 3$ to $C_0 = 16$:

Each block applies a normal 3×3 Conv, a 3×3 Conv with stride 2 to reduce spatial dimensions, and followed by a SiLU activate function. For the i-th convolutional block, the operations can be expressed as:

$$\mathbf{E}_{i+1} = \text{SiLU}(\text{Conv}_{i,\text{stride}=2}(\text{Conv}_i(\mathbf{E}_i))) \tag{15}$$

The number of channels for each block is as follows: $[C_0, C_1, C_2, C_3] = [16, 32, 64, 128]$

The output Conv layer maps the features from the last block to the target embedding dimension $C_{\text{out}} = 320$, expressed as:

$$\mathbf{E}_{\text{pose}} = \text{Conv}_{\text{out}}(\mathbf{E}_4) \tag{16}$$

1.4. Detailed Description of Neighbor Feature Centralization (NFC)

Step 1: Compute Distance Matrix Given all feature vectors in the gallery $\{\mathbf{z}_i\}_{i=1}^{N}$, our goal is to enhance each feature vector by aggregating features from its mutual nearest neighbors. Compute the pairwise distance matrix $\mathbf{D} = [d_{ij}]$ where d_{ij} represents the distance between features $\mathbf{z}_i$ and $\mathbf{z}_j$. To avoid self-matching, set the diagonal elements to a large constant, i.e.

$$d_{ii} = C, \quad \text{for } i = 1, 2, \ldots, N$$

Step 2: Find Top k_1 Nearest Neighbors For each feature $\mathbf{z}_i$, find its top k_1 nearest neighbors based on the distance matrix $\mathbf{D}$. Denote the set of indices of these neighbors as:

$$\mathcal{N}_i = \text{TopK}_{k_1}(\{d_{ij}\}_{j=1}^{N}) \tag{17}$$

Step 3: Identify Mutual Nearest Neighbors For each feature $\mathbf{z}_i$, identifies its mutual nearest neighbors by checking whether each neighbor in $\mathcal{N}_i$ also considers $\mathbf{z}_i$ as one of its top k_2 nearest neighbors. Specifically, for each $j \in \mathcal{N}_i$, checks if $i \in \mathcal{N}_j^{k_2}$, where $\mathcal{N}_j^{k_2}$ is the set of indices of the top k_2 nearest neighbors of $\mathbf{z}_j$. If this condition is satisfied, add j to the mutual nearest neighbor set $\mathcal{M}_i$:

$$\mathcal{M}_i = \{j \mid j \in \mathcal{N}_i,\ i \in \mathcal{N}_j^{k_2}\} \tag{18}$$

Step 4: Feature Centralization Enhancement Then it could centralize each feature vector $\mathbf{z}_i$ by aggregating the features of its mutual nearest neighbors:

$$\mathbf{z}_i^{\text{centralized}} = \mathbf{z}_i + \sum_{j \in \mathcal{M}_i} \mathbf{z}_j \tag{19}$$

This aggregation reduces feature noise and improves discriminability by incorporating information from similar features.

2. Experiments Supplementary

2.1. Data Cleansing

Training an effective generative model requires high-quality data support. In current ReID (Person Re-Identification) datasets, there are many low-quality images, and removing them can help reduce interference to the model. In our experiments, we found two main issues that need to be addressed: **Extremely Low-quality Images**: The dataset contains images with such low resolution that even the human

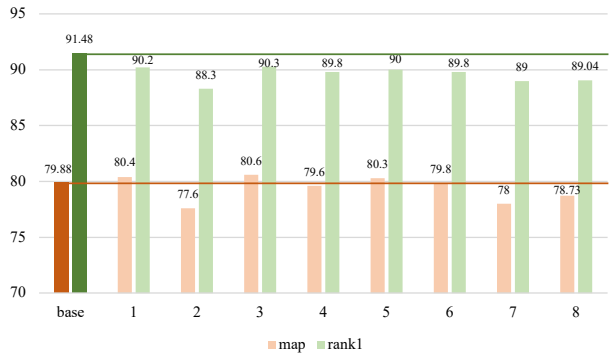

Figure 1. ReID results with images generated with the same pose on Market1501.

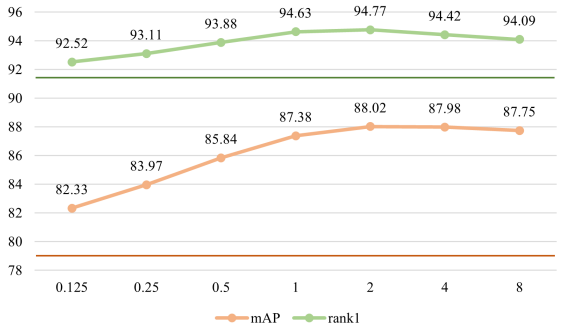

Figure 2. Impact of the quality coefficient η with TransReID on Market1501. The dark color lines are the baseline.

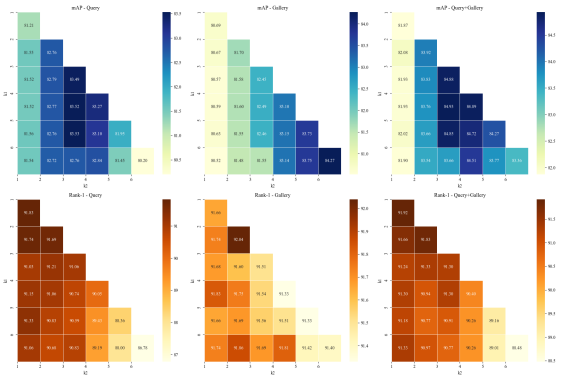

Figure 3. k_1/k_2 analysis of Neighbor Feature Centralization (NFC) with TransReID on Market1501 without re-ranking.

eye cannot recognize them as a "person". **Pose Estimation Failures**: The pose estimation model inevitably fails to detect pedestrian poses in some images.

2.1.1. Extremely Low-quality Images

To address this, manual filtering is impractical. Therefore, we designed an automated filtering algorithm. We leverage normal distribution of feature vector, if the feature on the edge of the distribution, largely due to the data itself is out of the distribution of its identity, and it can be picked up.

Let $\mathbf{f}_i \in \mathbb{R}^d$ denote the feature vector of the i-th sample of a particular identity, where d is the feature dimension. The mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are computed as follows:

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{f}_i, \quad \boldsymbol{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{f}_i - \boldsymbol{\mu})(\mathbf{f}_i - \boldsymbol{\mu})^\top \tag{20}$$

where N is the number of samples for a given ID.

To detect outliers, we compute the Mahalanobis distance d_i of each feature vector $\mathbf{f}_i$ from the mean vector $\boldsymbol{\mu}$, defined as:

$$dis_i = \sqrt{(\mathbf{f}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{f}_i - \boldsymbol{\mu})} \tag{21}$$

Given that the feature vectors are assumed to follow a multivariate normal distribution, we use quantiles of the Mahalanobis distance to filter out outliers. Specifically, we define a lower bound Q_p and an upper bound Q_{1-p} based on the p-th and $(1-p)$-th quantiles, respectively. Samples with distances outside this range are considered outliers and are removed, and can get a set S_i^{ref} for i^{th} ID:

$$S_i^{\text{ref}} = \{\mathbf{x}_i \mid dis_i \in [Q_p, Q_{1-p}]\} \tag{22}$$

2.1.2. Pose Filtering for 'Failed' Pose Estimation

We designed a pose filtering algorithm called PoseValid to eliminate cases where pose extraction has "completely failed." This algorithm checks the validity of the pose keypoints based on factors such as occlusion, keypoint positions, angles, and limb proportions, then get the set of valid poses.

$$S_i^{\text{trg}} = \{\mathbf{x}_i \mid \text{PoseValid}(\mathbf{x}_i) \text{ and } dis_i \in [Q_p, Q_{1-p}]\} \tag{23}$$

where the pose detector in this paper uses pretrained model of DWpose[26]. Given a set of keypoints representing a pose, we normalize the pose using the following steps:

1. Compute the body height (h):
   Calculate the Euclidean distance between the Neck (keypoint 1) and the Left Hip (keypoint 11):
   $$h = \|\mathbf{k}_{\text{Neck}} - \mathbf{k}_{\text{LHip}}\|$$
2. Translate the pose:
   Shift all keypoints so that the Neck is at the origin:
   $$\mathbf{k}_i' = \mathbf{k}_i - \mathbf{k}_{\text{Neck}}$$
3. Scale the pose:
   Divide each keypoint by the body height to normalize the size:
   $$\mathbf{k}_i^{\text{normalized}} = \frac{\mathbf{k}_i'}{h}$$

Then, the filtering process of PoseValid function evaluates the validity of pose keypoints by applying constraints on limb lengths, symmetry, and keypoint positions.

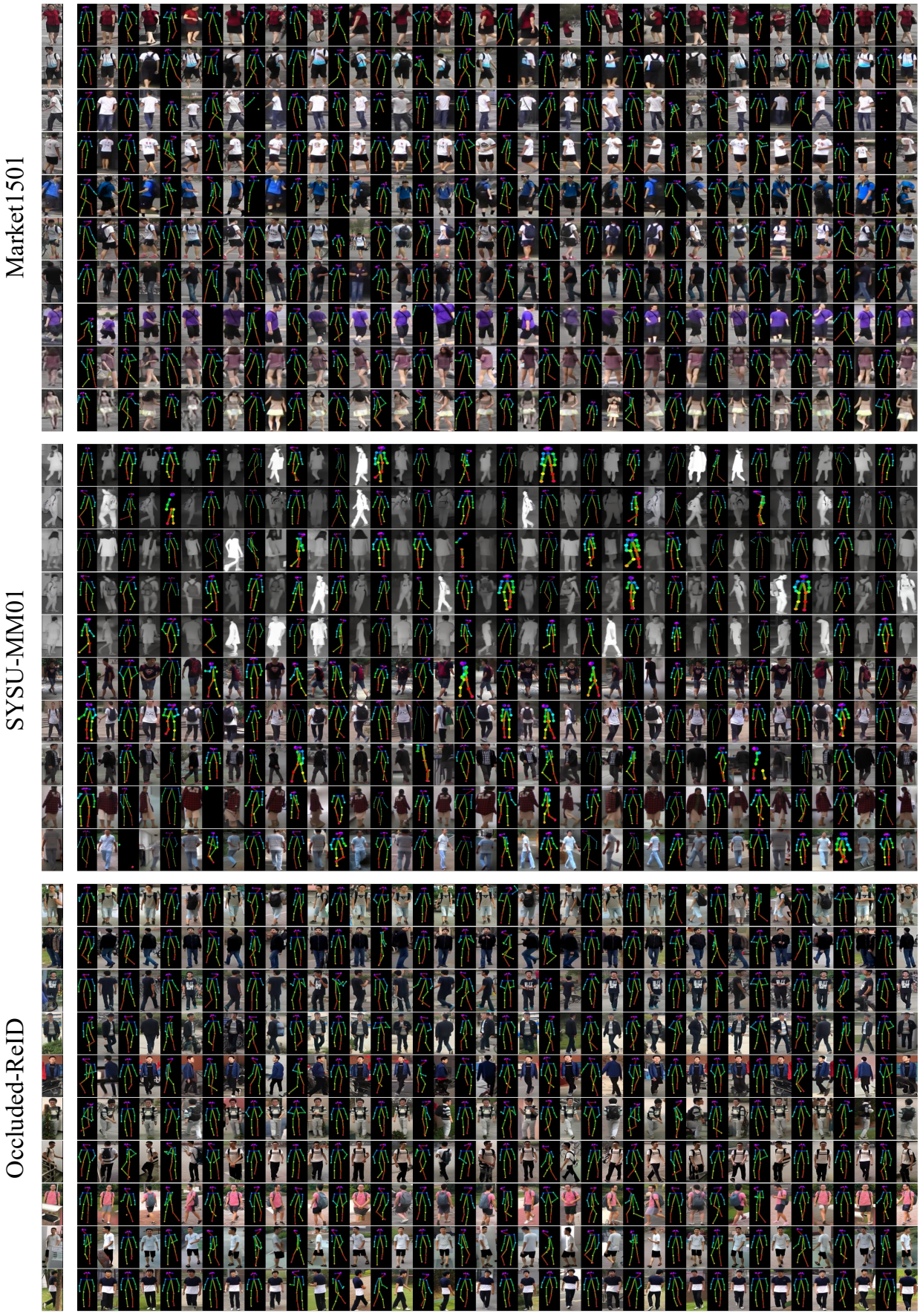

Figure 4. More random generated images on three datasets.

Figure 5. Some outliers detected via the mechanism formulated as Equation.22 on Market1501 and SYSU-MM01 with quartile 0.005.

+Ours	+Rerank	mAP	Rank1
✗	✗	79.88	91.48
✗	✓	89.56	92.07
✓	✗	90.39	94.74
✓	✓	**92.79**	**94.83**

Table 1. Compared to k-reciprocal rerank with official settings on Market1501 (k_1=20,k_2=6).

Methods	mAP	Rank1
TransReID on MSMT17	67.80	85.33
+ours	74.06	86.55

Table 2. Experiment on MSMT17 with TransReID and their official weights.

2.2. Generation quality and Pose Representation Study

To assess the quality of the generated images, we replaced the real images in the dataset with images of the same pose and performed inference validation. The results, as shown in Fig.1, indicate that the original model still successfully matches pedestrians without significant performance degradation. Even with all images in the same pose, the model can effectively differentiate between individuals. This suggests that our generated images are of high quality, retaining the main characteristics of the original images without notably impacting the ReID model. Moreover, we found that pedestrians walking at an angle have higher distinguishability compared to other poses (front, back, and side views), which are more representative of their identities.

2.3. More Random Generation

We provide additional randomly generated images in Fig.4 from Market-1501, SYSU-MM01 and Occluded-ReID datasets.

Methods	Market1501		Occluded-reID	
	Rank-1	mAP	Rank-1	mAP
BoT[15]	94.5	85.9	58.4	52.3
PCB[18]	93.8	81.6	-	-
VGTri[25]	-	-	81.0	71.0
PVPM[4]	-	-	66.8	59.5
HOReID[19]	94.2	84.9	80.3	70.2
ISP[32]	95.3	88.6	-	-
PAT[12]	95.4	88.0	81.6	72.1
TRANS[6]	95.2	88.9	-	-
CLIP[10]	95.7	89.8	-	-
SOLIDER[1]	96.9	93.9	-	-
SSGR[24]	96.1	89.3	78.5	72.9
FED[21]	95.0	86.3	86.3	79.3
BPBreid[16]	95.7	89.4	82.9	75.2
PFD[20]	95.5	89.7	83.0	81.5
KPR_{IN}[17]	95.9	89.6	85.4	79.1
KPR_{SOL}[17]	96.62	93.22	84.83	82.6
CLIP+ours	97.3	94.9	-	-
KPR_{IN}+ours	-	-	91	89.34

Table 3. Comparisons with state-of-the-art methods on Market1501 and Occluded-reID.

2.4. Collaborate with Re-ranking

Since our method does not change the features' original distribution, it could collaborate post-processing strategies like rerank, as shown in Tab.1.

2.5. Results on MSMT17 with TransReID

We conduct a simple experiment on MSMT17 dataset with with TransReID and their official pre-trained weights. As shown in Tab.2.

2.6. Comparisons with state-of-the-art methods on three ReID benchmarks

Comparison on three ReID benchmarks. Since Our method can be applied to any baseline, we choose three methods from three benchmarks which have the official codes and pre-trained weights. With our method, we achieve the new SOTA in three benchmarks, as shown in Fig.3 and Fig.4.

2.7. Analysis on quality coefficient η of Generation Model

Fig.2 illustrates the effect of adjusting the coefficient η on the performance of the ReID model. To evaluate this impact, we gradually increased the value of η and observed changes on the mAP and Rank-1 metrics.

As the value of η increases, the performance of the ReID model improves, reaching an optimal point. At $\eta = 2$, both mAP and Rank-1 achieve their maximum values of 88.02% and 94.77%, respectively. However, further increasing η

Methods	All-Search		Indoor-Search	
	mAP	Rank-1	mAP	Rank-1
PMT[14]	66.13	67.70	77.81	72.95
MCLNet [5]	61.98	65.40	76.58	72.56
MAUM [13]	68.79	71.68	81.94	76.9
CAL[22]	71.73	74.66	83.68	79.69
SAAI(w/o AIM) [2]	71.81	75.29	84.6	81.59
SEFL[3]	72.33	77.12	82.95	82.07
PartMix[9]	74.62	77.78	84.38	81.52
MID [7]	59.40	60.27	70.12	64.86
FMCNet [29]	62.51	66.34	74.09	68.15
MPANet [23]	68.24	70.58	80.95	76.74
CMT [8]	68.57	71.88	79.91	76.90
protoHPE [28]	70.59	71.92	81.31	77.81
MUN [27]	73.81	76.24	82.06	79.42
MSCLNet [31]	71.64	76.99	81.17	78.49
DEEN [30]	71.80	74.70	83.30	80.30
CIFT [11]	74.79	74.08	85.61	81.82
SAAI+ours	76.44	79.33	86.83	84.2

Table 4. Comparison with state-of-the-art methods on SYSU-MM01 without re-ranking.

beyond this point leads to a slight decline in performance. It is easy to find that using generated images to centralize features is effective. However, considering the quality of the generated image, direct adding, although also effective, may not always achieve the best results. Therefore adjusting η according to the generation quality of the model in this dataset can better centralize the features.

2.8. Analysis on k_1/k_2 of Neighbor Feature Centralization

We conducted a detailed analysis of different k_1 and k_2 combinations, evaluating the results of feature centralization enhancement separately on the Query and Gallery sets, as well as the combined effect (as shown in the Fig.3). The selection of these two parameters primarily depends on the number of potential positive samples within the set (adjusting k_1) and the confidence in feature associations (adjusting k2). Overall, medium parameter combinations (k_1 and k_2 in the range of 2-4) provide relatively optimal performance.

References

[1] Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15050–15061, 2023. 5

[2] Xingye Fang, Yang Yang, and Ying Fu. Visible-infrared person re-identification via semantic alignment and affinity inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11270–11279, 2023. 6

[3] Jiawei Feng, Ancong Wu, and Wei-Shi Zheng. Shape-erased feature learning for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22752–22761, 2023. 6

[4] Shang Gao, Jingya Wang, Huchuan Lu, and Zimo Liu. Pose-guided visible part matching for occluded person reid. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11744–11752, 2020. 5

[5] Xin Hao, Sanyuan Zhao, Mang Ye, and Jianbing Shen. Cross-modality person re-identification via modality confusion and center aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16403–16412, 2021. 6

[6] Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021. 5

[7] Zhipeng Huang, Jiawei Liu, Liang Li, Kecheng Zheng, and Zhengjun Zha. Modality-adaptive mixup and invariant decomposition for rgb-infrared person re-identification. In *AAAI Conference on Artificial Intelligence*, 2022. 6

[8] Kongzhu Jiang, Tianzhu Zhang, Xiang Liu, Bingqiao Qian, Yongdong Zhang, and Feng Wu. Cross-modality transformer for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 480–496. Springer, 2022. 6

[9] Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, and Kwanghoon Sohn. Partmix: Regularization strategy to learn part discovery for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18621–18632, 2023. 6

[10] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: exploiting vision-language model for image re-identification without concrete text labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1405–1413, 2023. 5

[11] Xulin Li, Yan Lu, Bin Liu, Yating Liu, Guojun Yin, Qi Chu, Jinyang Huang, Feng Zhu, Rui Zhao, and Nenghai Yu. Counterfactual intervention feature transfer for visible-infrared person re-identification. In *European conference on computer vision*, pages 381–398. Springer, 2022. 6

[12] 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. 5

[13] Jialun Liu, Yifan Sun, Feng Zhu, Hongbin Pei, Yi Yang, and Wenhui Li. Learning memory-augmented unidirectional metrics for cross-modality person re-identification. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19344–19353, 2022. 6

[14] Hu Lu, Xuezhang Zou, and Pingping Zhang. Learning progressive modality-shared transformers for effective visible-infrared person re-identification. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1835–1843, 2023. 6

[15] 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. 5

[16] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person re-identification. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1613–1623, 2023. 5

[17] Vladimir Somers, Alexandre Alahi, and Christophe De Vleeschouwer. Keypoint promptable re-identification. In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 5

[18] 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. 5

[19] Guan'an Wang, Shuo Yang, Huanyu Liu, Zhicheng Wang, Yang Yang, Shuliang Wang, Gang Yu, Erjin Zhou, and Jian Sun. High-order information matters: Learning relation and topology for occluded person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6449–6458, 2020. 5

[20] Tao Wang, Hong Liu, Pinhao Song, Tianyu Guo, and Wei Shi. Pose-guided feature disentangling for occluded person re-identification based on transformer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2540–2549, 2022. 5

[21] Zhikang Wang, Feng Zhu, Shixiang Tang, Rui Zhao, Lihuo He, and Jiangning Song. Feature erasing and diffusion network for occluded person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4754–4763, 2022. 5

[22] Jianbing Wu, Hong Liu, Yuxin Su, Wei Shi, and Hao Tang. Learning concordant attention via target-aware alignment for visible-infrared person re-identification. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11088–11097, 2023. 6

[23] Qiong Wu, Pingyang Dai, Jie Chen, Chia-Wen Lin, Yongjian Wu, Feiyue Huang, Bineng Zhong, and Rongrong Ji. Discover cross-modality nuances for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4330–4339, 2021. 6

[24] Cheng Yan, Guansong Pang, Jile Jiao, Xiao Bai, Xuetao Feng, and Chunhua Shen. Occluded person re-identification with single-scale global representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11875–11884, 2021. 5

[25] Jinrui Yang, Jiawei Zhang, Fufu Yu, Xinyang Jiang, Mengdan Zhang, Xing Sun, Ying-Cong Chen, and Wei-Shi Zheng. Learning to know where to see: A visibility-aware approach for occluded person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11885–11894, 2021. 5

[26] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 3

[27] Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, and Guoying Zhao. Modality unifying network for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11185–11195, 2023. 6

[28] Guiwei Zhang, Yongfei Zhang, and Zichang Tan. Protohpe: Prototype-guided high-frequency patch enhancement for visible-infrared person re-identification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 944–954, 2023. 6

[29] Qiang Zhang, Changzhou Lai, Jianan Liu, Nianchang Huang, and Jungong Han. Fmcnet: Feature-level modality compensation for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7349–7358, 2022. 6

[30] Yukang Zhang and Hanzi Wang. Diverse embedding expansion network and low-light cross-modality benchmark for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2153–2162, 2023. 6

[31] Yiyuan Zhang, Sanyuan Zhao, Yuhao Kang, and Jianbing Shen. Modality synergy complement learning with cascaded aggregation for visible-infrared person re-identification. In *European Conference on Computer Vision*, pages 462–479. Springer, 2022. 6

[32] 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. 5