Se-ReID: Spatially Enhanced Representation Learning for Scalable Person Re-identification

***Abstract*—Person Re-Identification (ReID) struggles with discriminative feature learning due to extreme intra-class variance and ambiguous boundary samples. Existing metric losses are often constrained by local mini-batch mining or rigid distance margins that ignore contextual data structures. To address these issues, we propose Se-ReID, a unified framework that enhances feature space representation through instance-level and centroid-level innovations. At the instance level, we introduce TriHard+ Loss with dynamic routing to prevent manifold collapse, alongside an alternative TriWeight Loss utilizing hard-adapted soft weighting to preserve dense intra-class structures. At the centroid level, we propose CentroidM Loss, which leverages learnable global proxies to transcend mini-batch limitations and effectively soften inter-class boundaries. These core metric modules are further supported by 1st & 2nd order mask techniques to eliminate sampling bias, and a streamlined cross-camera centroid retrieval strategy to filter gallery noise. Extensive experiments demonstrate that Se-ReID achieves remarkable performance on standard benchmarks (Market1501 and DukeMTMC-ReID) without relying on ReRank. Notably, it yields state-of-the-art (SOTA) results when integrated with the SOLIDER Transformer baseline, confirming its robust effectiveness and broad applicability across diverse architectures.improvements on MNIST. The code will be released.**

***Index Terms*—Extract features, hard and plain samples, hard-adapted weight, instance and centroid level, relative and absolute relation, reinforce boundary judgment**

I. Introduction

Unlike classic image classification (e.g., ImageNet), Person ReID suffers from severe label space explosion and extreme intra-class variance, leading to highly overlapping feature distributions even after model convergence (Fig.1a). Ambiguous positive pairs and highly similar negative pairs (Figs 1e/f) further hinder precise feature boundary establishment (Fig.1b). Discriminative feature learning [1]–[3] is key for intra-class compactness and inter-class separation [4] in ReID. Classic methods like Center Loss [5] cluster same-identity features, while Triplet Loss [6] and its variants enforce positive-negative margins. Recent work (e.g., MSML [7]) targets hard samples near boundaries, and Circle Loss [8] refines feature learning via boundary mining and weight adaptation.

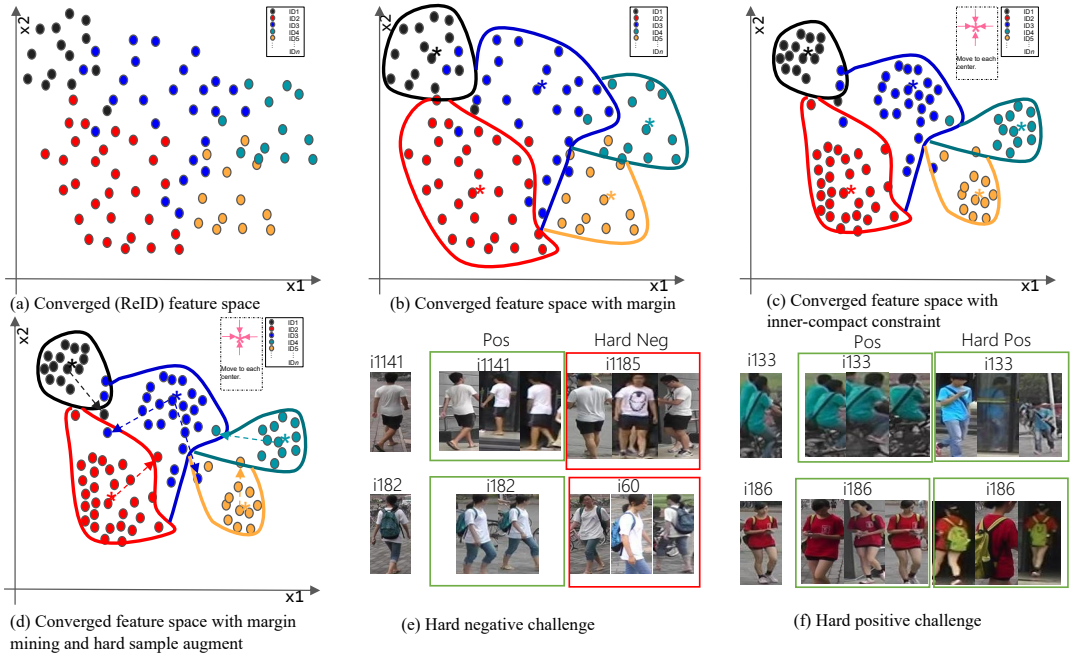
This paper focuses on intra-class compactness and inter-class separation in feature learning. For intra-class compactness, we enhance supervision with methods like Center Loss to narrow same-class feature distances (Fig.1c). For inter-class separation, we strengthen boundary judgment and hard sample training by exposing misclassified hard samples more, proposing Centroid Migration/Margin Mining Loss (CentroidM Loss) to adjust relationships between boundary instances and positive/hard negative centers—pushing hard negative centers away and pulling positive centers closer to boundary instances (Fig.1d). To prevent manifold collapse caused by extreme hard samples, we introduce TriHard+ Loss equipped with difficulty-aware dynamic routing and geometric constraints. Additionally, to avoid over-mining that misjudges ambiguous positives, we propose TriWeight Loss featuring a hard-adapted soft weighting mechanism. Finally, to capture true distributions, we analyze positional relationships from absolute and relative perspectives at instance and centroid levels, developing 1st & 2nd order masks integrated into all loss functions.

We hypothesize that the challenges could be addressed by a loss function approach based on an enhanced feature space representation. The proposed Se-ReID, based on spatially enhanced distributions, provides an effective framework that reflects a systematic methodology. Specifically, our contributions contain:

- We propose two alternative instance-level metric losses to explicitly enhance spatial representation. Specifically, **TriHard+ Loss** introduces difficulty-aware dynamic routing and spatial geometric constraints to prevent manifold collapse caused by extreme hard samples, while **TriWeight Loss** employs hard-adapted soft weighting to preserve dense intra-class structures without ignoring global distributions.
- We construct a global centroid-level metric constraint, **CentroidM Loss**, which synergizes absolute and relative spatial relationships alongside Center Loss. This architecture transcends the limited scope of mini-batch mining, effectively refining inter-class boundaries and augmenting global hard sample discrimination.
- We design **1st & 2nd order mask techniques** and seamlessly integrate them across all loss functions. This mechanism explicitly eliminates interference from repeated samples during distance computation, ensuring that the feature learning process is driven by unbiased, truthful data distributions.
- Extensive experiments demonstrate that Se-ReID achieves remarkable performance on standard benchmarks (Market1501 and DukeMTMC-ReID) without relying on ReRank. **Notably, it establishes state-of-the-art (SOTA) results when integrated with the Transformer baseline**, validating its robust effectiveness across diverse architectures, alongside highly interpretable feature clustering improvements on MNIST.

II. Related works

The Person ReID community is currently very active. Specifically, our research focuses on effective metric and representation losses. A concise overview is provided below.



(a) Converged (ReID) feature space (b) Converged feature space with margin (c) Converged feature space with inner-compact constraint

(d) Converged feature space with margin mining and hard sample augment (e) Hard negative challenge (f) Hard positive challenge

Fig. 1: ReID challenges and basic responses. (a) presents basic convergence but slightly muddled features. (b) marks the margin in (a). (c) adds constraints on inner compactness. (d) emphasises on inter class relations, hard samples and margin mining. (e) and (f) respectively represent hard negative and positive challenges.

A. ReID concerned Loss

FaceNet [6], a seminal face recognition method, replaces Softmax with Triplet Loss, optimizing feature distances between anchor-positive and anchor-negative pairs to enforce margin separation. Subsequent improvements include TriHard Loss [9], which constructs triplets by selecting the hardest positive/negative samples within batches, and Quadruplet Loss [10] that extends to quadruples while balancing absolute and relative distance constraints. Margin Sample Mining Loss (MSML) [7] further develops hard sample mining by widening the gap between positive pair upper bounds and negative pair lower bounds. Recent advances focus on distribution optimization: [4] introduces rank-in-rank loss for class imbalance mitigation, [2] leverages viewpoint synthesis with contrastive learning for invariant features, and [3] reduces background interference via consistency constraints and object-centric refinement.

B. Related Visual Recognition Methods

Beyond metric losses, several face recognition-oriented loss functions share conceptual alignment with our approach. Center Loss [5] learns class-specific centers as learnable parameters to enforce intra-class compactness. Sphere Loss [11] addresses class imbalance through weight normalization, emphasizing angular feature optimization for spatial robustness. AdaFace [12] introduces adaptive margins based on sample quality to enhance recognition flexibility. These methods collectively emphasize feature space regularization and adaptive learning mechanisms, resonating with our design philosophy.

While sharing the core objective of feature space regularization, Se-ReID departs from existing paradigms through: (I) High-order Structural Constraints that model the (a, p, n) triad with explicit angular constraints to prevent manifold collapse; (II) Difficulty-aware Adaptive Weighting via dynamic routing based on precise topological difficulty; (III) Global Boundary Mining using learnable centroids to overcome the myopic bottleneck of mini-batch sampling; and (IV) Distributional Fidelity, employing 1st and 2nd-order masks to guarantee an unbiased feature space without disrupting natural distributions.

III. Method

In this section, we describe the details of Se-ReID shown in Fig.2. It is worth noting that we take the instance level and centroid level constraints and retrieval. For constraints, the difference is that the centroid level constrains the centroid relationship (global distribution), while the instance level constrains the individual. For retrieval, the difference is that each item of gallery consists of an individual or centroid (details in **??**).

A. Difficulty-Aware Dynamic Routing for Representation

Our metric loss builds upon hard sample triplets but profoundly enhances the structural constraints (see Fig.3). While standard TriHard Loss [9] works on batches with P identities (each with K instances) by mining the hardest positive p and hardest negative n for an anchor a, it uses only two pairwise combinations ($a \sim p$, $a \sim n$). This unidirectional repulsion minimizes $d_{a,p}$ and maximizes $d_{a,n}$, but fundamentally ignores the structural correlation between p and n.

In crowded high-dimensional feature spaces, blindly pushing n away from a without global structural awareness often triggers a "spatial seesaw effect": n rotates around a and collapses into the feature territory of p, forming a new hard sample relative to p. To explicitly prohibit this, we deem the triad (a, p, n) as a complete structural unit and introduce a

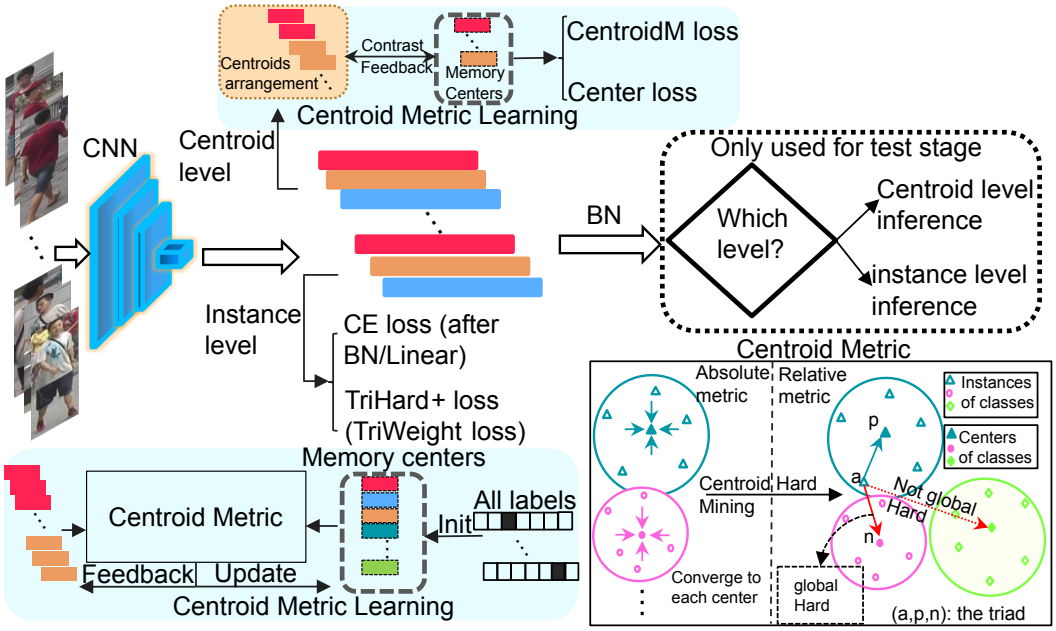


Fig. 2: Se-ReID baseline: To highlight our loss function (not engineering tuning or large models), we use a simple common backbone (ResNet50). For Centroid Metric Learning, memory centers (prototypes) are used to: (1) aggregate samples toward their positive prototype; (2) refine decision boundaries by mining relative positions between boundary samples and positive/difficult-negative prototypes.

symmetric repulsion term $d_{p,n}$ alongside an angular geometry constraint.

However, naively enforcing $d_{p,n} > d_{a,p}$ across all triplets risks undermining the hard-mining principle, as the network would waste gradients pushing apart inherently safe samples. To reconcile this, we propose a **Difficulty-Aware Dynamic Routing** mechanism (dubbed TriHard+ Loss). By dynamically evaluating whether n poses a greater threat to a or to p, we adaptively route the gradient penalties. The threat logits are defined as:

$$\mathcal{T}_{an} = s(-d_{a,n})^t, \quad \mathcal{T}_{pn} = s(-d_{p,n})^t, \tag{1}$$

where s and t are scaling and exponential factors (defaulted to 1 and 3). The dynamic routing weights are computed via Softmax:

$$w_{an} = \frac{e^{\mathcal{T}_{an}}}{e^{\mathcal{T}_{an}} + e^{\mathcal{T}_{pn}}}, \quad w_{pn} = \frac{e^{\mathcal{T}_{pn}}}{e^{\mathcal{T}_{an}} + e^{\mathcal{T}_{pn}}}. \tag{2}$$

Subsequently, we formulate the structural TriHard+ Loss using these dynamic weights and a Pythagorean angular constraint:

$$L_{th+} = \frac{1}{PK} \sum_{a \in batch} \Big(w_{an}[d_{a,p} - d_{a,n} + \alpha]_+ + w_{pn}[d_{a,p} - d_{p,n} + \alpha]_+ \Big) + \lambda L_{ang}, \; s.t. flag(a,p,n) =' real' \tag{3}$$

$$L_{ang} = \frac{1}{PK} \sum_{a \in batch} [d_{a,n}^2 + d_{a,p}^2 - d_{p,n}^2]_+, \quad s.t. flag(a,p,n) =' real' \tag{4}$$

where λ is the angular weight coefficient (set to 0.1), and $flag = 'real'$ means only using non-biased samples by mask (details in Eq.17). L_{ang} mathematically forces the angle $\angle(p,a,n)$ to be obtuse by ensuring $d_{p,n}$ remains the longest side of the triad. During regular optimization, $w_{an} \approx 1$, maintaining standard hard mining. However, if n collapses towards p ($d_{p,n}$ shrinks), w_{pn} sharply increases, instantly triggering the symmetric penalty to kick n into an orthogonal dimension. This guarantees spatial equilibrium.

B. CentroidM Loss

Center Loss enhances feature discrimination by enforcing intra-class compactness via learnable class centers, but it relies on absolute distance boundaries and ignores inter-class correlations. Since data labels lack relational semantics, the model struggles to distinguish visually similar classes, requiring a balance between discriminative power and generalization [13].

Centroid Migration/Margin Mining Loss (CentroidM Loss) is proposed to soften the clustering boundaries of various classes. On the premise of ensuring that different classes can be separated, it aims to retain both absolute and relative correlations to learn better comprehensive discrimination ability for difficult boundary samples.

However, traditional metric learning methods (e.g., TriHard Loss [6]) are inherently constrained by the mini-batch mechanism. They perform Local Mining, identifying hard negatives only among the limited samples ($B \times K$) in the current iteration, thereby ignoring the vast majority of negative samples in the entire dataset. To address this, CentroidM Loss (Fig.4) leverages learnable class centers as global prototypes to effectively compress the global distribution of the entire dataset. When an instance x_i computes loss against negative centers C_n, it performs Global Hard Negative Mining against the global distribution, transcending the limitations of the mini-batch with efficient computation.

Driven by this global perspective, we consider the CentroidM Loss with relative distance constraints and adopt the form of the triplet. As with the Center Loss, we create a learnable center parameter (which is shared with the Center

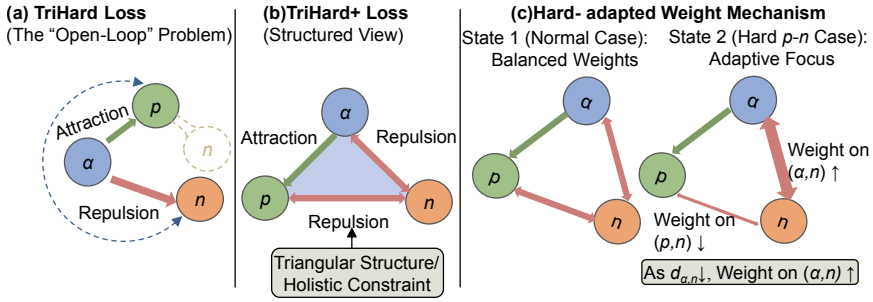


Fig. 3: Comparison of metric constraints. (a) Traditional TriHard: Focuses solely on anchor-centered distances, leaving a "blind spot" between positive and negative samples. (b) TriHard+ Structure: Enforces a strict spatial geometric constraint by introducing an explicit angular penalty on $\angle(p, a, n)$, fundamentally preventing the negative sample from collapsing into the positive's feature space. (c) Adaptive Weighting: As α encroaches on n (State 2), the mechanism adaptively amplifies α-n repulsion (thicker arrow) while relaxing p-n weight to resolve hard overlapping regions.

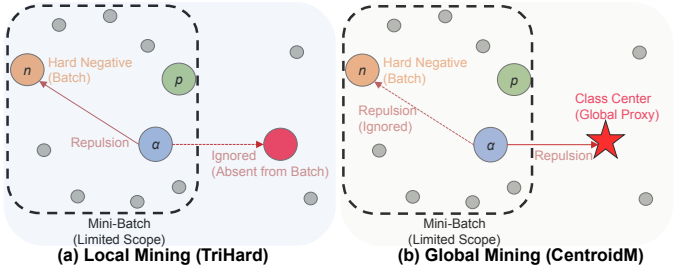


Fig. 4: Mining scope comparison. (a) Local Mining: Restricted to in-batch samples, often ignoring the true hardest negatives (dashed line). (b) Global Mining: Uses center proxies to transcend batch limitations, enabling effective optimization against the global data distribution.

Loss), where each term of the center parameter represents a center feature of a class to be learned. CentroidM Loss consists of two terms; the primary one is defined as follows:

$$L_{centr\_1} = \frac{1}{P \times K} \sum_{a \in batch} (\max_{p \in A} d_{a,c_p} - \min_{n \in B} d_{a,c_n} + \alpha)_+, \quad s.t. flag(a) = 'real', \tag{5}$$

where a stands for anchor, $P \times K$ is the number of a, c stands for the center to be learned. A/B represents positive/negative center sets. The basic workflow is presented in Algorithm 1. For classes with insufficient samples, repeated sampling can bias feature distribution learning. To mitigate this, we generate synthetic samples via random tensor generation (labeled as $fake$) and exclude fake inputs using *1st & 2nd-order mask* (in Eq.17) during CentroidM Loss computation. Additionally, to keep consistency, We further transform $Centr\_1$ to hard-adapted weight form as same as proposed $TriHard_+$. It reformulates the original negative pairs' distance as an attention-weighted softmax mechanism composed of two terms (d_{a,c_p}, d_{c_n,c_p}), thereby softening the decision boundary.

The second item of CentroidM Loss takes the following form:

$$L_{\text{centr\_2}} = \sum_i D(f(i), f_c(i)). \tag{6}$$

Eq.6 encourages samples to converge toward their respective class centers, and this part can be replaced with Center Loss (In this study, Center=Centr_2). The total CentroidM Loss is usually written as (λ_0, λ_1 is used to trade-off):

$$L_{\text{centr\_m}} = \lambda_1 L_{\text{centr\_1}} + \lambda_0 L_{\text{centr\_2}}. \tag{7}$$

CentroidM Loss is often used in conjunction with Cross-entropy Loss and Center Loss in representation learning tasks. In this case, centr_2 is omitted. Total losses in the following form.

$$L_{\text{total}} = L_{\text{cls}} + \lambda_0 L_{\text{ct}} + \lambda_1 L_{\text{centr\_m}}, \tag{8}$$

ct means Center Loss. λ_0 and λ_1 (small positive scalars) constrain CentroidM Loss's influence on backbone parameters during backpropagation, suppressing noise during updates. Meanwhile, center parameters are updated more aggressively by resetting their optimization weights to 1, ensuring stronger spatial regularization.

For a certain sample term of the first term ($L_{\text{centr\_1}}$) of the CentroidM Loss (assuming that the loss of this term is not equal to 0), we calculate the gradient value:

$$\frac{\partial L_{\text{centr\_1}}}{\partial \text{c}_\text{n}} = 2[f(x^a) - \text{c}_\text{n}]. \tag{9}$$

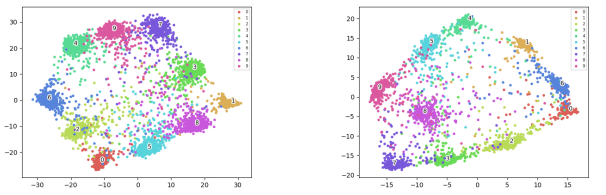
$$\begin{aligned} \frac{\partial L_{\text{centr\_1}}}{\partial f(x^a)} &= 2[f(x^a) - \text{c}_\text{p}] - 2[f(x^a) - \text{c}_\text{n}] \\ &= 2[\text{c}_\text{n} - \text{c}_\text{p}]. \end{aligned} \tag{10}$$

$$\frac{\partial L_{\text{centr\_1}}}{\partial \text{c}_\text{p}} = 2[\text{c}_\text{p} - f(x^a)]. \tag{11}$$

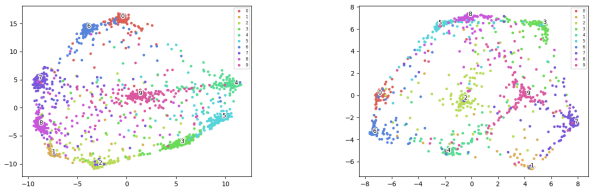
Adaptive Center Adjustment: Centers shift slightly toward class boundaries (controlled by margin α and weights λ_0,λ_1) , effectively encoding hard sample features into the feature space. This serves as an implicit feature-level augmentation, intensifying training and enhancing discrimination for challenging samples. The refined feature distribution ultimately strengthens model robustness and attribution capability.

Furthermore, the gradients of $L_{ct} + L_{centr\_1}$ with respect to $f(x_i)$ and updating c_j are computed as:

$$\begin{aligned} \frac{\partial(\lambda_0 L_{ct} + \lambda_1 L_{centr\_1})}{\partial(f(x_i^t))} &= 2\lambda_0 \frac{\sum_{i=1}^{B} I(y_i^t = j)(f(x_i^t) - c_j^t)}{\varepsilon + \sum_{i=1}^{B} I(y_i^t = j)} \\ &+ 2\lambda_1 (I(d_{a_i^t, c_{p^*}^t} - d_{a_i^t, c_{n^*}^t} + \alpha > 0) \cdot (\frac{I(y_i^t = j)}{\varepsilon + \sum_{i=1}^{B} I(y_i^t = j)} \\ &\cdot (f(x_i^t) - c_j^t) + \frac{I(y_i^t \neq j) \cdot (c_j^t - f(x_i^t))}{\varepsilon + \sum_{i=1}^{B} I(y_i^t \neq j)})). \end{aligned} \tag{12}$$



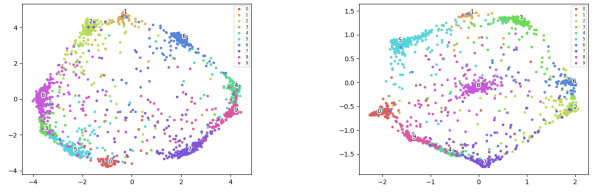
(a) $\lambda_0 = 0.001$ (b) $\lambda_0 = 0.01$



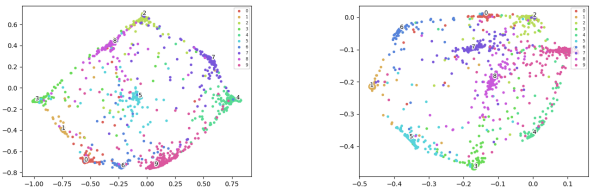
(c) $\lambda_0 = 0.1$ (d) $\lambda_0 = 1$

Fig. 5: Distribution of 6-layer-net features supervised by Cross-entropy Loss (coefficient=1) and Center Loss (coefficient=λ_0). Numbers 0-9 denote MNIST class labels and their feature centers.

$$\frac{\partial(\lambda_0 L_{ct} + \lambda_1 L_{centr\_1})}{\partial c_j^t} = 2\lambda_0 \frac{\sum_{i=1}^{B} I(y_i^t = j)(c_j^t - f(x_i^t))}{\varepsilon + \sum_{i=1}^{B} I(y_i^t = j)}$$
$$+ 2\lambda_1 (\sum_{i=1}^{B} I(d_{a_i^t, c_{p^*}^t} - d_{a_i^t, c_{n^*}^t} + \alpha > 0)$$
$$\cdot (\frac{I(y_i^t = j)(c_j^t - f(x_i^t))}{\varepsilon + \sum_{i=1}^{B} I(y_i^t = j)} + \frac{I(y_i^t \neq j)(f(x_i^t) - c_j^t)}{\varepsilon + \sum_{i=1}^{B} I(y_i^t \neq j)}). \tag{13}$$



(a) $\lambda_0 = 0.001, \lambda_1 = 0.0005$ (b) $\lambda_0 = 0.01, \lambda_1 = 0.005$



(c) $\lambda_0 = 0.1, \lambda_1 = 0.05$ (d) $\lambda_0 = 1, \lambda_1 = 0.5$

Fig. 6: Distribution of Using the same configuration as Figure 5 but adding "$\lambda_1 \times$Centr_1 Loss". The scale range of coordinate space is compressed, but the relative relationship of feature distribution is more clear.

In Eq.12 and Eq.13, for clarity of presentation, we compute the gradients based on the randomly sampled triplet loss, without hard mining. ϵ avoids division by zero. Each sub-polynomial (triplet-form) activates only when (negative pair distance $-$ positive pair distance) $< \alpha$. $I(\cdot)$ is an indicator function (1 if true, 0 otherwise).

In the experimental configuration corresponding to Figures 5 and 6, Tables V and VI, a 6-layer convolutional neural network is built with a feature vector dimension of 2 at the output of the network, followed by a 2×10 classifier, to do the representation learning task. MNIST dataset is used to train the classification task (representation learning). Cross-entropy Loss is used by default and not mentioned repeatedly. Fig.5 shows the representation of features in the feature space under the condition of Center Loss. Fig.6 show the representation of features in the feature space under the condition of Center Loss + Centr_1 Loss. λ_0 and λ_1 are the weighting coefficients of the two losses. (When just Center Loss, just the λ_0 is used.) After centr_1 loss is added, the scale range of coordinate space is compressed, but the relative relationship of feature distribution is more clear, exactly, the relative distribution within the class is more compact, and the distribution between the classes is more distinguishable.

Algorithm 1 Discriminative Feature Learning with CentroidM Loss

Require: Training data $\{x_i\}$; initial parameters: Convolution(θ_C), Linear(W), $\{c_j\}$; hyperparameters λ_1, λ_2, learning rate u^t

Ensure: Optimized parameters θ_C, W, c_j

$t \leftarrow 0$

while not converged **do**

$t \leftarrow t + 1$

Compute $L_{\text{total}}^t = L_{\text{cls}}^t + \lambda_0 L_{\text{ct}}^t + \lambda_1 L_{\text{centr\_1}}^t$

Compute $\frac{\partial L_{\text{total}}^t}{\partial x_i^t}$ for each i

Compute $\Delta c_j^t = \frac{\partial(\lambda_0 L_{ct} + \lambda_1 L_{centr\_1})}{\partial c_j^t}$ via Eq.13

Update $W^{t+1} = W^t - u^t \cdot \frac{\partial L_{\text{cls}}^t}{\partial W^t}$

Update $c_j^{t+1} = c_j^t - \Delta c_j^t$

Update $\theta_C^{t+1} = \theta_C^t - u^t \cdot \sum_i \frac{\partial L_{\text{total}}^t}{\partial x_i^t} \cdot \frac{\partial x_i^t}{\partial \theta_C^t}$

end while

C. TriWeight Loss

A similar idea to that of compacting distribution is that we also adopt a hard-adapted weight strategy for the batch all samples of Triplet Loss instead of hard mining. The most similar to our study is the method [14], but our method outperforms it in terms of sampling strategy, weighting detail, etc. The limited number of instances per identity in ReID datasets makes the sampling strategy crucial. Many existing methods that rely on resampling or first-order masking (which only masks individual samples) can introduce bias into the data distribution. We use the *1st & 2nd order masks*, to avoid data bias. TriWeight Loss is proposed as follows, allowing the model to adaptively mine margin and ensure inner-compact according to accurate distribution.

$$W_{ij}^n = \frac{e^{s[-d'_{a_i,n_j} - \max\limits_{n_k \in N_i}(-d'_{a_i,n_k})]^t}}{\sum\limits_{n_k \in N_i} e^{s[-d'_{a_i,n_k} - \max\limits_{n_k \in N_i}(-d'_{a_i,n_k})]^t}} \tag{14}$$
$$s.t.\ flag(a) \wedge flag(n) =' real'.$$

$$W_{iq}^p = \frac{e^{s[d'_{a_i,p_q} - \max\limits_{p_l \in P_i}(d'_{a_i,p_l})]^t}}{\sum\limits_{p_l \in P_i, a_i \neq p_l} e^{s[d'_{a_i,p_l} - \max\limits_{p_l \in P_i}(d'_{a_i,p_l})]^t}} \tag{15}$$
$$s.t.\ flag(a) \wedge flag(p) =' real'.$$

$$L_{TriW} = \sum_{i}^{B} \Big[\sum_{x_q^p \in P_i} W_{iq}^p ||f(x_i^a) - f(x_q^p)||_2^2 - \sum_{x_j^n \in N_i} W_{ij}^n ||f(x_i^a) - f(x_j^n)||_2^2 + \alpha \Big]_+ \quad s.t.\ flag(a,p,n) =' real'. \tag{16}$$

The formulations in Equations (14) to (16) share the hard-adapted weighting mechanism. Specifically, a, p, and n refer to the anchor, positive, and negative samples, respectively. The subscripts i, q, and j denote the indices of the anchor, the positive sample, and the specific negative sample currently being evaluated. Meanwhile, k^* and l^* represent the index sets of all available negative and positive samples, respectively. N_i and P_i are the corresponding negative and positive sample sets for anchor a_i. To avoid data bias, the constraint $flag(a,p,n) = 'real'$ dictates that all involved samples must be valid; any sample failing this condition is discarded.

The exponential factor t and temperature scaling coefficient s jointly regulate this hard-adapted weighting mechanism. Specifically, the parameter t is typically set as a positive odd integer (defaulted to 3) to strictly preserve the sign of the relative distance differences, ensuring that the direction of the gradient remains consistent with the relative sample hardness.

Regarding the coefficient s, it determines the sensitivity of the attention distribution. Since the input term of the exponential function naturally falls into the negative domain (where exponential outputs are ≤ 1), increasing a positive s sharply accelerates the curve's decay. This effectively suppresses the weights of common, uninformative pairs while maintaining high weights for the hardest pairs (where the input approaches 0), thereby widening the margin between critical boundary samples and ordinary ones. Conversely, a negative s would counter-productively emphasize easy samples and ignore hard constraints, which we found empirically leads to convergence failure. Therefore, while the theoretical effective range is constrained to $s \in [0, +\infty]$ to ensure the model strictly prioritizes intra-class compactness and inter-class separation, we empirically find that setting $s \in [0.5, 1.5]$ yields the optimal performance.

D. Getting an accurate distribution: 1st & 2nd order masks

Due to insufficient instances, ReID batch sampling process often uses resampling technology, but it also brings the risk of deviation in data distribution. To this end, we design 1st & 2nd order mask modules (FSM) to guarantee the $flag$ of the sample is $real$. The first-order masks (masking individual) are mainly used to deal with representation learning, while the second-order masks (masking pairs or triplets) are used for metric learning (as below, for batch-input x/mask with shapes $[N, C]$/$[N]$). (Illustration in Appendix.)

$$\begin{aligned} MASK^{1d}(x, mask) &= x \circ mask, \\ MASK^{2d}(x, mask) &= (mask \otimes mask) \circ MM(x). \end{aligned} \tag{17}$$

Where $\circ$:Hadamard product, $\otimes$:Outer Product, and MM:Metric Matrix ($MM_{ij} = d(x_i, x_j)$, MM is a symmetrical matrix).

TABLE I: Complexity summary for various methods. RS means random sample.

	Triplet(RS)	TriHard	TriHard+	CentroidM	TriWeight
$\mathcal{O}(\bullet)$	$B \cdot d$	$B^2 \cdot d$	$B^2 \cdot d$	$B \cdot C \cdot d$	$B^2 \cdot d$

E. Complexity Analysis of Methods

The complexity here may refer to two aspects: computational complexity and algorithmic complexity. First, it is noteworthy that the proposed 1st & 2nd order masking techniques, designed to eliminate interference from repeated samples during distance computation, introduce negligible computational overhead. Consequently, their impact on complexity is omitted in subsequent analyses. For clarity, we formalize key symbols: B denotes the batch size, C represents the number of distinct classes within a batch, and d signifies the embedding dimension, which directly governs the computational cost of pairwise (as well as the triplet) distance calculation.

Batch-wise computational complexity analysis (Tab.I). Random sampling is computationally efficient ($O(B \cdot d)$) but often ineffective, as many triplets already satisfy the margin constraint and thus provide negligible gradients. Hard negative mining dynamically selects the hardest positive and negative samples by comparing all sample pairs, resulting in $O(B^2 \cdot d)$ complexity. Difficulty-Aware Dynamic Routing (TriHard+ Loss) builds upon this by introducing dynamic routing weights and a spatial angular constraint. Since these structural penalties are computed utilizing the already extracted triad distances, they incur negligible $O(B)$ element-wise overhead, strictly maintaining the same overall complexity of $O(B^2 \cdot d)$. Furthermore, the $\mathcal{O}(B \cdot C \cdot d)$ computational cost of CentroidM derives from calculating the instance-to-centroid distance matrix, which is fully shared with Center Loss. By explicitly reusing this matrix, Centr_1 introduces a global distribution perspective with negligible marginal overhead, thereby avoiding any exacerbation of the $\mathcal{O}(B^2 \cdot d)$ bottleneck in the joint training system. TriWeight Loss computes all negative pairs, resulting in $O(B^2 \cdot d)$ complexity, with additional overhead for adaptive weights calculated across all negative pairs, giving a total complexity of $O(B^2 \cdot (e + f + d))$, where e/f, and d represent the costs of number-wise multiplication/addition and vector-wise distance calculation, respectively (so $f \ll e \ll d$).

Algorithm complexity analysis. Random Sampling: high algorithmic complexity due to inefficient gradient updates from non-informative triplets. Hard Mining & TriHard+: lower algorithmic complexity—while hard mining accelerates convergence by prioritizing meaningful gradients, TriHard+ further prevents spatial collapse via dynamic routing, yielding highly robust and efficient gradient trajectories. CentroidM: Reduces noise via global class prototypes, lowering algorithmic complexity (fewer iterations needed for stable convergence). TriWeight: Maximizes algorithmic efficiency by fine-tuning all informative pairs.

IV. Experiments

Aligned with standard ReID baselines [15], [16] (derived from [17]), we employ ResNet50 pretrained on ImageNet as our primary CNN backbone. Additionally, we adopt the SOLIDER model [18] as a Transformer-based backbone

specifically for comparisons with state-of-the-art methods. Throughout our experiments, we prioritize fair ablation studies over engineering-driven performance optimization. Bold values in the results denote optimal metrics.

Loss Functions: The baseline integrates Softmax Loss and Center Loss, enhanced by our proposed methods: TriHard+ Loss (i.e. hard-adapted weight), CentroidM Loss, and TriWeight Loss.

Optimization: Training uses the Adam optimizer [19] with an initial learning rate of 3.5×10^{-4}, reduced by a factor of 0.1 at epochs 40 and 70 (Max-epochs: 150), alongside a weight decay of 5×10^{-4}. Following [15], center parameters in CentroidM Loss and Center Loss are optimized via SGD [20] with a fixed learning rate of 0.5.

Evaluation: Adhering to cross-camera protocols [21], we evaluate both instance-level and centroid-level retrieval to ensure fairness and validity. For the *instance-level* retrieval, every gallery sample acts as a candidate. Conversely, for the *centroid-level* evaluation (inspired by [16]), we enforce a strict cross-camera protocol to establish unique centroid candidates. As formulated in Eq. 18, instead of a simple global average, the centroid candidate of each class is derived by averaging all valid gallery samples that are cross-camera relative to the query. Meanwhile, duplicate statistical counting is explicitly filtered out to ensure the absolute uniqueness of each centroid individual.

$$c_{p_{k,j}} = \frac{1}{|G_{k,*}/G_{k,j}|} \sum_{x_i \in G_{k,*}/G_{k,j}} f(x_i), \quad s.t.\ unique(G_{k,*}/G_{k,j}), \tag{18}$$

By dynamically constructing a query-specific global prototype for each identity, the evaluation paradigm inherently shifts from a traditional one-to-many positive matching to a simplified one-to-one retrieval. Since each query now corresponds to only a single ground-truth positive centroid, and the feature averaging effectively mitigates extreme visual variations, the intrinsic matching difficulty is considerably reduced. Consequently, the absolute performance gains at the centroid level may appear more marginal compared to instance-level metrics. Nevertheless, this streamlined one-to-one matching mechanism yields a significant practical advantage, achieving an average 8.2× speedup over standard instance retrieval (Tab.II).

TABLE II: Memory consumption and time overhead

DataSet	Retrieval level	Gallery size	Embeddings filesize(MB)	Eval time (s)
Market-1501	Instance	16k	120	2.48
	Centroid	0.75k	6	0.42
DukeMTMC-ReID	Instance	17k	140	1.89
	Centroid	1.1k	9	0.18

A. Ablation Study and Comparison with State-of-the-Art

We conduct comprehensive ablation studies primarily on Market1501 and DukeMTMC-ReID to validate the effectiveness of each proposed module. The evaluations are systematically decomposed into the following aspects, leading to a direct comparison with state-of-the-art methods.

Fig. 7: The 3 typical centroid level retrieval results on Market1501. Green/red borders denote correct/incorrect retrievals respectively.

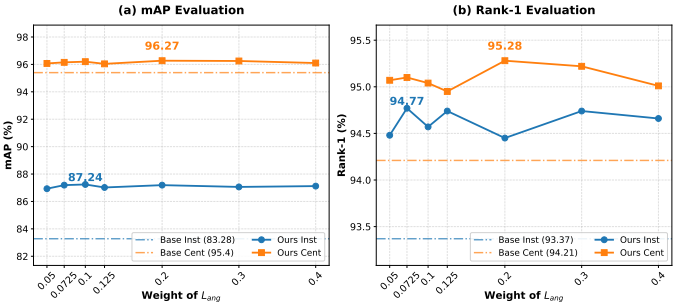


Fig. 8: Parameter sensitivity analysis on the weight of the proposed L_{ang}. The evaluation is conducted on both instance-level and centroid-level representations. The dashed lines indicate the performance of the baseline model, while the solid lines represent our method. The peak values are explicitly annotated.

Effectiveness of Difficulty-Aware Dynamic Routing (TriHard+) and TriWeight: To mitigate the rigid hard-only mining strategy of standard Triplet Loss, we introduce TriHard+ and TriWeight. As shown in Tables III and IV, substituting standard TriHard with our TriHard+ yields solid centroid-level mAP improvements, reaching 96.27% on Market1501 and 91.59% on DukeMTMC-ReID, alongside significant instance-level mAP gains of +3.96% and +3.94% over the baseline.

In addition, Fig.8 illustrates the parameter sensitivity of the L_{ang} weight from 0.05 to 0.4. Our method consistently outperforms the baseline across all settings. Notably, the instance-level mAP achieves a remarkable +3.96% improvement (from 83.28% to 87.24%) at the weight of 0.1, while the centroid-level mAP peaks at 96.27% at 0.2. The performance remains robust within the range of [0.0725, 0.2]. A larger weight ($\geq$ 0.3) leads to slight degradation due to the over-penalization of the competitive mechanism disrupting the main feature learning. Consequently, we set the weight to 0.1 for all subsequent experiments.

Furthermore, TriWeight Loss (Tables VII and VIII) effectively leverages all pairs within the batch via soft attention, driving the centroid mAP to an impressive 96.29% and 92.27% respectively. Concurrently, TriWeight provides robust instance-level mAP boosts of +3.15% and +3.09% over the baseline, explicitly outperforming the standard TriHard strategy by margins of +0.97% and +1.35% on the two datasets. These consistent high-level gains demonstrate that structurally penalizing the p-n distance and adaptively weighting hard boundaries successfully prevents manifold collapse.

Impact of CentroidM Loss and Masking Strategy: The CentroidM loss is designed to enforce global hard mining by utilizing learnable class centers. Both visual representations

(Figures 5 and 6) and quantitative evaluations (Tab.V) on MNIST confirm that integrating the $Centr\_1$ Loss (CentroidM Loss) tightly compresses intra-class features while pushing inter-class boundaries apart, leading to clear performance improvements. Tab.VI investigates the hyperparameter ratio ($\lambda_1 : \lambda_0$), confirming its standalone effectiveness. Furthermore, to address the distribution disruption caused by conventional resampling, our 1st & 2nd order masks effectively eliminate repeated sample interference. As presented in Tab.IX, they steadily boost the baseline performance without destructive data manipulation.

Comprehensive Synergy of the Framework: The synergistic effects of our proposed modules are detailed in Tables X and XI. For centroid-level retrieval, the integration of adaptive instance-level mining and global centroid regulation (TriWeight + CentroidM) yields the optimal performance. Notably, this combination achieves a remarkable centroid-level mAP of 96.74% and Rank-1 accuracy of 95.66% on Market1501, along with a peak mAP of 92.99% on DukeMTMC-ReID. For instance-level retrieval, the best performance is obtained by combining TriHard+ and CentroidM, demonstrating the complementary strengths of structured hard mining and global centroid constraints. Together, these results verify that local metric learning and global centroid regularization mutually enhance feature discriminability and robustness in different evaluation settings.

Furthermore, we visualize the three typical centroid-level retrieval results on Market1501 in Fig. 7. Analyzing the failure cases (indicated by red bounding boxes) row by row reveals several key misleading factors. In the first row, extreme similarity in clothing styles—such as light-colored dresses or white tops paired with dark lower wear—dominates the feature representation, leading to identity confusion. The second row demonstrates that congruent pedestrian postures, coupled with similar local visual attributes like shorts and exposed bare legs, can also easily misguide the model. Finally, the third row highlights the negative impact of visually prominent environmental obstacles; specifically, the presence of bicycles introduces severe background bias that overrides fine-grained identity cues.

Comparison with State-of-the-Art Methods: Finally, we benchmark our framework against recent State-of-the-Art (SOTA) methods in Tab.XII. To comprehensively evaluate our method, we categorize the comparisons by CNN and Transformer backbones. Under the standard instance-level protocol, our Se-ReID_I built on the ResNet50 baseline achieves a highly competitive 87.8% mAP on Market1501, outperforming strong CNN competitors like CTF [22] (87.7%) and CAAO [23] (87.3%). When integrated with the modern SOLIDER-tiny baseline, Se-ReID_I establishes new SOTA instance-level performance, pushing the Market1501 mAP to 92.1% (surpassing CLIP-ReID [24] at 90.5% and PFD [25] at 89.6%) and demonstrating robust results on DukeMTMC-ReID. Furthermore, driven by the spatially enhanced features, our centroid-based retrieval pipeline (Se-ReID_C) exhibits overwhelming superiority. It reaches a striking 98.1% mAP on Market1501 and 96.9% mAP on DukeMTMC-ReID under the Transformer architecture. It is worth emphasizing that Se-ReID_C not only achieves top-tier scores by naturally filtering out gallery noise, but also provides an $8.2\times$ inference speedup (Tab.II). Overall, the framework maintains robust baseline scalability across diverse architectures without relying on complex attention mechanisms or computationally expensive ReRank techniques.

TABLE III: C-l (centroid-level) results with TriHard+ Loss

Loss	Market1501				DukeMTMCReID			
	mAP	r=1	r=5	r=10	mAP	r=1	r=5	r=10
TriHard+	**96.27**	**95.28**	**97.57**	**98.75**	**91.59**	**89.23**	**95.2**	**96.68**
TriHard	95.79	94.71	97.06	98.6	91.46	89.09	94.88	96.36
baseline	95.4	94.21	97.12	98.72	90.6	87.66	94.79	96.23

TABLE IV: I-l (instance level) results with TriHard+

Loss	Market1501				DukeMTMCReID			
	mAP	r=1	r=5	r=10	mAP	r=1	r=5	r=10
TriHard+	**87.24**	**94.57**	**98.34**	**99.02**	**77.62**	**87.57**	**94.08**	**96.01**
TriHard	85.46	93.79	97.89	98.96	75.42	86.13	93.85	95.51
baseline	83.28	93.37	97.95	98.67	73.68	85.09	93.01	95.75

TABLE V: Comparison with/without CentroidM (Centr_1) on MNIST

$\lambda_0, \lambda_1 \backslash Accuracy$	Center Loss	Centr_1 Loss
0.001,0.0005	95.96	96.93
0.01,0.005	95.85	97.17
0.1,0.05	96.83	97.80
1,0.5	**97.05**	**98.00**

TABLE VI: Ablation study of CentroidM loss (CentroidM = centr_1 loss + center loss) on Market1501. By tuning the weight ratio of centr_1 loss to center loss, we observe more pronounced gains at the instance level. This validates our conjecture that the simplified centroid retrieval task benefits less from margin mining constraints.

	Centroid			Instance		
λ_1:λ_0	mAP	r=1	r=5	mAP	r=1	r=5
base	95.4	94.21	97.12	83.28	93.37	97.95
0.1	95.77	94.39	97.12	84.33	93.47	97.92
0.25	95.79	94.27	96.88	84.23	93.5	97.89
0.5	95.88	94.45	96.91	84.17	93.59	98.04
0.75	**96.22**	94.74	97.24	84.4	93.82	98.07
1	95.97	94.42	**97.33**	84.4	93.88	98.16
1.25	95.89	94.42	97.12	84.13	93.44	97.89
1.5	95.84	**94.86**	97.03	84.22	93.35	97.92
1.75	95.75	94.45	97.3	84.1	93.62	**98.19**
2	95.73	94.57	97.03	84.13	93.59	98.04
$+\infty$	95.7	94.15	97.24	**84.41**	**93.91**	98.07

TABLE VII: C&I-l (Centroid and instance level) experiment with TriWeight on Mk (Market1501)

Parameters	Centroid				Instance			
	mAP	r=1	r=5	r=10	mAP	r=1	r=5	r=10
TriWeight	**96.29**	**95.37**	**97.57**	**98.78**	**86.43**	**93.85**	**98.34**	**98.96**
TriHard	95.79	94.71	97.06	98.6	85.46	93.79	97.89	**98.96**
baseline	95.4	94.21	97.12	98.72	83.28	93.37	97.95	98.67

TABLE VIII: C&I-l results with TriWeight on Dk(DukeMTMCReID).

Parameters	Centroid				Instance			
	mAP	r=1	r=5	r=10	mAP	r=1	r=5	r=10
TriWeight	**92.27**	**89.9**	**95.74**	**96.86**	**76.77**	**86.98**	**94.25**	**96.27**
TriHard	91.46	89.09	94.88	96.36	75.42	86.13	93.85	95.51
baseline	90.6	87.66	94.79	96.23	73.68	85.09	93.01	95.75

TABLE IX: Ablation of 1st & 2nd order masks on Market1501.

1st order	2nd order	R1	mAP
		93.5	85.66
✓		94.34	86.13
	✓	94.38	86.21
✓	✓	94.97	87.81

TABLE X: C-l comprehensive ablation study. The losses are appended to the baseline.

Loss	Market1501			DukeMTMCReID		
	mAP	r=1	r=5	mAP	r=1	r=5
baseline	95.4	94.21	97.12	90.6	87.66	94.79
CentroidM	96.22	94.74	97.24	91.04	88.06	94.7
CentroidM W/O Center	95.7	94.15	97.24	90.71	88.02	94.52
TriHard$_+$	96.27	95.28	97.57	91.59	89.23	95.2
TriHard$_+$ CentroidM	96.76	95.36	97.78	91.98	89.21	95.77
TriWeight	96.29	95.37	97.57	92.27	**89.9**	95.74
TriWeight CentroidM	**96.74**	**95.66**	**97.98**	**92.99**	89.72	**96.38**

TABLE XI: I-l comprehensive ablation study (add losses).

Loss	Market1501			DukeMTMCReID		
	mAP	r=1	r=5	mAP	r=1	r=5
baseline	83.28	93.37	97.95	73.68	85.09	93.01
CentroidM	84.4	93.82	98.07	74.44	85.5	93.76
CentroidM W/O Center	84.41	93.91	98.07	74.48	85.05	93.18
TriHard$_+$	87.24	94.57	98.34	77.62	87.57	94.08
TriHard$_+$ CentroidM	**87.81**	**94.97**	**98.95**	**78.24**	**88.13**	**94.78**
TriWeight	86.43	93.85	98.34	76.77	86.98	94.25
TriWeight CentroidM	87.21	94.36	98.55	77.48	86.95	94.48

TABLE XII: Comparison with state-of-the-art methods. Methods are categorized by CNN and Transformer backbones. For fair comparison, results are reported under instance-level retrieval (denoted by "-I"). Bold and underline indicate the best and second-best performance, respectively. Centroid-level results ("-C") are provided for reference only.

Method	Market1501		DukeMTMC-ReID	
	mAP	r=1	mAP	r=1
TCA-ReID [26] (CNN)	86.3	94.3	75.8	86.4
TAPA [27]	86.8	94.7	75.8	86.4
OSNet [28]	84.9	94.8	73.5	88.6
CDNet [29]	83.7	93.7	73.9	86.7
StrongBase [30]	85.9	94.5	76.4	86.4
CAAO [23]	87.3	95.1	77.5	88.9
CTF [22]	87.7	94.8	74.9	87.4
FD-GAN [31]	77.7	90.5	64.5	80.0
PCB [32]	71.4	92.3	66.1	81.8
DSR [33]	64.3	83.6	—	—
FPR [34]	86.6	95.4	78.4	88.6
SBFR-baseline_I [15](ResNet50)	83.3	93.4	73.7	85.1
+Se-ReID_I	87.8	95.0	78.2	88.1
+Se-ReID_C	96.7	95.7	93.0	89.7
CLIP-ReID [24] (Trans.)	90.5	95.4	**83.1**	90.8
VT-ReID [35]	88.1	93.8	79.2	**92.6**
TransReID [36]	88.8	95.0	81.8	90.4
PAT [37]	88.0	95.4	78.2	88.8
PFD [25]	89.6	95.5	82.2	90.6
FED [38]	86.3	95.0	78.0	90.6
SOLIDER-tiny_I [18](Trans.)	<u>91.6</u>	<u>96.1</u>	81.8	91.1
+Se-ReID_I†	**92.1**	**96.7**	<u>82.8</u>	<u>91.4</u>
+Se-ReID_C†	98.1	96.8	96.9	94.8

V. CONCLUSION

We propose Se-ReID, a novel framework that effectively models both relative and absolute feature relationships through complementary loss functions. At the instance level, we design two alternative metric losses: TriHard+ Loss introduces difficulty-aware dynamic routing and spatial geometric constraints to prevent manifold collapse, while TriWeight Loss employs hard-adapted soft weighting to preserve dense intra-class structures without ignoring global distributions. At the centroid level, the parameter-efficient CentroidM loss enhances inter-class boundary discrimination and enables global hard mining, while sharing parameters with the Center loss. To guarantee an unbiased feature space, 1st and 2nd order masks are introduced to eliminate repeated sample interference without disrupting the natural data distribution. Furthermore, a cross-camera centroid gallery reconstruction scheme is established, which naturally filters out gallery noise and substantially boosts retrieval efficiency.

The framework demonstrates strong performance on standard benchmarks including Market1501 and DukeMTMC-ReID for both centroid and instance retrieval tasks. Specifically, on the Market1501 and DukeMTMC-ReID datasets, our proposed variants yield maximum improvements of 1.36% and 2.39% in centroid-level mAP, while the instance-level mAP surges by up to 4.53% and 4.56% over the CNN baseline. Furthermore, our approach achieves state-of-the-art (SOTA) performance on the Transformer baseline, validating its effectiveness and broad applicability. Although training is computationally more intensive due to the integration of multiple joint losses, inference efficiency is drastically improved (achieving an $8.2\times$ speedup) through the proposed centroid-based retrieval. Note that the high performance of centroid re-

trieval is mathematically expected, as prototype averaging acts as a spatial low-pass filter that inherently suppresses instance-level noise. Finally, current limitations in handling occlusions and cross-modal scenarios present promising directions for future research.

REFERENCES

[1] M. Zhang, Y. Xiao, F. Xiong, S. Li, Z. Cao, Z. Fang, and J. T. Zhou, "Person re-identification with hierarchical discriminative spatial aggregation," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 516–530, 2022.

[2] Z. Li, Y. Shi, H. Ling, J. Chen, B. Liu, R. Wang, and C. Zhao, "Viewpoint disentangling and generation for unsupervised object re-id," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 5, pp. 1–23, 2024.

[3] P. zhang, Y. Wang, Y. Liu, Z. Tu, and H. Lu, "Magic tokens: Select diverse tokens for multi-modal object re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

[4] X. Xu, X. Yuan, Z. Wang, K. Zhang, and R. Hu, "Rank-in-rank loss for person re-identification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 2s, pp. 1–21, 2022.

[5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*. Springer, 2016, Conference Proceedings, pp. 499–515.

[6] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, Conference Proceedings, pp. 815–823.

[7] Q. Xiao, H. Luo, and C. Zhang, "Margin sample mining loss: A deep learning based method for person re-identification," *arXiv preprint arXiv:1710.00478*, 2017.

[8] Y. Sun, L. Zheng, and Y. Yang, "Circle loss: A unified perspective of pair similarity optimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 6398–6407.

[9] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[10] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 403–412.

[11] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, Conference Proceedings, pp. 212–220.

[12] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 750–18 759.

[13] C. Yang, Q. Wu, J. Wang, and J. Yan, "Graph neural networks are inherently good generalizers: Insights by bridging gnns and mlps," *arXiv preprint arXiv:2212.09034*, 2022.

[14] M. Ye, J. Shen, G. Lin, T. Xiang, and S. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2021.

[15] M. Wieczorek, A. Michalowski, A. Wroblewska, and J. Dabrowski, "A strong baseline for fashion retrieval with person re-identification models," in *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part IV 27*. Springer, 2020, Conference Proceedings, pp. 294–301.

[16] M. Wieczorek, B. Rychalska, and J. Dabrowski, "On the unreasonable effectiveness of centroids in image retrieval," in *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28*. Springer, 2021, Conference Proceedings, pp. 212–223.

[17] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2019.

[18] W. Chen, X. Xu, J. Jia, H. Luo, Y. Wang, F. Wang, R. Jin, and X. Sun, "Beyond appearance: a semantic controllable self-supervised learning framework for human-centric visual tasks," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[19] Z. Zhang, "Improved adam optimizer for deep neural networks," in *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*. Ieee, 2018, Conference Proceedings, pp. 1–2.

[20] N. S. Keskar and R. Socher, "Improving generalization performance by switching from adam to sgd," *arXiv preprint arXiv:1712.07628*, 2017.

[21] G. Wang, J. Lai, P. Huang, and X. Xie, "Spatial-temporal person re-identification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 8933–8940.

[22] A. Zhang, Y. Gao, Y. Niu, W. Liu, and Y. 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*, 2021, pp. 598–607.

[23] C. Zhao, Z. Qu, X. Jiang, Y. Tu, and X. Bai, "Content-adaptive auto-occlusion network for occluded person re-identification," *IEEE Transactions on Image Processing*, vol. 32, pp. 4223–4236, 2023.

[24] S. Li, L. Sun, and Q. Li, "Clip-reid: exploiting vision-language model for image re-identification without concrete text labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, Conference Proceedings, pp. 1405–1413.

[25] T. Wang, H. Liu, P. Song, T. Guo, and W. Shi, "Pose-guided feature disentangling for occluded person re-identification based on transformer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2540–2549.

[26] X. Teng, L. Lan, J. Zhao, X. Li, and Y. Tang, "Highly efficient active learning with tracklet-aware co-cooperative annotators for person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 11, pp. 15 687–15 700, 2024.

[27] P. Wang, F. Su, Z. Zhao, Y. Zhao, and N. V. Boulgouris, "Gareid: Grouped and attentive high-order representation learning for person re-identification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 3, pp. 3990–4004, 2025.

[28] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3702–3712.

[29] H. Li, G. Wu, and W.-S. Zheng, "Combined depth space based architecture search for person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6729–6738.

[30] H. Luo, W. Jiang, Y. Gu, F. Liu, X. Liao, S. Lai, and J. Gu, "A strong baseline and batch normalization neck for deep person re-identification," *IEEE Transactions on Multimedia*, vol. 22, no. 10, pp. 2597–2609, 2019.

[31] Y. Ge, Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang *et al.*, "Fd-gan: Pose-guided feature distilling gan for robust person re-identification," *Advances in neural information processing systems*, vol. 31, 2018.

[32] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. 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)*, 2018, pp. 480–496.

[33] L. He, J. Liang, H. Li, and Z. Sun, "Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7073–7082.

[34] L. He, Y. Wang, W. Liu, H. Zhao, Z. Sun, and J. Feng, "Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8450–8459.

[35] S. Xiang, C. Liu, J. Ruan, S. Cai, S. Du, and D. Qian, "Vt-reid: Learning discriminative visual-text representation for polyp re-identification," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 3170–3174.

[36] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "Transreid: Transformer-based object re-identification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, Conference Proceedings, pp. 15 013–15 022.

[37] Y. Li, J. He, T. Zhang, X. Liu, Y. Zhang, and F. 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*, 2021, pp. 2898–2907.

[38] Z. Wang, F. Zhu, S. Tang, R. Zhao, L. He, and J. Song, "Feature erasing and diffusion network for occluded person re-identification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4754–4763.