Revisiting Attention in the Dark for Low-Light Person Re-Identification

Xiang Guo[1], **Ruimin Hu**[1,2*], **Dongliang Zhu**[1], **Mei Wang**[1]

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University
[2]School of Cyber Science and Engineering, Wuhan University, Wuhan, China
nanqiaobei@163.com, {hrm, zhudongliang,dr.mei.wang}@whu.edu.cn

Abstract

Person re-identification (Re-ID) under extremely low-light conditions suffers from severe image degradation, which significantly impairs the extraction of identity-discriminative features. Existing methods struggle to recover semantic information that is obscured under poor illumination. To better understand this problem, we conduct a comprehensive analysis of the semantic modeling behavior of Re-ID models in low-light settings. For the first time, we investigate the norm distributions of Query (Q), Key (K), and Value (V) vectors within the attention module and observe that, as illumination decreases, the norm of Query vectors in pedestrian regions drops significantly. This leads to dispersed attention and degraded feature representations. To address this issue, we propose a novel framework named Norm-Ratio Attention and Semantic Recovery Distillation Network (NRSRD), which consists of two key components: a Norm-Ratio Attention Module (NRA) and a Semantic Recovery Distillation Module (SRD). The former dynamically adjusts attention responses based on the ratio of K/Q vector norms, enhancing structural region perception while suppressing background interference. The latter transfers discriminative semantic knowledge from high-illumination auxiliary data to the low-light model, compensating for the semantic degradation caused by poor lighting. Extensive experiments on multiple publicly available low-light Re-ID benchmarks demonstrate the effectiveness and superiority of the proposed method.

Code — https://github.com/nanqiaobei/NRSRD

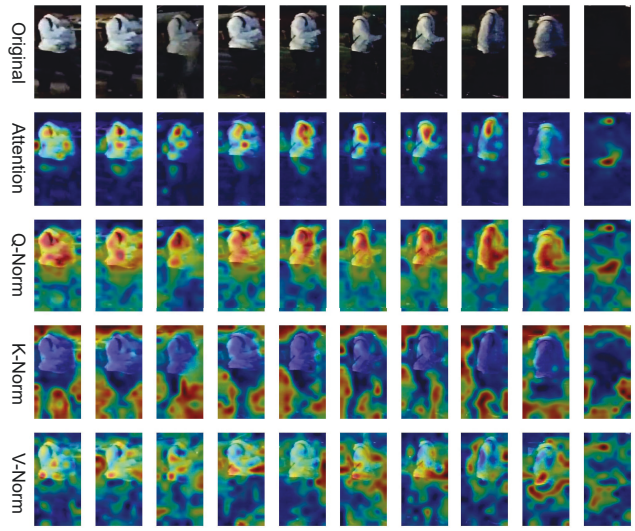


Figure 1: Visualization of attention maps and $Q/K/V$ vector norms from the final Transformer block of TransReID. As the illumination level decreases, the model's attention responses become significantly more dispersed, with a notable reduction in the Q-norm within pedestrian regions. This indicates weakened activation in discriminative areas, ultimately leading to attention degradation and reduced identity feature representation.

Introduction

Person re-identification (Re-ID) aims to accurately recognize the same individual across non-overlapping camera views, and it has been widely applied in practical scenarios such as public security and intelligent surveillance (Ye et al. 2021; Leng et al. 2025; Qin et al. 2024; Liu et al. 2024a). However, due to complex environmental and geographical factors, Re-ID systems often encounter significant illumination variations in real-world deployments (Li et al. 2024; Liu et al. 2024b; Chen et al. 2025; Huang et al. 2025). Existing methods primarily focus on modeling under normal or well-lit conditions, while overlooking the adverse effects of extreme lighting changes on image quality and recognition performance (Dai, Lu, and Li 2025; Pang et al. 2025). Although recent studies have begun to address the impact of illumination on Re-ID and have proposed certain mitigation strategies, pedestrian images captured under extremely low light conditions often suffer from severe degradation, including intense noise, low contrast, loss of fine details, and structural blurring. These degradations significantly obscure identity-relevant features and severely hinder effective feature extraction and matching (Hu et al. 2024; Shi et al. 2024; Li, Chen, and Ye 2024; Cui et al. 2024). Therefore, it is imperative to develop dedicated person Re-ID methods tailored for extremely low-light environments.

Under low-light conditions, pedestrian images suffer from

*Corresponding author

noise, structural blur, and texture loss, severely hindering discriminative feature extraction. To reduce noise interference, Zhang *et al.* (Zhang, Yuan, and Wang 2019) combined a denoising network with a Re-ID model to improve identity representation. While effective in suppressing noise, this approach neglects the "semantic drowning" effect caused by low illumination i.e., the loss of discriminative information due to insufficient brightness. To address this limitation, Lu *et al.* (Lu et al. 2023) proposed an illumination distillation framework to recover semantic content by leveraging the complementarity between original and enhanced images. However, it does not explicitly address noise artifacts in low-light conditions. Subsequently, Zhao *et al.* (Zhao et al. 2025) introduced a multi-branch EDA framework that jointly optimizes image enhancement and noise suppression modules, achieving impressive performance on several low-light Re-ID benchmarks. Furthermore, Lu *et al.* (Lu et al. 2025) developed a collaborative enhancement network that jointly models the interaction between illumination restoration and person re-identification tasks, enabling the cooperative optimization of these two heterogeneous objectives. While prior methods seek to restore identity cues under low illumination via enhancement strategies, their gains are mostly limited to the visual domain. Relying on image-level self-supervised signals, they struggle to recover high-level semantics severely degraded by extreme lighting (Zheng et al. 2023; Wang et al. 2020). This prompts a key question: *Can we recover discriminative semantics directly from low-light images without explicit enhancement?*

To investigate this problem, we conduct an in-depth analysis of the semantic modeling mechanism of the TransReID model under low-light conditions. As shown in Figure 1, with the gradual decrease in ambient illumination, the number of critical attention regions attended by the model significantly diminishes, leading to a progressive weakening of the extracted identity features.furthermore, Inspired by (Kobayashi et al. 2020) we analyze the $L2$-norm distributions of the Q, K, and V vectors within the attention module and examine their correspondence to semantic regions in the image. The results reveal that Q-norms are primarily concentrated in pedestrian structural regions, K-norms are strongly associated with background areas, and V-norms exhibit a distribution pattern similar to Q, mainly responding to semantically relevant regions. Notably, as illumination decreases, the Q-norms in pedestrian regions show a significant downward trend, this phenomenon indicates that pedestrian structural regions tend to exhibit low activation (i.e., small Q-norms) under low-light conditions, making it difficult for the attention mechanism to effectively model and extract semantic information from these regions. This serves as a deeper reason for the performance degradation of the model under extreme lighting conditions. Therefore, a key challenge lies in restoring the activation strength of Q vectors in structural regions under low-light scenarios, thereby guiding the model to refocus on discriminative semantic areas and ultimately improving recognition performance.

To effectively address this issue, we propose a novel framework named **Norm-Ratio Attention and Semantic Recovery Distillation Network** (NRSRD), which consists of two key components: a Norm-Ratio Attention Module (NRA) and a Semantic Recovery Distillation Module (SRD). These two modules are designed to enhance attention modeling and recover semantic information, respectively. Specifically, we introduce an attention modulation mechanism based on the ratio of Query (Q) to Key (K) vector norms. Since Q-norms effectively indicate pedestrian structural regions while K-norms are predominantly distributed in background areas, the K/Q norm ratio can be leveraged to suppress the attention contribution of background tokens, thereby enhancing the model's focus on human body regions. Moreover, considering that certain identity-related semantic cues may be irrecoverable at the image level under low-light conditions, we design a Semantic Recovery Distillation Module. This module distills discriminative semantic knowledge from high-illumination auxiliary data that are highly relevant to the target domain and transfers it to the low-light model. By sharing the representation space with the backbone network and employing class-specific and class-agnostic knowledge transfer, this module effectively restores the semantic representation capacity of structural regions degraded by poor lighting, thereby enhancing recognition performance in extremely low-light scenarios. our approach significantly improves Re-ID performance on nighttime datasets. The main contributions of this work are summarized as follows:

- We reveal that Q-norms highlight pedestrian structures, K-norms focus on backgrounds, and V-norms align with Q, attending to semantic regions. To the best of our knowledge, this is the first work to reveal the norm behavior of query, key, and value vectors in the final attention layer of TransReID, offering new insights into person re-identification.
- We propose an efficient Norm-Ratio Attention that suppresses background interference and enhances attention to discriminative structures.
- We propose a Semantic Recovery Distillation Module that leverages knowledge distillation to transfer latent semantics from well-lit data, compensating for information lost under low-light conditions.

Related Works

Person Re-Identification in Nighttime

Although person Re-ID has advanced under standard lighting, its performance in low-light conditions remains limited due to reduced visibility, noise, color distortion, and structural blur, which hinder identity-discriminative feature extraction. To tackle low-light challenges, researchers have explored various enhancement methods to improve image clarity and model robustness. Zhang *et al*. (Zhang, Yuan, and Wang 2019) proposed an end-to-end framework combining denoising and Re-ID, which effectively suppresses noise but overlooks semantic degradation caused by poor illumination. To address illumination degradation, Lu *et al*. (Lu et al. 2023) proposed a light distillation framework leveraging complementary information from original and relit images to recover lost content. While effective in enhancing brightness and detail, it ignores the pervasive noise in

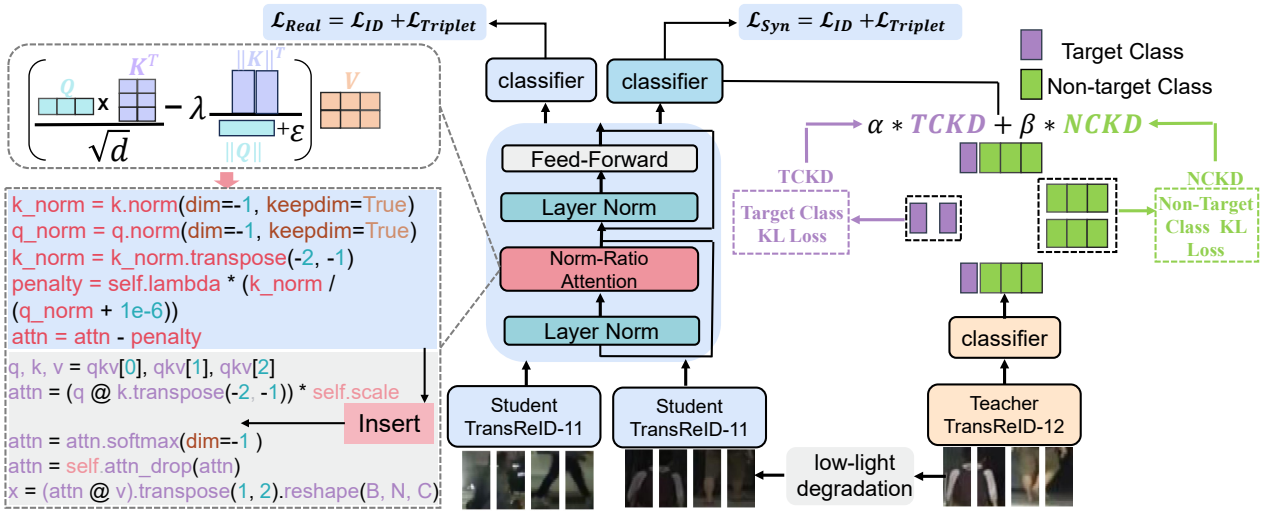


Figure 2: Overview of our framework. In the figure, "TransReID-11" refers to the first 11 layers of the TransReID model, while "TransReID-12" denotes the complete architecture. The full model comprises two core modules: (1) the Norm-Ratio Attention (NRA) module (left part), and (2) the Semantic Recovery Distillation (SRD) module (right part).

low-light images. Zhao *et al*. (Zhao et al. 2025) designed a multi-branch enhancement–denoising–alignment framework. By jointly modeling the enhancement and denoising processes and introducing a feature alignment module, the framework achieves dual optimization of image quality and identity preservation. Furthermore, Lu *et al*. (Lu et al. 2025) proposed a multi-domain collaborative network that jointly models relighting and Re-ID tasks, enabling feature-level interaction to enhance semantic discriminability and improve recognition performance under low-light conditions. *In contrast to prior methods that rely on illumination enhancement, we directly recover semantic information suppressed by low-light conditions to improve Re-ID task performance*.

Attention Analysis

Since the introduction of Transformers (Vaswani et al. 2017), self-attention has become a fundamental module in deep learning, enabling global context modeling via Query-Key similarity and Value aggregation. With Vision Transformers (ViT) (Dosovitskiy et al. 2020; Han et al. 2022; Khan et al. 2022), attention mechanisms have largely replaced convolutions in vision tasks. However, the internal behavior of attention remains insufficiently understood (Dong, Cordonnier, and Loukas 2021; Yeh et al. 2023). Recent efforts have sought to interpret attention via visualization and statistics, aiming to reveal whether models attend to semantically meaningful regions.

Early analyses focused on attention map (Zhang and Xiao 2019; Xue et al. 2022; Rassin et al. 2023). Chefer *et al*. (Chefer, Gur, and Wolf 2021) introduced gradient-based methods to trace multi-layer contributions, while Abnar *et al*. (Abnar and Zuidema 2020) proposed metrics to assess attention flow across layers. Touvron *et al*. (Touvron et al. 2021) studied the evolution of attention with network depth, and Paul *et al*. (Paul and Chen 2022) examined its robustness and discriminative capacity. Wang *et al*. (Wang et al. 2022) addressed token over-smoothing by restoring local features. Beyond attention weights, Kobayashi *et al*. (Kobayashi et al. 2020) emphasized the importance of Value vector norms in attention outputs, offering a norm-based perspective. In person Re-ID, TransReID (He et al. 2021) visualized attention focusing on human structures, demonstrating its impact on identity discrimination. However, most ReID studies still center on attention weights, with limited exploration of the representational behavior of Q, K, and V vectors (Xu et al. 2018; Sheng et al. 2023).

To bridge this gap, we perform a norm-based analysis of Q, K, and V. We observe that: (1) Q-norms align with pedestrian structure, concentrating on keypoints like body contours; (2) K-norms emphasize background regions, reflecting their role in context aggregation; and (3) V-norms mirror Q, highlighting semantic areas.

Method

In this section, we present our proposed model for nighttime person re-identification, termed NRSRD. The overall architecture is illustrated in Figure 2. The innovations of this model lie primarily in two aspects: (1) **Norm-Ratio Attention** (NRA), in which we modify the attention mechanism in the final layer of TransReID by introducing a penalty term based on the ratio of vector norms, thereby forming a norm-ratio constrained attention mechanism; and (2) **Semantic Recovery Distillation** (SRD), which leverages knowledge distillation to effectively recover crucial semantic information that is lost under low-light conditions.

Norm-Ratio Attention Module

The multi-head self-attention mechanism in TransReID degrades significantly under low-light conditions, exhibiting diffuse or misaligned attention that impairs feature extraction. We attribute this to the suppressed activation of Query

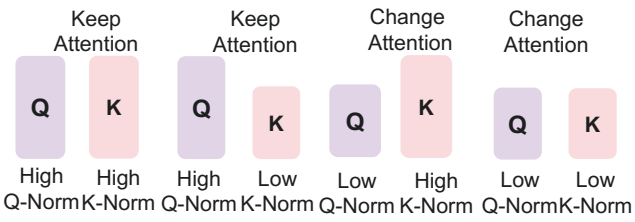


Figure 3: Q–K norm patterns in low-light TransReID.

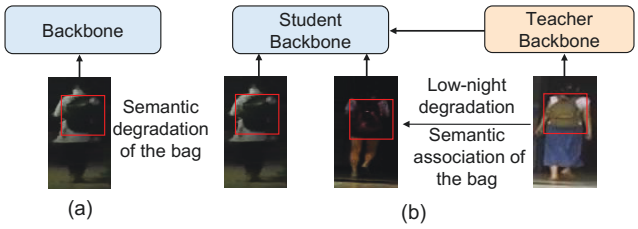


Figure 4: Motivation for using semantic recovery distillation

vectors in structural regions, as indicated by reduced Q-norms. This polarization weakens the representation of key semantic areas, resulting in semantic degradation.

To address this challenge, we reshape the attention distribution by modulating the relative contributions of the Q and K vector norms. As illustrated in Figure 3, certain combinations of Q–K norm values can severely disrupt the attention mechanism. In particular, a low Q-norm paired with a high K-norm tends to produce misleading attention links, while a low Q-norm with a low K-norm requires enhancing the expressiveness of Q to better recover structural semantics. Motivated by these observations, we introduce a penalty-based attention modulation strategy that amplifies the impact of Q-norms and suppresses the influence of K-norms. This design encourages the model to prioritize semantically meaningful regions even under poor illumination conditions. Building on this idea, we propose the **Norm-Ratio Attention** to effectively enhance feature discrimination in low-light scenarios. Specifically, we start from the standard attention formulation:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}}\right)\boldsymbol{V}. \quad (1)$$

Then, an extended formulation incorporating a norm-based penalty term:

$$\boldsymbol{\gamma} = \lambda \cdot \frac{\|\boldsymbol{K}\|}{\|\boldsymbol{Q}\| + \epsilon}, \quad (2)$$

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d}} - \boldsymbol{\gamma}\right)\boldsymbol{V}, \quad (3)$$

where $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{N\times d}$ denote the query, key, and value matrices, and d is the feature dimension. The term $\boldsymbol{\gamma}$ is the norm-based penalty, $\|\cdot\|$ denotes the ℓ_2 norm, λ is a learnable scaling factor, and $\epsilon = 10^{-6}$ ensures numerical stability. This mechanism enhances the role of query norms while suppressing the influence of key norms. It effectively reduces spurious attention, especially when low-norm queries attend to high-norm keys, often corresponding to background tokens. In contrast, meaningful connections such as high-query to high-key or high-query to low-key remain unaffected. Connections between low-norm queries and low-norm keys are also adaptively modulated. Overall, this strategy reshapes attention distribution without harming valid semantic relations and can be easily applied to the last attention block of TransReID.

Semantic Recovery Distillation Module

Low-light conditions significantly suppress the norm of Query vectors in structural regions, leading to partial loss of semantic information. To address this, we propose a knowledge distillation strategy as an alternative to traditional illumination enhancement, aiming to recover semantics degraded by low-light conditions. As illustrated in Figure 4, we hypothesize that certain discriminative cues present in well-lit images from external domains are missing in the target domain. Through knowledge distillation, we transfer these cues to effectively compensate for semantic degradation under poor illumination.

Specifically, we select high-illumination images X^H from an external domain and synthesize their low-light counterparts X^L via a physics-based degradation model. The original high-light image X is fed into a pre-trained teacher model to obtain the class probability distribution $\boldsymbol{P}^H = [p_1^H, p_2^H, \ldots, p_t^H, \ldots, p_C^H] \in \mathbb{R}^{1\times C}$, Simultaneously, the synthetic low-light image X^L is fed into the student model, yielding a corresponding prediction $\boldsymbol{P}^L = [p_1^L, p_2^L, \ldots, p_t^L, \ldots, p_C^L] \in \mathbb{R}^{1\times C}$. where $p_i^H(p_i^L)$ denotes the predicted probability of the i-th class and C is the total number of classes.

Inspired by (Zhao et al. 2022), we further incorporate a decoupled distillation strategy that separates category-relevant and category-irrelevant information, aiming to enhance the model's capability in capturing both discriminative semantics and shared background structures across domains:

$$\mathcal{L}_{\text{KD}}^{\text{decoupled}} = \alpha\text{TCKD} + \beta\text{NCKD}, \quad (4)$$

Here, $\alpha = 1.0$ and $\beta = 0.8$ are hyper-parameters. Target Class Knowledge Distillation (TCKD) transfers Class-specific semantic knowledge from daytime ReID datasets to student model, while Non-target Class Knowledge Distillation (NCKD) fundamentally enables effective logit distillation by aligning class-agnostic relational patterns. Specifically. TCKD quantifies the distribution similarity between teacher and student models for target-class probabilities, and can be formulated as follows:

$$\text{TCKD} = p_t^H \log\left(\frac{p_t^H}{p_t^L}\right) + p_{\backslash t}^H \log\left(\frac{p_{\backslash t}^H}{p_{\backslash t}^L}\right), \quad (5)$$

and NCKD measures the similarity between the teacher's and student's predicted probabilities over the non-target

Method	Venue	Night600				$\text{RGBNT201}_{\text{rgb}}$			
		mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
IDE+ (Zheng, Zheng, and Yang 2017)	TOMM 2017	3.3	7.75	20.0	27.4	19.4	17.2	30.3	37.4
PCB (Sun et al. 2018)	ECCV 2018	6.0	12.8	26.47	35.1	17.8	14.3	28.8	35.8
ABD-Net (Chen et al. 2019)	ICCV 2019	7.2	14.3	29.59	40.0	19.0	15.6	26.3	35.8
BoT (Luo et al. 2019)	CVPR 2019	5.4	11.2	22.60	30.6	21.9	19.3	32.8	43.5
AGW (Ye et al. 2021)	TPAMI 2021	6.1	12.5	24.40	30.9	23.7	21.1	37.2	49.3
TransReID (He et al. 2021)	ICCV 2021	8.4	16.0	31.1	39.9	36.1	34.7	53.60	63.2
IDF (Lu et al. 2023)	TMM 2023	9.2	17.2	34.4	43.8	38.0	38.4	53.7	63.2
EAD (Zhao et al. 2025)	AAAI2025	11.6	23.8	-	-	-	-	-	-
CNet (Lu et al. 2025)	TIFS2025	**13.3**	<u>25.0</u>	<u>43.0</u>	<u>51.7</u>	<u>55.3</u>	<u>57.4</u>	<u>73.6</u>	**81.5**
NRSRD(Ours)		<u>12.5</u>	**27.2**	**43.9**	**51.8**	**58.9**	**58.9**	**74.0**	<u>80.6</u>

Table 1: Performance comparison on two low-light person ReID datasets: Night600 and $\text{RGBNT201}_{\text{rgb}}$. We report mAP, Rank-1, Rank-5, and Rank-10 accuracy (%). Best results are shown in bold.

classes, and can be formulated as follows:

$$\text{NCKD} = p_t^H \sum_{i=1, i\neq t}^{C} \hat{p}_i^H \log\left(\frac{\hat{p}_i^H}{\hat{p}_i^L}\right). \tag{6}$$

Here, we use p_t to denote the predicted probability of the *target class*, and $p_{\setminus t}$ to represent the predicted probabilities over all *non-target classes*. Specifically, p_t^H and p_t^L denote the probabilities of the target class predicted by the teacher (high-light) and student (low-light) models, respectively, while $p_{\setminus t}^H$ and $p_{\setminus t}^L$ denote the predicted distributions over the non-target classes from the teacher $\hat{p}_i^H = p_i^H / p_{\setminus t}^H$ and student models $\hat{p}_i^L = p_i^L / p_{\setminus t}^L$ according to the formulation in (Zhao et al. 2022), respectively.

Optimization

The first component targets real low-light images from the target domain, where the student model is trained using cross-entropy and triplet losses (Hermans, Beyer, and Leibe 2017) to enhance its discriminative ability. The second component uses synthetic low-light data, with the student model optimized by the same losses while establishing correspondence with the teacher model. The third component comes from the semantic recovery distillation module, where the teacher model transfers the semantic knowledge lost under low-light conditions to the student model through a distillation loss. The overall objective can be formally expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{real}} + \mathcal{L}_{\text{syn}} + \mathcal{L}_{\text{KD}}^{decoupled}, \tag{7}$$

where $\mathcal{L}_{\text{real}}$ denotes the loss on real low-light images using the student model, defined as:

$$\mathcal{L}_{\text{real}} = \lambda_1 \mathcal{L}_{\text{ID}}^{\text{real}} + +\lambda_2 \mathcal{L}_{\text{Triplet}}^{\text{real}}, \tag{8}$$

and $\mathcal{L}_{\text{syn}}$ represents the loss on synthetic low-light images using the student model, defined as:

$$\mathcal{L}_{\text{syn}} = \lambda_3 \mathcal{L}_{\text{ID}}^{\text{syn}} + \lambda_4 \mathcal{L}_{\text{Triplet}}^{\text{syn}}, \tag{9}$$

The total loss function$\mathcal{L}_{\text{total}}$ can be rewritten as:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{ID}}^{\text{real}} + \lambda_2 \mathcal{L}_{\text{Triplet}}^{\text{real}} + \lambda_3 \mathcal{L}_{\text{ID}}^{\text{syn}} + \lambda_4 \mathcal{L}_{\text{Triplet}}^{\text{syn}} \\ + \lambda_5 \mathcal{L}_{\text{KD}}^{decoupled}.\end{aligned} \tag{10}$$

where the hyper-parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are set to 0.5, 0.5, 0.5, 0.5, 0.02, respectively.

Experiment

Datasets and Evaluation Settings

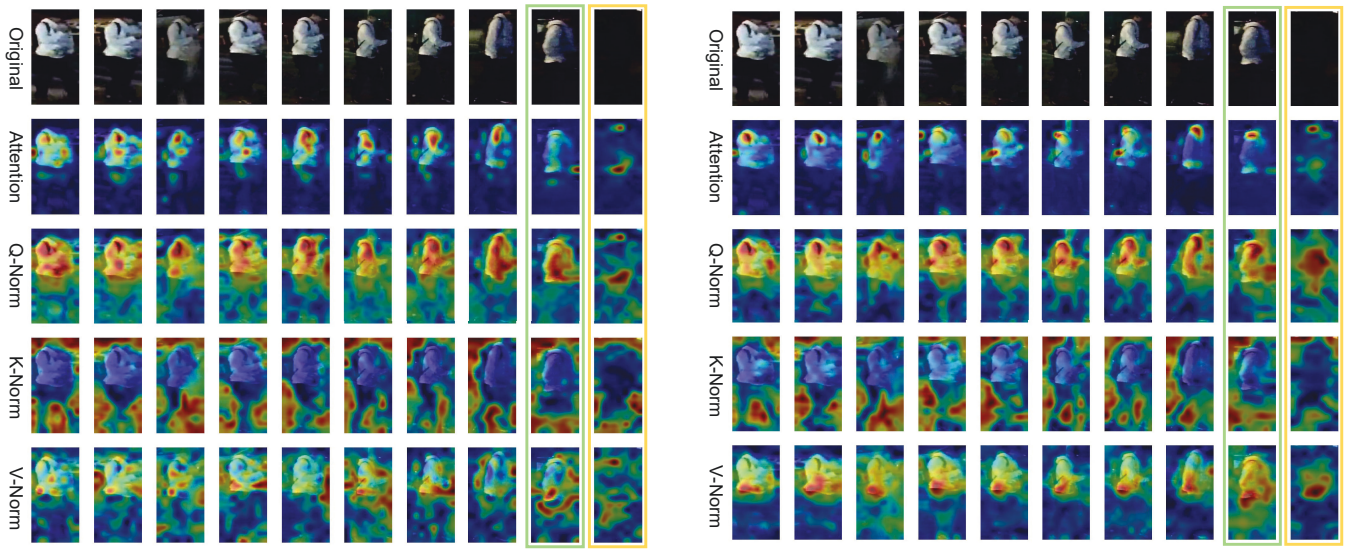
Night600 (Lu et al. 2023). The dataset contains 28,813 nighttime images of 600 identities captured from 8 non-overlapping surveillance cameras. It is evenly split, with 300 identities (14,462 images) for training and 300 for testing. The test set includes a query set of 2,180 images and a gallery of 14,351 images.

$\text{RGBNT201}_{\text{rgb}}$ (Zheng et al. 2021). This nighttime RGB subset of a multimodal ReID dataset contains 4,787 low-light images of 201 identities, with 171 used for training and 30 for testing. Captured under challenging nighttime conditions, it serves as a valuable benchmark for evaluating ReID performance in low-light scenarios.

Implementation Details. We use TransReID-SSL (Luo et al. 2021) as the backbone for both student and teacher models. Images are resized to 256×128, and training runs for 120 epochs with a batch size of 64 on a single NVIDIA A100 GPU. SGD with 0.9 momentum is used, with an initial learning rate of 0.0004, a 20-epoch warm-up, and cosine decay scheduling. The additional dataset used in this work is Syn_Dark, which contains both normally illuminated images and synthetically generated low-light counterparts. It is available from the source provided in (Lu et al. 2025).

Performance Evaluation

We evaluate state-of-the-art ReID methods on the nighttime datasets *Night600* and *RGBNT201_{rgb}*, with results shown in Table 1. (1) On *Night600*, traditional methods perform poorly under extreme low-light conditions; IDE+ and PCB achieve only 3.3% and 6.07% mAP, respectively, with low Rank-1 scores. More advanced architectures, such as BoT and AGW, show clear gains, while transformer-based TransReID further improves mAP and Rank-1 to 8.4% and 16.0%, respectively. Illumination-aware methods such as IDF, EAD, and CNet achieve superior results, with CNet performing best. Our method further improves performance,



(a) Baseline (b) Baseline+NRA

Figure 5: Under low-light conditions, we compare (a) the Baseline and (b) the Baseline with NRA in terms of attention maps and the norm distributions of the Q, K, and V vectors.

slightly lower than CNet in mAP but surpassing it in Rank-1, Rank-5, and Rank-10, with a notably higher Rank-1. (2) To evaluate the generalization ability of our method, we further conduct experiments on the $RGBNT201_{rgb}$ dataset. Our method achieves 58.9% in both mAP and Rank-1, outperforming CNet (55.3% mAP and 57.4% Rank-1). These results indicate that our method can effectively uncover semantic cues suppressed by poor illumination, even without explicit illumination enhancement, thereby demonstrating its robustness and effectiveness.

Ablation Study

Baseline	NRA	MDL	SRD	mAP	Rank-1	Rank-5
✓				9.5	20.0	36.1
✓	✓			10.4	21.8	39.0
✓		✓		10.9	24.2	41.8
✓	✓	✓		12.2	27.1	41.8
✓			✓	11.4	25.3	41.4
✓	✓		✓	**12.5**	**27.2**	**43.9**

Table 2: Module Ablation Results on Night600.

Ablation Study on Different Modules: To demonstrate the effectiveness of the modules in our proposed method, we conducted a series of ablation experiments on **Night600**. The rank-1, rank-5, rank-10, mAP accuracies (%) are reported in Table 2. For a more comprehensive ablation of the SRD module, we incorporate a multi-domain learning (MDL) strategy, which involves training the student model on a combination of synthetic low-light and target-domain samples. By comparing three pairs of experiments (Baseline+NRA vs. Baseline, Baseline+NRA+MDL vs. Baseline+MDL, and Baseline+NRA+SRD vs. Baseline+SRD), we observe consistent performance improvements with the addition of the NRA module, demonstrating its effectiveness. Furthermore, when comparing Baseline+MDL and Baseline+SRD, our SRD module achieves superior results across all metrics, indicating its advantage over standard multi-domain learning. Finally, by integrating both the NRA and SRD modules, our method achieves the best overall performance.

Visualization Analysis of the Effectiveness of the NRA Module: As illustrated in the Figure 5, the proposed NRA method leads to more focused and consistent attention distributions. On the one hand, as highlighted by the green box, previously incorrect attention regions have been effectively corrected. On the other hand, the comparison within the yellow box demonstrates that, even under extremely low-light conditions, the model is able to accurately focus its attention on key human body regions. Furthermore, the distribution of high-magnitude Q vectors increases significantly, while high-magnitude K vectors in background regions are notably reduced. These observations further validate the effectiveness of the proposed NRA method in enhancing attention mechanisms under low-light scenarios.

Validations and Analyses

Analysis of the Layer-wise Norm Behavior of TransReID's Q, K, and V Vectors: To gain a deeper understanding of the physical behavior of the Q, K, and V vectors, we analyze their norm distributions across different Transformer layers in the TransReID model. Specifically, we visualize the relationship between the vector norm of each patch and its distance to the image center, as shown in Figure 6. (1) We observe that among all layers, only the final layer (Block 12) exhibits a clear spatial distribution pattern: Q vectors with higher norms are primarily concentrated near the image

center, typically corresponding to pedestrian body regions, while K vectors with higher norms tend to appear farther from the center, often corresponding to background areas. This indicates that Q attends more to foreground structural regions, whereas K encodes more background information. Notably, this behavior is unique to the final attention layer and is not observed in earlier layers.

(2) Additionally, we identify an interesting trend under low-light conditions: in the first six layers, Q norms are generally higher than K norms, while in the latter six layers, K norms surpass those of Q. This pattern reveals a dynamic shift in the model's representation behavior across layers, reflecting how the Transformer gradually adjusts its focus between foreground and background features as depth increases.

Validation across different TransReID architectures: To further assess the effectiveness of the proposed Norm-Ratio Attention (NRA) when integrated into different backbone architectures, particularly under low-light conditions, we conduct empirical studies on several TransReID-SSL variants (Luo et al. 2021). As shown in Table 3, the integration of our method into ViT-S/16+ICS (Instance Center Structure) and ViT-B/16+ICS leads to performance improvements across various metrics, indicating that the proposed approach exhibits strong adaptability and effectiveness across different structures. In contrast, the performance

Backbone	NRA	mAP	Rank-1	Rank-5
ViT-S/16		8.5	18.8	34.2
ViT-S/16	✓	8.2	18.1	33.2
Gain (ViT-S/16)		**-0.3**	**-0.7**	**-1.0**
ViT-S/16+ICS		8.3	17.4	34.5
ViT-S/16+ICS	✓	8.7	18.5	34.8
Gain (ViT-S/16+ICS)		**+0.4**	**+1.1**	**+0.3**
ViT-B/16+ICS		9.5	20.0	36.1
ViT-B/16+ICS	✓	10.4	21.8	39.0
Gain (ViT-B/16+ICS)		**+0.9**	**+1.8**	**+2.9**

Table 3: NRA on Different Baselines (Night600).

on ViT-S/16 without ICS shows a decline, primarily due to the absence of the ICS module. ICS effectively mitigates the domain gap between models pre-trained under normal lighting and images captured under low-light conditions. Without it, our norm-ratio mechanism becomes more susceptible to large-scale illumination variations, thus resulting in suboptimal performance.

Validation under normal illumination conditions: To verify the generality of the following observations, (1) the

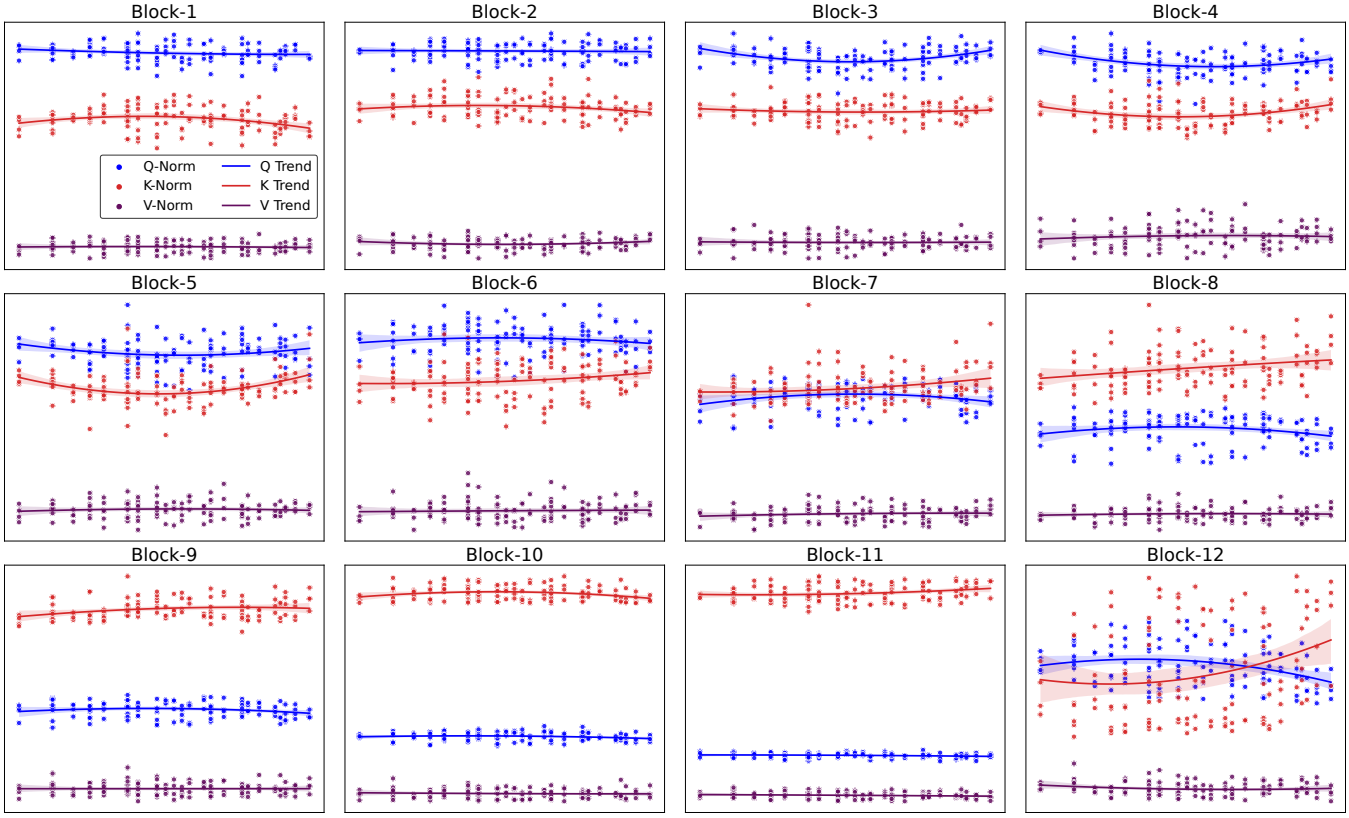


Figure 6: Layer-wise visualization of Q, K, and V vector norms relative to spatial distance from the image center. Each point denotes a patch token, with the x-axis showing distance to the center and the y-axis representing the vector norm.

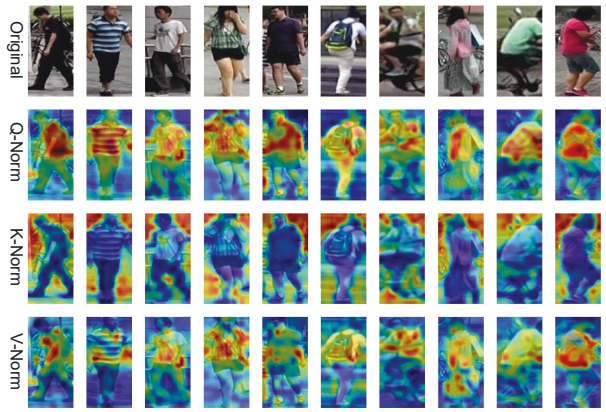


Figure 7: Visualization of Q/K/V Norm Distributions under Normal illumination

Method		Market1501			MSMT17		
Baseline	NRA	mAP	Rank-1	Rank-5	mAP	Rank-1	Rank-5
✓		93.1	96.6	98.9	75.1	89.6	94.7
✓	✓	93.1	96.8	99.0	73.4	89.1	94.4
Gain		**+0.0**	**+0.2**	**+0.1**	**-1.7**	**-0.5**	**-0.3**

Table 4: Ablation of NRA on Market1501 and MSMT17.

L2-norm distribution of Q vectors is closely aligned with pedestrian structural regions, (2) the K vector norms predominantly respond to background areas, and (3) the V vector norms exhibit a similar distribution to Q, mainly focusing on pedestrian semantic regions, as well as to evaluate the general applicability of our proposed norm ratio-based attention mechanism, we conducted experiments on two representative person re-identification datasets under normal illumination: Market1501 (Zheng et al. 2015) and MSMT17 (Wei et al. 2018). Figure 7 shows the visualization results for 10 randomly selected individuals from the Market1501 dataset. The Q, K, and V vectors consistently exhibit similar physical behaviors across both datasets, further supporting the universality of the observed patterns under different data conditions. Furthermore, we evaluated the performance of our proposed method under normal lighting conditions. As shown in Table 4, our approach improves various evaluation metrics on the Market1501 dataset, demonstrating its effectiveness in relatively stable illumination scenarios. However, on the MSMT17 dataset, a performance drop was observed across multiple metrics.

We attribute this to the differing illumination variation across datasets: Market1501 has consistent lighting, while MSMT17 shows greater fluctuation. This suggests our method performs better under stable lighting but may be less robust in highly variable illumination scenarios.

Conclusion

To better understand how existing Re-ID models operate under low-light conditions, this work conducts a systematic analysis of the norm distributions of the Q, K, and V vectors in the TransReID model. Our investigation reveals a consistent pattern in the final Transformer layer: the Q vector norms are predominantly concentrated in pedestrian structural regions, while the K vector norms tend to cluster in peripheral background areas. The V vector norms exhibit a distribution similar to that of Q, also responding to semantic pedestrian regions. Motivated by this, we propose a NRA to counteract the suppression of Q norms in low-light, enhancing semantic representation. We also introduce a SRD that transfers semantic knowledge from external domains to address illumination-induced degradation. Extensive experiments demonstrate the effectiveness and generalizability of our method. Future work will explore integrating illumination enhancement to restore regions with weakened Q norms for better semantic recovery.

Acknowledgements

This work is partially supported by the National Nature Science Foundation of China (No. 62572359, U22A2035, U1736206, U1803262), and the National Social Science Fund of China (No. 19ZDA113).

References

Abnar, S.; and Zuidema, W. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

Chefer, H.; Gur, S.; and Wolf, L. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 782–791.

Chen, S.; Ye, M.; Dong, X.; and Du, B. 2025. Perception Assisted Transformer for Unsupervised Object Re-Identification. *IEEE Transactions on Image Processing*.

Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; and Wang, Z. 2019. Abd-net: Attentive but diverse person re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8351–8361.

Cui, Z.; Zhou, J.; Wang, X.; Zhu, M.; and Peng, Y. 2024. Learning continual compatible representation for re-indexing free lifelong person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16614–16623.

Dai, W.; Lu, L.; and Li, Z. 2025. Diffusion-based Synthetic Data Generation for Visible-Infrared Person Re-Identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11185–11193.

Dong, Y.; Cordonnier, J.-B.; and Loukas, A. 2021. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, 2793–2803. PMLR.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Han, K.; Wang, Y.; Chen, H.; Chen, X.; Guo, J.; Liu, Z.; Tang, Y.; Xiao, A.; Xu, C.; Xu, Y.; et al. 2022. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1): 87–110.

He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; and Jiang, W. 2021. Transreid: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15013–15022.

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

Hu, B.; Liu, J.; Zheng, Y.; Zheng, K.; and Zha, Z.-J. 2024. Exert Diversity and Mitigate Bias: Domain Generalizable Person Re-identification with a Comprehensive Benchmark. *International Journal of Computer Vision*, 132(11): 5124–5150.

Huang, Y.; Huang, Y.; Zhang, Z.; Wu, Q.; Zhong, Y.; and Wang, L. 2025. CSFRNet: Integrating Clothing Status Awareness for Long-Term Person Re-identification. *International Journal of Computer Vision*, 133(3): :3180—-3202.

Khan, S.; Naseer, M.; Hayat, M.; Zamir, S. W.; Khan, F. S.; and Shah, M. 2022. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s): 1–41.

Kobayashi, G.; Kuribayashi, T.; Yokoi, S.; and Inui, K. 2020. Attention is not only a weight: Analyzing transformers with vector norms. *arXiv preprint arXiv:2004.10102*.

Leng, J.; Kuang, C.; Li, S.; Gan, J.; Chen, H.; and Gao, X. 2025. Dual-Space Video Person Re-identification. *International Journal of Computer Vision*, 133(6): 3667–3688.

Li, C.; Chen, S.; and Ye, M. 2024. Adaptive high-frequency transformer for diverse wildlife re-identification. In *European Conference on Computer Vision*, 296–313. Springer.

Li, H.; Chen, J.; Zheng, A.; Wu, Y.; and Luo, Y. 2024. Day-night cross-domain vehicle re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12626–12635.

Liu, M.; Bian, Y.; Liu, Q.; Wang, X.; and Wang, Y. 2024a. Weakly supervised tracklet association learning with video labels for person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5): 3595–3607.

Liu, M.; Wang, F.; Wang, X.; Wang, Y.; and Roy-Chowdhury, A. K. 2024b. A two-stage noise-tolerant paradigm for label corrupted person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7): 4944–4956.

Lu, A.; Li, C.; Zha, T.; Wang, X.; Tang, J.; and Luo, B. 2025. Nighttime Person Re-Identification via Collaborative Enhancement Network with Multi-domain Learning. *IEEE Transactions on Information Forensics and Security*.

Lu, A.; Zhang, Z.; Huang, Y.; Zhang, Y.; Li, C.; Tang, J.; and Wang, L. 2023. Illumination distillation framework for nighttime person re-identification and a new benchmark. *IEEE Transactions on Multimedia*, 26: 406–419.

Luo, H.; Gu, Y.; Liao, X.; Lai, S.; and Jiang, W. 2019. 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*, 0–0.

Luo, H.; Wang, P.; Xu, Y.; Ding, F.; Zhou, Y.; Wang, F.; Li, H.; and Jin, R. 2021. Self-supervised pre-training for transformer-based person re-identification. *arXiv preprint arXiv:2111.12084*.

Pang, Z.; Wang, J.; Zhao, L.; and Wang, C. 2025. Identity-Clothing Similarity Modeling for Unsupervised Clothing Change Person Re-Identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 19251–19260.

Paul, S.; and Chen, P.-Y. 2022. Vision transformers are robust learners. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 36, 2071–2081.

Qin, Y.; Chen, Y.; Peng, D.; Peng, X.; Zhou, J. T.; and Hu, P. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27197–27206.

Rassin, R.; Hirsch, E.; Glickman, D.; Ravfogel, S.; Goldberg, Y.; and Chechik, G. 2023. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36: 3536–3559.

Sheng, H.; Wang, S.; Chen, H.; Yang, D.; Huang, Y.; Shen, J.; and Ke, W. 2023. Discriminative feature learning with co-occurrence attention network for vehicle ReID. *IEEE transactions on circuits and systems for video technology*, 34(5): 3510–3522.

Shi, J.; Yin, X.; Chen, Y.; Zhang, Y.; Zhang, Z.; Xie, Y.; and Qu, Y. 2024. Multi-memory matching for unsupervised visible-infrared person re-identification. In *European Conference on Computer Vision*, 456–474. Springer.

Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; and Wang, S. 2018. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European conference on computer vision (ECCV)*, 480–496.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 32–42.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, P.; Zheng, W.; Chen, T.; and Wang, Z. 2022. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962*.

Wang, Y.; Cao, Y.; Zha, Z.-J.; Zhang, J.; and Xiong, Z. 2020. Deep degradation prior for low-quality image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11049–11058.

Wei, L.; Zhang, S.; Gao, W.; and Tian, Q. 2018. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 79–88.

Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; and Ouyang, W. 2018. Attention-aware compositional network for person

re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2119–2128.

Xue, X.; Zhang, C.; Niu, Z.; and Wu, X. 2022. Multi-level attention map network for multimodal sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 5105–5118.

Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; and Hoi, S. C. 2021. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6): 2872–2893.

Yeh, C.; Chen, Y.; Wu, A.; Chen, C.; Viégas, F.; and Wattenberg, M. 2023. Attentionviz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics*, 30(1): 262–272.

Zhang, J.; Yuan, Y.; and Wang, Q. 2019. Night person reidentification and a benchmark. *IEEE Access*, 7: 95496–95504.

Zhang, W.; and Xiao, C. 2019. PCAN: 3D attention map learning using contextual information for point cloud based retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12436–12445.

Zhao, B.; Cui, Q.; Song, R.; Qiu, Y.; and Liang, J. 2022. Decoupled knowledge distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11953–11962.

Zhao, Y.; Ruan, W.; Li, H.; and Ye, M. 2025. NightReID: A Large-Scale Nighttime Person Re-Identification Benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10519–10527.

Zheng, A.; Wang, Z.; Chen, Z.; Li, C.; and Tang, J. 2021. Robust multi-modality person re-identification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3529–3537.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, 1116–1124.

Zheng, N.; Huang, J.; Zhou, M.; Yang, Z.; Zhu, Q.; and Zhao, F. 2023. Learning semantic degradation-aware guidance for recognition-driven unsupervised low-light image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 3678–3686.

Zheng, Z.; Zheng, L.; and Yang, Y. 2017. A discriminatively learned cnn embedding for person reidentification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 14(1): 1–20.