

Multi-Run Statistical Evaluation of Retinex-Net, SNR-Aware, and RetinexFormer for Nighttime Vehicle Re-Identification

Muhammad Rafif^{1,*}, Radius Tanone²

¹ Informatics Engineering, Faculty of Information Technology, Universitas Kristen Satya Wacana, Salatiga, Indonesia

² Artificial Intelligence Research Center, Universitas Kristen Satya Wacana, Salatiga, Indonesia

Email: ^{1,*}672023001@student.uksw.edu, ²radius.tanone@uksw.edu

Correspondence Author Email: 672023001@student.uksw.edu

Received: 31/08/2026, Revised: 17/09/2026, Accepted: 21/09/2026, Published: 22/09/2026

Abstract—Nighttime vehicle re-identification (Re-ID) is challenging because low illumination, sensor noise, and motion blur can obscure identity-related visual features in surveillance images. Comparative studies of low-light enhancement (LLE) as a preprocessing step for Re-ID often rely on a single training run, making it difficult to distinguish genuine method effects from training stochasticity. This study evaluates Retinex-Net, SNR-Aware, and RetinexFormer as LLE preprocessing methods for TransReID with a ViT-Base/16 backbone on a stratified nighttime subset of the DN-348 dataset. The subset contains 8,255 images from 150 training identities and 120 shared gallery/query identities, with nighttime degradation simulated using gamma darkening, Poisson-Gaussian noise, and Gaussian blur. Each scenario was trained using three independent random seeds, with performance summarized across runs using mean Average Precision (mAP) and Cumulative Matching Characteristic (CMC) at Rank-1, Rank-5, and Rank-10. Additional statistical validation was performed on a representative run using bootstrap confidence intervals and paired Wilcoxon signed-rank tests with Holm-Bonferroni correction. Image quality was additionally assessed using PSNR, SSIM, and LPIPS, while SIE and JPM were examined through a factorial ablation. RetinexFormer achieved the highest mean mAP of $48.90 \pm 0.11\%$, followed by SNR-Aware at $48.69 \pm 0.21\%$ and the baseline at $48.51 \pm 0.33\%$. However, neither RetinexFormer nor SNR-Aware showed a statistically significant improvement over the baseline, whereas Retinex-Net performed significantly worse. RetinexFormer also produced the best image-quality metrics, indicating that stronger visual restoration does not necessarily yield better downstream Re-ID. These findings demonstrate the importance of multi-run and statistically validated evaluation when assessing LLE methods for nighttime vehicle Re-ID. Given that none of the evaluated LLE methods yielded a statistically significant mAP improvement over the no-enhancement baseline, these results indicate that LLE should not be adopted by default as a preprocessing step for nighttime vehicle Re-ID and should only be deployed after scenario-specific multi-run statistical validation.

Keywords: Vehicle Re-Identification; Low-Light Enhancement; Transreid; Multi-Run Evaluation; Statistical Validation

1. INTRODUCTION

Vehicle re-identification (Re-ID) is an important component of intelligent transportation and urban traffic surveillance because it matches the same vehicle across images captured by different cameras [1]. Under nighttime conditions the task becomes considerably more difficult, because limited illumination lowers the signal-to-noise ratio and obscures the visual characteristics used to establish identity, so that Re-ID performance decreases substantially compared with daytime conditions [2], with comparable low-illumination degradation also reported for nighttime person Re-ID, motivating dedicated identity-discovery mechanisms for darkness-affected imagery [3]. This is particularly relevant to high-ISO surveillance imagery, where insufficient illumination is usually accompanied by sensor noise and motion blur. Poor-quality visual inputs can also interfere with subsequent image-processing tasks, motivating the use of image-to-image translation and enhancement approaches to improve the suitability of images for further analysis [4]. A low-light enhancement (LLE) module intended to support downstream Re-ID is therefore required, rather than one that only improves the perceptual appearance of an image.

Several LLE approaches use different underlying architectures. Retinex-Net decomposes an image into reflectance and illumination components through a two-stage convolutional network [5], SNR-Aware routes high-SNR regions to a convolutional branch and low-SNR regions to a transformer branch [6], and RetinexFormer embeds Retinex theory in a single-stage transformer with illumination-guided attention [7]. All three perform strongly on general low-light enhancement benchmarks such as LOL-v1 [8]. Beyond these three supervised methods, other LLE paradigms have also been proposed, including zero-reference curve estimation [9] and unsupervised GAN-based enhancement [10], reflecting substantial architectural diversity within the field. Broader benchmarking studies further indicate that the relative ranking of LLE methods can vary considerably across datasets and evaluation protocols [11]. Recent surveys corroborate this rapid diversification of LLE research across transformer-based, diffusion-based, and lightweight architectures [12]–[15], while dedicated benchmarking efforts continue to evaluate LLE methods on increasingly realistic outdoor and street-level imagery [16]. In parallel, transformer-based vehicle Re-ID has continued to develop rapidly [17], further motivating a controlled, statistically validated comparison of LLE preprocessing within this architecture family. Nevertheless, their effectiveness as preprocessing for Vision Transformer-based vehicle Re-ID has not been sufficiently investigated under controlled nighttime conditions, and improvements in image quality do not necessarily translate into better downstream Re-ID because enhancement may alter visual characteristics that are important for identity matching.

A second issue concerns the reliability of comparative evaluations. Existing LLE-for-Re-ID comparisons commonly report a single training run per scenario, which makes it difficult to separate a genuine method effect from

stochastic variation caused by weight initialization, batch sampling order, and random augmentation. Methodological studies have shown that variation between independent training runs can be comparable to the performance differences that are sometimes interpreted as methodological improvement [18]. Related concerns regarding non-determinism and the need for significance testing when comparing learning algorithms have also been raised in other subfields of deep learning [19]. Recent methodological recommendations therefore emphasize repeated training with different random seeds and the reporting of uncertainty rather than a single performance value [20]. Recent large-scale reproducibility audits of machine learning benchmarks reinforce this concern, showing that reported performance gains are frequently confounded by unreported sources of variance such as random seeds and evaluation configuration [21].

The contribution of the Re-ID architecture itself must also be considered. TransReID incorporates Side Information Embedding (SIE) and the Jigsaw Patch Module (JPM), which are designed to improve Re-ID performance [22]. When these components are used without separate evaluation, an observed difference may originate from the architecture rather than from the LLE preprocessing method, so component-level analysis is necessary to isolate the actual contribution of image enhancement.

Two research gaps therefore follow. First, LLE preprocessing for nighttime vehicle Re-ID has not been evaluated through repeated independent training runs with statistical validation, so reported differences may not reflect the actual effect of the enhancement method. Second, the contribution of Re-ID architectural components, particularly SIE and JPM in TransReID, has not been isolated from the effect of LLE preprocessing. Addressing both gaps is necessary for a more reliable assessment of LLE methods in nighttime vehicle Re-ID.

To address them, this study conducts a controlled comparative evaluation of Retinex-Net, SNR-Aware, and RetinexFormer as preprocessing methods for nighttime vehicle Re-ID using TransReID with a ViT-Base/16 backbone. Four scenarios are evaluated under an identical protocol, consisting of a baseline without LLE and three LLE-based scenarios, each trained with three independent random seeds. The evaluation is complemented by bootstrap confidence intervals and paired Wilcoxon signed-rank tests with Holm-Bonferroni correction at the per-query level, and by an ablation of SIE and JPM on the baseline scenario.

The objective is to provide a statistically validated comparison of LLE methods for nighttime vehicle Re-ID that accounts for variation across independent training runs and for the possible contribution of Re-ID architectural components. The main contribution is therefore a multi-run and statistically validated benchmark: unlike evaluations that rely on a single training result, this study combines independent random seeds, uncertainty estimation, paired statistical testing, and multiple-comparison correction, while the explicit ablation of SIE and JPM separates the contribution of image enhancement from that of the Re-ID architecture. The findings are expected to clarify the practical effectiveness and statistical reliability of LLE preprocessing for nighttime vehicle Re-ID and to highlight the risk of overstating the benefit of an enhancement method evaluated from a single training run.

2. RESEARCH METHODOLOGY

This study uses an experimental, comparative, and controlled research design with a multi-run evaluation framework. Four preprocessing scenarios, namely a baseline without LLE, Retinex-Net, SNR-Aware, and RetinexFormer, are evaluated using the same downstream Re-ID architecture and training protocol, each trained independently with three random seeds (42, 123, and 456), resulting in twelve main training runs. In addition, twelve supplementary training runs are conducted to examine the contribution of the architectural components of TransReID. All main scenarios are trained for 60 complete epochs without early stopping so that the training budget is identical across experimental conditions. The research object of this study is the downstream performance of nighttime vehicle re-identification after low-light enhancement preprocessing. The study focuses on three LLE methods, namely Retinex-Net, SNR-Aware, and RetinexFormer, which are evaluated using TransReID with a ViT-Base/16 backbone on a stratified nighttime subset of the DN-348 dataset. The study examines whether LLE preprocessing provides a consistent and statistically supported benefit to vehicle Re-ID performance. The contribution of the Side Information Embedding (SIE) and Jigsaw Patch Module (JPM) components is additionally examined through a factorial ablation under the baseline configuration.

2.1 Research Stages

The overall research stages are presented in Figure 1. The workflow consists of dataset preparation, stratified subset construction, image degradation, LLE preprocessing, TransReID training, performance evaluation, statistical validation, and architectural ablation. A nighttime subset is first selected while maintaining sufficient image representation for each vehicle identity, and the selected images are then degraded through a three-stage pipeline that simulates nighttime surveillance conditions. The degraded images are processed under the four experimental scenarios and used as inputs for TransReID training under an identical configuration. Re-ID performance is then evaluated using mAP and CMC, after which statistical validation and component ablation are conducted. Because all four scenarios receive the same degraded images, the observed differences can be associated primarily with the preprocessing method.

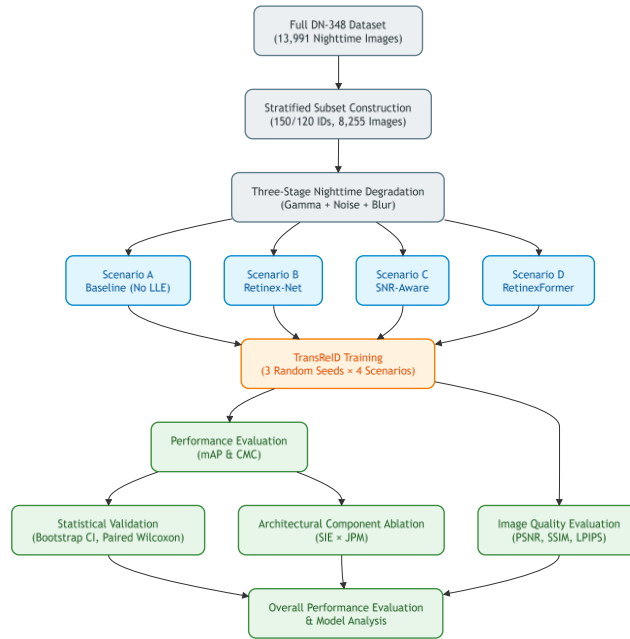


Figure 1. End-to-end experimental workflow from dataset preparation to model evaluation and analysis

2.2 Dataset Preparation and Subset Construction

This study uses the Day-Night Cross-domain Vehicle Re-identification (DN-348) dataset [23], whose file naming format follows the Market-1501 convention [24]. Unlike widely used vehicle Re-ID benchmarks such as VeRi-776, which predominantly contain daytime imagery [25], DN-348 explicitly provides nighttime data, making it suitable for the controlled nighttime evaluation conducted in this study. To reduce computational requirements while maintaining representation across vehicle identities, a subset of the nighttime images is selected using stratified sampling, as summarized in Table 1. The gallery and query sets share the same 120 identities, whereas the training identities do not overlap with the gallery or query identities.

The identities are initially selected randomly using a subsampling seed of 2026, which is kept different from the seeds used for the main training experiments. The number of images per identity is then limited to maintain a balanced representation and ensure that each identity contains sufficient samples for PK sampling during triplet-loss training. The resulting subset statistics are presented in Table 1.

Table 1. Statistics of the nighttime DN-348 subset used in the study

Split	Full Population (IDs / Images)	Selected Subset (IDs / Images)	Image Cap/ID
Train	200 / 10,022	150 / 5,961	40
Gallery	147 / 3,528	120 / 1,934	20
Query	147 / 441	120 / 360	8

2.3 Nighttime Image Degradation

To simulate the characteristics of high-ISO nighttime surveillance imagery, all selected images undergo a three-stage degradation process that is applied consistently to each image so that the experimental input can be reproduced: gamma darkening to reduce illumination, heteroscedastic Poisson-Gaussian noise to simulate high-ISO sensor noise, and Gaussian blur to represent motion-induced degradation.

The first stage applies gamma darkening to the original image according to Equation (1):

$$I_{\text{stage1}}(x, y) = \left(\frac{I(x, y)}{255} \right)^{\gamma} \times 255 \quad (1)$$

where $I(x, y)$ represents the original pixel intensity at coordinates (x, y) , and γ denotes the gamma parameter controlling the degree of darkening.

The second stage introduces heteroscedastic Poisson-Gaussian noise to the darkened image, as formulated in Equation (2):

$$I_{\text{stage2}}(x, y) = I_{\text{stage1}}(x, y) + a \cdot \sqrt{I_{\text{stage1}}(x, y)} \cdot N_1(0, 1) + b \cdot N_2(0, 1) \quad (2)$$

where a represents the scale of the Poisson component, b represents the standard deviation of the Gaussian component, and $N_1(0, 1)$ and $N_2(0, 1)$ denote independent standard normal random variables.

The final stage applies Gaussian blur to the noisy image, as shown in Equation (3):

$$I_{\text{stage3}}(x, y) = I_{\text{stage2}}(x, y) * G_{\sigma}(x, y) \quad (3)$$

where G_{σ} denotes a two-dimensional Gaussian kernel with standard deviation σ , and $*$ represents the convolution operation.

For each image, the gamma parameter γ is randomly selected from the set $\{2.0, 2.5, 3.0\}$. The Poisson scale parameter is set to $a = 0.03$, while the Gaussian standard deviation is set to $b = 0.012$. The Gaussian blur parameter σ is randomly selected from the interval $[0.5, 1.0]$. These parameters are applied consistently according to the predefined degradation procedure to produce nighttime images with reduced illumination, sensor noise, and motion blur. The resulting degraded images are then used as identical inputs for the four experimental scenarios, namely the baseline, Retinex-Net, SNR-Aware, and RetinexFormer.

2.4 Low-Light Enhancement Methods

Three LLE methods are evaluated as preprocessing modules: Retinex-Net [4], SNR-Aware [5], and RetinexFormer [6]. Each method uses its official pretrained weights without additional retraining, so that the comparison reflects the original enhancement capability of each method. Retinex-Net decomposes an input image into reflectance and illumination maps using DecomNet and enhances the illumination component with RelightNet; SNR-Aware estimates the signal-to-noise ratio from the luminance component and directs high-SNR regions to a convolutional branch and low-SNR regions to a transformer branch with masked attention; RetinexFormer applies illumination-guided multi-head self-attention within a single-stage architecture. The pretrained weights and principal configurations are summarized in Table 2.

Table 2. Configurations of the three low-light enhancement methods

Method	Pretrained Weights	Main Configuration
Retinex-Net	9200.tar (Decom + Relight)	DecomNet + RelightNet, two-stage CNN
SNR-Aware	LOLv1.pth	$n_f = 64$, CNN + Transformer branches
RetinexFormer	LOL_v1.pth	$n_{feat} = 40$, stage = 1, num_blocks = [1, 2, 2]

2.5 TransReID Configuration and Training Procedure

The downstream Re-ID model is based on TransReID [22] using a Vision Transformer ViT-Base/16 backbone [26] pretrained on ImageNet-21k, with an input resolution of 256×128 pixels, SIE enabled for two cameras, JPM activated during training with $K = 4$, and a BNNeck layer before the classifier. The training objective combines cross-entropy loss with label smoothing and hard-mined triplet loss. Optimization uses AdamW with a layer-wise learning-rate decay of 0.9, a 10-epoch warm-up followed by cosine decay, weight decay of 0.05, and stochastic depth with a drop-path rate of 0.1. The batch size is 96, consisting of 24 identities with four instances per identity, and the model is trained for 60 epochs without early stopping using BF16 mixed precision on an NVIDIA A100-SXM4 GPU with 40 GB of VRAM. Random horizontal flipping, random cropping, and random erasing are applied during training, while flip test-time augmentation is used during inference. The complete training configuration is presented in Table 3.

Table 3. Training hyperparameters of the TransReID model

Component	Value
Backbone	ViT-Base/16 (pretrained on ImageNet-21k)
Input image size	256×128
SIE / JPM	Enabled, 2 cameras / Enabled, $K = 4$
Number of epochs	60 (full training, without early stopping)
Batch size	96 (24 identities $\times$ 4 instances)
Optimizer	AdamW + layer-wise LR decay (0.9)
Peak learning rate	3.5×10^{-5} (10-epoch warm-up + cosine decay)
Weight decay	0.05
Drop path	0.1
Loss function	Cross-entropy (label smoothing = 0.1) + triplet loss (margin = 0.3)
Training precision	BF16 (NVIDIA A100, without gradient scaling)
Multi-run seeds	42, 123, 456 (3 runs/scenario)

2.6 Performance Evaluation and Statistical Validation

Re-ID performance is evaluated according to the Market-1501 evaluation protocol [24]. The primary metric is mean Average Precision (mAP), complemented by the Cumulative Matching Characteristic (CMC) at Rank-1, Rank-5, and Rank-10. Feature vectors are L2-normalized before calculating the cosine distance between query and gallery representations, and all metrics are calculated on the 360 query images of the selected nighttime subset.

In addition, the quality of the enhanced images is evaluated using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS) [27] relative to

the corresponding clean images [7]. These image-level metrics are calculated on the query subset to examine the relationship between enhancement quality and downstream Re-ID performance.

For the multi-run evaluation, performance is reported as the mean and standard deviation obtained from the three independent seeds of each scenario. A representative run using seed 42 is analyzed further: a 95% bootstrap confidence interval is calculated at the query level [28], followed by an examination of confidence-interval overlap between scenarios. Paired Wilcoxon signed-rank tests are then applied to per-query Average Precision values for all six pairwise comparisons, with Holm-Bonferroni correction [29], and rank-biserial correlation as the effect size.

2.7 Architectural Component Ablation

An additional ablation experiment is conducted to examine the individual contribution of SIE and JPM independently of LLE preprocessing. The ablation uses the baseline scenario without LLE and follows a 2×2 factorial configuration consisting of SIE enabled or disabled and JPM enabled or disabled. To maintain consistency with the multi-run statistical framework applied to every other scenario in this study, the ablation is trained using the same three independent seeds (42, 123, and 456) used for the main experiments, rather than a single seed. Performance for each ablation variant is reported as the mean and standard deviation of mAP and CMC across the three seeds, following the identical procedure used for the main Re-ID results in Section 2.6.

The ablation variants are trained using the same 60-epoch budget and evaluated under the same Re-ID protocol as the main experiments, so that the contribution of the TransReID architectural components can be examined separately from the effects of image enhancement.

2.8 Computational Environment and Reproducibility

All experiments are performed on an NVIDIA A100-SXM4 GPU with 40 GB of VRAM using BF16 mixed precision and deterministic execution settings. Computational efficiency is evaluated by measuring the number of parameters and the inference latency of TransReID and each LLE component at an input resolution of 256×128 pixels, over 50 iterations after 10 warm-up iterations, with CUDA synchronization at every iteration. Fixed configurations, independent seeds, identical training budgets, and deterministic settings are used to improve the reproducibility of the results.

3. RESULTS AND DISCUSSION

3.1 Degradation Validation

Before evaluating the low-light enhancement (LLE) methods, the degradation procedure was examined. The three-stage process reduced the mean pixel intensity from 54.6 to 17.1 in the training set, from 56.9 to 17.8 in the gallery set, and from 60.9 to 19.3 in the query set, corresponding to reductions of 68.8%, 68.8%, and 68.4% respectively. The gamma parameters were distributed evenly across the 8,255 selected images, with 32.9% at $\gamma = 2.0$, 32.8% at $\gamma = 2.5$, and 34.3% at $\gamma = 3.0$, so that the evaluation is not dominated by a single degradation level. More importantly, the same degraded images were provided to the baseline, Retinex-Net, SNR-Aware, and RetinexFormer, which ensures that the subsequent differences in Re-ID performance are associated with the preprocessing strategy rather than with differences in the input degradation.

3.2 Main Re-ID Performance Across Independent Runs

Table 4 presents the Re-ID performance obtained from three independent training runs using random seeds 42, 123, and 456. RetinexFormer achieved the highest mean mAP of $48.90 \pm 0.11\%$, followed by SNR-Aware at $48.69 \pm 0.21\%$, the baseline at $48.51 \pm 0.33\%$, and Retinex-Net at $45.65 \pm 0.51\%$. The improvement of RetinexFormer over the baseline is small, whereas Retinex-Net shows a considerably larger decrease, indicating that the three LLE methods do not have equivalent effects on downstream vehicle Re-ID.

Table 4. Re-ID Performance Across Three Independent Training Runs

Scenario	mAP (%)	Rank-1 (%)	Rank-5 (%)	Rank-10 (%)
Baseline	48.51 ± 0.33	77.41 ± 1.25	89.54 ± 0.32	92.50 ± 0.28
Retinex-Net	45.65 ± 0.51	73.24 ± 1.37	87.04 ± 0.58	91.20 ± 1.40
SNR-Aware	48.69 ± 0.21	77.22 ± 0.73	89.54 ± 0.85	92.31 ± 0.80
RetinexFormer	48.90 ± 0.11	74.72 ± 1.00	88.80 ± 1.28	91.76 ± 1.37

The pattern differs for Rank-1, where the baseline is highest at $77.41 \pm 1.25\%$, followed by SNR-Aware at $77.22 \pm 0.73\%$, RetinexFormer at $74.72 \pm 1.00\%$, and Retinex-Net at $73.24 \pm 1.37\%$. The method with the highest mAP therefore does not achieve the highest top-ranked accuracy: RetinexFormer improves the aggregate mAP but lowers Rank-1, so both mAP and CMC-based metrics are required to interpret the downstream effect of LLE. The magnitude of the differences is also relevant, because RetinexFormer improves the mean mAP by less than one percentage point relative to the baseline while Retinex-Net decreases it by several percentage points, which suggests

that Retinex-Net affects the identity representation more strongly, whereas SNR-Aware and RetinexFormer remain close to the baseline.

To provide a more direct comparison with the baseline, the differences in mAP and Rank-1 across the three LLE scenarios are summarized in Table 5.

Table 5. Differences in mAP (Δ) and Rank-1 Across Scenarios Relative to the Baseline (Calculated from the Multi-Run Mean)

Scenario	Δ mAP (pp)	Δ Rank-1 (pp)
B – Retinex-Net	–2.86	–4.17
C – SNR-Aware	+0.18	–0.19
D – RetinexFormer	+0.39	–2.69

As shown in Table 5, Retinex-Net produces the largest decrease in performance, with a reduction of 2.86 percentage points in mAP and 4.17 percentage points in Rank-1 relative to the baseline. SNR-Aware produces a small increase of 0.18 percentage points in mAP with an almost unchanged Rank-1 of –0.19 percentage points, whereas RetinexFormer produces the largest mAP improvement of 0.39 percentage points but loses 2.69 percentage points in Rank-1. The effect of LLE therefore depends on the evaluation metric: the numerical gain of RetinexFormer is limited to mAP and does not extend to top-ranked retrieval, while Retinex-Net declines consistently in both metrics.

3.3 Confidence-Interval and Statistical Analysis

To determine whether the observed differences were sufficiently robust, the uncertainty of the results was evaluated using 95% bootstrap confidence intervals. Figure 2 presents the mAP values and their corresponding intervals. The intervals for the baseline, SNR-Aware, and RetinexFormer overlap substantially, indicating that their numerical differences are small relative to the uncertainty of the evaluation, whereas Retinex-Net is clearly separated from the other scenarios, which is consistent with its substantially lower mAP in the multi-run results.

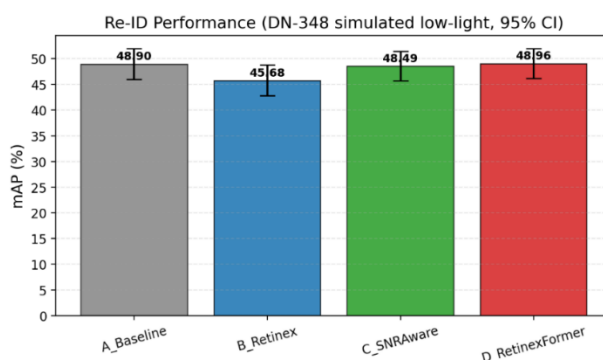


Figure 2. Mean Average Precision with 95% Bootstrap Confidence Intervals Across Experimental Scenarios

The confidence intervals therefore qualify the numerical ranking obtained from Table 4. Although RetinexFormer achieves the highest mean mAP, its overlap with the baseline indicates that the 0.39-percentage-point advantage should not be interpreted as a meaningful improvement, and SNR-Aware shows an even smaller separation from the baseline.

The paired Wilcoxon signed-rank test was then applied to the per-query Average Precision values across the six possible scenario pairs. Holm-Bonferroni correction was used to account for multiple comparisons, while the rank-biserial correlation was used to describe the effect size. The statistical analysis shows that the comparisons involving Retinex-Net are statistically significant, whereas the comparisons among the baseline, SNR-Aware, and RetinexFormer are not statistically significant.

Table 6. Pairwise Statistical Comparison of Per-Query Average Precision

Comparison	Median AP Difference	Holm-Adjusted p-value	Rank-Biserial Correlation	Significance
Baseline vs. Retinex-Net	–0.0165	3.4×10^{-6}	–0.304	Significant
Baseline vs. SNR-Aware	–0.0006	1.000	–0.039	Not significant
Baseline vs. RetinexFormer	–0.0023	1.000	–0.001	Not significant
Retinex-Net vs. SNR-Aware	+0.0151	6.3×10^{-6}	+0.293	Significant
Retinex-Net vs. RetinexFormer	+0.0243	3.0×10^{-6}	+0.307	Significant
SNR-Aware vs. RetinexFormer	0.0000	1.000	+0.016	Not significant

The comparison between the baseline and RetinexFormer is particularly important because RetinexFormer has the highest numerical mAP. It produces a median AP difference of –0.0023, a Holm-adjusted p-value of 1.000, and a rank-biserial correlation of –0.001, indicating a negligible effect, and the same interpretation applies to SNR-Aware.

These findings directly address the methodological gap identified in the Introduction, because a comparison based on a single training result could have presented the highest numerical mAP of RetinexFormer as evidence of superiority, whereas independent training runs, confidence intervals, and paired statistical testing show that this advantage is not statistically reliable.

3.4 Image Quality and Its Relationship with Re-ID

To examine whether improvements in visual quality correspond to improvements in downstream Re-ID performance, the enhanced images were evaluated using PSNR, SSIM, and LPIPS on the 360 query images, using the corresponding clean DN-348 images as references. Table 7 summarizes the results, where higher PSNR and SSIM values and lower LPIPS values indicate greater similarity to the reference images.

Table 7. Image Quality of LLE Results on the Query Subset ($n = 360$) Using Clean DN-348 Images as References

Method	PSNR (dB) $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Retinex-Net	14.65 ± 1.76	0.5332 ± 0.1445	0.3302 ± 0.1034
SNR-Aware	17.67 ± 3.39	0.6908 ± 0.1346	0.2019 ± 0.0959
RetinexFormer	20.02 ± 3.50	0.7317 ± 0.1097	0.1870 ± 0.0779

As shown in Table 7, RetinexFormer achieves the best values for all three metrics, SNR-Aware ranks second, and Retinex-Net is the weakest. The image-quality ranking is therefore clear, but it does not correspond to the downstream results, because RetinexFormer improves the mean mAP by only 0.39 percentage points and the difference is not statistically significant. An image can thus become more similar to its clean reference without providing additional identity-discriminative information to the Re-ID model. Retinex-Net shows a different pattern, since its poor reconstruction quality is accompanied by a substantial decrease in Re-ID performance, which indicates that severe degradation of the reconstructed image does affect identity-related features. Taken together, PSNR, SSIM, and LPIPS should not be interpreted as direct predictors of vehicle identification performance, and downstream task evaluation is required alongside conventional image-quality metrics.

3.5 Ablation of SIE and JPM

In addition to evaluating the LLE methods, an ablation experiment was conducted to examine the contribution of Side Information Embedding (SIE) and the Jigsaw Patch Module (JPM) in the baseline TransReID configuration. The experiment used a 2×2 factorial design. Following the revised protocol in Section 2.7, all four variants are trained using the same three independent seeds (42, 123, 456) used in the main experiments, and results are reported as mean $\pm$ standard deviation. The complete configuration with both SIE and JPM achieved a mean mAP of $48.51 \pm 0.33\%$ (identical to the A_Baseline scenario reported in Table 4), while removing SIE produced $48.86 \pm 0.69\%$, removing JPM produced $48.87 \pm 0.55\%$, and removing both components produced $48.71 \pm 0.17\%$.

Table 8. Factorial Ablation Results of SIE $\times$ JPM on the Baseline Scenario (Mean $\pm$ SD, $N = 3$ Seeds)

Variant	SIE	JPM	mAP (%)	Δ mAP Relative to Full (pp)
Full (SIE + JPM)	Enabled	Enabled	48.51 ± 0.33	0.00
Without SIE	Disabled	Enabled	48.86 ± 0.69	+0.35
Without JPM	Enabled	Disabled	48.87 ± 0.55	+0.36
Without SIE & JPM	Disabled	Disabled	48.71 ± 0.17	+0.20

SIE incorporates camera-related information into the Re-ID representation. In the selected subset, the training and gallery images originate from one camera, whereas the query images originate from another camera, which may limit the usefulness of camera-specific information during inference. With the ablation now repeated across three independent seeds, the mAP change associated with removing SIE (+0.35 pp) and JPM (+0.36 pp) is smaller than, or comparable to, the corresponding across-seed standard deviations (0.69 pp and 0.55 pp, respectively). This indicates that neither component produces a change that is clearly distinguishable from ordinary run-to-run training variability, so these results should be interpreted as showing that SIE and JPM have, at most, a small and statistically inconclusive individual effect on this baseline configuration, rather than as evidence of a genuine superiority or inferiority of either component.

The ablation nevertheless matters for interpreting the LLE comparison, because it shows that the downstream architecture can itself influence Re-ID performance. Without this examination, small differences between preprocessing scenarios could be attributed entirely to LLE, so the architectural configuration should be considered when interpreting the effect of image enhancement.

3.6 CMC and Per-Query Performance Analysis

Figure 3 further examines the difference between the baseline and RetinexFormer at the individual-query level. Across the 360 query images, RetinexFormer achieves higher Average Precision for 170 queries, while the baseline performs better for 188 queries, with identical AP values for the remaining two queries. This relatively balanced distribution indicates that the improvement obtained by RetinexFormer is not consistent across individual vehicle instances.

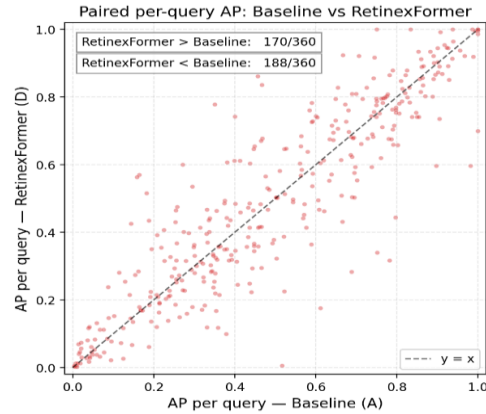


Figure 3. Per-Query Average Precision Comparison Between Baseline and RetinexFormer

The per-query distribution explains why the overall mAP difference remains small: some vehicle instances benefit from the enhancement while others perform better without it, so an aggregate improvement in mAP does not imply that retrieval improves for every vehicle identity.

The retrieval behavior across different ranks is illustrated by the Cumulative Matching Characteristic (CMC) curves in Figure 4. Retinex-Net remains below the other scenarios at the lower retrieval ranks, particularly from Rank-1 to approximately Rank-9, whereas the baseline, SNR-Aware, and RetinexFormer remain relatively close, although their relative positions vary across ranks.

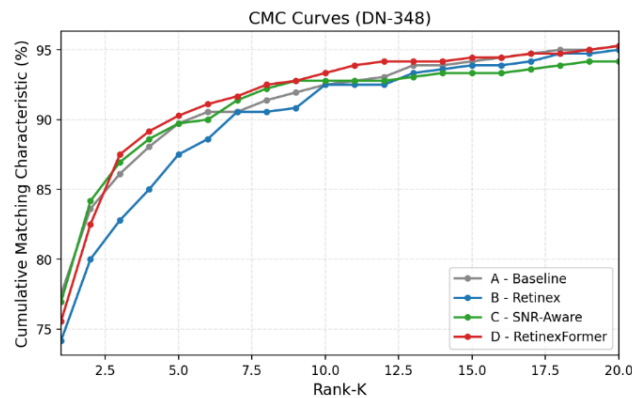


Figure 4. Cumulative Matching Characteristic (CMC) Curves from Rank-1 to Rank-20 on the DN-348 Nighttime Subset

To further quantify the query-level effects of each LLE method, the changes in Average Precision relative to the baseline were categorized as improved, stable, or degraded using an absolute AP difference threshold of 0.10. The resulting distribution is presented in Table 9.

Table 9. Per-Query Analysis of Each LLE Method Relative to the Baseline ($|\Delta AP| = 0.10$ Threshold)

LLE Method	Improved (%)	Stable (%)	Degraded (%)
Retinex-Net	10.3	68.1	21.7
SNR-Aware	18.3	63.1	18.6
RetinexFormer	16.9	67.5	15.6

Retinex-Net improves only 10.3% of the queries while degrading 21.7%, the highest proportion of degraded queries among the three methods. SNR-Aware improves 18.3% and degrades 18.6%, whereas RetinexFormer improves 16.9% and degrades 15.6%. Neither distribution results in a statistically significant overall improvement, which indicates that the positive effects of LLE are concentrated in particular query instances rather than distributed across the dataset.

3.7 Computational Efficiency

The computational cost of applying LLE before Re-ID inference was also examined. As shown in Table 10, TransReID with the ViT-B/16 backbone contains 85.90 million parameters and requires an inference latency of 6.62 ± 0.34 ms per image. Retinex-Net is the lightest LLE method with 0.56 million parameters and 2.40 ± 0.19 ms per image, SNR-Aware contains 39.12 million parameters and requires 7.77 ± 0.38 ms, and RetinexFormer contains only 1.61 million parameters but exhibits the highest latency at 9.66 ± 0.20 ms per image.

Table 10. Computational Efficiency of the Evaluated Models

Component	Parameters (Million)	Inference Latency (ms/image)
TransReID (ViT-B/16)	85.90	6.62 ± 0.34
Retinex-Net	0.56	2.40 ± 0.19
SNR-Aware	39.12	7.77 ± 0.38
RetinexFormer	1.61	9.66 ± 0.20

Parameter count therefore does not directly determine inference latency, since RetinexFormer has considerably fewer parameters than SNR-Aware but is slower, indicating that computational cost also depends on the operations performed within the network. Because LLE is applied before the Re-ID model, this cost is added to the overall inference pipeline, so the selection of an enhancement method should consider image quality, Re-ID accuracy, and computational requirements jointly. Retinex-Net is advantageous when efficiency is prioritized but produces the weakest downstream performance, whereas RetinexFormer provides stronger restoration at a higher computational cost; because its Re-ID improvement is not statistically significant, its additional latency should be justified by application requirements rather than by mAP improvement alone.

Overall, the effectiveness of LLE for nighttime vehicle Re-ID cannot be determined from the highest numerical mAP alone. RetinexFormer obtains the highest mean mAP and the strongest image-quality results, yet its improvement over the baseline is only 0.39 percentage points and is not statistically significant, and the per-query analysis shows the baseline performing better for 188 queries against 170 for RetinexFormer. SNR-Aware remains close to the baseline, whereas Retinex-Net is clearly inferior, decreasing mAP by 2.86 and Rank-1 by 4.17 percentage points with the weakest image quality, and the statistical analysis confirms that the differences involving Retinex-Net are significant. Visual enhancement and identity retrieval are therefore related but distinct objectives.

The multi-run design is the main methodological contribution of this study, because a single-run comparison could have concluded that RetinexFormer was superior on the basis of its highest numerical mAP, whereas independent training runs, confidence intervals, and paired statistical testing show that this apparent advantage is not statistically supported. The ablation analysis further demonstrates that the downstream Re-ID architecture can be examined independently of LLE preprocessing. With the ablation now trained across the same three seeds used throughout this study, the mAP changes observed after disabling SIE and JPM (+0.35 pp and +0.36 pp, respectively) are of a similar magnitude to the corresponding across-seed standard deviations, indicating that neither component produces an effect that is clearly distinguishable from ordinary training variability on this baseline configuration. This result reinforces, rather than merely supports, the central argument of this study: without a multi-run evaluation design, a single-seed estimate could easily be mistaken for a genuine architectural effect when it in fact falls within normal run-to-run noise. This finding is consistent with the broader objective of separating the effect of image enhancement from other sources of variation in the experimental pipeline.

Taken together, LLE should not be regarded as an automatic performance-enhancing component for nighttime vehicle Re-ID, and an LLE method should be selected through a task-oriented evaluation that considers downstream Re-ID performance, statistical reliability, image quality, architectural interaction, and computational efficiency.

3.8 Discussion

The findings are compared with the previous studies reviewed in Section 1. The comparison is contextual rather than a direct performance ranking, because those studies differ in dataset, degradation assumption, downstream task, and evaluation protocol. Table 11 summarizes the comparison.

Table 11. Comparison Between the Findings of This Study and Those Reported in Previous Studies

Previous Study	Focus and Evaluation Protocol	Main Emphasis of the Previous Study	Position of the Present Findings
Zhao et al. [2]; Yuan et al. [3]	Nighttime person Re-ID benchmarks and identity-oriented architectures, generally reported from single training runs	Nighttime degradation substantially reduces identity discriminability	Generic LLE preprocessing does not recover this loss for vehicles, which supports identity-aware design rather than off-the-shelf enhancement
Wei et al. [4]	Retinex-Net; two-stage Retinex decomposition evaluated with image-level quality metrics	Improved illumination restoration over earlier Retinex-based approaches	Weakest restoration here (PSNR 14.65 dB) and the only method with a significant mAP decrease (−2.86 pp; $p = 3.4 \times 10^{-6}$)
Xu et al. [5]	SNR-Aware; SNR-guided CNN-transformer enhancement evaluated on image-quality benchmarks	Competitive restoration through region-adaptive processing	Intermediate restoration (PSNR 17.67 dB) but only +0.18 pp mAP, which is not significant ($p = 1.000$)

Cai et al. [6]	RetinexFormer; one-stage illumination-guided transformer evaluated on image-quality benchmarks	Leading restoration quality among Retinex-based methods	Best image quality (PSNR 20.02 dB, SSIM 0.7317, LPIPS 0.1870) but only +0.39 pp mAP and –2.69 pp Rank-1, without significance
Rasheed et al. [7]; Liu et al. [10]; Islam et al. [15]	Benchmarking of LLE methods, dominated by image-level quality assessment	Method ranking shifts across datasets and protocols; downstream evaluation is needed	Supplies a downstream case for vehicle Re-ID in which the image-quality ranking and the Re-ID ranking do not coincide
Yu et al. [16]; Li et al. [22]	Transformer-based vehicle Re-ID and day-night cross-domain vehicle Re-ID on DN-348	Cross-domain learning and feature design as the principal solution	Uses the same nighttime data but isolates preprocessing as one controlled factor with three seeds and per-query testing
Bouthillier et al. [17]; Henderson et al. [18]; Herrmann et al. [19]; Hochlehnert et al. [20]	Methodological studies of training variance, non-determinism, and reproducibility	Repeated runs, uncertainty reporting, and significance testing are required	Applies the protocol to nighttime vehicle Re-ID and shows that sub-one-point differences are indistinguishable from run-to-run noise
He et al. [21]	TransReID with SIE and JPM, reported from single runs on large-scale benchmarks	SIE and JPM improve Re-ID accuracy	SIE (+0.35 pp) and JPM (+0.36 pp) effects lie within the across-seed standard deviations (0.69 and 0.55 pp) on this subset
This study	Three LLE methods as preprocessing for TransReID (ViT-Base/16) on a nighttime DN-348 subset; three seeds, bootstrap CI, Wilcoxon with Holm-Bonferroni correction	Highest mean mAP of $48.90 \pm 0.11\%$ (RetinexFormer) against $48.51 \pm 0.33\%$ for the baseline	No LLE method provides a statistically significant improvement over the no-enhancement baseline, and the best-restoring method is not the best Re-ID method

Compared with previous work, the present findings agree at the image level but diverge at the task level. Nighttime Re-ID studies [2], [3] report a substantial loss of identity information under low illumination and respond with identity-oriented architectures rather than with generic enhancement. The image-quality ordering reported in the original publications of Retinex-Net [5], SNR-Aware [6], and RetinexFormer [7] is reproduced here, with PSNR values of 14.65, 17.67, and 20.02 dB respectively, although the absolute values are lower than those reported on benchmarks such as LOL-v1 [8] because the images are degraded by a three-stage nighttime pipeline and the pretrained weights are used without retraining. Benchmarking studies of low-light enhancement [8], [11], [16] have argued that method ranking shifts across datasets and protocols and that image-level metrics alone are insufficient. This study supplies the corresponding task-level evidence: RetinexFormer is the best method according to PSNR, SSIM, and LPIPS, yet it improves the mean mAP by only 0.39 percentage points, reduces Rank-1 by 2.69 percentage points, and returns a Holm-adjusted p-value of 1.000 against the baseline. Transformer-based vehicle Re-ID [17] and day-night cross-domain vehicle Re-ID on DN-348 [23] address nighttime degradation through feature design and cross-domain learning; the present results explain empirically why an enhancement-only pipeline is a weaker alternative, because RetinexFormer improves 16.9% of the queries while degrading 15.6%.

The results also provide a domain-specific instance of the methodological concerns raised in earlier work on benchmark variance [18], non-determinism [19], empirical practice in machine learning [20], and reproducibility [21], since the across-seed standard deviation of the baseline is 0.33 percentage points of mAP, which already exceeds the 0.18-point advantage of SNR-Aware and is comparable to the 0.39-point advantage of RetinexFormer. For the same reason the ablation does not contradict TransReID [22], whose SIE and JPM gains were reported from single runs on large-scale benchmarks, because on this nighttime subset the corresponding changes of +0.35 and +0.36 percentage points lie within the across-seed standard deviations of 0.69 and 0.55 percentage points. The comparison is nevertheless bounded by the supervised scope of the evaluated methods, since zero-reference [9], unsupervised generative [10], and diffusion-based [30] paradigms are not covered. The distinct contribution of this work is therefore not a new enhancement method but statistically validated evidence that, for nighttime vehicle Re-ID, a higher restoration quality does not imply a higher identification accuracy.

4. CONCLUSION

This study evaluated three low-light enhancement (LLE) methods, namely Retinex-Net, SNR-Aware, and RetinexFormer, as preprocessing approaches for nighttime vehicle re-identification using a controlled multi-run

experimental design. RetinexFormer achieved the highest mean mAP of $48.90 \pm 0.11\%$ across three independent seeds, followed by SNR-Aware at $48.69 \pm 0.21\%$, the baseline at $48.51 \pm 0.33\%$, and Retinex-Net at $45.65 \pm 0.51\%$; however, neither RetinexFormer nor SNR-Aware provided a statistically significant improvement over the baseline, whereas Retinex-Net performed significantly worse, which demonstrates that a higher point estimate obtained from a single or limited number of training runs should not be interpreted as evidence of genuine improvement without considering training variability and statistical uncertainty. Image quality did not translate into downstream identification either, because RetinexFormer consistently achieved the best PSNR, SSIM, and LPIPS results while improving retrieval for only 16.9% of the queries and degrading 15.6%, and SNR-Aware improved 18.3% while degrading 18.6%, indicating that LLE can benefit individual vehicle queries while simultaneously reducing the discriminative information required by others and making image-level quality metrics insufficient as standalone indicators of Re-ID effectiveness. The ablation analysis of Side Information Embedding (SIE) and the Jigsaw Patch Module (JPM), repeated across the same three independent seeds used for the main experiments, shows that removing SIE changes the mean mAP from $48.51 \pm 0.33\%$ to $48.86 \pm 0.69\%$, removing JPM changes it to $48.87 \pm 0.55\%$, and removing both components changes it to $48.71 \pm 0.17\%$; these differences are comparable to, or smaller than, the corresponding across-seed standard deviations, indicating that SIE and JPM have, at most, a small and statistically inconclusive individual effect on this baseline configuration. From a practical perspective, Retinex-Net provides the lowest parameter count and inference latency but also the weakest downstream performance, whereas RetinexFormer provides stronger image restoration at a higher computational cost, so the selection of an LLE method should consider downstream recognition performance, statistical reliability, image quality, and computational efficiency jointly. Given that none of the evaluated LLE methods produced a statistically significant improvement over the no-enhancement baseline, this study recommends that future researchers adopt, as a minimum evaluation protocol for nighttime vehicle Re-ID and comparable settings, at least three independent training runs per scenario with explicit seed control, mean $\pm$ standard deviation reporting across runs, a bootstrap or paired non-parametric significance test with correction for multiple comparisons, and architectural ablation studies trained under the same multi-run protocol as the main experiments, since a single-run or single-seed estimate at any stage of the pipeline may be statistically indistinguishable from ordinary training variability. The findings are limited by the use of a selected subset of 8,255 images from the 13,991 images available in the nighttime population, a single Re-ID backbone, and one vehicle Re-ID dataset, and by a camera configuration that constrains the general interpretation of SIE. Future research should therefore validate the findings on the full DN-348 dataset and additional nighttime vehicle Re-ID benchmarks, investigate end-to-end joint training between LLE and Re-ID, and further evaluate recent zero-reference and diffusion-based enhancement methods to provide broader evidence regarding the relationship between low-light restoration and downstream vehicle re-identification.

5. DECLARATIONS

5.1 Author Contributions

Conceptualization: M.R. and R.T.;

Methodology: M.R. and R.T.;

Software: M.R.;

Validation: M.R. and R.T.;

Formal Analysis: M.R. and R.T.;

Investigation: M.R.;

Resources: M.R. and R.T.;

Data Curation: M.R.;

Writing Original Draft Preparation: M.R.;

Writing Review and Editing: M.R. and R.T.;

Visualization: M.R.;

Supervision: R.T.;

All authors have read and agreed to the published version of the manuscript.

5.2 Data Availability

The DN-348 dataset used in this study is publicly available from its original authors. The stratified nighttime subset, the degradation configuration, and the evaluation scripts generated during this study are available on request from the corresponding author.

5.3 Funding

No specific grant from governmental, private, or nonprofit funding organisations was received for this study..

5.4 Institutional Review Board Statement

Not applicable.

5.5 Informed Consent Statement

Not applicable.

5.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.7 Generative AI and AI-assisted Technologies in The Writing Process

The authors used Generative AI to improve the writing clarity and language of this paper. The authors reviewed and edited the AI-assisted content and take full responsibility for the content of the final publication.

REFERENCES

- [1] A. Amiri, A. Kaya, and A. S. Kececi, "A Comprehensive Survey on Deep-Learning-Based Vehicle Re-Identification: Models, Data Sets, and Challenges," *arXiv preprint arXiv:2401.10643*, 2024, doi: 10.48550/arXiv.2401.10643.
- [2] Y. Zhao, W. Ruan, H. Li, and M. Ye, "NightReID: A Large-Scale Nighttime Person Re-Identification Benchmark," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 10, pp. 10519–10527, 2025, doi: 10.1609/aaai.v39i10.33142.
- [3] X. Yuan, Y. He, and G. Hao, "Identity Hides in Darkness: Learning Feature Discovery Transformer for Nighttime Person Re-Identification," *Sensors*, vol. 25, no. 3, Art. no. 862, 2025, doi: 10.3390/s25030862.
- [4] R. Tanone, "Conditional GAN Approach for Image to Image Translation on Potato Leaf," *International Journal of Information Technology and Business*, vol. 6, no. 1, pp. 17–20, 2023, doi: 10.24246/ijiteb.612023.17-20.
- [5] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep Retinex Decomposition for Low-Light Enhancement," in *British Machine Vision Conference (BMVC)*, 2018, doi: 10.48550/arXiv.1808.04560.
- [6] X. Xu, R. Wang, C.-W. Fu, and J. Jia, "SNR-Aware Low-Light Image Enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 17693–17703, 2022, doi: 10.1109/CVPR52688.2022.01719.
- [7] Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang, "Retinexformer: One-Stage Retinex-Based Transformer for Low-Light Image Enhancement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 12470–12479, 2023, doi: 10.48550/arXiv.2303.06705.
- [8] M. T. Rasheed, D. Shi, and H. Khan, "A Comprehensive Experiment-Based Review of Low-Light Image Enhancement Methods and Benchmarking Low-Light Image Quality Assessment," *Signal Process.*, vol. 204, Art. no. 108821, 2023, doi: 10.1016/j.sigpro.2022.108821.
- [9] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1780–1789, 2020, doi: 10.48550/arXiv.2001.06826.
- [10] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "EnlightenGAN: Deep Light Enhancement Without Paired Supervision," *IEEE Trans. Image Process.*, vol. 30, pp. 2340–2349, 2021, doi: 10.1109/TIP.2021.3051462.
- [11] J. Liu, D. Xu, W. Yang, M. Fan, and H. Huang, "Benchmarking Low-Light Image Enhancement and Beyond," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1153–1184, 2021, doi: 10.1007/s11263-020-01418-8.
- [12] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-Light Image and Video Enhancement Using Deep Learning: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9396–9416, 2022, doi: 10.1109/TPAMI.2021.3126387.
- [13] Z. Tian, P. Qu, J. Li, Y. Sun, G. Li, Z. Liang, and W. Zhang, "A Survey of Deep Learning-Based Low-Light Image Enhancement," *Sensors*, vol. 23, no. 18, Art. no. 7763, 2023, doi: 10.3390/s23187763.
- [14] J. Guo, J. Ma, Á. F. García-Fernández, Y. Zhang, and H. Liang, "A Survey on Image Enhancement for Low-Light Images," *Heliyon*, vol. 9, no. 4, Art. no. e14558, 2023, doi: 10.1016/j.heliyon.2023.e14558.
- [15] J. Ye, C. Qiu, and Z. Zhang, "A Survey on Learning-Based Low-Light Image and Video Enhancement," *Displays*, vol. 81, Art. no. 102614, 2024, doi: 10.1016/j.displa.2023.102614.
- [16] M. T. Islam, I. Alam, S. S. Woo, S. Anwar, I. K. H. Lee, and K. Muhammad, "LoLI-Street: Benchmarking Low-Light Image Enhancement and Beyond," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, pp. 1250–1267, 2024, doi: 10.1007/978-981-96-0917-8_20.
- [17] Z. Yu, Z. Huang, J. Pei, L. Tahsin, and D. Sun, "Semantic-Oriented Feature Coupling Transformer for Vehicle Re-Identification in Intelligent Transportation System," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 3, pp. 2803–2813, 2024, doi: 10.1109/TITS.2023.3257873.
- [18] X. Bouthillier et al., "Accounting for Variance in Machine Learning Benchmarks," in *Proc. Mach. Learn. Syst. (MLSys)*, vol. 3, pp. 747–769, 2021, doi: 10.48550/arXiv.2103.03098.
- [19] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, and D. Meger, "Deep Reinforcement Learning That Matters," in *Proc. AAAI Conf. Artif. Intell.*, pp. 3207–3214, 2018, doi: 10.48550/arXiv.1709.06560.
- [20] M. Herrmann et al., "Position: Why We Must Rethink Empirical Research in Machine Learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 235, pp. 18228–18247, 2024, doi: 10.48550/arXiv.2405.02200.
- [21] A. Hochlehnert, H. Bhatnagar, V. Udandara, S. Albanie, A. Prabhu, and M. Bethge, "A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Reproducibility," in *Proc. Conf. Language Modeling (COLM)*, 2025, doi: 10.48550/arXiv.2504.07086.
- [22] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-Based Object Re-Identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 14993–15002, 2021, doi: 10.1109/ICCV48922.2021.01474.



- [23] H. Li, J. Chen, A. Zheng, Y. Wu, and J. Luo, “Day-Night Cross-Domain Vehicle Re-Identification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 12626–12635, 2024. doi: 10.1109/CVPR52733.2024.01200.
- [24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable Person Re-Identification: A Benchmark,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 1116–1124, 2015. doi: 10.1109/ICCV.2015.133.
- [25] X. Liu, W. Liu, H. Ma, and H. Fu, “Large-Scale Vehicle Re-Identification in Urban Surveillance Videos,” in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, pp. 1–6, 2016. doi: 10.1109/ICME.2016.7553002.
- [26] A. Dosovitskiy et al., “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. doi: 10.48550/arXiv.2010.11929.
- [27] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.
- [28] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall, 1993.
- [29] S. Holm, “A Simple Sequentially Rejective Multiple Test Procedure,” *Scand. J. Stat.*, vol. 6, no. 2, pp. 65–70, 1979, doi: 10.2307/4615733.
- [30] H. Jiang, A. Luo, H. Fan, S. Han, and S. Liu, “Low-Light Image Enhancement With Wavelet-Based Diffusion Models,” *ACM Trans. Graph.*, vol. 42, no. 6, Art. no. 238, 2023, doi: 10.1145/3618373.