# Towards to Robust Visual Place Recognition for Mobile Robots with an End-to-end Dark-enhanced Net

Zhenyu Li, Tianyi Shang, Pengjie Xu, Zhaojun Deng, and Ruirui Zhang

***Abstract*— Recent years have witnessed a fast evolution and promising performance of the vision transformer (ViT)-based place recognizer, which aims at building a general system. State-of-the-arts (SOTA) can hardly carry on their superiority at low light so far, thereby considerably blocking the broadening of visual place recognition-related mobile robot applications. To perform robust visual place recognition in low-light scenes, this paper proposes an end-to-end trainable dark-enhanced Net, which tries to alleviate the impact of poor illumination and environmental noise. Specifically, a lightweight dark enhancement module, i.e.,** ResEM**, is firstly trained to efficiently improve image illumination quality by residual-based adversarial learning. A dual-level sampling pyramid transformer, i.e.,** DSPFormer**, is then constructed to extract discriminative features through aggregating reconstructed descriptors. Moreover, to improve the performance and reliability of place recognition, a re-ranking method based on cross-entropy loss is used for final place matching. To provide a comprehensive evaluation, we also build two challenging place benchmarks, namely** SimPlace **and** DarkPlace**. Evaluations of both the public benchmarks and the newly built benchmarks show that the task-inspired design enables the recognizer to achieve significant performance improvements in the nighttime for robot place recognition compared to other top-ranked place recognizers. Our code is available: https://github.com/CV4RA/Dark-enhanced-VPR-Net**

***Index Terms*— Mobile robot, Nighttime place recognition, Image enhancement, Dual-sampling pyramid transformer, challenging benchmarks.**

## I. Introduction

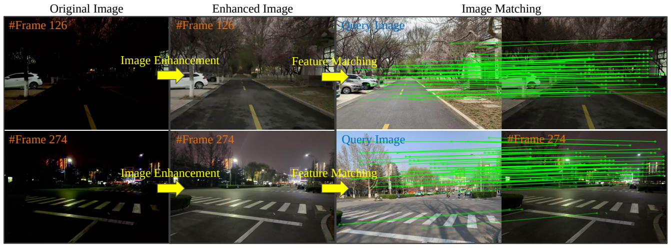


Fig. 1. The contrast results reveale some details in dark areas before and after image enhancement. The proposed end-to-end dark-enhanced Net performs feature matching and image matching successfully.

VISUAL Place Recognition (VPR) is a crucial technology used to solve the ”Where am I?” problem. Specifically, VPR determines whether the current view of a robot or mobile device is from a location that has been visited in the past. With the ability to visually recognize places, VPR can be used as a key component to handle many problems in computer vision and robotics, such as semantic-based image retrieval [1], closed-loop detection [2], and visual localization [3]. Recent research has shown that ViT models, designed for general visual learning tasks with large-scale data, can offer valuable feature representations for VPR challenges. However, these SOTA models often struggle to generalize effectively to low-illumination conditions due to the absence of low-illumination scenes in existing datasets. With the continuous advancement of machine learning and deep learning, various visual tasks are also being promoted to develop rapidly, such as research on basic models of spectral remote sensing [4], object detection [5], semantic segmentation [6], etc.

The main challenge of VPR in low-illumination conditions is the significant decline in the quality of visual information, which manifests as low image contrast, increased noise, and difficulty in discerning details, leading to challenges in feature extraction. Additionally, illumination issues can cause color distortion and strong shadows, leading to misleading information and obscuring true environmental features. Researching VPR, especially under low-illumination conditions, has multiple motivations.

Firstly, Autonomous vehicles, drones, and robots need to accurately recognize place and understand their surroundings under various lighting conditions to navigate effectively. Secondly, VPR in practical applications such as search and rescue,

This work was funded by the Shandong Provincial Natural Science Foundation (ZR2024QF284), the Qing Chuang Plan by the Department of Education of Shandong Province (24240904, 24240902), the Chinese Society of Construction Machinery Young Talent Lifting Project (CCMSYESS2023001), the Opening Foundation of Key Laboratory of Intelligent Robot (HBIR202301), the Open Project of Fujian Key Laboratory of Spatial Information Perception and Intelligent Processing (FKLSIPIP1027). Corresponding authors: Zhenyu Li.

Zhenyu Li and Ruirui Zhang are with the School of Mechanical Engineering, Qilu University of Technology (Shandong Academy of Sciences), Jinan 250353, China (e-mail: lizhenyu@qlu.edu.cn; rrzhang@qlu.edu.cn).

Tianyi Shang is with the Department of Electronic and Information Engineering, Fuzhou University, Fuzhou 350100, China (e-mail: 832201319@fzu.edu.cn).

Pengjie Xu is with the School of Mechanical Engineering, Shanghai Jiaotong University, Shanghai 200030, China (e-mail: xupengjie194105@sjtu.edu.cn).

Zhaojun Deng is with the School of Mechanical Engineering, Tongji University, Shanghai 201804, China (e-mail: dengzhaojun@tongji.edu.cn).

security monitoring, and surveillance must handle variable lighting conditions to enhance system reliability and safety. Furthermore, Research in VPR also advances artificial intelligence and robotics, making these technologies more intelligent and complex. Lastly, accurate VPR is crucial for tasks like environmental monitoring, urban planning, and mapping.

In low-illumination environments, the quality of the image will be seriously affected by, such as low contrast, large noise, loss of detail, etc., which will negatively affect the performance of the VPR task. Also, there is no VPR method specifically designed for low-illumination environments, making it difficult for even the currently top-ranked place recognizers to maintain their SOTA performance under low-illumination conditions. Therefore, the wide application of place recognition is hindered by light conditions so far. $How\ to\ alleviate\ the\ impact\ of\ poor\ illumination\ on\ mobile\ VPR$ ?

The limitations of dark VPR can be mitigated by enhancing low-light images before VPR, which inspires us to introduce an end-to-end dark enhancement Net for mobile robots. Considering the limited computing resources and complex working conditions of mobile robots, the enhancement modules designed must be lightweight, robust, easy to deploy, and, most importantly, facilitate place recognition. For traditional low-light enhancement methods, such as histogram equalization-based methods and Retinex model-based methods, the ideal assumption of considering the reflection component as the enhancement result does not always hold, especially considering various lighting properties. Noise is typically disregarded in Retinex models [7], thus it is preserved or even amplified in the enhanced results. As a result, finding an effective prior or regularization is challenging due to inaccurate priors or regularization that can result in artifacts and color bias in the enhancement outcomes. Naturally thinking, adversarial learning utilizes the ideal lighting scene from prior scenes to enhance image quality under low-light conditions, producing a more robust performance regardless of lighting conditions. To this purpose, a lightweight image enhancement module namely ResEM is proposed to facilitate robot VPR. Fig. 1 shows the contrast effect before and after image enhancement.

Although image enhancement can improve image quality, the inherent inferiority of the original image limits the effectiveness of the enhancement. It will be crucial to utilize visual encoders to accurately extract essential image information. $How\ to\ improve\ the\ discriminative\ representation\ of\ global\ descriptors\ in\ feature\ encoding$ ?

Further, this paper proposes a dual-level sampling pyramid transformer, DSPFomer, which introduces a same-level down-to-up sampling mechanism into the network structure and associates two-level attention to achieve more refined and flexible feature extraction and information processing, and effectively captures and utilizes input data, thereby improving the performance and generalization ability of the model. Fig. 2 exhibits the pipeline of the VPR framework appending ResEM and DSPFormer.

The main contributions of this paper are summarized below:

- A lightweight and efficient image enhancement module with skip-connection adversarial learning module, known as ResEM, is designed to mitigate the impact of poor illumination on robot VPR.
- A dual-level sampling pyramid transformer, named DSP-Fomer, is proposed to enhance the discriminative capability of visual features through an attention-association mechanism (symmetric upsampling to downsampling structure) for feature encoding.
- A re-ranking method based on cross-entropy loss is proposed for final place matching. Two challenging benchmarks, namely SimPlace and DarkPlace, have been developed to offer a comprehensive evaluation of VPR performance.

## II. Related Work

### A. Visual Place Recognition for Mobile Robots

Generally, the VPR approaches can be categorized as CNN-based approaches [8] and ViT-based approaches [9]. On account of the appealing VPR performance, ViT-based place recognizers become the current trend of VPR. VPR is essentially a type of image retrieval task. For example, Qin et al. [10] proposed a 3D CAD model retrieval method based on sketches and unsupervised mutation autoencoders. Hou et al. [11] proposed an efficient graph convolution network based on B-rep for 3D-CAD model classification and retrieval. Qin et al. [12] used deep residual learning to perform fine-grained leukocyte classification on microscopic images. Arandjelovic et al. [13] proposed a CNN-based weakly supervised place recognition framework, namely NetVLAD. By extracting patch-level features from NetVLAD residuals, Hausler et al. [14] provide a new method that combines the advantages of local descriptor methods and global descriptor methods, namely Patch-NetVLAD. VPR essentially belongs to the category of image retrieval and is a special two-classification problem. To address this problem, Cao et al. [15] studied two main types of image representation: global and local image features and their key contribution is to unify global and local features into a single deep model, thereby enabling accurate retrieval through effective feature extraction. The emergence of ViT is further promoting the continuous progress of various visual tasks. Zhu et al. [16] proposed a unified place recognition framework that uses a new transformer model, $R^2Former$, to handle retrieval and re-ranking. Keetha et al. [17] developed a general VPR framework, namely AnyLoc, capable of operating in both structured and unstructured environments without any retraining or fine-tuning. Ali-bey et al. [18] proposed a VPR method based on the holistic feature aggregation technique, which takes the feature maps from the pre-trained backbone as a set of global features. However, although these place recognizers perform well under normal lighting conditions, their robustness is greatly reduced under low-light conditions. $How\ can\ we\ improve\ the\ display\ quality\ of\ images\ so\ that\ mobile\ robots\ can\ more\ easily\ understand\ the\ scene$ ?

### B. Low-light Image Enhancement Methods

Images taken in low-light environments often suffer from complex degradation. Simply adjusting the lighting inevitably creates bursts of hidden noise and color distortion. To achieve satisfactory lighting, cleanliness, and realistic results from degraded inputs, Guo et al. [19] proposed a new framework inspired by the divide-and-conquer principle, which significantly alleviates the entanglement of degradation. Zhao et al. [20] proposed a unified deep framework for low-light image enhancement, which uses a generative strategy for Retinex decomposition, by which the decomposition is cast as a generative problem. *Deep learning-based methods have achieved remarkable success in image enhancement. However, are they still competitive in the absence of paired training data* ?

Recently, generative adversarial networks (GAN) have made breakthrough progress in the field of image enhancement. Jiang et al. [21] proposed an efficient unsupervised generative adversarial network, called Enlightenment GAN, which can be trained without low-light/normal-light image pairs. Jiang et al. [22] proposed a Stage Transformer Guided Network (STGNet), which can effectively handle the distribution of specific regions and enhance different low-light images. Niu et al. [23] proposed a defect image generation method with controllable defect area and intensity based on the GAN network. Yu et al. [24] proposed a new generative adversarial network (GAN), multi-granularity GAN (MGGAN), for wafer map augmentation and augmentation. Although GAN networks have progressed in many visual tasks, adding enhancement modules will inevitably increase the computational burden. *How to make the image enhancement model lightweight and achieve end-to-end visual task execution* ?

## III. Methodology

Aiming to address the low-light problem that frequently occurs in robot VPR, an efficient end-to-end trainable dark-enhanced Net is proposed. As shown in Fig. 2, inspired by ResNet, this paper first constructed a lightweight GAN network for image enhancement. Subsequently, the enhanced images are input into the dual-level sampling pyramid transformer module, utilizing pyramid attention to extract information from various feature scales. In decoding sampling, low-level information is enhanced into high-level information through the cross-attention mechanism, and VLAD is utilized to aggregate cascaded features for achieving global feature representation. This dual-level encoder-decoder architecture avoids connections and channel adjustments between different layers, which can reduce the amount of computing while optimizing the utilization of multi-layer feature information. Finally, a re-ranking module based on cross-entropy loss is further utilized to enhance the accuracy and efficiency of place matching.

### *A. Enhancement Module*

Our model utilizes the Wasserstein GAN framework [25], which comprises a generator network trained alongside a discriminator network, as shown in Fig. 3.

*1) Generator Network:* The generator network $\mathbf{G}(I, M)$ of the enhancement module takes an input comprising an image $I$ with pixel values in the range of $[-1, 1]$ and a binary mask $M$. Both inputs are of the same spatial dimension and are connected through channel direction. Both $I$ and $M$ consist of a known pixel area and an unknown pixel area. In contrast to the inpainting framework, the unknown region only shares its boundary with the known region on one side. $I$ is set to 0 in the unknown region while $M$ is set to 1 in the unknown region and is set to 0 in the known region. During training,

$$I = x \odot (1 - M) \tag{1}$$

Where $x$ is sampled from the real image distribution $\chi$, $\odot$ is the element-wise multiplication operator.

The output of $\mathbf{G}(I, M)$ has the same dimensions as the original input $I$. The final stage before inputting the discriminator $D$ is to replace the content synthesized by $\mathbf{G}$ in the unmasked region with the known input pixels:

$$\hat{x} = \mathbf{G}(I, M) \odot M + I \tag{2}$$

To fully restore discriminative image features, the generator utilizes an encoding-decoding structure design and incorporates residual connections to achieve a lightweight model.

For encoding processing, A multi-layer hierarchical convolutional network is utilized to extract detailed features. With a stride of 2, the input context is downsampled through all convolutional layers, transforming it into a latent representation that describes semantic features. To expedite convergence, batch normalization layers are placed sequentially after corresponding convolutional layers.

Instead of using a fractional-stepped transposed convolution method to upsample the latent feature representation, we adopt a subpixel convolution method to fill in missing pixels by reshuffling activations within the feature map. The calculation formula for the intermediate feature map is as follows:

$$\hat{x}^{(i+1)} = f^l x(i)) = \varphi(W_l * \hat{x}^i + b_l) \tag{3}$$

Where $f^l$ represents the sub-pixel convolution operation, $W_l$, $b_l$, and $*$ respectively represent the convolution kernel, convolution bias, and convolution operator in the sub-pixel convolution operation $f^l$, $\varphi$ is the periodic channel operator.

To enhance convergence during training and reuse of existing features extracted in the encoder, as well as to improve the coding efficiency of the generator, skip connections are established between the encoder and the decoder:

$$\hat{x}^{i+1} = \sigma(\mathbf{G}(\hat{x}^i) + \hat{x}^{N+1-i}) \tag{4}$$

Where $\sigma$ is the nonlinear activation function, $N$ is the layer numbers of the encoder or the decoder. The enhancement module also utilizes a discriminator to punish generators for producing low-resolution images and uses the error gradient provided by the discriminator to help the generator adjust its parameters.

*2) Discriminator Network:* In addition to the images being classified as real or fake, a one-hot class label $y$ is also passed to the discriminator, and $f_y$ maps $y$ to a vector of the same size as the output. The output of the discriminator is expressed as:

$$\mathbf{D}(\mathbf{G}, y) = f_\gamma(\gamma(\mathbf{G})) + (\gamma(\mathbf{G}), f_y(y)) \tag{5}$$

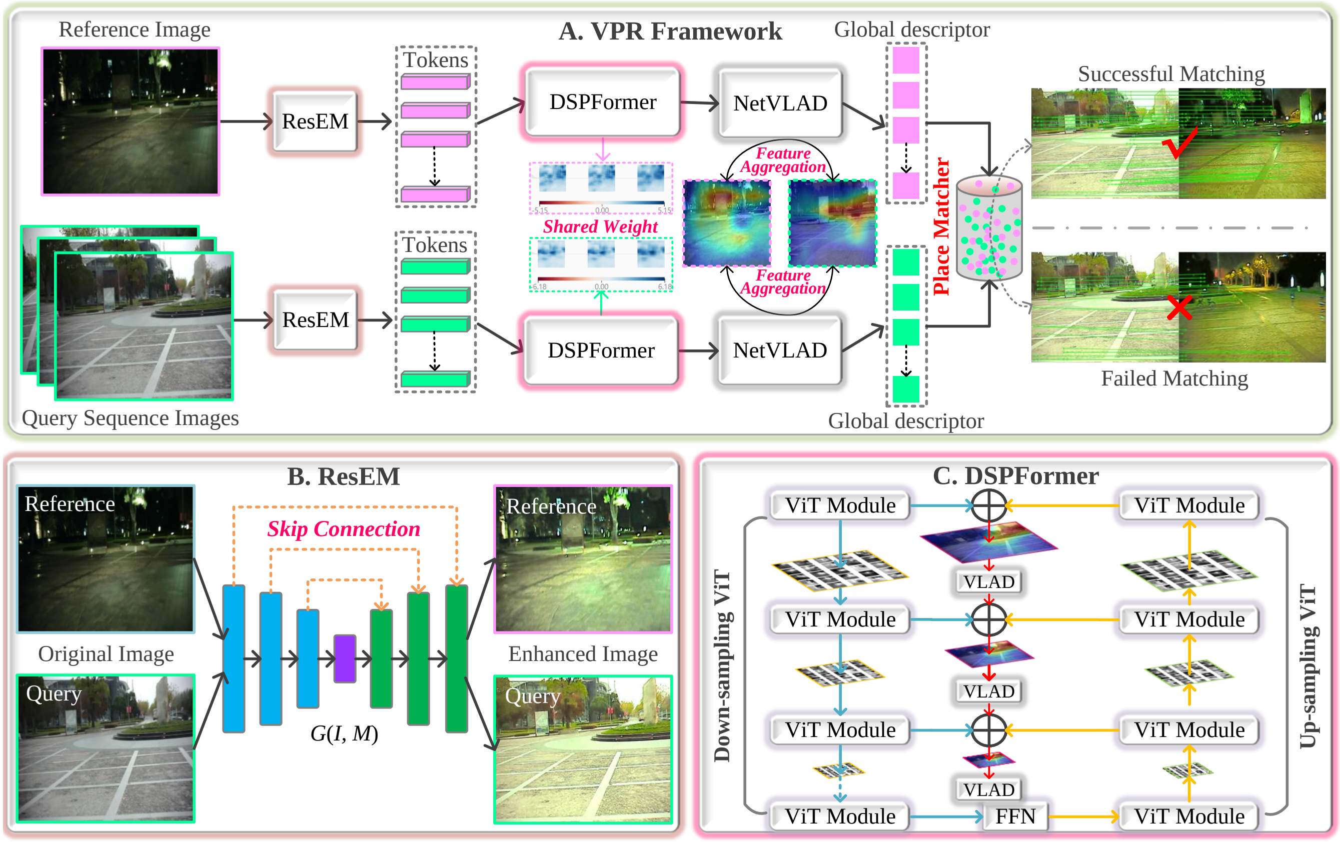


Fig. 2. Illustration of the proposed end-to-end trainable framework for mobile robot VPR. The VPR framework consists of a reinforcement module, an encoding module, and a matching module, connected in an end-to-end trainable manner. The original image is preprocessed using a lightweight GAN for image enhancement and a pyramid transformer network connected end-to-end to a dual-level sampling for feature encoding. At the backend of the VPR framework, a reranking network based on cross-entropy loss is used for the final place matching.

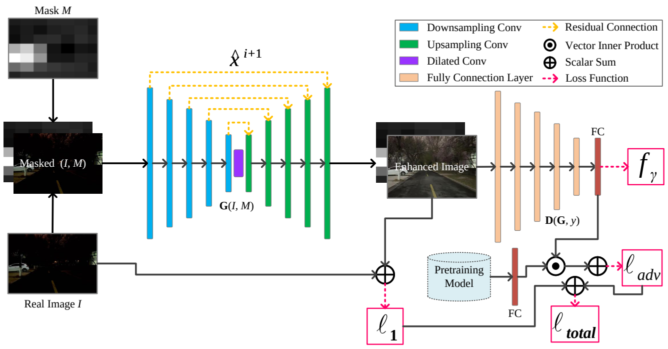


Fig. 3. Overveiw of residual GAN structure. We replace $\boldsymbol{y}$ with activations from the pre-trained image classification network ResNet50 in ImageNet.

Where $\gamma(\cdot)$ is a mapping function that converts feature maps into vectors, $f_\gamma(\cdot)$ is a learned linear function from the fully connected layer that maps this vector to a scalar.

*3) Training Loss:* Finally, the enhancement module is trained by a combination of reconstruction losses and adversarial losses. The reconstruction loss is optimized for coarse image consistency and implemented as an $\ell_1$ loss applied to the full output of $\mathbf{G}$:

$$\ell_1 = \parallel I - G(I, M) \parallel_1 \tag{6}$$

For adversarial loss that improves coarse predictions, we use the Wasserstein loss [26], [27]:

$$\begin{aligned} \ell_{<adv,\mathbf{D}>} &= \mathbb{E}[\mathrm{ReLU}(1 - \mathbf{G}(x, M, x)) + \\ &\quad \mathrm{ReLU}(1 + \mathbf{D}(\hat{x}, M, x))] \\ \ell_{<adv,\mathbf{G}>} &= \mathbb{E}[-\mathbf{D}(\hat{x}, M, x)] \end{aligned} \tag{7}$$

Finally, the total loss of our enhanced model is:

$$\ell_{total} = \ell_1 + \mu \ell_{<adv,\mathbf{G}>} \tag{8}$$

Where $\mu$ is the adversarial loss coefficient, we set $\mu = 0.9 \times 10^{-2}$ in our experiments.

## B. Dual-level Sampling Pyramid Transformer

The feature encoder is a dual-level sampling pyramid transformer, consisting of two interconnected stages of symmetric upsampling and downsampling. The two stages each contain $S$ stages, and each stage has several ViTs. In most cases, the backbone gradually upsamples and downsamples the input image, each producing $S + 1$ feature maps.

$$\begin{aligned} F(f_\gamma)^S &= F^0 + F^1, ..., +F^S \\ F^{'}(f_\gamma)^S &= F^{'0} + F^{'1}, ..., +F^{'S} \end{aligned} \tag{9}$$

The feature encoder is composed of a backbone, neck, and head. Up sampling and down sampling can share the same architecture and parameters, gradually aggregating the features

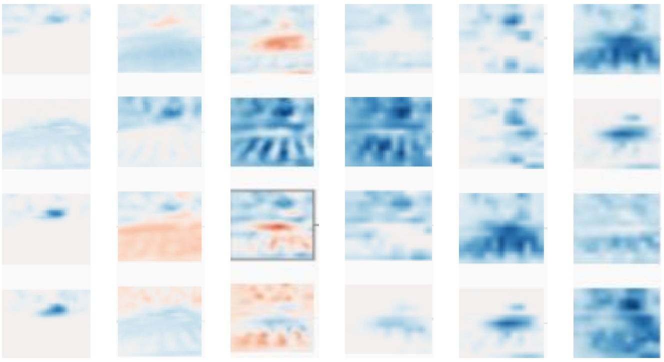


Fig. 4. Weight maps learned from a subset of 24 neurons in the first attention block. Blue corresponds to positive weights, and red corresponds to negative weights.

of the corresponding stages to reconstruct a new feature map. Therefore, the neck can be expressed as:

$$\begin{aligned} &F'^{S} = F^{S} \\ &F'^{s} = F^{s} + g^{s}\left(F'^{s+1}; \phi^{s}\right), \quad 1 \leqslant s < S \end{aligned} \tag{10}$$

Where $g^s(\cdot)$ is the upsampling processing, $\phi$ represents the layer-wise parameter, $F^S$ and $F'^S$ are the down-sampling feature maps and up-sampling feature maps.

We construct reconstruction loss to optimize a series of multi-stage feature reconstruction processes. We define that $H(\cdot)$ represents a few transformer blocks that reconstruct the original image from $F' = \sum_{S}^{s=1} F^s$. To gain the ability to capture multi-stage features, we add a VLAD module at each stage, namely $Net(\cdot)$. The loss function is expressed as:

$$\ell_{att} = \left\| X^0 - H^0(F') \right\| + \eta \textstyle\sum_{s=1}^{S} Net^s(\|X^s - H^s(F')\|) \tag{11}$$

Where $X^s$ is the the expected output at the $s^th$ decoder stage, $\eta$ is the feature reconstruction coefficient, $Net(\cdot)$ is the aggregation function. $\left\| X^0 - H^0(F') \right\|$ represents image reconstruction, $\sum_{s=1}^{S} Net^s(\|X^s - H^s(F')\|)$ represents feature reconstruction.

Fig. 4 illustrates a subset of the weights learned from the first layer, specifically 24 out of 400 neurons in the hidden layer. The weights of each unit have been reshaped to $16 \times 16$ to align with the spatial size of the feature maps from the backbone. As we can see, the hidden units learn a wide range of regional feature selections in cross-attention aggregation. Furthermore, we observed that some neurons focus on one or more small spots in the image, while others focus on the entire input. We believe that the combination of these cross-attention neurons can replace CNN and pyramid-deep model-based VPR algorithms.

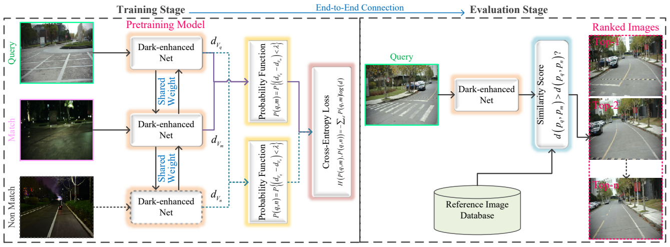


Fig. 5. Proposed reranking place matcher based on cross-entropy loss. Place matcher is connected in an end-to-end manner by a training phase and an evaluation phase.

## C. Place Matcher

We establish a cross-entropy loss network for place re-ranking, to predict the relative distance between the input query and the database reference to achieve metric learning, which is shown in Fig. 5. To this end, we retrieve two queries from the database simultaneously, forming triples $T(p, m, n)$. We define two probability functions $P(q, m)$ and $P(q, n)$ respectively, which respectively represent the metric probability of the distance between the query and the two references, and set the margin condition to $\lambda$. The cross-entropy of $P(q, n)$ represented by $P(q, m)$ can be calculated as follows:

$$H(P(q,m), P(q,n)) = -\textstyle\sum_d P(q,m) \log P(q,n) \tag{12}$$

Among them, the solutions to the two probability functions are as follows:

$$\begin{aligned} P(q,m) = P\{(d_{V_q} - d_{V_m}) < \lambda_1\} \\ P(q,n) = P\{(d_{V_q} - d_{V_n}) < \lambda_2\} \end{aligned} \tag{13}$$

## IV. Experimental Setup and Result Analysis

### A. Implementation Details

We train the dark-enhanced Net end-to-end using different learning rates and optimizers for different stages. The image enhancement module adopts a learning rate of $10^{-4}$, while the image encoding module adopts a learning rate of $1.8 \times 10^{-4}$ with a weight decay of 0.08. The batch size for both stages is 64, and the optimizers for both stages are Adam and AdamW respectively. We pre-train the model on the ImageNet-1k dataset, which contains training 1.8M images of 1000 classes, and do not use class labels during the pre-training stage. Each training image is preprocessed to $224 \times 224$ pixels and divided into $14 \times 14$ patches of size $16 \times 16$ pixels. The hyperparameters are set to $\mu = 0.01$, $\eta = 0.2$, $\lambda = 0.2$, $\lambda = 0.4$. We train all these models using $4 \times$ 12G NVIDIA GeForce RTX 3060.

### B. Benchmarks and Evaluation Metrics

*1) Datasets:* We train our model on the built DarkPlace dataset (including 1542 images with the size of $1920 \times 1080$), which covers a wide range of real-world scenarios for robot VPR, e.g., different viewpoints, and light changes. The model in urban scenarios (including KITTI_00 [28], Tokyo 27/4 [29], VPRICE [30], and Nordland [31] datasets) is further finetuned by the built SimPlace dataset (including 3709 images with the size of $2048 \times 1024$).

*2) Evaluation Metrics:* For evaluation, we followed previous research [13]-[18] and set the threshold for the correct place matching at 25 candidates. We report $Recall@N$ ($N$=1 to 25) as the evaluation metric. We also calculate costs (including enhancement, extraction, matching, and inference time), feature dimensions, GFLOPs, and memory footprint in detail to provide comprehensive metrics for model evaluation.

### C. Experimental Results

In this section, we compare the proposed dark-enhanced Net with previous SOTAs such as NetVLAD [13], Patch-NetVLAD [14], DELG [15], $R^2$Former [16], AnyLoc [17], and MixVPR [18] based on public benchmarks, e.g., KITTI_00, Tokyo 27/4, VPRICE, Nordland, and built benchmarks, e.g., SimPlace, DarkPlace. We also conducted ablation studies to verify the impact of each module or parameter of the proposed Net on the overall performance.

*1) Comparison with SOTAs:* As shown in Fig. 6 and Table I, the method proposed in this paper has a higher average recall rate than the SOTA methods across the four publicly available datasets. Thanks to the implementation of end-to-end image enhancement, our method further improved the average recall rate with absolute AR@1 improvement of 2.5%, 1.6%, 1.7%, and 1.8% on four public datasets respectively. Fig. 7 and Table II show comparisons of the precision-recall curves between the proposed dark-enhanced Net and SOTA methods on low-light challenging datasets (built SimPlace and DarkPlace datasets) with absolute AR@1 improvement of 24%, and 26% respectively. Thanks to the enhanced module this paper adopted, the proposed Net can still maintain performance equivalent to that under ideal illumination under extreme loss-light conditions, demonstrating strong robustness while other SOTA nets have lost performance under the same conditions.

We record the computation time of all methods on DarkPlace, as shown in Fig. 8. Due to the inclusion of the image enhancement module and the delay caused by end-to-end training, the computation time of the VPR model is large compared to other methods while our efforts to simplify the enhancement module.

In Fig. 9, we visualize the attention maps generated by the proposed DarkVPR and baseline methods. On the feature encoding, DarkVPR demonstrates advantages in detecting entire salient objects or regions in low-illumination datasets. This capability arises because DarkVPR compels the model to preserve richer visual features (multi-scale feature maps), thereby enhancing recognition results downstream.

*2) Ablation Study:* We perform ablation studies to validate the contribution of important modules and parameters to the overall model.

- Ablation Studies on the Image Enhanced Module
  The ablation studies on the image-enhanced module in Table III show the absolute AR@1 improvement of 26.7% and 25.3% on SimPlace and DarkPlace respectively, which confirms the effectiveness of Dark Enhancement for VPR missions in extreme low-light environments.
- Ablation Study on Key Components
  In Table IV, we utilize the "reranking" configuration as a baseline and systematically remove various components to assess the significance of each component in our approach. For simplicity, inspired by [16], we freeze the global retrieval module, and all models are trained using partial negative mining. We get the following conclusions: 1) When re-ranking is removed, VPR performance drops significantly, which shows that re-ranking

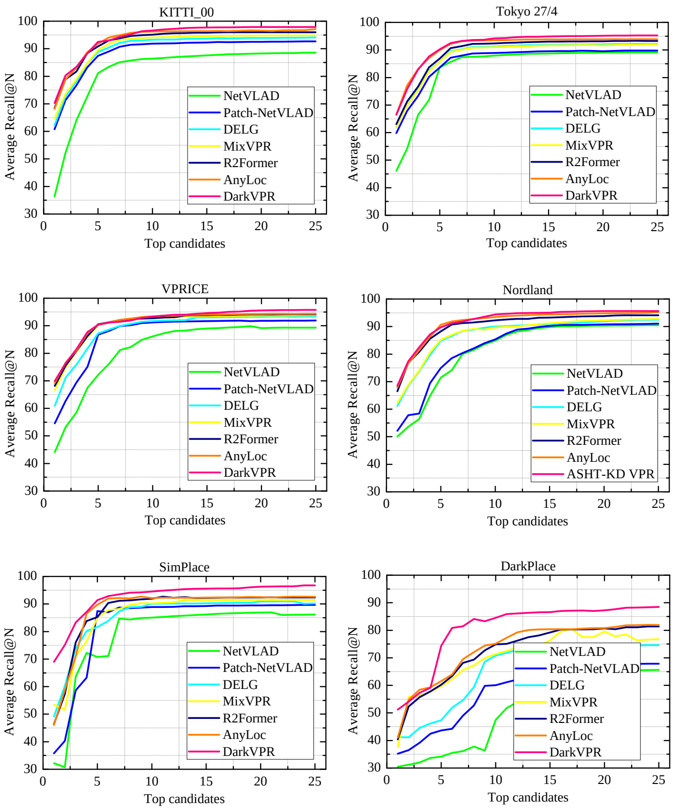


Fig. 6. Comparisons of the proposed dark-enhanced Net (presented in the figure as ”DarkVPR”) with SOTA methods on four public benchmarks: KITTI_00, Tokyo 27/4, VPRICE, Nordland, and also built SimPlace based on CARLAR simulator and DarkPlace based on real-world scenarios.

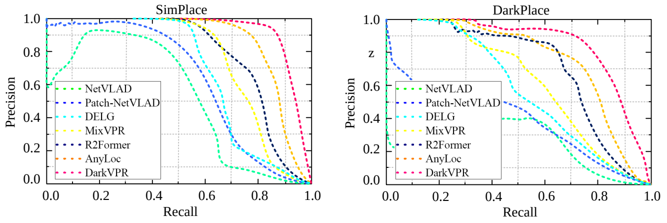


Fig. 7. Precision-recall curves of comparison methods on built SimPlace and DarkPlace datasets.

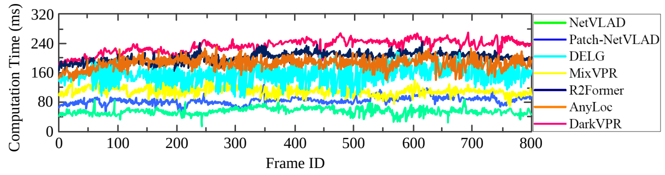


Fig. 8. Time evaluation on DarkPlace dataset.

TABLE I
COMPARISON OF THE PROPOSED METHOD WITH PREVIOUS SOTA RESULTS ON PUBLIC DATASETS.

| Methods | KITTI_00 | | Tokyo27/4 | | VPRICE | | Nordland | |
|---|---|---|---|---|---|---|---|---|
| | AR@1 | AR@5 | AR@1 | AR@5 | AR@1 | AR@5 | AR@1 | AR@5 |
| NetVLAD [13] | 54.7 | 63.3 | 15.6 | 22.9 | 50.3 | 60.5 | 58.7 | 61.6 |
| Patch-NetVLAD [14] | 73.6 | 79.4 | 50.2 | 58.4 | 79.2 | 83.6 | 78.7 | 83.4 |
| DELG [15] | 75.1 | 81.9 | 58.9 | 64.2 | 81.5 | 85.2 | 79.5 | 84.4 |
| MixVPR [18] | 76.2 | 82.3 | 60.7 | 64.9 | 81.7 | 86.0 | 81.2 | 85.1 |
| $R^2$Former [16] | 85.4 | 88.1 | 79.2 | 85.8 | 84.4 | 91.4 | 87.3 | 91.8 |
| AnyLoc [17] | 85.6 | 88.3 | 80.0 | 85.5 | 85.1 | 91.6 | 87.4 | 91.8 |
| DarkVPR | 87.8 | 90.3 | 81.3 | 88.4 | 86.6 | 94.3 | 89.0 | 92.3 |

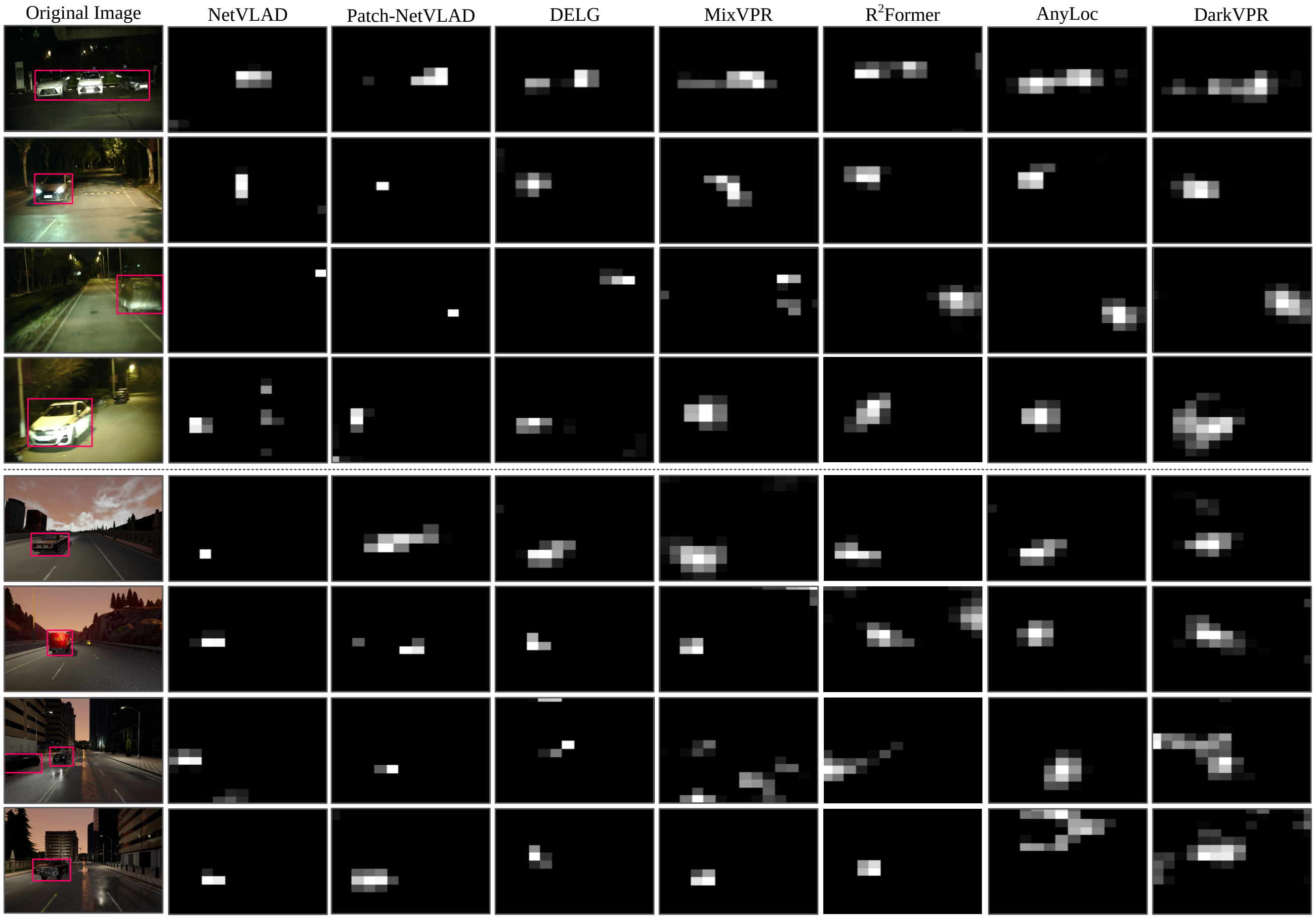


Fig. 9. A comparison between the attention maps generated by NetVLAD, Patch-NetVLAD, DELG, MixVPR, $\boldsymbol{R^2}$Former, AnyLoc, and the proposed DarkVPR. In each case, the red block indicates the query token, and the attention map (White regions) between the query and other tokens at the corresponding Net is shown. The examples comes from DarkPlace and SimPlace datasets.

TABLE II
COMPARISON OF THE PROPOSED METHOD WITH PREVIOUS SOTA RESULTS ON BUILT LOW-LIGHT SCENE DATASETS.

| Methods | SimPlace | | DarkPlace | |
|---|---|---|---|---|
| | AR@1 | AR@5 | AR@1 | AR@5 |
| NetVLAD [13] | 21.8 | 39.8 | 16.1 | 24.6 |
| Patch-NetVLAD [14] | 24.7 | 42.5 | 19.9 | 27.3 |
| DELG [15] | 46.2 | 61.7 | 38.5 | 47.5 |
| MixVPR [18] | 48.9 | 66.0 | 44.7 | 55.4 |
| $R^2$Former [16] | 62.6 | 71.4 | 56.2 | 66.8 |
| AnyLoc [17] | 64.8 | 75.2 | 58.0 | 73.6 |
| DarkVPR | 85.4 | 88.6 | 78.2 | 85.0 |

TABLE III
ABLATION STUDIES ON THE IMAGE ENHANCED MODULE ON DARKPLACE DATASET.

| Methods | SimPlace | | DarkPlace | |
|---|---|---|---|---|
| | AR@1 | AR@5 | AR@1 | AR@5 |
| Non-Dark-enhanced VPR | 62.6 | 73.5 | 58.4 | 72.2 |
| Dark-enhanced VPR | 85.4 | 88.6 | 78.2 | 85.0 |

can improve VPR matching accuracy. 2) We observe a negligible performance drop when positional embeddings are removed, which means that positional embeddings do

TABLE IV
ABLATION STUDY ON KEY COMPONENTS BASED ON DARKPLACE DATASET.

| | AR@1 | AR@5 | Latency per Query (ms) ↓ |
|---|---|---|---|
| No Reranking | 70.6 | 74.7 | 126.74 |
| Reranking (baseline) | 78.2 | 85.0 | 244.63 |
| Remove Positional Embedding | 76.8 | 83.9 | 241.90 |
| Remove Attention Selection | 73.2 | 79.7 | 226.35 |
| Remove Up Sampling | 72.4 | 81.6 | 186.51 |
| Remove Transformer Block-1 | 77.2 | 84.5 | 233.75 |
| Remove Transformer Block-2 | 74.5 | 82.0 | 221.68 |
| Remove Residual Connection | 78.0 | 84.8 | 284.85 |

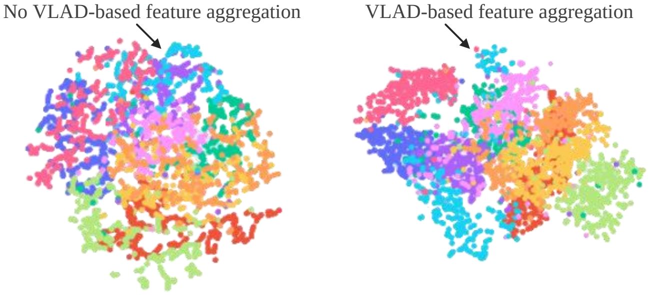


Fig. 10. Visualization of multi-stage attention feature aggregation.

not help much for low-light VPR tasks. 3) We can see that when upsampling is removed, the VPR performance drops significantly, which shows that gradient feature extraction and aggregation are very important for feature representation. 4) When we replace attention-based feature selection with random selection, the model becomes unstable, and the performance drops significantly. 5) The most significant component is the intermediate module, such as Transformer Block-2. Its removal leads to a notable performance decline, suggesting that feature correlation plays a significant role in the final decision. 6) We propose a lightweight illumination enhancement module to reduce the computational complexity of the model. From the Table, we can see that removing the residual connection in the illumination enhancement module does not affect performance, but it increases the inference time by $40ms$.

- Ablation Study on NetVLAD layer
  In our work, we utilize a NetVLAD layer to aggregate multi-stage attention features during up- and down-sampling. To explore the role of VLAD layers in feature representation, we utilize t-SN plots to visualize attention aggregation. Results in Fig. 10 show that feature aggregation based on VLAD can accurately classify and aggregate a large number of attention features in the image, while random aggregation methods without VLAD layers cannot achieve the classification and aggregation of features of varying importance.
- Ablation Study on Training Performance The benefits of more complete feature reconstruction and dark enhancement can be quantified using two-fold experiments as shown in Fig. 11. On the left, we observe that DarkVPR achieves better training results (i.e., lower loss values). Note that simply using a multi-layer visual Transformer (with multi-scale feature maps) does not improve the training results, which means that pre-training is the main contributor. In the right part, we show that better fine-tuning helps the downstream visual recognition task (success rate of place matching) converge faster and achieve higher upper bounds. From the comprehensive results, it can be seen that multi-scale feature reconstruction and dark enhancement are effective means to ensure the success of model pre-training and fine-tuning.

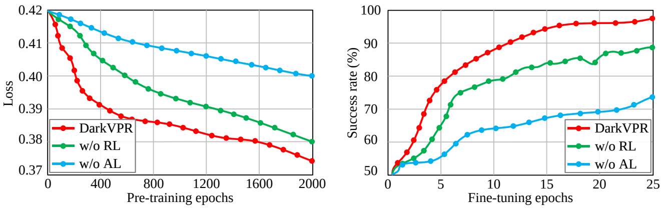


Fig. 11. Left: A comparison between the train loss generated by DarkVPR, the variant without reconstruction loss (w/o RL), without adversarial loss (w/o AL) in the pre-training stage. Right: A comparison between the success rate generated by DarkVPR, the variant without reconstruction loss (w/o RL), without adversarial loss (w/o AL) in the fine-tuning stage.

*3) Visualization:* Fig. 12 presents some detailed cases showcasing successfully matched place images. It can be observed from the examples that utilizing the image enhancement module can significantly enhance the image quality under low-light conditions, which is highly conducive to emphasizing the extracted attention features. Fig. 13 displays the rerecognition results of various methods in the low-light scenes. It can be observed that on overlapping paths, our method can successfully match features of the same places, while other methods exhibit more mismatches and dropped frames, which demonstrates the robustness of our method in extreme low-light conditions.

In addition, examples of loop closure detection and place recognition of the proposed method are provided in Fig. 14(a) and Fig. 14(b), which describe the results of loop closure for a query on the SimPlace dataset and DarkPlace dataset. It is noted that daytime scenes and nighttime scenes are considered as reference places and query places, respectively. Despite differences in orientation or localization, the top-1 candidate frames are successfully retrieved.

*4) Model Inference:* The Table V presents the number of parameters and inference speed for each key component of the model. It is evident that ResEM, which employs a lightweight GAN and residual connections, achieves a faster inference speed due to its shallow convolutional layers, reduced number of parameters, and lower computational complexity. In contrast, the DSPFormer module is designed to extract high-level features from the enhanced image by utilizing a multi-head self-attention mechanism and Transformer architecture, which typically results in higher computational complexity, as it necessitates multiple self-attention calculations and feature aggregation for each image block, thereby increasing the inference time.

## V. CONCLUSIONS

We propose an end-to-end dark-enhanced Net for mobile robot place recognition. We demonstrate for the first time that

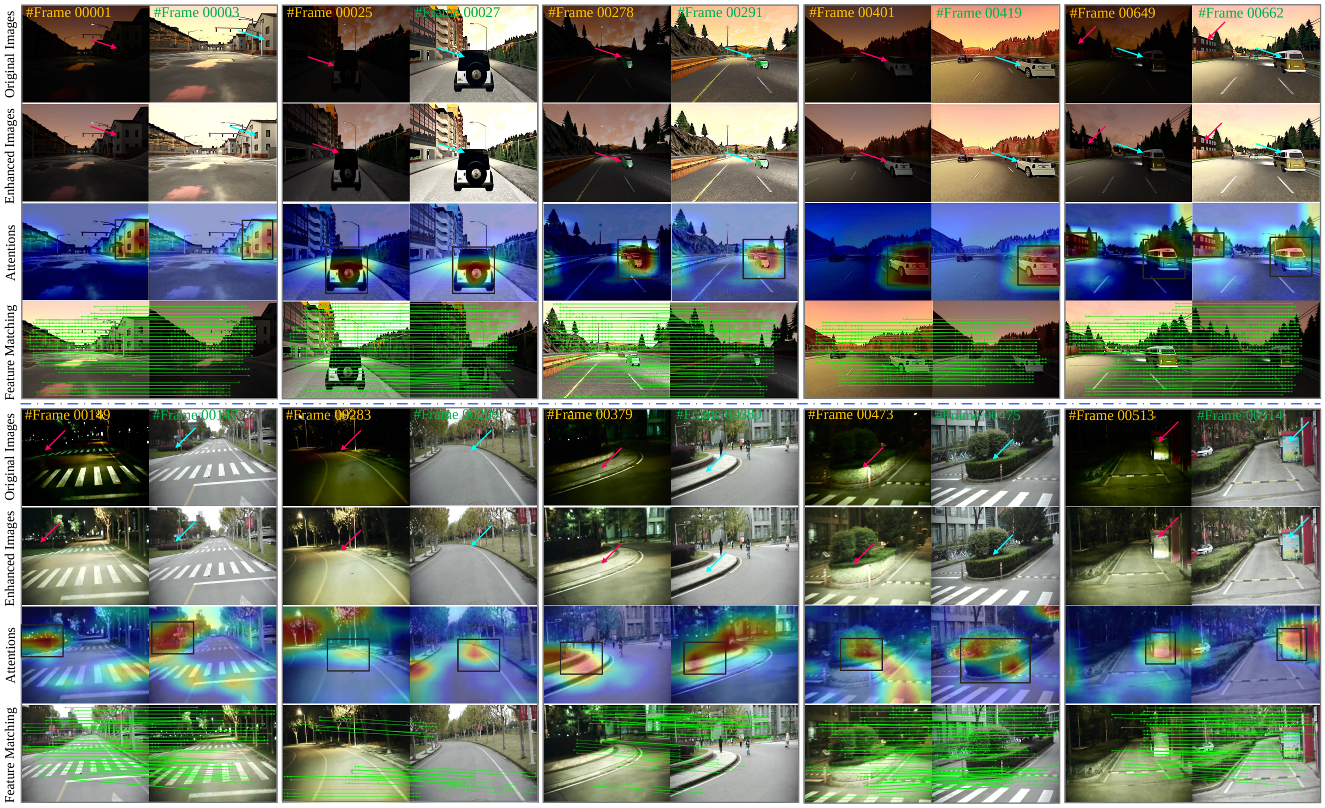


Fig. 12. Example visualization of attention aggregations and matched pairs based on the built SimPlace and DarkPlace datasets. Attention aggregation heatmap with arrows pointing to salient regions.

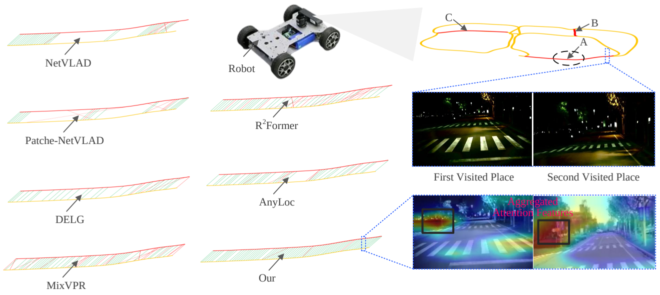
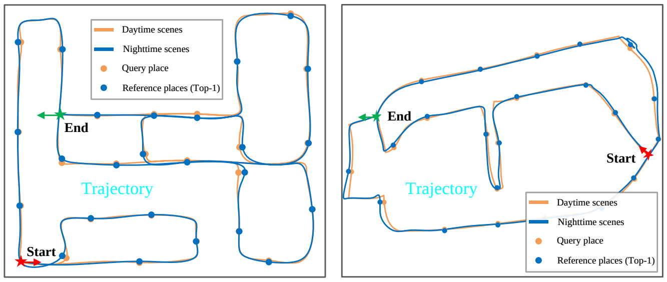


Fig. 13. Place re-recognition experiments for mobile robot in low-light conditions. We plot correct (green) and incorrect (red) matches with maximum accuracy ($100\%$) for all methods. The blue dotted box corresponds to the extracted scene instance, while the orange trajectory and red trajectory represent the two visits of the robot, respectively. White space with missing red and green matches indicates that the robot is dropping frames.

image enhancement far outperforms existing SOTA methods on low-light VPR tasks. Our dual-level down-sampling and up-sampling pyramid transformer technique can aggregate prominent features in complex scenes, enhancing robust feature encoding to more effectively capture the dynamic semantic properties of low-light virtual environments and real-world environments. We developed a place matcher using cross-entropy loss, and its impact was confirmed in ablation experiments. Besides, We build two new low-light scene benchmarks for challenging place recognition experiments, which provide more convenience for VPR research in extreme environments and also contribute to the development of VPR in the future.

TABLE V
INFERENCE AND PARAMETERS FOR EACH COMPONENT.

| Module | Inference Speed (ms/image) | Parameter (M) | GFLOPs |
|---|---|---|---|
| ResEM | 10 | 8.6 | 0.96 |
| DSPFormer | 48 | 74 | 19.8 |
| Place matcher | 8 | 1.6 | 0.85 |
| Total | 66 | 84.2 | 21.5 |



(a) SimPlace dataset (b) DarkPlace dataset

Fig. 14. The illustration of loop closure detection and place recognition of our method on SimPlace dataset and DarkPlace dataset, where the query, top-1 retrieval result and trajectory are given.

Although our method achieves SOTA results, it comes at the expense of increased computation time. In the future, we will focus on further simplifying the model and enhancing the computational efficiency of the overall model.

## References

[1] S. R. Dubey, ”A Decade Survey of Content Based Image Retrieval Using Deep Learning,” IEEE Transactions on Circuits and Systems for Video Technology, 2022, vol. 32, no. 5, pp. 2687-2704.

[2] K. A. Tsintotas, L. Bampis, A. Gasteratos, ”The Revisiting Problem in Simultaneous Localization and Mapping: A Survey on Visual Loop Closure Detection,” IEEE Transactions on Intelligent Transportation Systems, 2022, vol. 23, no. 11, pp. 19929-19953.

[3] L. Zhao X. Deng, X. Gui, J. Sun, T. Li, B. Zhang, ”Graph-Based Robust Localization of Object-Level Map for Mobile Robotic Navigation,” IEEE Transactions on Industrial Electronics, 2024, vol. 71, no. 1, pp. 697-707.

[4] D. Hong, B. Zhang, X. Li, Y. L, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, A. Plaza, ”SpectralGPT: Spectral Remote Sensing Foundation Model”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, pp. 1-18.

[5] J. Liu, W. Sun, C. Liu, X. Zhang and Q. Fu, ”Robotic Continuous Grasping System by Shape Transformer-Guided Multiobject Category-Level 6-D Pose Estimation”, IEEE Transactions on Industrial Informatics, 2023, vol. 19, no. 11, pp. 11171-11181.

[6] T. Jing, Q. -H. Meng and H. -R. Hou, ”SmokeSeger: A Transformer-CNN Coupled Model for Urban Scene Smoke Segmentation,” IEEE Transactions on Industrial Informatics, 2024, vol. 20, no. 2, pp. 1385-1396.

[7] E. H. Land, ”The Retinex Theory of Color Vision,” Scientific American, 1977, 237(6), pp. 108-129.

[8] S. Lowry, N. Sünderhauf, P. Newman, J. J. Leonard, D. Cox, P. Corke, M. J. Milford, ”Visual Place Recognition: A Survey. IEEE Transactions on Robotics, 2015, 32(1), pp. 1-19.

[9] S. Schubert, P. Neubert, S. Garg, M. Milford, T. Fischer, ”Visual Place Recognition: A Tutorial,” IEEE Robotics and Automation Magazine, 2023, pp. 1-15.

[10] F. Qin, Qiu S, S. Gao J. Bai, ”3D CAD Model Retrieval based on Sketch and Unsupervised Variational Autoencoder”, Advanced Engineering Informatics. 2022, vol. 51, pp. 1-13.

[11] J. Hou, C. Luo, F. Qin, Y. Shao, X. Chen, ”FuS-GCN: Efficient B-rep based Graph Convolutional Networks for 3D-CAD Model Classification and Retrieval”. Advanced Engineering Informatics. 2023, 56, pp. 1-12.

[12] F. Qin, N. Gao, Y. Peng, Z. Wu, S. Shen, Grudtsin A, ”Fine-grained leukocyte classification with deep residual learning for microscopic images”, Computer Methods and Programs in Biomedicine, 2018, 162, 1-10.

[13] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, J. Sivic, ”NetVLAD: CNN Architecture for Weakly Supervised Place Recognition”, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5297-5307.

[14] S. Hausler, S. Garg, M. Xu, M. Milford, T. Fischer, ”Patch-netvlad: Multi-scale Fusion of Locally-global Descriptors for Place Recognition”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14141-14152.

[15] B. Cao, A. Araujo, J. Sim, ”Unifying Deep Local and Global Features for Image Search” In Proceedings of the European Conference on Computer Vision, 2020, pp. 726-743.

[16] S. Zhu, L. Yang, C. Chen, M. Shah, X. Shen, H. Wang, ”R2former: Unified Retrieval and Reranking Transformer for Place Recognition”, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19370-19380.

[17] N. Keetha, A. Mishra, J. Karhade, K.M. Jatavallabhula, S. Scherer, M. Krishna, S. Garg, ”Anyloc: Towards universal visual place recognition”, IEEE Robotics and Automation Letters, 2023, pp.1-10.

[18] A. Ali-Bey, B. Chaib-Draa, P. Giguere, ”Mixvpr: Feature mixing for visual place recognition,” In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2998-3007.

[19] X. Guo, Q. Hu, ”Low-light image enhancement via breaking down the darkness”, International Journal of Computer Vision, 2023, vol. 131, no. 1, pp. 48-66.

[20] Z. Zhao, B. Xiong, L. Wang, Q. Ou, L. Yu, F. Kuang, ”RetinexDIP: A Unified Deep Framework for Low-Light Image Enhancement,” IEEE Transactions on Circuits and Systems for Video Technology, 2022, vol. 32, no. 3, pp. 1076-1088.

[21] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, Z. Wang, ”Enlightengan: Deep Light Enhancement Without Paired Supervision”, IEEE Transactions on Image Processing, 2021, 30, pp. 2340-2349.

[22] N. Jiang, J. Lin, T. Zhang, H. Zheng, T. Zhao, ”Low-Light Image Enhancement via Stage-Transformer-Guided Network,” IEEE Transactions on Circuits and Systems for Video Technology, 2023, vol. 33, no. 8, pp. 3701-3712.

[23] S. Niu, B. Li, X. Wang, Y. Peng, ”Region-and Strength-Controllable GAN for Defect Generation and Segmentation in Industrial Images,” in IEEE Transactions on Industrial Informatics, 2022, vol. 18, no. 7, pp. 4531-4541.

[24] J. Yu and J. Liu, ”Multiple Granularities Generative Adversarial Network for Recognition of Wafer Map Defects,” IEEE Transactions on Industrial Informatics, 2022, vol. 18, no. 3, pp. 1674-1683.

[25] P. Teterwak, A. Sarna, D. Krishnan, A. Maschinot, D. Belanger, C. Liu, W. T. Freeman, ”Boundless: Generative adversarial networks for image extension”, In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10521-10530.

[26] D. Tran, R. Ranganath, D. Blei, ”Hierarchical Implicit Models and Likelihood-free Variational Inference”, Advances in Neural Information Processing Systems, 2017, 30, pp. 1-11.

[27] P. Teterwak, A. Sarna, D. Krishnan, A. Maschinot, D. Belanger, C. Liu, W. T. Freeman, ”Boundless: Generative Adversarial Networks for Image Extension”, In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10521-10530.

[28] A. Geiger, P. Lenz, C. Stiller, R. Urtasun, ”Vision meets robotics: The kitti dataset,” The International Journal of Robotics Research, 2013, 32(11), pp. 1231-1237.

[29] A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla, ”24/7 place recognition by view synthesis,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1808–1817.

[30] T. Naseer, M. Ruhnke, C. Stachniss, L. Spinello, and W. Burgard, ”Robust Visual SLAM Across Seasons,” In IEEE/RSJ International Conference on Intelligent Robots and Systems, IEEE, 2015, pp. 2529–2535.

[31] N. S¨underhauf, S. Shirazi, F. Dayoub, B. Upcroft, and M. Milford, ”On the Performance of Convnet Features for Place Recognition,” In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2015, pp. 4297–4304.