DOI: 10.19884/j.1672-5220.202502014

EST-NSS: An Enhanced Swin Transformer for Nighttime Semantic Segmentation

TAO Ran[1], DING Yi[1*], LU Hailun[1], LU Xiaohui[1], YIN Wenxiu[1], ZHANG Qi[2]
1. School of Information and Intelligent Science, Donghua University, Shanghai 201620, China
2. Shanghai Moule Network Technology Co., Ltd., Shanghai 201620, China

Abstract: Although semantic segmentation of daytime images has achieved remarkable success, nighttime semantic segmentation still faces significant challenges due to the low brightness and contrast in nighttime images. To address these challenges, we proposed an enhanced Swin Transformer for nighttime semantic segmentation (EST-NSS) for nighttime urban streetscape segmentation. The EST-NSS utilizes Swin Transformer as the backbone network. In the image preprocessing stage, we designed a nighttime light enhancer (NLE), which integrates receptive-field attention convolution (RFAConv), attention mechanisms, and a Transformer to enhance image brightness, contrast, and semantic representation, thereby improving the model's perception of key regions. In the feature processing stage, we designed a dual-domain context unit (termed SCBlock), which integrates attention mechanisms and depthwise separable convolution to enhance multi-scale feature representations, thereby improving the model's accuracy in segmenting object boundaries. Extensive experimental results on the NightCity dataset show that our EST-NSS significantly enhances the performance of nighttime semantic segmentation. The mean intersection over union (mIoU) improved by 3.6 percentage points compared with Swin Transformer. Compared with the state-of-the-art models, the EST-NSS achieved superior performance in 15 out of 19 categories.

Keywords: semantic segmentation; Swin Transformer; nighttime light enhancement; feature enhancement
CLC number: TP391.4 **Document code**: A
Article ID: 1672-5220(2026)05-0131-12

0 Introduction

Semantic segmentation, a core task in computer vision, aims to divide an image into distinct and meaningful sections. Semantic segmentation is widely used in critical fields like autonomous driving[1] and medical image analysis[2-3]. Deep learning-based methods, leveraging the powerful feature extraction of convolutional neural networks (CNNs), have shown promising performance on daytime datasets like Cityscapes[4] and KITTI[5]. However, nighttime images suffer from low brightness and contrast, resulting in unclear segmentation content and a lack of clarity in edge details[6].

Due to the lack of pixel-wise annotated nighttime datasets, most existing methods[7-10] used unsupervised domain adaptation[11-12] to shift models trained on bright daytime settings to nighttime scenes. These methods include adapting daytime models through an intermediate domain[8,13] and training image transformation networks to convert daytime images to nighttime images[14-15]. However, these methods struggle when the domain gap is large and involve time-consuming preparations. The latest model based on unsupervised domain adaptation, prompt images guidance (PIG)[16], achieves promising results on the NightCity[17] dataset, with a mean intersection over union (mIoU) of 41.2%, but there is still potential for further improvement. The NightCity dataset, which contains 2 998 annotated nighttime training images, enables semantic segmentation in nighttime scenes using supervised learning methods.

Swin Transformer[18], as a classical semantic segmentation model, can generate multi-scale feature maps and preserve spatial positional information during the downsampling process. However, in the nighttime scene, Swin Transformer faces difficulties because of the low brightness and contrast in nighttime images.

First, low brightness leads to the loss of crucial information regarding object structure, color, and texture, making it difficult for the Swin Transformer to capture the feature relationships across spatial regions. Second, low contrast makes it difficult to accurately identify the actual location and boundaries of objects in the image, thereby making it harder to separate the objects from the background.

In this paper, we propose an enhanced Swin Transformer for nighttime semantic segmentation (EST-NSS), which adopts the Swin Transformer as the backbone network. To address the limitations of Swin Transformer in nighttime semantic segmentation, we design a nighttime light enhancer (NLE) and a dual-domain context unit (termed SCBlock). First, to enhance

Received date: 2025-02-20
Foundation item: Fundamental Research Funds for the Central Universities, China (No. 2232025H-03)
* Correspondence should be addressed to DING Yi, email: yiding@mail.dhu.edu.cn
Citation: Tao R, Ding Y, Lu H L, et al. EST-NSS: an enhanced Swin Transformer for nighttime semantic segmentation [J]. *Journal of Donghua University (English Edition)*, 2026, 43(5): 131-142.

image brightness and contrast while strengthening feature representation, we propose the NLE module by integrating receptive-field attention convolution (RFAConv)[19], attention mechanisms[20], and Transformer[21] architecture to preprocess images. Second, to further accurately segment the boundary of the object, we design SCBlock to enhance feature maps from different levels by integrating attention mechanisms and depthwise separable convolution[22] to reallocate the weights. Additionally, an online hard example mining (OHEM)[23] loss function is integrated to prioritize learning from challenging samples.

The main contributions to our work are summarized as follows.

1) We design a network EST-NSS for nighttime urban streetscape segmentation in an end-to-end manner. Extensive experimental results on the NightCity dataset show that our EST-NSS model significantly enhances the performance of nighttime semantic segmentation. The mIoU improves by 3.6 percentage points compared with Swin Transformer. Compared with the state-of-the-art models, EST-NSS achieves superior performance in 15 out of 19 categories.

2) We propose an NLE that integrates RFAConv, attention mechanisms, and a Transformer to enhance image brightness and contrast, as well as semantic representation during image preprocessing, thereby improving EST-NSS's recognition capability for the subsequent segmentation task.

3) We propose an SCBlock that integrates attention mechanisms and depthwise separable convolution to enhance multi-scale feature maps during feature processing, thereby optimizing the segmentation boundaries.

The remainder of this paper is as follows. In Section 1, we introduce related works, including nighttime semantic segmentation, RFAConv, attention mechanisms, depthwise separable convolution, and Swin Transformer. In Section 2, we provide a detailed explanation of the architecture and design of our EST-NSS. In Section 3, we present the dataset selection and experimental setup, followed by a comparison of experimental results from different methods. Finally, in Section 4, we summarize the contributions to this work.

1 Related Work

1.1 Nighttime semantic segmentation

Most existing research[24-28] focused on semantic segmentation in bright conditions, with limited attention to more challenging nighttime scenarios. Early studies[7-10] adapted models trained on daytime scenes by creating intermediate domains. Dai et al.[8] proposed a strategy to reduce the burden of manually annotating nighttime images by leveraging abundant daytime annotations. Subsequent research[13] explored training image transformation networks to generate synthetic datasets, and Sakaridis et al.[29] inferred real labels using cross-temporal correspondences, creating the DarkZurich dataset. But the DarkZurich dataset has fewer pixel-wise annotations. Meanwhile, these methods[14-15,30] require additional preprocessing, are time-consuming, and rely on intermediate transition stages, while struggling to generate semantically consistent transformed images when the gap between daytime and nighttime scenes is large. To address these issues, researchers have begun exploring optimization solutions. For example, Wu et al.[7] introduced DANNet, which adapts the Cityscapes dataset to the DarkZurich-D and DarkZurich-N datasets through multi-objective adjustment. Lee et al.[30] used GPS-guided learning to generate pseudo-labels for training segmentation networks. Xia et al.[31] noted that the low dynamic range of traditional frame-based cameras results in poorer nighttime image quality, characterized by reduced color contrast and detail, hindering effective boundary differentiation. Xie et al.[16] proposed a night-focused network (NFNet) that enhances unsupervised domain adaptation by leveraging prompt image guidance and pseudo-label fusion via domain similarity guidance. With the development of specialized datasets such as NightCity[17] and ACDC Night[32], image quality and scene diversity have significantly improved, providing a solid foundation for nighttime semantic segmentation methods.

1.2 RFAConv

RFAConv[19] is an innovative technique designed to overcome the limitations of existing spatial attention mechanisms, which struggle to capture spatial variations across different positions. Built upon the receptive-field attention (RFA) mechanism, RFAConv addresses these challenges by dynamically generating spatial features associated with the size of the convolutional kernels. The receptive field's spatial features are composed of non-overlapping sliding windows, each representing a receptive field block. Based on this structure, RFAConv is proposed as a novel convolutional operation. During its construction, group convolution is employed to efficiently extract spatial features from these receptive fields.

The process of RFAConv can be summarized as follows. First, the input feature map is divided into multiple non-overlapping sliding windows, each representing a receptive field block to model local spatial structures. Then, the receptive-field attention mechanism is used to dynamically generate spatial features for each window, capturing the importance and spatial variations across different regions. Finally, group convolution is applied in parallel within each receptive field block to efficiently extract features, thereby enhancing local representation and improving overall spatial awareness.

In our work, RFAConv is used to construct a convolutional pyramid for end-to-end image enhancement. RFAConv captures spatial features at different scales, effectively enhancing key regions of the image and providing stronger feature support for subsequent semantic segmentation tasks.

1.3 Attention mechanisms

Attention mechanisms[20] have become a key component in modern deep learning models, particularly in tasks such as image recognition and segmentation. The fundamental idea behind attention mechanisms is to allow the model to focus on the most relevant parts of the input, enhancing the representation of important features while suppressing irrelevant ones.

One common approach is channel attention, which focuses on capturing the importance of different channels in the feature map. This is achieved by applying global max pooling and global average pooling along the spatial dimensions, followed by a shared multilayer perceptron (MLP) to generate attention maps, which are then combined to calculate the channel attention weights. In addition to channel attention, spatial attention is also crucial for emphasizing spatial regions of interest. It operates by first applying global max pooling and global average pooling along the channel dimension, then concatenating the resulting maps along the channel axis and passing them through a convolutional layer to produce a combined feature map. The final spatial attention weights are derived by applying a Sigmoid activation function.

In our work, attention mechanisms are utilized to filter the extracted features, enabling the model to place greater focus on the important information.

1.4 Depthwise separable convolution

Depthwise separable convolution[22] is a widely adopted technique aimed at reducing the computational cost and parameter complexity of standard convolutional layers. Depthwise separable convolution factorizes a standard convolution into two separate operations: a depthwise convolution, which applies a single convolutional filter per input channel, and a pointwise convolution, which uses a 1×1 kernel to linearly combine the output of the depthwise stage. This decomposition significantly reduces both computation and model size. In semantic segmentation tasks, depthwise separable convolutions have been integrated into lightweight backbones and decoders to maintain high segmentation accuracy while reducing inference latency.

In our work, Depthwise separable convolution is employed to build an enhancement module, SCBlock, which reduces computational cost while effectively capturing local spatial features. This helps improve feature representation in challenging nighttime scenes.

1.5 Swin Transformer

Transformer[21] is a neural network architecture based on self-attention mechanisms, designed to effectively capture long-range dependencies in sequential data. Natural language processing was the first field to use the Transformer. The significant success of Transformers in the language domain has inspired researchers to explore their application in computer vision.

Vision Transformer (ViT), developed by Dosovitskiy et al.[33], demonstrated promising results in picture categorization. Inspired by ViT, researchers began creating a number of new paradigms based on its model to replace standard CNNs[34-35]. ViT struggles with demanding output-based vision tasks, such as object recognition and segmentation, because of the computing resource limits involved with processing high-definition pictures[36]. Therefore, Liu et al.[18] proposed the Swin Transformer, which uses a shifted-window strategy to constrain attentional computations to fixed-size windows. This approach employs sliding-window operations, including non-overlapping local windows and overlapping cross-windows, to enhance efficiency and effectiveness in processing visual data. These qualities enable the Swin Transformer to reduce computational complexity while increasing efficiency. In addition, the Swin Transformer exhibits a reduced dependency on large datasets compared to other ViT-based models, making it particularly suitable for nighttime urban streetscapes. Swin Transformer preserves sensitivity to image location information while reducing computational costs by performing multi-head self-attention computations within windowed partitions, as opposed to the more computationally intensive approach of Vision Transformers. The attention computation strategy significantly reduces the computational burden, while the shifting window mechanism allows interactions between adjacent windows, thereby enhancing global modeling capability.

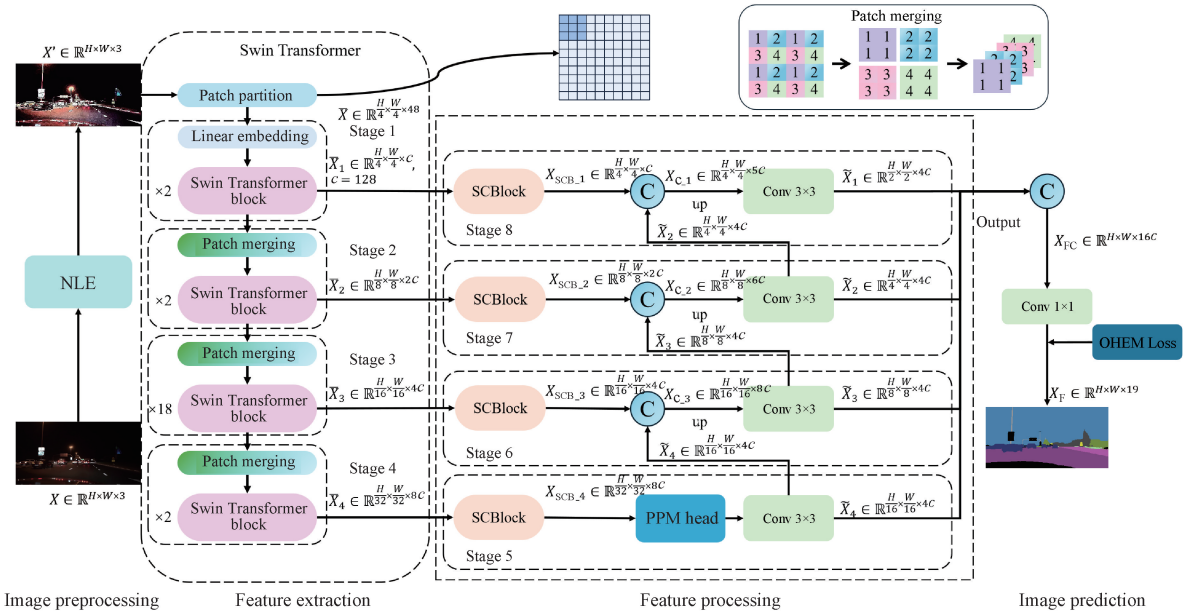
In our work, the Swin Transformer is used as the backbone network for its ability to generate hierarchical feature maps through CNN-like downsampling. Compared with Swin Transformer, EST-NSS structurally designs two key modules, NLE and SCBlock, specifically for nighttime semantic segmentation. NLE is designed by integrating Transformer, attention mechanisms, and RFAConv to enhance image brightness, contrast, and semantic representation, thereby avoiding the omission of important information during feature extraction. SCBlock is designed by integrating attention mechanisms and depthwise separable convolution to enhance feature maps at multiple scales. The fused multi-scale features enable the model to integrate contextual information across different levels, which helps further optimize the segmentation boundaries.

2 Methods

To address the low brightness and contrast of nighttime segmentation, we designed an EST-NSS model, which uses Swin Transformer as the backbone network. Additionally, an NLE is designed for image enhancement in the preprocessing stage, and an SCBlock is designed for feature enhancement in the feature processing stage, for better semantic segmentation of nighttime images.

2.1 Network architecture

The entire structure of the EST-NSS is shown in Fig. 1, and the symbols in Fig. 1 are defined in Table 1.



PPM—pyramid pooling module; Conv—convolution.

Fig. 1 General architecture of EST-NSS

Table 1 Definitions of symbols in Fig. 1

Symbol	Definition
$\boldsymbol{X}$	Input image, $\boldsymbol{X} \in \mathbb{R}^{H\times W\times 3}$
$\boldsymbol{X}'$	Image enhanced by NLE, $\boldsymbol{X}' \in \mathbb{R}^{H\times W\times 3}$
$\overline{\boldsymbol{X}}_i$ (i=1, 2, 3, 4)	Feature map output by the ith stage of Swin Transformer. Spatial size and channels: $i=1$: $\frac{H}{4}\times\frac{W}{4}\times C$ (here, $C=128$); $i=2$: $\frac{H}{8}\times\frac{W}{8}\times 2C$; $i=3$: $\frac{H}{16}\times\frac{W}{16}\times 4C$; $i=4$: $\frac{H}{32}\times\frac{W}{32}\times 8C$
$\boldsymbol{X}_i$ (i=1, 2, 3, 4)	Input feature map of the SCBlock at stage i, which is identical to the output of the Swin Transformer stage i, i. e., $\overline{\boldsymbol{X}}_i$. Its spatial resolution and channel dimension are the same as those of $\overline{\boldsymbol{X}}_i$
$\boldsymbol{X}_{\mathrm{SCB}_i}$ (i=1, 2, 3, 4)	Feature map enhanced by SCBlock at stage i, with the same spatial size and channels as $\overline{\boldsymbol{X}}_i$
$\tilde{\boldsymbol{X}}_i$ (i=1, 2, 3, 4)	Intermediate feature maps during decoder upsampling (e. g., $\tilde{\boldsymbol{X}}_4$ is the result of the bottom feature after PPM and a 3×3 convolution)
$\boldsymbol{X}_{\mathrm{C}_i}$ (i=1, 2, 3, 4)	Concatenated feature map at decoder stage i, formed by concatenating the upsampled feature map $\tilde{\boldsymbol{X}}_{i+1}$ with the SCBlock-enhanced feature map $\boldsymbol{X}_{\mathrm{SCB}_i}$
$\boldsymbol{X}_{\mathrm{FC}}$	Final concatenated (fused) feature map with a size of $H\times W\times 16C$
$\boldsymbol{X}_{\mathrm{F}}$	Final output feature map (after convolution, adjusting channels to 19 classes) of size $H\times W\times 19$
H	Height of the input image in pixels
W	Width of the input image in pixels
C	Base number of channels in stage 1, here, C=128 (i. e., the channel dimension of the Swin Transformer stage 1 output)
$\mathbb{R}$	The set of real numbers. It denotes the numeric space to which all feature map elements belong

As shown in Fig. 1, there are four steps in the EST-NSS: image preprocessing, feature extraction, feature processing, and image prediction.

At the first step, image preprocessing, an input image $\boldsymbol{X}$ is sent into the NLE module to enhance the visibility and semantic representation capabilities of the image in nighttime conditions. The output of the NLE module is denoted as $\boldsymbol{X}'$.

At the second step, feature extraction, a standard Swin Transformer[18] model is used as the backbone network. In this step, $\boldsymbol{X}'$ is first divided into multiple patch tokens by using patch partition, resulting in a feature map $\overline{\boldsymbol{X}} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 48}$, where each patch has a channel dimension of $4\times4\times3=48$.

In stage 1, $\overline{\boldsymbol{X}}$ is projected into a higher-dimensional space via linear embedding, and then processed by a Swin Transformer block, resulting in a feature map $\overline{\boldsymbol{X}}_1$. The channel dimension in stage 1 is denoted as C. In stage 2, patch merging layers are employed to down-sample the spatial resolution by a factor of two while simultaneously increasing the channel dimension, and the Swin Transformer block handles the spatial dependencies of the feature map. The subsequent stage 3 and stage 4 follow the same procedure.

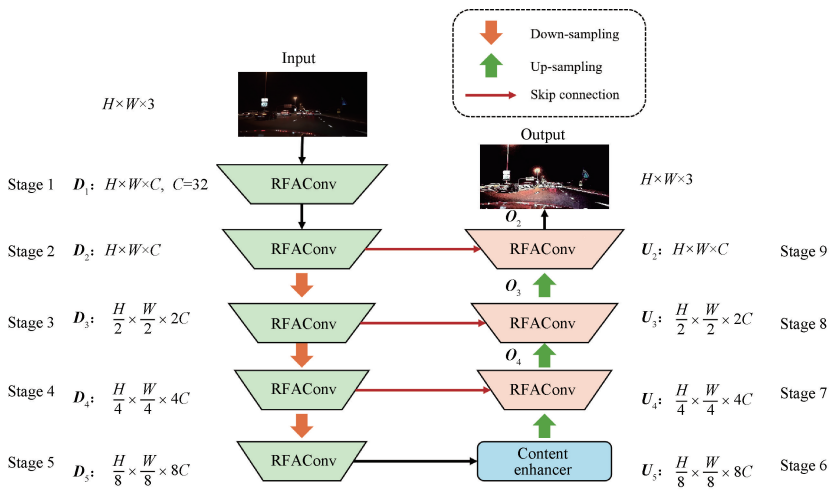
In the third step, feature processing, each of the four stages first applies the SCBlock for feature enhancement. The SCBlock does not alter the spatial resolution or the number of channels of the feature map. In stage 5, for the enhanced feature map of the bottom layer, $\boldsymbol{X}_{\text{SCB_4}}$, a PPM head performs multi-scale pooling to capture contextual information at various scales. Then, a 3×3 convolution is used to adjust the feature map, doubling the resolution while keeping the channel dimension fixed at $4C$, resulting in an intermediate feature map $\tilde{\boldsymbol{X}}_4$. In stage 6, upsampling is applied. The intermediate feature map $\tilde{\boldsymbol{X}}_4$ is concatenated with the feature map $\boldsymbol{X}_{\text{SCB_3}}$ from the previous layer, producing a feature map $\boldsymbol{X}_{\text{C_3}}$ that has the same spatial size but an increased channel dimension. The subsequent stage 7 and stage 8 follow the same procedure.

At the fourth step, image prediction, features from different levels are fused and upsampled to the original resolution, yielding the final concatenated feature map $\boldsymbol{X}_{\text{FC}}$. This is followed by a convolution operation to generate the final feature map $\boldsymbol{X}_{\text{F}}$, where the number of channels is adjusted to match the number of categories. The segmentation results are obtained using the OHEM loss function, which enhances the training process by focusing on hard examples.

2.2 NLE module

The NLE module integrates RFAConv, attention mechanisms, and Transformer architecture to preprocess images. Nighttime images often involve complex scenes, such as streets and intersections, requiring the network to capture a broader receptive field. The use of these techniques enhances contextual understanding and spatial feature attention within this receptive field, allowing the network to effectively extract meaningful feature information from larger areas.

The entire architecture of the NLE and its inner implementation are shown in Fig. 2.



$\boldsymbol{D}_i$ —the output feature map of the ith downsampling stage ($i=1, 2, 3, 4, 5$), corresponding to stages 1–5. The resolution is halved and the channel dimension is doubled at each stage ($i=2, 3, 4, 5$), while stage 1 adjusts the channel number from 3 to 32.

$\boldsymbol{U}_j$ — the output feature map of the jth upsampling stage ($j=2, 3, 4, 5$), corresponding to stages 9, 8, 7, and 6, respectively. The resolution is progressively restored and the channel dimension is halved at each stage. $\boldsymbol{O}_k$ — the fused feature map after concatenating $\boldsymbol{U}_k$ and $\boldsymbol{D}_k$ at the skip connection stage k, i.e., $\boldsymbol{O}_k = \text{Concat}(\boldsymbol{U}_k, \boldsymbol{D}_k)$.

Fig. 2 Architecture of NLE

As shown in Fig. 2, NLE is a convolutional network composed of RFAConv and the content enhancer, enabling NLE to perform end-to-end enhancement. NLE includes the content enhancer to improve the network's ability to gather data at the global and local scales. In the first RFAConv layer, the input image size is not altered but is used to adjust the channel number from 3 to 32 to capture low-level features, which correspond to stage 1. In stage 2, the second RFAConv layer further enhances the feature representation and semantic information while keeping the channel dimension unchanged.

In the following stages 3, 4, and 5, downsampling is used to compress feature maps and progressively enlarge the receptive field to capture global information. The image resolution is halved at each stage, and the channel dimension is doubled. The output resolutions of these stages are $\frac{H}{2}\times\frac{W}{2}\times 2C$, $\frac{H}{4}\times\frac{W}{4}\times 4C$, and $\frac{H}{8}\times\frac{W}{8}\times 8C$, respectively.

In stage 6, the feature map is sent to the content enhancer, which refines these features without altering the image size or the number of channels. Following this, in stages 7, 8, and 9, upsampling operations are performed, where the feature map size is restored to twice the original size, and the number of channels is halved.

Meanwhile, skip connections are introduced to connect the upsampled and downsampled feature maps. This approach helps preserve fine-grained spatial details and improves boundary accuracy, which is particularly critical for nighttime scenes with low brightness and contrast.

The connection between feature maps at each layer is expressed as

$$\boldsymbol{O}_k = \boldsymbol{U}_k \oplus \boldsymbol{D}_k,\ 2 \leqslant k \leqslant 4, \tag{1}$$

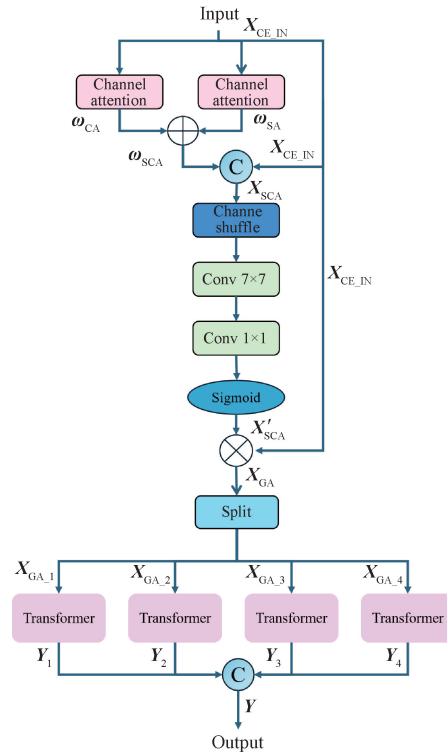
where $\boldsymbol{U}_k$ and $\boldsymbol{D}_k$ represent the feature maps produced during the upsampling and downsampling stages of the kth layer, respectively. These feature maps are concatenated to form the fusion result $\boldsymbol{O}_k$ at each layer. The final output is derived from the result of the last layer of upsampling.

The content enhancer module amplifies global features and filters essential content information using spatial and channel attention mechanisms. Additionally, it utilizes the self-attention mechanism of Transformers to capture and enhance features from various local regions. The structure of the content enhancer is illustrated in Fig. 3.

The input feature map $\boldsymbol{X}_{\text{CE_IN}}$ is put into the content enhancer in the last layer of the RFA convolutional network to extract content features, which contain most of the semantic information and contextual relationships of the image. The data are concatenated to produce the total attention weight $\boldsymbol{\omega}_{\text{SCA}}$:

$$\boldsymbol{\omega}_{\text{SCA}} = \boldsymbol{\omega}_{\text{SA}} + \boldsymbol{\omega}_{\text{CA}}, \tag{2}$$

where $\boldsymbol{\omega}_{\text{SA}}$ and $\boldsymbol{\omega}_{\text{CA}}$ represent the weights of spatial attention and channel attention, respectively.



$\boldsymbol{X}_{\text{CE_IN}}$ —the input feature map to the content enhancer within NLE, with a size of $\frac{H}{8}\times\frac{W}{8}\times 8C$; $\boldsymbol{X}_{\text{GA}}$ —the global attention weighted feature map.

Fig. 3 Structure of content enhancer

To calculate the initial weighted feature map $\boldsymbol{X}_{\text{SCA}}$, multiply $\boldsymbol{\omega}_{\text{SCA}}$ by the original feature map $\boldsymbol{X}_{\text{CE_IN}}$, as shown in Eq. (3).

$$\boldsymbol{X}_{\text{SCA}} = \boldsymbol{\omega}_{\text{SCA}} \times \boldsymbol{X}_{\text{CE_IN}}. \tag{3}$$

The module then uses Channel Shuffle to rearrange the feature map such that information travels between various channel groups based on $\boldsymbol{X}_{\text{SCA}}$. The restructured result is further broadened by a 7×7 convolution of the receptive field to capture global information. Then, the convolution result is processed by a 1×1 convolution layer, which restores the number of channels to the original while keeping the spatial resolution unchanged. Finally, the convolution result is passed through a Sigmoid activation function to generate the feature map $\boldsymbol{X}'_{\text{SCA}}$.

The restructured attention-enhancing feature map $\boldsymbol{X}'_{\text{SCA}}$ is expressed as

$$\boldsymbol{X}'_{\text{SCA}} = \sigma(F_{\text{Conv1}}(F_{\text{Conv7}}(F_{\text{Rearrange}}(\boldsymbol{X}_{\text{SCA}})))), \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function; $F_{\text{Conv1}}(\cdot)$ represents the convolution of 1×1; $F_{\text{Conv7}}(\cdot)$ represents the convolution of 7×7; $F_{\text{Rearrange}}(\cdot)$ means channel shuffle.

The feature map $\boldsymbol{X}'_{\text{SCA}}$ is concatenated with the original feature map $\boldsymbol{X}_{\text{CE_IN}}$ to obtain the global attention weighted feature map $\boldsymbol{X}_{\text{GA}}$:

$$\boldsymbol{X}_{\mathrm{GA}} = \boldsymbol{X}_{\mathrm{CE_IN}} \oplus \boldsymbol{X}'_{\mathrm{SCA}}, \tag{5}$$

where $\oplus$ denotes channel-wise concatenation of feature maps.

Subsequently, the feature map $\boldsymbol{X}_{\mathrm{GA}} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 8C}$ is split into four sections, each of which $\boldsymbol{X}_{\mathrm{GA}_i} \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C}$ is transmitted through the Transformer, as shown as Eq. (6), and employs the self-attention mechanism to improve local information extraction. In the end, the processed feature maps $\boldsymbol{Y}_i \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C}$ are concatenated to produce the global feature map $\boldsymbol{Y}$:

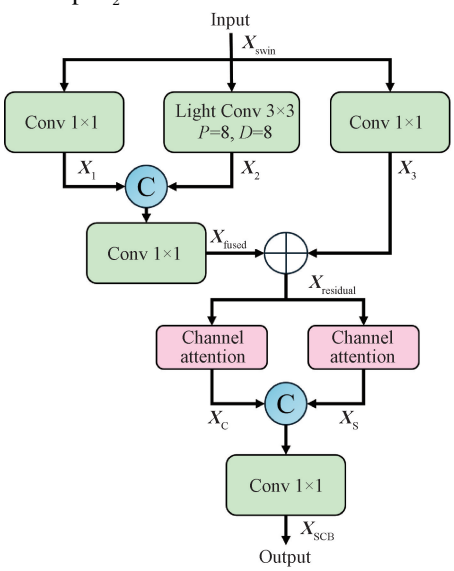
$$\boldsymbol{Y}_i = F_{\mathrm{Transformer}}(\boldsymbol{X}_{\mathrm{GA}_i}),\ 1 \leqslant i \leqslant 4, \tag{6}$$

$$\boldsymbol{Y} = \sum_{i=1}^{4} \boldsymbol{Y}_i,\ 1 \leqslant i \leqslant 4. \tag{7}$$

2.3 SCBlock

The SCBlock was designed to improve the network's ability to capture fine-grained features and redistribute feature weights, thereby strengthening the focus on key information while preserving the original semantic information by integrating attention mechanisms and depthwise separable convolution.

The implementation of SCBlock is depicted in Fig. 4. In Fig. 4, three branches, $\boldsymbol{X}_1$, $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$, are constructed for the input feature map $\boldsymbol{X}_{\mathrm{swin}} \in \mathbb{R}^{H\times W\times C}$. The $\boldsymbol{X}_{\mathrm{swin}}$ is inserted into two 1×1 normal convolutions to generate the feature maps, referred to as $\boldsymbol{X}_1$ and $\boldsymbol{X}_3$. Meanwhile, $\boldsymbol{X}_{\mathrm{swin}}$ is fed into a 3×3 LightConv to generate the feature map $\boldsymbol{X}_2$.



P—padding size (pixels); *D*—dilation rate.

Fig. 4 The implementation of SCBlock

This LightConv consists of two stages: first, a 1×1 convolution is applied to map the input features to different channels and perform nonlinear activation, enhancing the interaction between channels. Then, the depthwise separable convolution is used to extract spatial local structure information. This process is formulated as

$$\boldsymbol{X}_2 = F_{\mathrm{DSConv}}(F_{\mathrm{Conv1}}(\boldsymbol{X}_{\mathrm{swin}})), \tag{8}$$

where $F_{\mathrm{DSConv}}(\cdot)$ stands for depthwise separable convolution; $F_{\mathrm{Conv1}}(\cdot)$ stands for a 1×1 convolution.

The $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are concatenated along the channel dimension, and then a 1×1 convolution is applied to fuse the channels, resulting in the shallow fused feature map $\boldsymbol{X}_{\mathrm{fused}}$:

$$\boldsymbol{X}_{\mathrm{fused}} = F_{\mathrm{Conv1}}(\boldsymbol{X}_1 \oplus \boldsymbol{X}_2). \tag{9}$$

The shallow fused feature X_{fused} is combined with the third branch feature map X_3 through a residual connection to obtain the residual fused feature map X_{residual}:

$$\boldsymbol{X}_{\mathrm{residual}} = \boldsymbol{X}_{\mathrm{fused}} + \boldsymbol{X}_3. \tag{10}$$

In the attention enhancement stage, the residual fused feature map $\boldsymbol{X}_{\mathrm{residual}}$ is element-wise multiplied with the channel attention weights and spatial attention weights, respectively, to obtain the intermediate feature maps $\boldsymbol{X}_{\mathrm{C}}$ and $\boldsymbol{X}_{\mathrm{S}}$.

Finally, in the feature reconstruction stage, these two attention-enhanced feature maps are concatenated and passed through a 1×1 convolution to generate the final output $\boldsymbol{X}_{\mathrm{SCB}}$:

$$\boldsymbol{X}_{\mathrm{SCB}} = F_{\mathrm{Conv1}}(\boldsymbol{X}_{\mathrm{C}} \oplus \boldsymbol{X}_{\mathrm{S}}). \tag{11}$$

3 Experiments

3.1 Dataset and evaluation metrics

NightCity is a large-scale dataset specifically designed for nighttime urban street scenes, providing pixel-level annotations for supervised semantic segmentation tasks. The dataset includes 19 annotated classes, with a total of 2 998 images allocated for training and 1 299 images reserved for validation or testing purposes. The NightCity dataset is utilized for both model training and performance evaluation.

For all experiments, we adopt the mIoU as the evaluation metric, which measures the accuracy of the segmentation by calculating the ratio of the intersection of predicted and ground truth regions to the union of these regions. Specifically, for each semantic class, the intersection over union (IoU) is computed as the ratio of the intersection area to the union area between the predicted region and the ground-truth region. A higher value indicates better performance.

3.2 Experimental settings

We developed the EST-NSS model using PyTorch, utilizing a single RTX 4090D GPU, and based it on the Swin Transformer architecture. The network was trained using the AdamW optimizer with parameters $\beta \in \{0.90, 0.99\}$ and a weight decay of 0.1, starting with an initial learning rate of 6×10^{-5}. Our data augmentation strategy included random cropping to a size of 512 with a scale

range of 0.5 to 1.0, random scaling between 0.5 and 2.0, resizing images to 1 024 × 512 dimensions, and random horizontal flipping with a probability of 0.5.

3.3 Comparison with state-of-the-art methods

We evaluated the EST-NSS framework on the NightCity-val dataset and compared it with several state-of-the-art methods. We selected three mainstream semantic segmentation models: RefineNet[25], PSPNet[37], and DeepLabV2[24]. We also included the latest PIG[16] model based on unsupervised domain adaptation, which included networks based on the DAFormer[38] and HRDA[39] backbones. Additionally, we compared the transformer-based models: MaskFormer[26], Mask2Former[27] and SegFormer[28], as well as the baseline Swin Transformer[18]. The results are shown in Table 2; the best outcomes are shown in bold.

Table 2 Results of EST-NSS compared to those of other methods

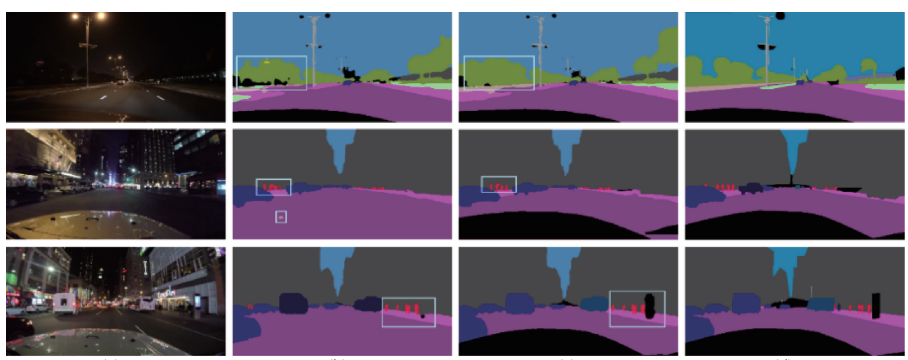
Method	IoU/%																			mIoU
	Road	SK	BD	Wall	Fence	Pole	TL	TS	VT	TR	Sky	Person	Rider	Car	Truck	Bus	Train	MC	BC	/%
RefineNet	75.2	35.1	50.2	25.3	28.6	32.4	27.7	29.1	41.6	10.8	1.3	33.9	**52.6**	58.3	8.1	0.7	30.3	**37.8**	11.4	31.1
PSPNet	77.3	28.6	53.7	17.9	26.5	29.3	27.4	27.1	56.1	17.9	24.1	34.7	27.4	51.8	2.3	0.0	64.2	26.5	13.8	31.9
DeepLabV2	79.7	31.8	56.0	18.3	27.1	24.4	16.3	27.5	41.3	16.2	21.0	34.6	39.8	60.1	7.9	0.1	63.8	34.5	9.2	32.1
DAFormer	85.8	25.7	70.4	19.1	10.3	31.2	18.4	28.4	16.9	4.0	15.5	49.6	9.3	69.8	61.2	50.7	20.1	16.5	26.2	33.1
HRDA	**88.6**	35.5	67.8	16.4	19.0	29.2	23.7	41.8	16.8	17.1	13.1	52.2	14.3	77.2	65.9	61.9	29.2	28.3	31.4	38.4
PIG (DAFormer)	86.9	27.2	75.9	16.1	11.1	32.0	20.0	41.3	47.0	11.3	82.1	48.1	6.8	73.6	62.8	61.2	36.0	20.5	18.7	41.0
PIG (HRDA)	87.1	26.9	75.1	14.3	12.2	26.8	23.0	44.7	46.8	11.1	81.8	46.5	2.8	69.8	65.5	52.2	33.5	30.9	32.2	41.2
MaskFormer	85.5	48.2	81.4	53.1	48.5	37.9	30.3	48.2	58.1	25.1	86.4	52.3	22.8	78.2	58.2	65.0	52.9	21.8	43.1	53.5
Mask2Former	86.3	51.0	82.4	55.8	51.1	**41.3**	31.6	50.2	58.6	23.0	86.5	54.0	19.7	80.4	60.5	64.8	46.2	23.0	40.0	53.0
SegFormer	85.8	48.6	82.0	54.0	50.2	34.8	29.5	50.8	59.6	23.7	86.3	52.5	17.1	79.0	64.5	70.9	57.3	26.3	39.0	53.3
Swin Transformer	86.1	51.8	82.6	57.0	52.2	36.5	32.9	51.7	60.4	25.4	86.6	53.2	19.1	80.6	69.5	71.7	56.4	25.5	34.1	54.4
EST-NSS (ours)	87.6	**56.1**	**84.2**	**63.6**	**59.8**	35.1	**33.0**	**57.9**	**63.1**	**32.4**	**87.6**	**53.6**	18.6	**81.5**	**70.4**	**78.1**	**77.9**	21.2	**39.6**	**58.0**

Note: SK, BD, TL, TS, VT, TR, MC, and BC represent sidewalk, building, traffic light, traffic sign, vegetation, terrain, motorcycle, and bicycle, respectively.

As shown in Table 2, early models such as RefineNet, PSPNet, and DeepLabV2 perform poorly and show limited effectiveness. Domain adaptation-based methods like DAFormer, HRDA, and PIG improve performance to some extent, but their effectiveness is constrained by the significant domain gap between source and target domains. Transformer-based models such as MaskFormer, Mask2Former, SegFormer, and Swin Transformer generally outperform both traditional and domain adaptation methods.

EST-NSS achieved an improvement of 3.6 percentage points on mIoU compared with the baseline and obtained the best performance in 15 out of 19 categories. For static classes like sidewalk, building, wall, and fence, EST-NSS achieved the highest mIoU, demonstrating its strong capability in understanding structured scenes. This is attributed to the NLE, which enhances brightness, contrast, and content while assigning attention weights within the receptive field, ensuring that the network effectively focuses on important regions of the image. For key categories such as person, car, truck, bus, and train, EST-NSS also achieved the highest mIoU. This verifies that SCBlock enhances the model's ability to capture critical information by reweighting feature maps. Specifically, SCBlock refines feature extraction through spatial and channel attention mechanisms, significantly improving segmentation performance. It assigns different weights to spatial positions and adjusts the importance of different channels, enabling the model to focus more precisely on crucial regions.

The visual comparison results between EST-NSS and the baseline are shown in Fig. 5.



(a) (b) (c) (d)

Fig. 5 Visual comparison of EST-NSS: (a) input; (b) Swin Transformer; (c) EST-NSS (ours); (d) ground truth

Compared with the baseline and ground truth in Fig. 5, EST-NSS shows more refined edges and detailed structures, particularly in the first row. It also achieves better segmentation of complex objects, such as crowds and billboards, as seen in the second and third rows. EST-NSS also appears with fewer misclassifications and missed classifications. This demonstrates EST-NSS's ability to capture intricate details and accurately delineate object boundaries, delivering superior performance in challenging scenes.

3.4 Ablation study

To examine the influence of each module on the network, we performed ablation experiments. This section presents an ablation study conducted on the NightCity dataset to evaluate the impact of each module in the EST-NSS architecture. Table 3 summarizes the results of the ablation experiments, showing the effect of adding NLE and SCBlock modules on segmentation performance.

Table 3 Ablation experiment results

Method	mIoU/%
Swin Transformer (baseline)	54.4
Swin Transformer+NLE	57.8
Swin Transformer+SCBlock	55.9
Swin Transformer+NLE+SCBlock	58.0

As shown in Table 3, the baseline Swin Transformer achieves an mIoU score of 54.4% on the NightCity dataset. Adding the NLE enhances images, leading to an improvement in mIoU to 57.8%. Incorporating the SCBlock independently improves the mIoU to 55.9%, as it helps refine fine-grained features and better redistribute feature weights. The combination of both NLE and SCBlock achieves the highest mIoU score of 58.0%, which is an increase of 3.6 percentage points compared with the baseline, demonstrating the complementary nature of these modules and their collective contribution to improved segmentation performance.

We evaluated loss functions for nighttime semantic segmentation using the NightCity dataset.

As shown in Fig. 6, cross entropy loss (CE) achieved an mIoU of 56.7% and struggled with low-contrast features. Dice loss (Dice) improved mIoU to 57.4% by handling class imbalance but faltered in ambiguous scenarios. OHEM outperformed others with an mIoU of 58.0% by focusing on hard samples and minority classes, enhancing fine-grained feature capture in complex nighttime scenes. The mIoU obtained with OHEM increased by 1.3 percentage points compared with CE, underscoring its effectiveness for nighttime semantic segmentation. The focal loss (Focal), which is sensitive to noise, achieved an mIoU of only 56.5%.

After appropriate combination of loss functions, the Focal + CE combination achieved an mIoU of 57.1%, which was comparable to the standalone Dice loss but slightly inferior. The Focal + Dice combination, emphasizing class imbalance handling, reached an mIoU of 57.6%, yet showed only marginal improvement over the standalone Dice loss.

Combining Focal loss and OHEM loss did not exhibit a synergistic effect; the performance dropped by 0.1 percentage points compared with using OHEM loss alone.

The results, summarized in Fig. 6, underline the importance of selecting an appropriate loss function. OHEM demonstrates its ability to balance the focus on hard-to-learn samples and minority classes.

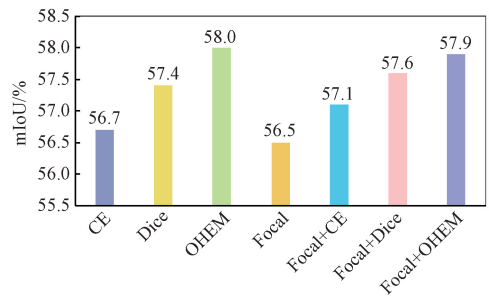


Fig. 6 Results of ablation experiments for loss functions

4 Conclusions

In this paper, we propose a novel nighttime urban streetscape image semantic segmentation network, EST-NSS (enhanced Swin Transformer for nighttime semantic segmentation). Specifically, during the image preprocessing stage, an image preprocessing module, NLE, is designed to enhance brightness, contrast, and semantic representation. Also, a feature enhancement module, SCBlock, is designed to enhance multi-scale feature representations. The whole network is trained in an end-to-end manner, and experimental results on the NightCity dataset show that EST-NSS outperforms existing methods for nighttime city street scenes. In comparison with state-of-the-art models, EST-NSS achieved the best results in 15 out of 19 categories. Additionally, through comparison experiments with different combinations of loss functions, we found that OHEM is more suitable for nighttime semantic segmentation. The NLE module, SCBlock module, and the choice of loss function designed in the proposed EST-NSS network provide valuable insights for nighttime semantic segmentation and similar tasks.

References

[1] Yang L N, Liang X D, Wang T R, et al. Real-to-virtual domain unification for end-to-end autonomous driving [C]// European Conference on Computer Vision. Cham: Springer, 2018: 553-570.

[2] Chen X, Williams B M, Vallabhaneni S R,

et al. Learning active contour models for medical image segmentation [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 11624-11632.

[3] Pang B, Liu G H. SGFNet: an attention-enhanced network for colonic polyp segmentation in endoscopic images[J]. *Journal of Donghua University (English Edition)*, 2026, 43(3): 101-112.

[4] Cordts M, Omran M, Ramos S, et al. The Cityscapes dataset for semantic urban scene understanding [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 3213-3223.

[5] Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite [C]// 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2012: 3354-3361.

[6] Liu W Y, Li W T, Zhu J K, et al. Improving nighttime driving-scene segmentation via dual image-adaptive learnable filters [J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023, 33(10): 5855-5867.

[7] Wu X Y, Wu Z Y, Ju L L, et al. A one-stage domain adaptation network with image alignment for unsupervised nighttime semantic segmentation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(1): 58-72.

[8] Dai D X, Van Gool L. Dark model adaptation: semantic image segmentation from daytime to nighttime [C]// 2018 21st International Conference on Intelligent Transportation Systems. Piscataway: IEEE, 2018: 3819-3824.

[9] Shen F Y, Pataki Z, Gurram A, et al. LoopDA: constructing self-loops to adapt nighttime semantic segmentation [C]// 2023 IEEE/CVF Winter Conference on Applications of Computer Vision. Piscataway: IEEE, 2023: 3255-3265.

[10] Wu X Y, Wu Z Y, Guo H, et al. DANNet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 15764-15773.

[11] Schölkopf B, Platt J, Hofmann T. Analysis of representations for domain adaptation [C]// Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference. Cambridge: The MIT Press, 2007: 137-144.

[12] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation [C]// Proceedings of the 32nd International Conference on Machine Learning (ICML). [S.l.]: JMLR.org, 2015: 1180-1189.

[13] Sakaridis C, Dai D X, Van Gool L. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, 44(6): 3139-3153.

[14] Romera E, Bergasa L M, Yang K L, et al. Bridging the day and night domain gap for semantic segmentation [C]//2019 IEEE Intelligent Vehicles Symposium. (IV). Piscataway: IEEE, 2019: 1312-1318.

[15] Sun L, Wang K W, Yang K L, et al. See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion [PP/OL]. arXiv (2019-08-16) [2025-02-01]. https://arxiv.org/abs/1908.05868#1.

[16] Xie Z F, Qiu R, Wang S, et al. PIG: prompt images guidance for night-time scene parsing [J]. *IEEE Transactions on Image Processing*, 2024, 33: 3921-3934.

[17] Tan X, Xu K, Cao Y, et al. Night-time scene parsing with a large real dataset [J]. *IEEE Transactions on Image Processing*, 2021, 30: 9085-9098.

[18] Liu Z, Lin Y T, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows [C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 9992-10002.

[19] Zhang X, Liu C, Yang D G, et al. RFAConv: innovating spatial attention and standard convolutional operation [PP/OL]. arXiv (2024-04-06) [2025-04-21]. https://arxiv.org/abs/2304.03198.

[20] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module [C]// European Conference on Computer Vision. Cham: Springer, 2018: 3-19.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. Cambridge: The MIT Press, 2017, 30: 5998-6008.

[22] Howard A G, Zhu M, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications [PP/OL]. arXiv (2017-04-17) [2025-04-21]. https://arxiv.org/abs/1704.04861.

[23] Shrivastava A, Gupta A, Girshick R. Training region-based object detectors with online hard example mining [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 761-769.

[24] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018,

40(4): 834-848.
[25] Lin G S, Milan A, Shen C H, et al. RefineNet: multi-path refinement networks for high-resolution semantic segmentation [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 5168-5177.
[26] Cheng B, Schwing A, Kirillov A. Per-pixel classification is not all you need for semantic segmentation [C]// Advances in Neural Information Processing Systems 34 (NeurIPS 2021). Red Hook: Curran Associates, Inc., 2021, 34: 17864-17875.
[27] Cheng B, Misra I, Schwing A G, et al. Masked-attention mask transformer for universal image segmentation[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 1290-1299.
[28] Xie E, Wang W, Yu Z, et al. SegFormer: simple and efficient design for semantic segmentation with transformers [C]//Advances in Neural Information Processing Systems 34 (NeurIPS 2021), 2021, 34: 12077-12090.
[29] Sakaridis C, Dai D X, Van Gool L. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation [C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 7373-7382.
[30] Lee H, Han C, Yoo J S, et al. GPS-GLASS: learning nighttime semantic segmentation using daytime video and GPS data [C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 4003-4012.
[31] Xia R H, Zhao C Q, Zheng M, et al. CMDA: cross-modality domain adaptation for nighttime semantic segmentation [C]//2023 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2023: 21515-21524.
[32] Sakaridis C, Dai D X, Van Gool L. ACDC: the adverse conditions dataset with correspondences for semantic driving scene understanding [C]/2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 10745-10755.
[33] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: Transformers for image recognition at scale [PP/OL]. arXiv (2020-10-22)[2025-04-21]. https://arxiv.org/abs/2010.11929.
[34] Chen H T, Wang Y H, Guo T Y, et al. Pre-trained image processing transformer [C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 12294-12305.
[35] Wolf T, Debut L, Sanh V, et al. Transformers: state-of-the-art natural language processing [C/OL]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2020: 38-45.
[36] Touvron H, Cord M, Douze M, et al. Training data-efficient image transformers & distillation through attention [C]// Proceedings of the 38th International Conference on Machine Learning. [S.l.]: PMLR, 2021, 139: 10347-10357.
[37] Zhao H S, Shi J P, Qi X J, et al. Pyramid scene parsing network [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 6230-6239.
[38] Hoyer L S, Dai D X, Van Gool L. DAFormer: improving network architectures and training strategies for domain-adaptive semantic segmentation [C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 9914-9925.
[39] Hoyer L S, Dai D X, Van Gool L, et al. HRDA: context-aware high-resolution domain-adaptive semantic segmentation [C]//European Conference on Computer Vision. Cham: Springer, 2022: 372-391.

EST-NSS：一种增强的夜间语义分割 Swin Transformer

陶　然[1]，丁　一[1]，卢海伦[1]，芦晓慧[1]，尹文秀[1]，张　琦[2]
1. 东华大学 信息与智能科学学院，上海 201620
2. 上海谋乐网络科技有限公司，上海 201620

摘　要：尽管白天图像的语义分割已经取得了显著成功，但由于夜间图像亮度和对比度较低，夜间图像语义分割仍然面临重大挑战。为此，本文提出了一种用于夜间城市街景图像分割的增强型夜间语义分割 Swin Transformer(EST-NSS)。EST-NSS 以 Swin Transformer 为骨干网络。在图像预处理阶段，我们设计了夜间光增强模块(nighttime light enhancer，NLE)。该模块融合了感受野注意力卷积(receptive-field attention convolution，RFAConv)、注意力机制和 Transformer，以增强图像亮度、对比度和语义表示，从而提升模型对关键区域的感知能力。在特征处理阶段，我们设计了双域上下文单元(称为 SCBlock)。该模块融合了注意力机制和深度可分离卷积，以增强多尺度特征表示，从而提高模型分割物体边界的精度。在 NightCity 数据集上的大量实验结果表明，EST-NSS 显著提升了夜间语义分割的性能。其 mIoU 较 Swin Transformer 提高了 3.6 个百分点。与现有先进模型相比，EST-NSS 在 19 个类别中的 15 个类别上取得了最佳性能。

关键词：语义分割；Swin Transformer；夜间光照增强；特征增强