Enhancing Nighttime Semantic Segmentation with Visual-Linguistic Priors and Wavelet Transform

Jianhou Zhou[1] , **Xiaolong Zhou**[2] , **Sixian Chan**[3]* and **Zhaomin Chen**[4] , **Xiaoqin Zhang**[3]
[1]Hangzhou Dianzi University
[2]Quzhou University
[3]Zhejiang University Of Technology
[4]Wenzhou University
naiive_jou@hdu.edu.cn, xiaolong@ieee.org, sxchan@zjut.edu.cn, chenzhaomin123@gmail.com,
zhangxiaoqinnan@gmail.com

Abstract

Nighttime semantic segmentation is a critical yet challenging task in autonomous driving. Most existing methods are designed for daytime scenarios, resulting in poor nighttime performance due to texture loss and decreased object visibility. Low-light enhancement was applied before segmentation but failed to recover nighttime-specific details, introducing noise or losing delicate structures. Recent work shows that large-scale image-text pairs can effectively leverage natural language priors to guide visual representation, achieving remarkable performance across various downstream visual tasks. However, effectively employing visual-linguistic priors for nighttime semantic segmentation remains underexplored. To address these issues, we propose Text-WaveletFormer, a novel end-to-end framework that integrates text prompts and wavelet-based texture enhancement. Specifically, to compensate for the low recognizability of objects in nighttime scenes, we design a Text-Image Fusion Module (TIFM) to incorporate textual priors to improve nighttime object recognition. In addition, to alleviate the lack of texture details in nighttime conditions, we introduce a Wavelet Guided Texture Amplifier Module (WTAM) to fuse wavelet and raw image features via cross-attention, restoring low-light details. Finally, extensive experiments on benchmarks including NightCity, NightCity-fine, BDD100K, and CityScapes demonstrate our method's superior performance over existing approaches.

1 Introduction

In recent decades, autonomous driving has emerged as a frontier in computer vision, intelligent transportation, and robotics, attracting significant attention. Semantic segmentation, a core technology for perceiving driving scenes [Li *et al.*, 2021; Schutera *et al.*, 2020], is crucial for enabling autonomous driving. However, existing segmentation methods [Chen *et al.*, 2021; Yan *et al.*, 2022] are designed mainly for the day. In low-light environments, these methods fail to address the unique challenges of nighttime settings effectively due to insufficient illumination and the loss of fine details. Although some studies have explored segmentation for nighttime scenes [Xie *et al.*, 2023; Liu *et al.*, 2024], they remain hindered by issues such as detail loss and poor object distinguishability. Therefore, robust segmentation methods tailored for nighttime scenarios are crucial to achieving reliable autonomous driving throughout the day.

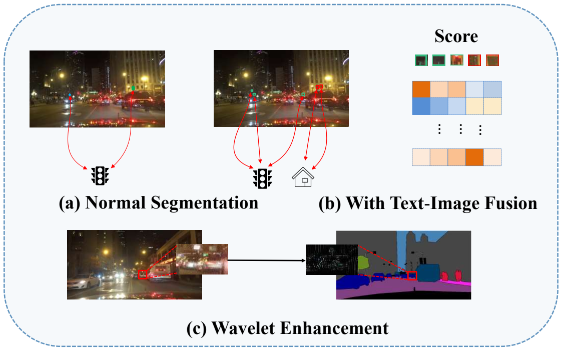

Figure 1: Illustration of our motivation. (a) Conventional segmentation methods relying solely on visual information may fail to match objects accurately under low light. (b) We use nighttime categories as textual prompts to improve object recognition. (c) Wavelet transform is applied to enhance texture details in low-light images.

Let us rethink the challenges faced in nighttime semantic segmentation: (1) Conventional segmentation methods typically use multi-scale visual features extracted by the backbone network as input for the segmentation decoder. However, in low-light or underexposed regions, relying solely on visual features often leads to inaccurate object recognition (as shown in Fig.1 (a), where low-light visual similarity can result in incorrect object matching). This limitation highlights that visual feature extraction alone is insufficient to address the complexity of object recognition. (2) Reduced ambient light in nighttime environments significantly diminishes the discernibility of textures and other details (as shown in Fig.1 (c), where the details of the target texture are difficult to iden-

*Corresponding Author

tify). This poses a challenge for networks to capture critical visual elements. Without accurate texture information, it becomes difficult to perceive foreground objects with distinct semantics in nighttime scenes effectively.

Conventional low-light segmentation methods rely on prior assumptions, multi-stage training, and complex tuning processes, making it challenging to comprehensively address the complexity of nighttime scenes. For example, [Elmahdy *et al.*, 2024] combined a relighting model with a high-resolution network to handle complex nighttime illumination conditions. [Liu *et al.*, 2024]focused on challenging categories by employing a dual-stream network and adaptive probability fusion mechanism. [Chen *et al.*, 2023] designed an adaptive weighted down-sampling layer and interference suppression learning to mitigate the impact of low-light noise on features. Although these methods alleviate some challenges in nighttime scene segmentation, they often require complex tuning during training and struggle to achieve accurate object matching in nighttime scenarios.

Recently, with the emergence of large-scale image-text pretraining models such as CLIP [Radford *et al.*, 2021], the semantic association capability between text-image pairs has provided new solutions for dense prediction tasks (e.g., image segmentation and object detection). [Rao *et al.*, 2022] extended CLIP's image-text matching to pixel-text matching, significantly improving the performance of segmentation models. [Liu *et al.*, 2025] combined linguistic information with visual features, enabling models to identify arbitrary target regions based on textual descriptions. These studies demonstrate that introducing vision-language priorities can significantly enhance the models' ability to understand and offer new perspectives for recognition in nighttime scenarios.

On the other hand, introducing frequency domain information complements the representation of texture details. [Xie *et al.*, 2023] analyzed the differences in the frequency domain distributions between daytime and nighttime images, pointing out that the high-frequency information of nighttime images is lost due to insufficient illumination. [Yang *et al.*, 2024] utilized Haar wavelet transform to decompose features into low-frequency and high-frequency components, which were then combined with spatial features to improve segmentation performance for remote sensing images. These studies indicate that frequency information can effectively compensate for deficiencies in spatial features, offering insights for enhancing texture details in low-light scenarios.

In summary, to address the challenges in semantic segmentation tasks under nighttime scenarios, we propose a Text-WaveletFormer, composed of the Text-Image Fusion Module (TIFM), and the Wavelet-Guided Texture Amplifier Module (WTAM). (1) For the TIFM, we leverage textual descriptions of nighttime categories to provide semantic supplementation and localization support for image segmentation and achieve accurate object matching in low-light environments. It effectively enhances the model's ability to understand target regions. (2) For the WTAM, to capture the weakened texture detail information of the original image, we introduce an attention mechanism to compute the weights for detail enhancement. The local detail features extracted through wavelet transforms are fused with the target features of the original image, thereby amplifying subtle but critical visual features in the original image. It enables the model to capture key information in the foreground more sensitively in low-light conditions. By addressing accurate object recognition and texture detail enhancement, our approach effectively analyzes and segments targets in nighttime scenes, achieving promising results. Our contributions are summarized as follows:

- We propose a novel nighttime semantic segmentation framework, Text-WaveletFormer, which overcomes the limitations of relying solely on low-light image enhancement before segmentation.
- To address the challenge of low object discernibility in nighttime scenes, we introduce the TIFM to integrate textual prompts with visual information, leveraging prior knowledge to enhance the model's ability to recognize objects in nighttime environments.
- To alleviate the issue of texture detail loss under low-light conditions, we design the WTAM to utilize wavelet transforms and attention mechanisms to capture high-frequency components of wavelet-transformed images, thereby enhancing texture details in the image.
- Extensive experiments on various challenging benchmarks show that the proposed method outperforms state-of-the-art nighttime semantic segmentation.

2 Related Works

Semantic segmentation is one of the core tasks in computer vision, aiming to classify each pixel in an image for precise visual understanding. Early models, such as FCN [Long *et al.*, 2015], U-Net [Ronneberger *et al.*, 2015], and SegNet [Badrinarayanan *et al.*, 2017], introduced skip connections and encoder-decoder architectures, significantly improving segmentation accuracy. Later models, including DeepLab [Chen *et al.*, 2014; Chen *et al.*, 2017] and UperNet [Liu *et al.*, 2020], employed dilated convolutions and feature pyramids, further enhancing feature learning and context aggregation capabilities. In recent years, Transformer-based frameworks, such as MaskFormer[Li *et al.*, 2022] and Mask2Former [Cheng *et al.*, 2022], shifted the focus from pixel-level to mask classification, significantly improving instance segmentation accuracy. However, semantic segmentation tasks still face challenges under low-light conditions.

To address this, nighttime semantic segmentation has become a critical research direction. Early studies primarily focused on unsupervised domain adaptation [Wu *et al.*, 2021; Gao *et al.*, 2022; Liu *et al.*, 2023] to bridge the gap between daytime and nighttime data. With the introduction of the NightCity [Tan *et al.*, 2021] dataset, research attention has gradually shifted towards fully supervised learning [Liu *et al.*, 2024; Wei *et al.*, 2023]. Despite these advancements, most existing methods still rely on enhancing the visual quality of nighttime images (e.g., illumination adjustment) before applying general segmentation architectures, which do not fully address the inherent challenges of nighttime scenes. Therefore, this paper proposes a novel semantic segmentation network that leverages category-level text prompts to enhance

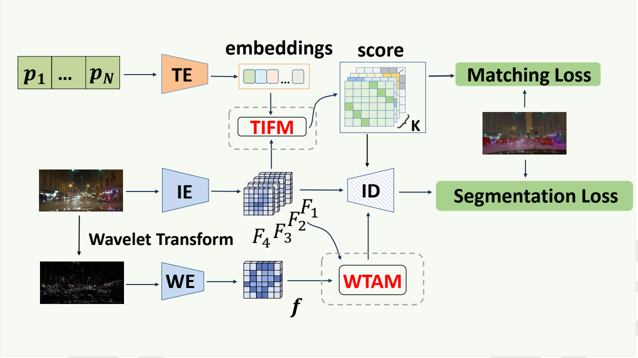

Figure 2: The architecture of the proposed Text-WaveletFormer framework, including a TIFM for accurate target recognition and localization, and a WTAM for enhancing texture details.Where TE, IE, and WE represent the Text Encoder, Image Encoder, and Wavelet Image Encoder, respectively, and ID represents the Image Decoder.

object recognition under low-light conditions, while also improving target texture details to tackle the challenges of nighttime semantic segmentation effectively.

3 Method

3.1 Overview

As illustrated in Fig.2, we propose a novel nighttime semantic segmentation framework, Text-WaveletFormer, which overcomes the limitations of traditional low-light image segmentation through TIFM and WTAM. TIFM aligns category-level text embeddings with multi-scale image features, thereby enabling text semantics-guided segmentation. WTAM leverages multi-scale and wavelet texture features, enhancing image detail reconstruction through multi-head attention.

3.2 Text-Image Fusion

Recent studies [Rao *et al.*, 2022; Zhou *et al.*, 2022a] show that textual information can improve the understanding of images by the model. However, directly transferring CLIP's knowledge to segmentation tasks in nighttime scenarios is prone to environmental noise, resulting in inaccurate segmentation. To address this, we propose the TIFM, as shown in Fig.3, to better leverage language priors and improve segmentation reliability under nighttime conditions.

Given the high-level features $F_4 \in \mathbb{R}^{H_4 \times W_4 \times C_4}$ extracted from the fourth stage of the backbone, where H_4, W_4, and C_4 represent the height, width, and number of channels of the feature map, the TIFM first applies global average pooling to obtain the global feature $\overline{F_4} \in \mathbb{R}^{1 \times C_4}$. This global feature is then concatenated with F_4 along the channel dimension to form $[\overline{F_4}, F_4]$, which is subsequently passed through an MHSA layer for processing, producing the output $[\overline{Z}, Z]$, where $\overline{Z} \in \mathbb{R}^{1 \times C_4}$ and $Z \in \mathbb{R}^{H_4 \times W_4 \times C_4}$.

We construct text prompts for K nighttime semantic segmentation classes. Drawing from DenseCLIP [Rao *et al.*, 2022] and CoOp[Zhou *et al.*, 2022b], we introduce learnable text contexts to enhance category prompts. Backpropagation-driven context optimization improves downstream task transferability, thus modifying text encoder input as follows:

$$[p, e_k], \quad 1 \leq k \leq K \tag{1}$$

where $p \in \mathbb{R}^{N \times C}$ represents the learnable text context, and $e_k \in \mathbb{R}^{C}$ denotes the embedding of K-th class name. The category-specific text features are then extracted using the CLIP text encoder.

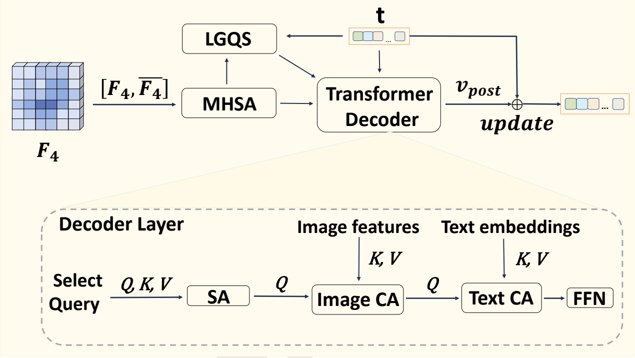

Figure 3: The structure of proposed TIFM. It enhances model's ability to comprehend target regions in nighttime scenarios by facilitating interaction between textual prompts and visual information.

Next, text descriptions provide semantic and localization support for image segmentation, while visual context enhances the accuracy of text representations. Through the cross-attention mechanism in the Transformer decoder, we explore the interaction between vision and language. By introducing a language-guided query selection mechanism, we precisely localize visual objects based on input text, selecting the most relevant visual information as decoder queries. Specifically, given image features $X_I \in \mathbb{R}^{N_I \times d}$ and text features $X_T \in \mathbb{R}^{N_T \times d}$, the Top_N^q operation is applied along the last dimension to select the most relevant features:

$$I_{Nq} = \text{Top}_N^q \left(\text{Max}^{-1} \left(X_I X_T^T \right) \right) \tag{2}$$

The selected feature information is then combined with the visual and textual features and fed into the decoder, where the cross-attention mechanism refines the visual cues:

$$v_{\text{post}} = \text{TransDecoder} \left(I_{Nq}, t, \left[\overline{Z}, Z \right] \right) \tag{3}$$

Furthermore, after identifying the visual cues most relevant to the textual features, the textual features are updated via residual connections:

$$t \leftarrow t + \gamma \cdot v_{\text{post}} \tag{4}$$

where $\gamma \in \mathbb{R}^{C}$ is a learnable parameter that controls the scaling of the residuals. γ is initialized with a very small value(e.g., 10^{-4}) to preserve the linguistic priors in the textual features as much as possible. Finally, the compatibility score between pixels and text is computed using the language-compatible visual features Z and the updated textual features t:

$$\text{score} = \hat{Z} \cdot \hat{t}^T \tag{5}$$

where $\hat{Z}$ and $\hat{t}$ are the L2-norm normalized results of the visual features Z and textual features t along the channel dimension. The score describes the degree of pixel-text matching and can be regarded as a weakly supervised segmentation

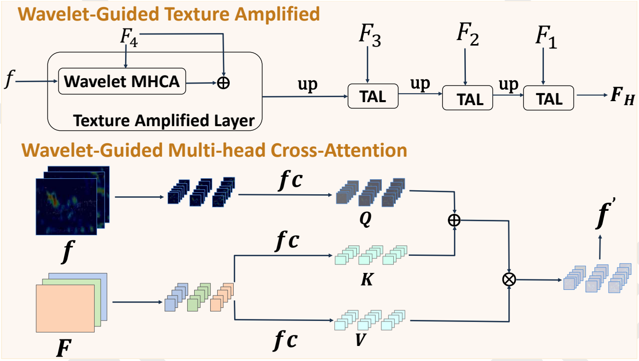

Figure 4: The proposed WTAM is specifically designed to enhance weakened texture details in low-light environments.

result at a low resolution. Therefore, it can be used to calculate an auxiliary segmentation loss. Additionally, the score map can be concatenated with the final feature map to explicitly integrate visual-linguistic priors.

3.3 Wavelet-Guided Texture Amplifier

Since complex details such as edges and textures can be captured through wavelet transform, particularly in low-light or complex background conditions, we design the WTAM to enhance the suppressed texture details in pixel-level features at different resolution scales, as shown in Fig. 4. Specifically, during the image preprocessing stage, given a nighttime image $x \in \mathbb{R}^{H\times W\times 3}$, a two-dimensional discrete wavelet transform (DWT) is performed on each channel $c \in \{R, G, B\}$:

$$(A_c, (H_c, V_c, D_c)) = \text{DWT}(x_c) \tag{6}$$

where H_c, V_c, and D_c represent the horizontal, vertical, and diagonal high-frequency components, respectively, capturing details in different directions. To enhance texture, the low-frequency components A_R, A_G, and A_B from the three channels are averaged to form a new low-frequency component A_{avg}, scaled by a weight λ. Simultaneously, for each channel's high-frequency components, a weight μ amplifies their contributions:

$$A_{\text{avg}} = \lambda \cdot \frac{A_R + A_G + A_B}{3} \tag{7}$$

$$H'_c = \mu \cdot H_c, \quad V'_c = \mu \cdot V_c, \quad D'_c = \mu \cdot D_c \tag{8}$$

So far, the weighted low-frequency and high-frequency components are combined, and an inverse wavelet transform is performed to reconstruct the image for each channel:

$$x'_c = \text{DWT}^{-1}\left(A_{\text{avg}}, (H_c, V_c, D_c)\right) \tag{9}$$

The reconstructed results of the three channels are then combined to produce the final processed image $x' \in \mathbb{R}^{H\times W\times 3}$. The reconstructed image is subsequently fed into a lightweight encoder to extract texture features f.

To obtain fine-grained target information with richer texture details, we first extract multi-scale features from the backbone network at different resolution levels $\{F_1, F_2, F_3, F_4\}$, along with the texture features f. Subsequently, the WTAM is applied at each resolution level to enhance the suppressed texture details in the pixel-level features. Inspired by the complementarity of frequency and spatial features[Xie *et al.*, 2023; Yang *et al.*, 2024], we explore the potential of wavelet transform to amplify texture details for nighttime image segmentation. Given image features $F \in \mathbb{R}^{c_F\times h\times w}$ from the backbone network and texture features $f \in \mathbb{R}^{c_f\times h\times w}$ from the lightweight encoder, we introduce an attention mechanism to compute enhancement weights, fusing wavelet-derived local details with the original image features to restore degraded details in low-exposure conditions better. First, two convolutional layers map F and f to the same dimension c, then reshape them into F_c and f_c for attention input. The query comes from reshaped texture features f_c, while the key and value are derived from the reshaped image features F_c. Formally:

$$Q = W_Q \cdot f_c, \quad K = W_K \cdot F_c, \quad V = W_V \cdot F_c \tag{10}$$

where $F_c, f_c \in \mathbb{R}^{h\times w\times c}$ are the reshaped versions of the inputs, and W_Q, W_K, and $W_V \in \mathbb{R}^{c\times d}$. In the typical attention mechanism, the attention score between the query and the key is computed using a dot product, and the attention weights are obtained through softmax:

$$\text{Attention}(Q, K) = \text{softmax}\left(\frac{Q\cdot K^T}{\sqrt{d}}\right) \tag{11}$$

Unlike standard attention mechanisms that capture limited textural details, we design a wavelet-guided attention mechanism that more finely integrates wavelet textures into target features, effectively amplifying subtle yet crucial visual characteristics of the original image.

$$\text{Attention}(Q, K) = \sum_{c=1}^{d}\left(\frac{(Q' + K')^2}{\sqrt{d}}\right) \tag{12}$$

where $Q', K' \in \mathbb{R}^{hw\times hw\times c}$ are the repeated and expanded versions of Q and K along the second and first dimensions, respectively. Finally, the customized attention weights are applied to the values V, and the result is reshaped to obtain the amplified output:

$$f' = \text{reshape}\left(\text{Attention}(Q, K)\cdot V\right) \tag{13}$$

The wavelet-guided texture amplification process is performed on features at different scales and resolutions. Textures are transferred across resolution scales through upsampling and skip connections until the final high-resolution feature F_H is obtained. The texture information derived from the wavelet transform can extract weakened detail information in nighttime scenes more effectively, enabling the segmentation model to flexibly and coherently interpret low-light images. In this way, the WTAM can capture texture information from coarse to fine, facilitating accurate delineation of objects and regions with varying levels of complexity.

3.4 Training and Inference

For training, we follow the configuration of previous work [Cheng *et al.*, 2022], using cross-entropy loss to constrain

Method	**Backbone**	**Parameters**	**NightCity**	**NightCity-fine**
NightCity[Tan *et al.*, 2021]	ResNet101	84.6M	51.8	55.9
PSPNet[Zhao *et al.*, 2017]	ResNet101	88.3M	46.3	49.5
DeepLabV3+[Chen *et al.*, 2018]	ResNet101	60.1M	54.7	58.8
DANet[Fu *et al.*, 2019]	ResNet101	76.3M	56.0	59.3
NightLab[Deng *et al.*, 2022]	ResNet101	98.5M	55.9	62.3
Mask2former[Cheng *et al.*, 2022]	ResNet101	63.7M	58.9	61.5
DTP[Wei *et al.*, 2023]	ResNet101	63.9M	57.6	60.4
ours	ResNet101	95.2M	**60.9** (**+2.0**)	**63.3** (**+1.8**)
UPerNet[Liu *et al.*, 2020]	Swin-Base	102.5M	57.7	60.5
UPer-Swin[Liu *et al.*, 2021]	Swin-Base	119.9M	58.4	61.1
NightLab[Deng *et al.*, 2022]	Swin-Base	242.4M	59.8	62.3
Mask2former[Cheng *et al.*, 2022]	Swin-Base	107.8M	61.0	63.6
DTP[Wei *et al.*, 2023]	Swin-Base	122.5M	61.2	64.2
ours	Swin-Base	167.6M	**62.8** (**+1.6**)	**65.4** (**+1.2**)

Table 1: Comparison of methods on NightCity and NightCity-fine datasets. Improvements over previous methods highlighted.

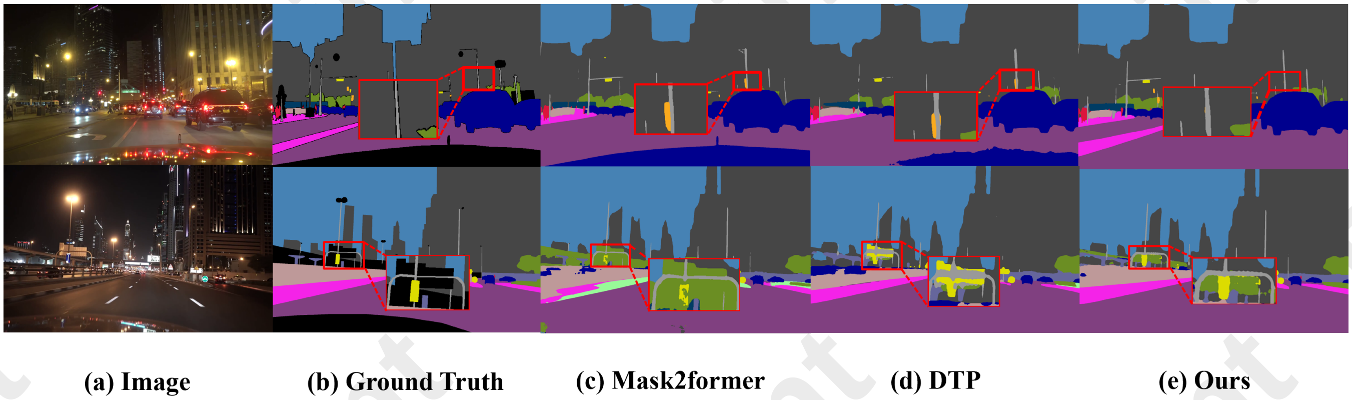
(a) Image **(b) Ground Truth** **(c) Mask2former** **(d) DTP** **(e) Ours**

Figure 5: Qualitative comparison of our Text-WaveletFormer and other methods on the NightCity-fine dataset.

the instance classification scores, and a linear combination of binary cross-entropy loss and Dice loss to constrain the mask predictions. Additionally, the text-image matching score can be considered as weakly supervised segmentation results at low resolution, and during training, cross-entropy loss is used to compute the loss between it with the segmentation labels. This auxiliary segmentation loss is computed only during the training phase.

Method	**Backbone**	**Train on NightCity & CityScapes**	
		NightCity	**CityScapes**
NightCity[Tan *et al.*, 2021]	ResNet101	**53.9**	**76.9**
DeepLabV3+[Chen *et al.*, 2018]	ResNet101	**59.0**	**73.6**
Mask2former[Cheng *et al.*, 2022]	ResNet101	**59.7**	**79.9**
DTP[Wei *et al.*, 2023]	ResNet101	**59.9**	**75.2**
ours	ResNet101	**61.9** (**+2.0**)	**81.2** (**+1.3**)
UPer-Swin[Liu *et al.*, 2021]	Swin-Base	**59.7**	**76.0**
NightLab[Deng *et al.*, 2022]	Swin-Base	**60.2**	**77.1**
Mask2former[Cheng *et al.*, 2022]	Swin-Base	**62.3**	**83.3**
DTP[Wei *et al.*, 2023]	Swin-Base	**63.3**	**78.3**
ours	Swin-Base	**64.1** (**+0.8**)	**84.3** (**+1.0**)

Table 2: Comparison of results on the NightCity and CityScapes datasets. Note that the training process (NightCity and CityScapes) refers to the use of both training sets.

4 Experiments

4.1 Dataset

Following previous practices[Deng *et al.*, 2022; Wei *et al.*, 2023; Cheng *et al.*, 2022], we evaluate the nighttime semantic segmentation performance of our method on four datasets: NightCity, NightCity-fine, CityScapes, and BDD100K.

NightCity [Tan *et al.*, 2021], the largest nighttime semantic segmentation dataset, contains 4,297 real nighttime images, with 2,998 for training and 1,299 for validation. All images are 1024 × 512 in resolution, with 19 categories consistent with CityScapes. It provides valuable annotations for nighttime scene research.

NightCity-fine [Wei *et al.*, 2023] is an improved version of NightCity, correcting annotation errors in both the training and validation sets, with human annotator assistance, resulting in more accurate labels and better evaluation.

CityScapes [Cordts *et al.*, 2016], an autonomous driving dataset, contains daytime images from 50 cities, with 2,975 training and 500 validation images, all at 2048 × 1024 resolution. Following previous setups, as shown in Tab.2, we use only the training set to assist in the training process, serving

Method	**Backbone**	**BDD100K-night**	
		Train on B-N	**Train on B-N & B-D**
NightCity[Tan *et al.*, 2021]	ResNet101	28.4	39.7
DeepLabV3+[Chen *et al.*, 2018]+	ResNet101	30.1	43.4
NightLab[Deng *et al.*, 2022]	ResNet101	31.3	45.1
Mask2former[Cheng *et al.*, 2022]	ResNet101	31.7	50.2
DTP[Wei *et al.*, 2023]	ResNet101	31.4	47.5
Learning Nightime[Liu *et al.*, 2024]	ResNet101	31.3	-
ours	ResNet101	33.6 (**+1.9**)	53.1 (**+2.9**)
UPer-Swin[Liu *et al.*, 2021]	Swin-Base	31.7	48.0
NightLab[Deng *et al.*, 2022]	Swin-Base	35.4	50.4
Mask2former[Cheng *et al.*, 2022]	Swin-Base	34.5	55.1
DTP[Wei *et al.*, 2023]	Swin-Base	36.9	53.1
Learning Nightime[Liu *et al.*, 2024]	Swin-Base	36.6	-
ours	Swin-Base	38.6 (**+1.7**)	57.5 (**+2.4**)

Table 3: Comparison of results under different training processes on the BDD100K dataset. B-N denotes the BDD100K-night training set, while B-N & B-D represents the entire BDD100K training set.

as a reference for the effectiveness of our method.

BDD100K [Yu *et al.*, 2020] is a large-scale driving dataset with various weather conditions, including nighttime (B-N) and daytime (B-D) scenes. We use a subset, BDD100K-night, for supplementary experiments, with 314 nighttime images for training and 31 for validation. The complementary dataset is BDD100K-day.

4.2 Implementation Details

To ensure a fair comparison of model performance, we adopt the commonly used evaluation metric in image segmentation tasks—mean Intersection over Union (mIoU). Our model is implemented using the MMSegmentation framework. During training, we apply the recommended preprocessing methods from MMSegmentation, including mean normalization, random scaling, and flipping, on the NightCity, NightCity-fine, and CityScapes datasets. We use the default CityScapes 90k training configuration with a batch size of 16. For wavelet image reconstruction, $\lambda = 0.1$ and $\mu = 10$, with their effects on segmentation accuracy tested in the ablation experiments. To handle size variations during inference, we use input image versions rescaled by factors of [0.5, 0.75, 1.0, 1.25, 1.5, 1.75]. Additionally, we apply horizontal flipping and average the predictions from all augmented versions.

4.3 Main Results

Quantitative Evaluations. Our method demonstrated superior performance across multiple benchmark datasets, achieving state-of-the-art results on NightCity, NightCity-fine, and BDD100K-night. As shown in Tab.1, using the ResNet101 backbone, our method achieved mIoU scores of 60.9 and 63.3 on the NightCity and NightCity-fine datasets, respectively, improving over existing methods by 2.0 and 1.8%. With the Swin-Base backbone, performance further improves to 62.8 and 65.4. Additionally, as shown in Tab.2, when trained jointly on the NightCity and CityScapes datasets, our method achieved 64.1 mIoU on NightCity and 84.3 on CityScapes, improving by 0.8 and 1.0% over existing methods. These results highlighted our model's ability to generalize across different lighting conditions, effectively integrating knowledge from both day and night images to enhance segmentation performance. Finally, as shown in Tab.3, on the BDD100K-night dataset, our method obtained 38.6 mIoU when trained only on nighttime data (B-N), improving by 1.7 percentage points over existing methods and 57.5 mIoU when trained on the full BDD100K dataset (B-N & B-D), outperforming other methods by 2.4%. These results highlighted the robustness and generalization of our approach in handling complex nighttime segmentation tasks and its ability to generalize across different lighting conditions.

Components		**Train on B-N & B-D**	**Train on NightCity**
WTAM	**TIFM**		
✗	✗	55.1	61.0
✓	✗	56.9 (**+1.8**)	62.1 (**+1.1**)
✗	✓	56.3 (**+1.2**)	62.2 (**+1.2**)
✓	✓	57.5 (**+2.4**)	62.8 (**+1.8**)

Table 4: Ablation of main components on NightCity and BDD100k-night.

Text Prompt	**NightCity-fine**
CLIP	64.8
CLIP+ours	65.4 (**+0.6**)
BLIP	63.9
BLIP+ours	65.0 (**+1.1**)

Table 5: Ablation study on the effectiveness of our text prompt strategy.

Qualitative Results. As shown in Fig. 5, our method exhibited strong segmentation performance across various nighttime scenes. Specifically, the proposed Text-WaveletFormer performed well in most cases, especially in low-light areas, where subtle objects like ”traffic signs” and ”traffic lights” were still accurately detected. In the first row of Fig. 5, other methods misclassified the window next to the pole as a traffic light, while our method correctly identified it. In the second row of Fig. 5, other methods struggled to maintain the integrity of the ”traffic sign” area, whereas our method succeeded.

4.4 Ablation Study

Study on the WTAM. To assess the contribution of Wavelet Enhancement, we conducted experiments on BDD100K-night and NightCity datasets. As shown in Tab.4, without the WTAM, the baseline performance was 55.1 on BDD100K-night and 61.0 on NightCity. After applying the WTAM, the scores increased to 56.9 (↑1.8) and 62.2 (↑1.2), respectively, demonstrating its effectiveness in improving performance on low-light datasets and enhancing segmentation accuracy in complex environments.

Contribution of the TIFM. We further analyzed the contribution of the TIFM module to overall performance. As shown in Tab.4, experiments with and without the TIFM, combined with the WTAM, highlighted its impact. Without the TIFM, the performance improvement was limited, even with the WTAM: on BDD100K-night, the score is 56.9 (↑1.8), and on NightCity, it is 62.1 (↑1.1). With TIFM, performance improves further: on BDD100K-night, the score rises to 57.5 (↑2.4), and on NightCity, to 62.8 (↑1.8). These results underscored the significant contribution of the TIFM module. When combined with the WTAM, the TIFM led to substan-

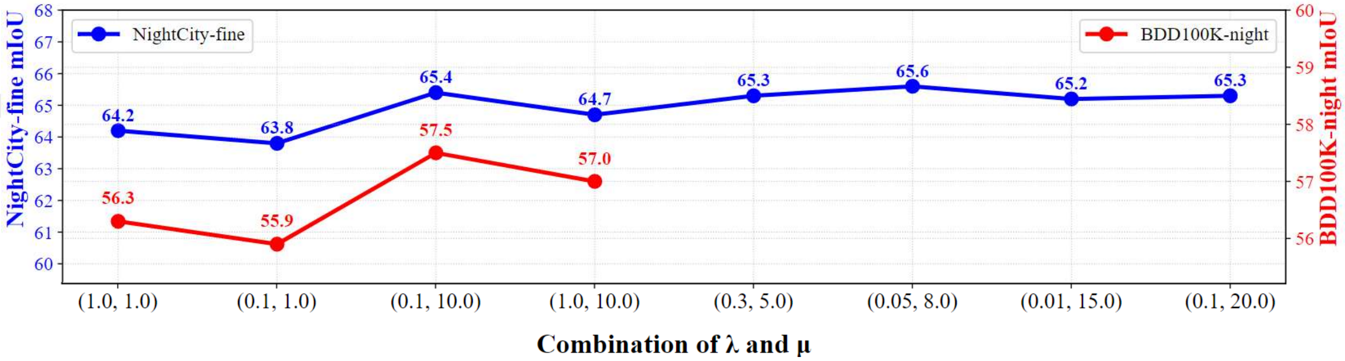

Figure 6: Selection and sensitivity analysis of hyperparameters λ and μ in wavelet transform image reconstruction.

Enhancing Strategies	NightCity-fine
Fourier	64.2
Canny	64.0
HFC	63.9
Sobel	64.4
Wavelet	65.4

Table 6: Ablation study of different enhancement strategies in the texture enlargement module on the Nightcity-fine dataset.

tial performance gains, demonstrating strong synergy on both datasets.

5 Discussion

Does the TIFM better utilize language priors? Previous experiments have demonstrated the effectiveness of our proposed Text-Image Fusion module. However, since DenseCLIP has successfully applied language priors from text encoders to general segmentation tasks, it is important to assess whether this method's applicability is restricted in specific domains, such as nighttime scenes. Specifically, we investigate whether DenseCLIP's direct transfer to nighttime scenes is sufficient and whether the TIFM module better leverages language priors in such scenarios. To answer this, we conducted experiments comparing different models' performance on the NightCity-fine dataset, as shown in Tab.5. We incorporated text encoder knowledge from the base models of CLIP and BLIP and optimized them with the TIFM. Results show that the mIoU score of the CLIP base model improved from 64.8 to 65.4 (+0.6) after adding OURS, while the BLIP base model's mIoU score increased from 63.9 to 65.0 (+1.1). These results indicate that the Text-Image Fusion module better utilizes language priors to nighttime scenes, significantly enhancing model performance. This further validates the effectiveness and applicability of our approach.

Is the WTAM the only effective method for amplifying texture detail information? To assess whether wavelet enhancement was the sole effective method for texture detail amplification, we comparatively analyzed various enhancement strategies on the NightCity-fine dataset, as depicted in Tab.6. Fourier, Canny, and HFC (Gaussian filtering) methods demonstrated comparable performance, yielding mIoU values of 64.2, 64.0, and 63.9, respectively. Sobel marginally improved performance, achieving 64.4. The wavelet method significantly outperformed alternative strategies, scoring 65.4, thereby validating WTAM's efficacy in capturing multi-scale texture details, rendering it particularly well-suited for complex nighttime scenarios.

Analysis of hyperparameters. When selecting λ and μ, we hypothesized the importance of high-frequency information ($\mu > 1$) while reducing the contribution of low-frequency information ($\lambda < 1$), as low-frequency information is somewhat redundant but not entirely negligible. Consequently, we empirically chose the values 0.1 and 10. In preliminary experiments, we attempted to set λ and μ as learnable parameters, but this increased computational overhead without significant performance improvements. Further analysis, as illustrated in Fig. 6, indicates that the combination of hyperparameters that aligns with our assumptions performs better, and minor deviations of λ and μ within the assumed range typically do not substantially impact model performance, demonstrating a lack of sensitivity to hyperparameter selection, consistent with our initial hypothesis.

6 Conclusion

In this work, we proposed a novel nighttime semantic segmentation method, Text-WaveletFormer, which integrated text prompts and wavelet transforms to enhance segmentation performance in low-light environments. The TIFM enabled the model to better understand target regions in nighttime scenes, while the WTAM enhanced high-frequency texture details, thereby improving the quality of visual information. Extensive experimental results demonstrated that Text-WaveletFormer outperformed existing methods in nighttime semantic segmentation tasks, highlighting the potential of text prompts for improving scene understanding in low-light conditions. We hope that Text-WaveletFormer will inspire further research in the field of nighttime semantic segmentation. Future work may involve extending our methods to broader visual degradation scenarios, such as haze or other extreme weather conditions. Explore how language priors can enhance the model's understanding and perception of complex visual environments.

Acknowledgments

This work is partially supported by the the National Key Research and Development Program Project of China (Grant No. 2024YFC3306902), National Natural Science Foundation of China under (Grant No. U24A20242, 61906168, 62272267), the Fundamental Research Funds for the Provincial Universities of Zhejiang(Grant No.RF-A2024013), Zhejiang Provincial Natural Science Foundation of China (Grant No. LY23F020023, LZ23F020001), Construction of Hubei Provincial Key Laboratory for Intelligent Visual Monitoring of Hydropower Projects (Grant No. 2022SDSJ01), and the Hangzhou AI major scientific and technological innovation project (Grant No. 2022AIZD0061).

References

[Badrinarayanan *et al.*, 2017] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[Chen *et al.*, 2014] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1412.7062*, 2014.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Florian Schroff, Hartwig Adam, Jonathan Huang, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):835–846, 2017.

[Chen *et al.*, 2018] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[Chen *et al.*, 2021] Huaian Chen, Yi Jin, Guoqiang Jin, Changan Zhu, and Enhong Chen. Semisupervised semantic segmentation by improving prediction confidence. *IEEE Transactions on Neural Networks and Learning Systems*, 33(9):4991–5003, 2021.

[Chen *et al.*, 2023] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023.

[Cheng *et al.*, 2022] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[Deng *et al.*, 2022] Xueqing Deng, Peng Wang, Xiaochen Lian, and Shawn Newsam. Nightlab: A dual-level architecture with hardness detection for segmentation at night. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16938–16948, 2022.

[Elmahdy *et al.*, 2024] Sarah Elmahdy, Rodaina Hebishy, and Ali Hamdi. Rhrsegnet: Relighting high-resolution night-time semantic segmentation. In *2024 Intelligent Methods, Systems, and Applications (IMSA)*, pages 456–461. IEEE, 2024.

[Fu *et al.*, 2019] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

[Gao *et al.*, 2022] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. Cross-domain correlation distillation for unsupervised domain adaptation in nighttime semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9913–9923, 2022.

[Li *et al.*, 2021] Guofa Li, Yifan Yang, Xingda Qu, Dongpu Cao, and Keqiang Li. A deep learning based image enhancement approach for autonomous driving at night. *Knowledge-Based Systems*, 213:106617, 2021.

[Li *et al.*, 2022] Junnan Li, Wei Yang, Xuming He, et al. Maskformer: Masked image modeling for visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14813–14823, 2022.

[Liu *et al.*, 2020] Xialei Liu, Yuqian Li, Hongwei Wang, Ping Luo, and Shaozi Li. Upernet: Unified perceptual parsing for scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6568–6577, 2020.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[Liu *et al.*, 2023] Wenyu Liu, Wentong Li, Jianke Zhu, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Improving nighttime driving-scene segmentation via dual image-adaptive learnable filters. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10):5855–5867, 2023.

[Liu *et al.*, 2024] Wenxi Liu, Jiaxin Cai, Qi Li, Chenyang Liao, Jingjing Cao, Shengfeng He, and Yuanlong Yu. Learning nighttime semantic segmentation the hard way.

ACM Transactions on Multimedia Computing, Communications and Applications, 20(7):1–23, 2024.

[Liu *et al.*, 2025] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2025.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Rao *et al.*, 2022] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18082–18091, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[Schutera *et al.*, 2020] Mark Schutera, Mostafa Hussein, Jochen Abhau, Ralf Mikut, and Markus Reischl. Night-to-day: Online image-to-image translation for object detection within autonomous driving by night. *IEEE Transactions on Intelligent Vehicles*, 6(3):480–489, 2020.

[Tan *et al.*, 2021] Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson WH Lau. Night-time scene parsing with a large real dataset. *IEEE Transactions on Image Processing*, 30:9085–9098, 2021.

[Wei *et al.*, 2023] Zhixiang Wei, Lin Chen, Tao Tu, Pengyang Ling, Huaian Chen, and Yi Jin. Disentangle then parse: Night-time semantic segmentation with illumination disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21593–21603, 2023.

[Wu *et al.*, 2021] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021.

[Xie *et al.*, 2023] Zhifeng Xie, Sen Wang, Ke Xu, Zhizhong Zhang, Xin Tan, Yuan Xie, and Lizhuang Ma. Boosting night-time scene parsing with learnable frequency. *IEEE Transactions on Image Processing*, 32:2386–2398, 2023.

[Yan *et al.*, 2022] Haotian Yan, Chuang Zhang, and Ming Wu. Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention. *arXiv preprint arXiv:2201.01615*, 2022.

[Yang *et al.*, 2024] Yunsong Yang, Genji Yuan, and Jinjiang Li. Sffnet: A wavelet-based spatial and frequency domain fusion network for remote sensing segmentation. *arXiv preprint arXiv:2405.01992*, 2024.

[Yu *et al.*, 2020] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[Zhou *et al.*, 2022a] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.

[Zhou *et al.*, 2022b] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.