Exploring Reliable Matching with Phase Enhancement for Night-time Semantic Segmentation

Yuwen Pan[1⋆], Rui Sun[1⋆], Naisong Luo[1⋆], Tianzhu Zhang [1,2†], and Yongdong Zhang[1,3]

[1] University of Science and Technology of China
[2] Deep Space Exploration Laboratory
[3] State Key Laboratory of Communication Content Cognition, People's Daily Online
{panyw,issunrui,lns6}@mail.ustc.edu.cn, {tzzhang,zhyd73}@ustc.edu.cn

Abstract. Semantic segmentation of night-time images holds significant importance in computer vision, particularly for applications like night environment perception in autonomous driving systems. However, existing methods tend to parse night-time images from a day-time perspective, leaving the inherent challenges in low-light conditions (such as compromised texture and deceiving matching errors) unexplored. To address these issues, we propose a novel end-to-end optimized approach, named NightFormer, tailored for night-time semantic segmentation, avoiding the conventional practice of forcibly fitting night-time images into day-time distributions. Specifically, we design a pixel-level texture enhancement module to acquire texture-aware features hierarchically with phase enhancement and amplified attention, and an object-level reliable matching module to realize accurate association matching via reliable attention in low-light environments. Extensive experimental results on various challenging benchmarks including NightCity, BDD and Cityscapes demonstrate that our proposed method performs favorably against state-of-the-art night-time semantic segmentation methods.

Keywords: Night-time Semantic Segmentation · Segmentation · Phase Enhancement

1 Introduction

Semantic segmentation is a critical computer vision task essential for scene parsing and image processing. While existing works predominantly focus on visual perception in daytime scenarios [13, 24, 46], practical applications, such as autonomous driving [12, 27], demand robust solutions for challenging night-time environments. Yet the specific environmental characteristics in night scenes such as low visibility and poor exposure render universally generalized image methods

† Corresponding author
⋆ Equal contribution

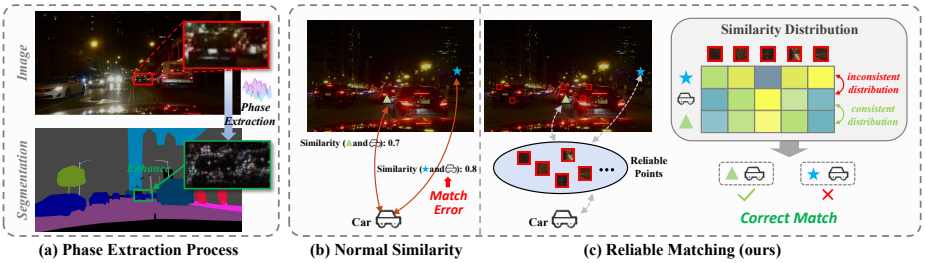

Fig. 1: Illustration of our motivation. (a) Due to poor lighting conditions and blurred details at night, we utilize Fourier phase decomposition to amplify texture information in night images. (b) Normal similarity paradigm tends to directly calculate the similarity between pixels and prototypes, which may lead to mismatching errors due to deceiving surroundings. (c) We propose the reliable attention with adaptively selected reliable points as bridge to calculate similarity rather than direct semantic-pixel matching, achieving more accurate correlation.

impractical for night-time [9,32], necessitating the development of segmentation methods tailored explicitly for night scenarios.

With significant advances in deep learning (DL), DL-based methods have ushered in new research directions for night-time semantic segmentation. Due to the initial scarcity of night-time image datasets, early approaches [25, 28, 39] adopt unsupervised domain adaptation to extend the scene understanding capabilities with the aid of daytime image datasets. However, significant challenges arise due to the substantial disparity in illumination and exposure conditions between night-time and day-time images, leaving a formidable domain gap that poses difficulties in effective bridging. Recently, in order to compensate for the lack of night image datasets, Nightcity [32] introduces a comprehensive night-time image dataset, elevating the research paradigm from the realm of unsupervised methods to a fully-supervised domain. Several following approaches strategically prioritize the enhancement of night-time images to simulate day-time conditions, avoiding a direct confrontation with the inherent challenges posed by night-time scenarios. Specifically, NightLab [7] migrates the distribution of night-time images to the daytime domain by collecting nighttime-daytime pairs to supervise the proposed light enhancement module before the segmentation process, but inevitably requiring additional data. DTP [38] proposes a light-disentangled strategy that decouples the night-time images into light-invariant and light-specific information and then segments disentangled images, but achieving an optimal disentanglement strategy demands intricate manual tuning. These methods, which rely on supplementary data [7] or manually tuned parameters [38] during training, lack a guarantee of generalizability across diverse scenarios and datasets, and the two-stage manner falls short of ensuring end-to-end optimization. Hence, the exploration of a comprehensive segmentation paradigm tailored for night-time scenarios becomes imperative to address the existing limitations.

To effectively parse the night-time images, we aim to adopt an end-to-end optimized methodology from a new perspective, avoiding the previous practice of forcibly fitting a distribution from night-time to day-time. However, it is non-trivial to achieve this goal without careful consideration of inherent challenges posed by night-time images: (1) Owing to the diminished ambient lighting in nocturnal settings, the discernibility of textures and other intricate details is considerably compromised (the texture of the car is difficult to identify as shown in the left of Fig. 1 (a)), posing a challenge for the network to capture these crucial visual elements. Without accurate texture information, effective perception of foregrounds with distinct semantics in night scenes becomes unattainable. *How to capture texture information in low-light environments* is imperative for comprehensive understanding of night scenes. (2) Traditional transformer-based methods [3, 42] normally acquire similarities directly between pixel and learnable prototypes/classifiers. However, harsh night scenes present challenges with blurred foreground and background contours in dark/underexposed areas. Discerning subtle differences becomes arduous, the normal similarity calculation may lead to erroneous matching (the prototype *car* is deceived by the tricky background $\star$ in Fig. 1 (b)) resulting in mismatches during association matching. *How to resolve association matching errors in night scenes* to achieve object-level perception is highly expected.

In this paper, to address the inherent problems in night scenes, we propose NightFormer customized for night-time semantic segmentation, consisting of a pixel-level texture enhancement module and an object-level reliable matching module. (1) In the pixel-level texture enhancement module, to further capture the image details and texture information, we use the phase operation of Fourier transform to focus on the details of the texture in the night scene as shown in Fig. 1 (a). To efficiently integrate the extracted texture information into the target features, we propose an amplified attention mechanism to hierarchically explore inconspicuous objects of all scales in adverse conditions. (2) In the object-level reliable matching module, we first propose a set of prototypes to capture semantic information in the night scenes. And in order to realize accurate association matching in low-light environments, we further design a reliable attention mechanism to adaptively select the dependable key points as the medium, and use them as the bridge to obtain their relationship with the prototypes and the similarity distribution of night-time image features, since similar pixels and prototypes exist in a near-consistent distribution as shown in Fig. 1 (c). In this way, we are able to achieve effective parsing and segmentation of the night scene in both pixel and object-level manners with desirable results.

To sum up, our contributions can be summarized as follows:

- We propose NightFormer specially designed for semantic segmentation in night-time scenes, offering a novel perspective without forcing the night-time images to fit the distribution of the day-time domain.
- We design a pixel-level texture enhancement module to acquire phase-enhanced features via hierarchical amplified attention, and an object-level reliable

matching module to realize accurate association matching via reliable attention in low-light environments.
– Extensive experimental results on various challenging benchmarks including NightCity, BDD and Cityscapes demonstrate that our proposed method performs favorably against state-of-the-art night-time semantic segmentation methods.

2 Related Work

2.1 Semantic Segmentation

Semantic segmentation is a fundamental task in computer vision aimed at understanding and parsing visual information at the pixel level [2, 17, 19–21, 29, 30, 35–37, 41, 43]. With the development of deep learning, Fully Convolutional Networks (FCNs [15]) emerged as pioneering architectures, allowing end-to-end pixel-wise classification. Subsequent models, including U-Net [24] and SegNet [1], introduced innovations like skip connections and encoder-decoder architectures to enhance segmentation accuracy. In well-illuminated scenarios, state-of-the-art methods [18,31,34] leverage powerful deep neural networks such as DeepLab [2] and PSPNet [47]. These models benefit from large-scale datasets, enabling effective feature learning and context aggregation for precise semantic understanding in daytime environments. Though achieving promising results, these methods cannot generalize well in night scenarios due to challenging illumination environments. In this paper, we propose a customized semantic segmentation network to address inherent issues in night-time images.

2.2 Night-time Semantic Segmentation

Night-time semantic segmentation, a critical aspect in computer vision for applications such as autonomous driving in low-light conditions, has garnered significant attention. Early works [10, 25, 26, 28, 39] tend to bridge the gap between daytime and night-time conditions, leveraging unsupervised domain adaptation, with daytime image datasets to enhance the segmentation capabilities in night scenes. Recently, with the introduction of a large night-time dataset proposed by NightCity [32], the task of night-time semantic segmentation has shifted from domain adaptation to a fully-supervised approach. NightLab [7] introduces a dual-level architecture using a light-enhanced module to output augmented images and a hard region detector to optimize difficult part recognition, prompting the vision system but with redundant modules. Furthermore, DTP [38] parses the night scenes by adopting a lighting-disentangled strategy to enhance existing day-time methods for night-time segmentation, which alleviates the limitations of tough illumination conditions at night time to a certain extent. Notably, existing methods tend to map or decouple night-time images into the daytime domain before generating the results with a universal segmentation architecture, leaving the inherent challenges in night scenes unexplored. In this paper, we aim to address the task of night-time semantic segmentation essentially by enhancing texture details and achieving reliable matching in low-light conditions.

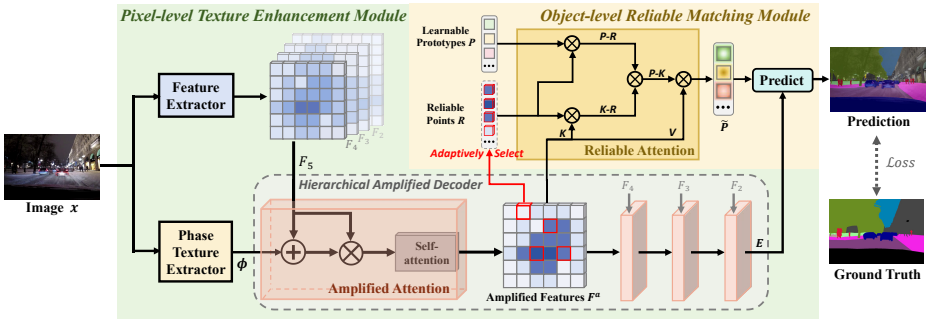

Fig. 2: Framework of our proposed NightFormer. It includes a pixel-level texture enhancement module (Sec. 3.2) to hierarchically aggregate phase texture into target information with amplified attention and an object-level reliable matching module (Sec. 3.3) to realize accurate matching between prototypes and pixels with reliable attention.

3 Method

In this section, we first present the overview of the proposed NightFormer in Sec. 3.1. Then we describe the details of the pixel-level texture enhancement module in Sec. 3.2 and the object-level reliable matching module in Sec. 3.3. Finally, in Sec. 3.4, the training and inference procedure are discussed.

3.1 Overview

As shown in Fig. 2, given an input image $x \in \mathbb{R}^{H\times W\times 3}$, where H and W refer to the height and width of the input, respectively. The pixel-level texture enhancement module extracts the hierarchical pixel embeddings with an image encoder and the texture-aware features with a Fouier-based encoder, the following hierarchical amplified decoder combines them later through amplified attention of multiple scales. The enhanced features F^a are then fed into the object-level reliable matching module to realize accurate perception with reliable points and update the learnable prototypes by empowering them with semantic-aware abilities. Finally, the updated prototypes are interacted with restored high-level features to generate the final segmentation result.

3.2 Pixel-level Texture Enhancement Module

In order to disentangle pixel-level texture and target information fused in the pixel feature, we first model the representation of texture details by exploring Fourier frequency domain. Then we amplify the latent texture information in the acquired pixel-level features with the hierarchical amplified decoder.

Phase Texture Extraction. Existing methods leveraging Fourier Transform typically focus on either optimizing computational costs [5] or transferring styles

between different domains [44]. In contrast, we aim to explore the potential benefits of the Fourier spectrum in segmentation. In the frequency domain, it is known that the phase component of Fourier spectrum preserves high-level statistics information, which contains essential information about the image structure and texture [44]. Therefore, we can utilize the phase of Fourier spectrum to enhance blurred /compromised details in night-time images. Specifically, in the phase texture extractor, given a night-time image $x \in \mathbb{R}^{H\times W\times 3}$, we apply a two-dimensional Fourier transform $\mathcal{F}(x)$ as:

$$\mathcal{F}(x)_{u,v} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j} e^{-J2\pi\left(\frac{ui}{H}+\frac{vj}{W}\right)}, \tag{1}$$

where J refers to the imaginary unit. Then we can acquire the corresponding amplitude $\mathcal{A}$ and phase Φ as:

$$\mathcal{A}(x)_{u,v} = \left|\mathcal{F}(x)_{u,v}\right|, \tag{2}$$

$$\Phi(x)_{u,v} = \arg(\mathcal{F}(x)_{u,v}) = \arctan\left[\frac{\mathrm{Im}(\mathcal{F}(x)_{u,v})}{\mathrm{Re}(\mathcal{F}(x)_{u,v})}\right], \tag{3}$$

where Im and Re represent the imaginary and real part of $\mathcal{F}(x)$, respectively. In this way, to finally generate our phase texture map, we fix the amplitude to an average constant c^a and apply inverse Fourier transform to acquire the phase texture map as

$$\overline{\Phi}(x) = \mathcal{F}^{-1}[\Phi(x)_{u,v} e^{-Jc^a}], \tag{4}$$

where $\mathcal{F}^{-1}$ denotes the inverse Fourier transformation. Then, we can obtain the phase characteristic ϕ through a light-weight encoder (*e.g.*, ResNet-18 [11]), preparing for the following amplification of extracted pixel-level features.

Hierarchical Amplified Decoder. To acquire fine-grained target information with more texture details, we first obtain the features from different stages of the backbone network $\{F_5, F_4, F_3, F_2\}$, and then generate the corresponding amplified pixel-level features through the amplified attention mechanism at each stage.

Inspired by the success of Transformer architecture in discovering local regions, we further explore the potential of attention mechanism in enhancing phase texture information for night-time segmentation. Given an image feature $\mathbf{F} \in \mathbb{R}^{h\times w\times c}$ extracted from the feature extractor, we amplify the phase characteristics (*i.e.*, texture information) ϕ to better restore compromised details in undesirable exposure conditions through a novel **amplified attention** mechanism. First, we employ two convolution layers to map $\mathbf{F}$ and ϕ to the same dimension C, then we can obtain $\overline{\mathbf{F}} \in \mathbb{R}^{h\times w\times C}$ and $\overline{\phi} \in \mathbb{R}^{h\times w\times C}$, respectively. Rather than affinities as vanilla transformer attention mechanism [33], we custom design the amplified attention to integrate phase texture into target features in a finer manner, prompting the vision system to highlight salient regions in

night-time images. Specifically, we acquire the *amplified map* $\mathbf{A} \in \mathbb{R}^{h \times w}$ from pixel-level features and phase characteristic as:

$$\mathbf{A}_{i,j} = \sum_{c=1}^{C} (\overline{\mathbf{F}}_{i,j,c} + \overline{\phi}_{i,j,c})^2, \tag{5}$$

where i, j, and c are the index of height, width, and channel, respectively. Finally, we can get the amplified pixel features $\mathbf{F}^a$ by weighting the *amplified map* $\mathbf{A}$ to original image features in a pixel-wise manner as:

$$\mathbf{F}^a = \mathbf{F} \circ \mathbf{A}, \tag{6}$$

where $\circ$ denotes the element-wise product. Besides, we use a self-attention layer for further aggregating target information across different pixels.

The amplification process is repeated with upsampling operations until the final high-resolution feature $\mathbf{E}$ is obtained. Compared with boundary detection, utilizing texture information obtained from phase can more effectively mine the detailed information in night scenes, guiding the model to parse these scenes in a unified and flexible manner. In this way, the hierarchical amplified decoder is able to capture texture information ranging from coarse to fine details, beneficial for accurately delineating objects and regions with varying complexities. Experiments show that this multi-level amplified design contributes considerable performance gain to final semantic segmentation as shown in Tab. 5.

3.3 Object-level Reliable Matching Module

To effectively aggregate target information with different semantics, we learn a set of prototypes $\mathbf{P} = \{\boldsymbol{p}_n\}_{n=1}^{N}$, where $\boldsymbol{p}_n \in \mathbb{R}^{1 \times L}$ and N denotes the number of prototypes. The prototypes, *i.e.*, learnable query vectors, can absorb class-wise knowledge via cross-attention in a dynamic manner [3, 4, 22]. They can evolve into a compact and distinct representation for each semantic class and serve as anchors around vision features, facilitating more effective aggregation. For each layer in the object-level reliable matching module, these learnable prototypes are first fed into a self-attention layer, where all keys, queries and values arise from initial prototypes to incorporate the local context in night-time images. As the surrounding backgrounds can be deceptive due to the tough illumination environments at night, the direct similarity calculation is susceptible to background pixel interference, resulting in erroneous segmentation. Thus we design a novel reliable attention mechanism to find related reliable points as bridge to acquire more accurate correlations.

Reliable Attention Mechanism. Given the amplified feature $\mathbf{F}^a$ derived from the pixel-level texture enhancement module, the queries arise from the prototypes $\mathbf{P}$, and keys and values arise from the input features $\tilde{\mathbf{F}}^a = [\boldsymbol{f}_1^a; \boldsymbol{f}_2^a; ...; \boldsymbol{f}_{hw}^a]$ (flattened $\mathbf{F}^a$). Formally,

$$\mathbf{Q}_n = \boldsymbol{p}_n \mathbf{W}^Q, \mathbf{K}_m = \boldsymbol{f}_m^a \mathbf{W}^K, \mathbf{V}_m = \boldsymbol{f}_m^a \mathbf{W}^V, \tag{7}$$

where $n \in [1, \ldots, N]$, $m \in 1, 2, ..., hw$ and $\mathbf{W}^Q \in \mathbb{R}^{C \times C_k}$, $\mathbf{W}^K \in \mathbb{R}^{C \times C_k}$, $\mathbf{W}^V \in \mathbb{R}^{C \times C_v}$ are linear projections. We then can obtain the similarity between $\mathbf{Q}_n$ and $\mathbf{K}_m$ as:

$$sim_{n,m} = \frac{\exp\left(\beta_{n,m}\right)}{\sum_{m=1}^{hw} \exp\left(\beta_{n,m}\right)}, \beta_{n,m} = \frac{\mathbf{Q}_n \mathbf{K}_m^\top}{\sqrt{C_k}}. \tag{8}$$

Direct similarity calculation is unreliable due to the high similarity between foreground pixels with different semantics in night scenes, especially in under-exposed regions. We aim to find a reliable medium that can be used as a basis to build a matching bridge between prototypes and pixels. The reliable score for each pixel can be obtained via the weighted sum over all similarities as:

$$score_m = \sum_{n=1}^{N} sim_{n,m}, m \in 1, 2, ..., hw, \tag{9}$$

where the top-K pixels are selected with the largest correlations of semantics to be reliable points $\mathbf{R}$. The sum of similarities between a pixel and all prototypes $score_m$ can be regarded as an aggregated measure of confidence/reliability. Each similarity score indicates the degree to which a pixel's feature vector aligns with a prototype. By summing these similarities, we capture the overall confidence/reliable score (how well a pixel's feature vector is semantically consistent with multiple prototypes), which implies that the pixel tends to be well-represented within the semantic space spanned by the prototypes, making it a reliable candidate for accurate discrimination. The corresponding features of reliable points $\mathbf{R}$ can be denoted as $\mathbf{F}^R = \{\boldsymbol{f}_n^R\}_{k=1}^K$. Then we calculate the prototype-reliable similarity and the pixel-reliable similarity respectively as same as Eq.(8):

$$\begin{aligned} Sim_n^q &= \text{softmax}(\frac{(\boldsymbol{p}_n \mathbf{W}^Q)(\mathbf{F}^R \mathbf{W}^K)^\top}{\sqrt{C_k}}), \\ Sim_m^k &= \text{softmax}(\frac{(\boldsymbol{f}_m^a \mathbf{W}^Q)(\mathbf{F}^R \mathbf{W}^K)^\top}{\sqrt{C_k}}). \end{aligned} \tag{10}$$

And then, we can obtain the similarity between prototypes and pixels with reliable points as:

$$Sim_{n,m}^{qk} = Sim_n^q (Sim_m^k)^\top, \tag{11}$$

which is used to acquire more accurate correlations. Finally, the updated prototypes can be acquired by blending values with the reliable similarity $Sim_{n,m}^{qk}$ as:

$$\tilde{\boldsymbol{p}}_n = \sum_{m=1}^{hw} Sim_{n,m}^{qk} \mathbf{V}_m, \tag{12}$$

and following general transformer pipeline [33], we equip updated prototypes with self-attention and FFN at the output of the reliable attention. In this way, the learnable prototypes $\mathbf{P}$ are modified in the object-level reliable matching module via reliable attention and finally evolve into reliable classifiers $\tilde{\mathbf{P}}$.

Table 1: Comparisons of existing night-time semantic segmentation methods and general segmentation approaches on the NightCity [32] and NightCity-fine [38] datasets. The best results are shown in **bold**.

Method	Backbone	Parameters	NightCity	NightCity-fine
NightCity [32]	ResNet101	84.6M	51.8	55.9
PSPNet [47]	ResNet101	88.3M	46.3	49.5
DeeplabV3+ [2]	ResNet101	60.1M	54.7	58.8
DANet [8]	ResNet101	76.3M	56.0	59.3
NightLab [7]	ResNet101	98.5M	55.9	62.3
DTP [38]	ResNet101	63.9M	57.6	60.4
NightFormer (ours)	ResNet101	58.4M	**59.8**(↑2.2)	**62.8**(↑2.4)
UPerNet [41]	Swin-Base	102.5M	57.7	60.5
UPer-Swin [14]	Swin-Base	119.9M	58.4	61.1
NightLab [7]	Swin-Base	242.4M	59.8	62.3
DTP [38]	Swin-Base	122.5M	61.2	64.2
NightFormer (ours)	Swin-Base	111.9M	**63.5**(↑2.3)	**65.9**(↑1.7)

3.4 Training and Inference

With the final high-resolution features $\mathbf{E} \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$ and the learned prototypes $\tilde{\mathbf{P}} \in \mathbb{R}^{N \times C}$ as classifiers, we can finally obtain the segmentation map as:

$$\mathbf{M} = \mathbf{E} \times \tilde{\mathbf{P}}^{\top}. \tag{13}$$

For better training our network, we use the conventionally used loss paradigm [7, 32], including the dice loss and binary cross-entropy loss to supervise mask prediction and the cross-entropy loss for mask recognition.

4 Experiments

In this section, we will first introduce the datasets used in our work in Sec. 4.1. And the implementation details are shown in Sec. 4.2. In Sec. 4.3, we illustrate the specific metric for better evaluation of our method. Then we further analyze the main results including quantitative evaluations and qualitative results in Sec. 4.4. Finally, we ablate the effectiveness of our method in Sec. 4.5.

4.1 Dataset

Following the conventional practice [32,38], there are 4 datasets included to evaluate the night-time semantic segmentation performance of our method: NightCity [32], NightCity-fine [38], CityScapes [6], and BDD100K [45].
NightCity [32] is a dataset containing 4,297 real night-time images, divided into 2,998 train images and 1,299 val images with pixel-level semantic annotations. The labels align with Cityscapes [6] including 19 classes.

Table 2: Comparisons of results on the NightCity [32] and CityScapes [6] datasets. Please note that the training procedure (NightCity&CityScapes) includes both training sets.

Method	Backbone	Trained on NightCity&CityScapes	
		NightCity	CityScapes
NightCity [32]	ResNet101	53.9	76.9
DeeplabV3+ [2]	ResNet101	59.0	73.6
DTP [38]	ResNet101	59.9	75.2
NightFormer (ours)	ResNet101	**62.2**(↑2.3)	**80.3**(↑5.1)
UPer-Swin [14]	Swin-Base	59.7	76.0
NightLab [7]	Swin-Base	60.2	77.1
DTP [38]	Swin-Base	63.3	78.3
NightFormer (ours)	Swin-Base	**65.4**(↑2.1)	**82.1**(↑3.8)

Table 3: Comparisons of results on the BDD100K dataset of different training procedures. B-N denotes the BDD100K-night [7] training set, B-N&B-D denotes the whole BDD100K [45] training set.

Method	Backbone	Trained on B-N	Trained on B-N&B-D
		BDD100K-night	
NightCity [32]	ResNet101	28.4	39.7
DeeplabV3+ [2]	ResNet101	30.1	43.4
NightLab [7]	ResNet101	31.3	45.1
NightFormer (ours)	ResNet101	**32.8**(↑1.5)	**47.5**(↑2.4)
UPer-Swin [14]	Swin-Base	31.7	48.0
NightLab [7]	Swin-Base	35.4	50.4
DTP [38]	Swin-Base	36.9	53.1
NightFormer (ours)	Swin-Base	**38.2**(↑1.3)	**55.4**(↑2.3)

NightCity-fine [38] is a refined version of NightCity with the same images. It identifies mislabeled validation images and re-annotates more accurate labels with the assistance of human labelers for better evaluation.
CityScapes [6] is a commonly used benchmark dataset for semantic segmentation tasks. Please note that urban images in this dataset are mostly in daytime scenes. Following the previous setting [32, 38], we only use the training set of CityScapes to aid the training procedure as a reference to the validity of our method as shown in Tab. 2.
BDD100K [45] is a large-scale, diverse driving dataset with different weather conditions including night-time (B-N) and day-time (B-D). For our supplementary experiments, we only utilize the night images in BDD100K as BDD100K-night following the setting in [7] with a validation set of 58 images, while the other 7000 images are used for training due to the scarcity of night-time images.

4.2 Implementation Details

We adopt Pytorch [23] and Detectron2 [40] to implement the proposed method. 4 NVIDIA GeForce RTX 3090 GPUs are used for training. We consider both ResNet101 and Swin-Base image backbones following [7,38] for better evaluation. The extractor of phase texture is ResNet-18 [11], which is a light-weight CNN encoder. During the training stage, our model is trained with a batch size of 16, using the Adam optimizer [16] with an initial learning rate of 0.0001 for the first 60,000 iterations and 0.00001 for the last 20,000 iterations. The input image is resized to the resolution of $512 * 1024$. We set the number of learnable prototypes as $N = 32$, and the number of reliable points as $K = 40$. We ablate the effects of these super-parameters in detail in our ablation studies as shown in Fig. 4.

4.3 Metric

For a fair comparison, we adopt the metric of mean intersection over union (mIOU), which is a commonly used evaluation metric for image segmentation tasks. It measures the similarity between the predicted segmentation mask and

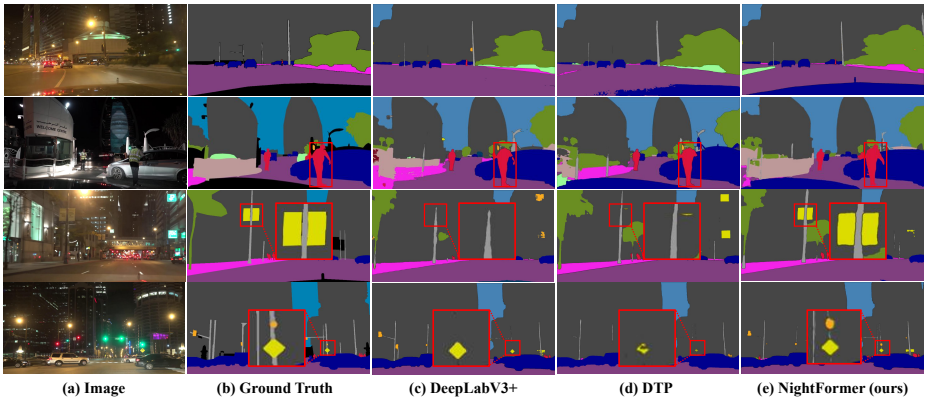

Fig. 3: Comparison of qualitative results of our NightFormer and other methods on the NightCity [32] dataset.

the ground truth mask by computing the ratio of the intersection area to the union area of the two masks for each object class.

4.4 Main Results

Quantitative Evaluations. Our method demonstrates superior performance in night-time semantic segmentation, outperforming state-of-the-art methods, as illustrated in Tab. 1. Evaluation on both the NightCity and NightCity-fine datasets reveals compelling results: a mIOU of 59.8 and 62.8 based on ResNet101, and 63.5 and 65.9 based on Swin-Base. These achievements surpass all existing methods with the same backbone. In supplementary experiments, as shown in Tab. 2, our NightFormer further enhances scene comprehension when applied to both NightCity and CityScapes datasets. It is significant that our method exhibits superior performance not only in night images (NightCity [32] dataset) but also demonstrates substantial improvements on the CityScapes dataset across various illumination conditions. This underscores our network's capability to assimilate knowledge from both night and day images jointly, achieving better information capture from both domains. Besides, as shown in Tab. 3, whether training on night-only images (B-N) or on a mixed night-day image set (B-N&B-D), the results on BDD100k-night [7] further validate our effectiveness.

Qualitative Results. As shown in Fig. 3, NightFormer shows promising segmentation performance in diverse night-time scenes. In specific, our method performs great in most scenarios, especially in the low-light regions. Even with some annotation errors in the ground truth set of NightCity [32], our NightFormer still generates promising results for inconspicuous objects such as "traffic sign" and "traffic light" as shown in the last row in Fig. 3. It can also be observed that in the per-class IOU demonstration in Fig. 5, applying our module designs bring

significant performance improvements for the categories "traffic sign", "person" and "truck".

4.5 Ablation Study

Does the Fourier phase extraction contribute to parsing the night-time images? Yes. As shown in the ablation experiments on main components of our NightFormer in Tab. 4, the incorporation of Fourier phase extraction yields a discernible performance improvement, elevating the mIOU from 61.1 to 63.5 on the NightCity dataset. The inherent low visibility in night-time images can lead to the loss of texture details in certain targets. And the application of Fourier phase extraction proves instrumental in effectively capturing intricate textures. In Fig. 5, we present the detailed per-class IOU comparisons corresponding to Tab. 4. Notably, our approach exhibits the most significant improvement in recognizing small targets (*e.g.*, *traffic sign* and *traffic light*). The phase enhancement promotes the understanding of visual cues, leveraging amplified attention to enrich the model's perceptual capabilities.

Table 4: Ablation of main components on NightCity [32] and BDD100k-night [7].

Main Components		NightCity	BDD100k-night
Phase enhancement	Reliable matching		
×	×	57.8	49.6
✓	×	60.4	52.1
×	✓	61.1	53.4
✓	✓	**63.5**	**55.4**

Is the hierarchical amplified decoder effective for night-time segmentation? Yes. As shown in Tab. 5, substantial performance improvements are evident on both datasets when employing multi-level features compared to utilizing only the bottom embedding. This underscores the effectiveness of our hierarchical fusion design. The rationale behind this success lies in the fact that amplified features at different levels are well-suited for capturing textures at diverse scales. The incorporation of fused hierarchical features proves instrumental in effectively parsing various types of targets in various night-time scenes.

Can the reliable attention improve model performance? Yes. The design of our reliable attention mechanism proves immensely beneficial to the model, particularly when complemented by phase extraction, as shown in Tab. 4. Please note that the absence of reliable matching denotes the utilization of a commonly employed cross-attention mechanism. The adaptive selection of reliable points within this design holds pivotal importance for pixel-level precision in low-visibility or under-exposed regions, which effectively addresses mismatching issues arising from the conventional direct similarity calculation. As shown in Fig. 6, by leveraging the reliable attention mechanism, our learnable prototypes effectively cluster high responses in the foreground area while minimizing attention to irrelevant regions. In contrast, without reliable matching, prototypes tend to adsorb information from irrelevant areas and activate deceiving regions, leading to sub-optimal results. The chosen reliable points also influence the comparison of similarity distributions, as illustrated in Fig. 4 (b).

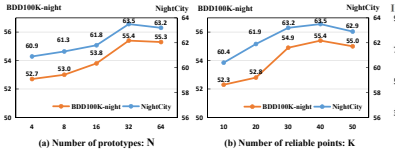

Fig. 4: Ablation of N and K on NightCity [32] and B-N [7].

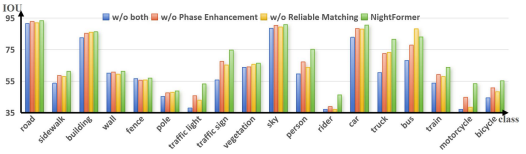

Fig. 5: Demonstration of per-class IOU in Tab. 4 on the NightCity [32] dataset.

Table 5: Ablation of different hierarchical fusion designs in the fusion layer on both NightCity [32] and BDD100K-night [7] datasets.

Fusion Strategy	NightCity	BDD100k-night
$\{F_5\}$	61.5	53.7
$\{F_5, F_4\}$	61.8	53.8
$\{F_5, F_4, F_3\}$	62.9	54.6
$\{F_5, F_4, F_3, F_2\}$	**63.5**	**55.4**

Table 6: Ablation of different enhancing operations in the pixel-level enhancement module on both NightCity [32] and BDD100K-night [7] datasets.

Enhancing Operation	NightCity	BDD100k-night
✘	61.1	53.4
Canny Operator	61.9	54.3
Sobel Operator	62.1	54.4
Fourier Phase	**63.5**	**55.4**

How much does the hyper-parameters affect the model performance? As illustrated in Fig. 4, we conducted ablation experiments focusing on the number of prototypes N and reliable points K. In the object-level reliable matching module, the prototype plays a pivotal role in clustering similar semantics. The selection of a specific number of prototypes is crucial for an effective semantic matching process. The performance peaks when the number of prototypes is set to 32. Deviating towards too many or too few prototypes results in a drop in mIOU performance. In the reliable attention mechanism, the number of reference points, denoted as K, determines how many reliable foreground pixels are selected to establish the correlation between semantics and pixels. Ablating the number of reference points, as depicted in Fig. 4 (b), reveals that performance reaches its top when $K = 40$, signifying that this number is sufficient for achieving the necessary correction. Too many reliable points may overly emphasize model discrimination and include irrelevant background information, ultimately disrupting the similarity distribution.

Is Fourier phase extraction the only way to amplify texture information? In our supplementary experiments, we extended our texture enhancement approaches to include other operations, such as the Canny operator and Sobel operator. As shown in Tab. 6, when applying traditional algorithms, we observed tiny improvement in performance, but not as promising as the enhancement achieved through phase extraction. We attribute this phenomenon to the fact that certain operations may extract evident contours or textures in a conventional manner. However, this type of information has already been effectively parsed during the encoding of the original image. The repetitive inclusion of con-

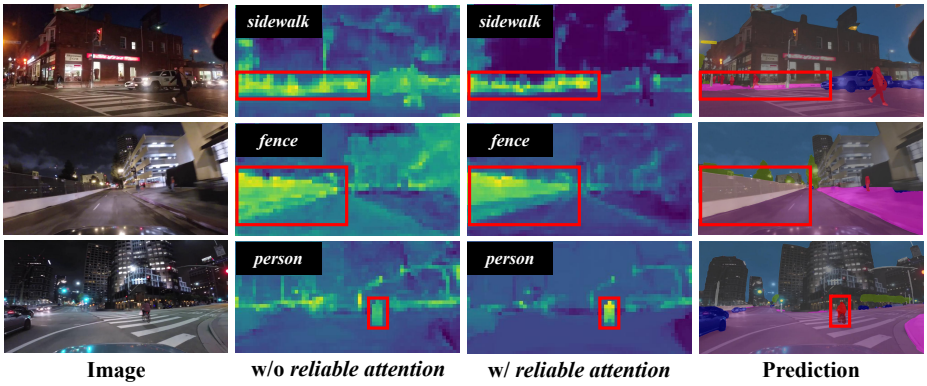

Fig. 6: Activation maps of different semantics with and without reliable attention.

tour information does not contribute significantly. Besides, inherent challenges in night scenes, *e.g.*, abundant noise and poor lighting, result in ambiguous boundary predictions due to deceptive regions. In contrast, Fourier frequency domain decomposition has the capability to disentangle essential information within different domains for nocturnal images and implicitly enhances visual perception, making more substantial contributions to amplified pixel-level features.

5 Conclusion

In this paper, we propose NightFormer for night-time semantic segmentation. Rather than forcing night-time images to conform to the day-time distribution, we aim to efficiently parse night-time scenes by addressing intrinsic challenges such as compromised texture details and mismatching errors. Specifically, we design a pixel-level texture enhancement module that hierarchically aggregates phase texture with amplified attention and an object-level reliable matching module that accurately matches semantics to pixels using reliable attention. Extensive experimental results demonstrate the effectiveness of our proposed NightFormer in night-time semantic segmentation. Exploring the applicability of our method in more challenging scenarios is a promising direction for the future work.

Acknowledgements

This work was partially supported by the National Nature Science Foundation of China (Grant 12150007, 62121002, 62071122), and Youth Innovation Promotion Association CAS.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence **39**(12), 2481–2495 (2017)
2. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
3. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
4. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in neural information processing systems **34**, 17864–17875 (2021)
5. Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. Advances in Neural Information Processing Systems **33**, 4479–4488 (2020)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
7. Deng, X., Wang, P., Lian, X., Newsam, S.: 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. pp. 16938–16948 (2022)
8. Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3146–3154 (2019)
9. Fu, X., Zeng, D., Huang, Y., Liao, Y., Ding, X., Paisley, J.: A fusion-based enhancing method for weakly illuminated images. Signal Processing **129**, 82–96 (2016)
10. Gao, H., Guo, J., Wang, G., Zhang, Q.: 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. pp. 9913–9923 (2022)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Li, G., Yang, Y., Qu, X., Cao, D., Li, K.: A deep learning based image enhancement approach for autonomous driving at night. Knowledge-Based Systems **213**, 106617 (2021)
13. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1925–1934 (2017)
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

16. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
17. Luo, N., Pan, Y., Sun, R., Zhang, T., Xiong, Z., Wu, F.: Camouflaged instance segmentation via explicit de-camouflaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17918–17927 (2023)
18. Luo, N., Sun, R., Pan, Y., Zhang, T., Wu, F.: Electron microscopy images as set of fragments for mitochondrial segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3981–3989 (2024)
19. Mai, H., Sun, R., Wang, Y., Zhang, T., Wu, F.: Pay attention to target: Relation-aware temporal consistency for domain adaptive video semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4162–4170 (2024)
20. Mai, H., Sun, R., Zhang, T., Wu, F.: Rankmatch: Exploring the better consistency regularization for semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3391–3401 (2024)
21. Mai, H., Sun, R., Zhang, T., Xiong, Z., Wu, F.: Dualrel: Semi-supervised mitochondria segmentation from a prototype perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19617–19626 (2023)
22. Pan, Y., Luo, N., Sun, R., Meng, M., Zhang, T., Xiong, Z., Zhang, Y.: Adaptive template transformer for mitochondria segmentation in electron microscopy images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21474–21484 (2023)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
24. Ronneberger, O., Fischer, P., Brox, T.: 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. pp. 234–241. Springer (2015)
25. Sakaridis, C., Dai, D., Gool, L.V.: Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7374–7383 (2019)
26. Sakaridis, C., Dai, D., Van Gool, L.: Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(6), 3139–3153 (2020)
27. Schutera, M., Hussein, M., Abhau, J., Mikut, R., Reischl, M.: 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)
28. Song, C., Wu, J., Zhu, L., Zhang, M., Ling, H.: Nighttime road scene parsing by unsupervised domain adaptation. IEEE transactions on intelligent transportation systems **23**(4), 3244–3255 (2020)
29. Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., Zhang, Y.: Lesion-aware transformers for diabetic retinopathy grading. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10938–10947 (2021)
30. Sun, R., Luo, N., Pan, Y., Mai, H., Zhang, T., Xiong, Z., Wu, F.: Appearance prompt vision transformer for connectome reconstruction. In: IJCAI. pp. 1423–1431 (2023)

31. Sun, R., Wang, Y., Mai, H., Zhang, T., Wu, F.: Alignment before aggregation: trajectory memory retrieval network for video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1218–1228 (2023)
32. Tan, X., Xu, K., Cao, Y., Zhang, Y., Ma, L., Lau, R.W.: Night-time scene parsing with a large real dataset. IEEE Transactions on Image Processing **30**, 9085–9098 (2021)
33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
34. Wang, Y., Luo, N., Zhang, T.: Focus on query: Adversarial mining transformer for few-shot segmentation. Advances in Neural Information Processing Systems **36**, 31524–31542 (2023)
35. Wang, Y., Sun, R., Luo, N., Pan, Y., Zhang, T.: Image-to-image matching via foundation models: A new perspective for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3952–3963 (2024)
36. Wang, Y., Sun, R., Zhang, T.: Rethinking the correlation in few-shot segmentation: A buoys view. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7183–7192 (2023)
37. Wang, Y., Sun, R., Zhang, Z., Zhang, T.: Adaptive agent transformer for few-shot segmentation. In: European Conference on Computer Vision. pp. 36–52. Springer (2022)
38. Wei, Z., Chen, L., Tu, T., Ling, P., Chen, H., Jin, Y.: Disentangle then parse: Night-time semantic segmentation with illumination disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21593–21603 (2023)
39. Wu, X., Wu, Z., Guo, H., Ju, L., Wang, S.: 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. pp. 15769–15778 (2021)
40. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. `https://github.com/facebookresearch/detectron2` (2019)
41. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: Proceedings of the European conference on computer vision (ECCV). pp. 418–434 (2018)
42. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems **34**, 12077–12090 (2021)
43. Xiong, G., Wang, Y., Li, Z., Yang, W., Zhang, T., Zhou, X., Zhang, S., Zhang, Y.: Aggregation and purification: Dual enhancement network for point cloud few-shot segmentation. In: Elkind, E. (ed.) Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. International Joint Conferences on Artificial Intelligence Organization (8 2024)
44. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4085–4095 (2020)
45. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)

46. Yuan, Y., Huang, L., Guo, J., Zhang, C., Chen, X., Wang, J.: Ocnet: Object context network for scene parsing. arXiv preprint arXiv:1809.00916 (2018)
47. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)