Contents lists available at ScienceDirect

Displays

journal homepage: www.elsevier.com/locate/displa

DDFL: Dual-Domain Feature Learning for nighttime semantic segmentation☆

Xiao Lin [a,b,c], Peiwen Tan [a], Zhengkai Wang [a], Lizhuang Ma [d,e], Yan Li [a,*]

[a] *College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Artificial Intelligence Education Research Institute, Shanghai 200234, China*

[b] *Shanghai Engineering Research Center of Intelligent Education and Big Data, Shanghai Normal University, Shanghai 200234, China*

[c] *The Research Base of Online Education for Shanghai Middle and Primary Schools, Shanghai 200234, China*

[d] *College of Computer Science and Technology, East China Normal University, Shanghai 200062, China*

[e] *The Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China*

ARTICLE INFO

Keywords:
Semantic segmentation
Exposure correction
Frequency domain features
Dual domain fusion

ABSTRACT

Nighttime semantic segmentation has been playing a critical role in intelligent transportation, building safety and urban management. However, nighttime scenes present some challenges such as complex structures, multiple light sources, uneven lighting and blurry image noise, which severely degrade the segmentation quality of nighttime images. To address these challenges, we propose a Dual-Domain Feature Learning (DDFL) model for nighttime semantic segmentation. Our approach introduces three innovative ideas. First, we establish an exposure correction module to address the impact of lighting differences on the model's learning, so as to maximally restore the pixel distortion and blurry areas caused by artificial light in nighttime scenes. Second, we incorporate frequency domain information into the nighttime segmentation task to give the model stronger discrimination ability. Finally, we introduce a dual-domain fusion module to complement the information of learning from the spatial and frequency domains in a cross-fusion manner, enabling the network to perceive semantic information while preserving details. The proposed model was experimentally tested on the Nightcity, Nightcity+ and BDD100k datasets. Our results demonstrate that our model outperforms mainstream models, achieving mIoU scores of 56.73%, 57.41% and 28.97%, respectively, under different lighting, image exposure levels, and resolutions. These results show that our model is capable of segmenting nighttime scenes efficiently in a high-quality way.

1. Introduction

Nighttime semantic segmentation has broad applications in various fields such as intelligent transportation [1], security monitoring [2–4], modern healthcare [5,6] and autonomous driving [7,8], providing strong support for the construction of smart cities. Recent efforts [9] have shown that computer vision techniques, deep learning in particular, are possible to achieve precise segmentation and recognition of different objects in nighttime scenes, such as vehicles, pedestrians, and buildings.

Thus, there exists a growing interest in deep-learning methods for nighttime semantic segmentation. Domain adaptation methods [10–12] in nighttime semantic segmentation tasks focused on reducing the differences between daytime and nighttime domains, for example, Wu et al. [10] used a generator and two discriminators to construct an adversarial network that made the intensity distribution and segmentation results of images from different domains similarly, thereby assisting nighttime segmentation with daytime segmentation. Besides, it is also possible to learn about the variance in diff-domain-illumination from another perspective, such as Gao et al. [11] proposed that the same class had differential invariance and used correlation distillation to build a cross-domain adaptive framework that extracted content and style knowledge from the feature space, calculated the difference in illumination between the two domains, and uncovered their inherent differences. [12] further divided the domain adaptation process into two steps to achieve gradual domain adaptation from easy to difficult, reducing distribution divergence in the nighttime domain itself. However, these methods require additional computational resources to achieve domain adaptation, and the training process is complex. Most importantly, these domain adaptation methods ignore the intrinsic differences between complex and exposed nighttime scenes, leading to largely segmentation offset in complex nighttime image exposure.

Another line of research captured some important information in nighttime images to assist in segmentation. Tan et al. [9] introduced an EGNet to predict the exposure location and used the exposure guidance

☆ This paper was recommended for publication by Guangtao Zhai.

* Corresponding author.
E-mail address: yanli@shnu.edu.cn (Y. Li).

https://doi.org/10.1016/j.displa.2024.102685
Received 17 June 2023; Received in revised form 28 February 2024; Accepted 20 March 2024
Available online 26 March 2024

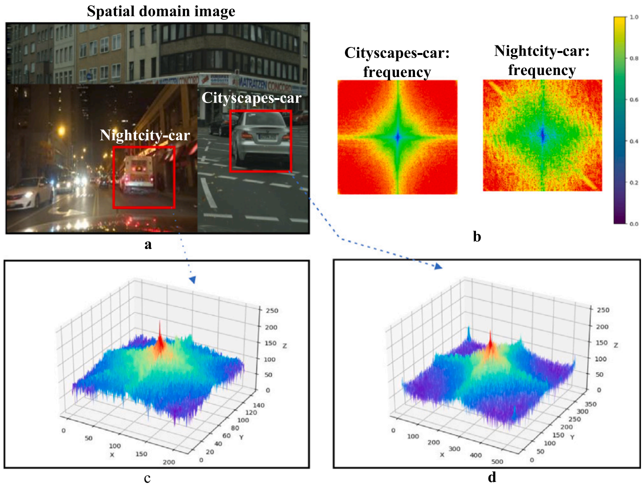

Fig. 1. Comparison between image spatial domain and frequency domain. a: the mixed spatial domain image from Cityspace and Nightcity; b: the 2D top view of the frequency domain features of the Cityscapes-car and Nightcity-car; c and d: the 3D view of the frequency domain images of the Cityscapes-car and Nightcity-car.

layer generated by the prediction results to guide the segmentation flow, effectively distinguishing areas with insufficient/excessive exposure. [13] designed a feature transfer layer that adaptively integrates edge prior knowledge to achieve edge-guided segmentation. Though the simple use of special information contained in nighttime images can reduce the computational requirements to some extent, artificial light in nighttime scenes still affects the imaging clarity of different objects, especially when the clarity is low, the model's credibility in extracting nighttime information is also low.

In the era of big data, leveraging multiple modalities enables models to extract valuable information more rapidly. At present, many fields have harnessed cross-modal information, such as image-text hash retrieval [14,15], thermal-rgb semantic segmentation, video caption [16] and so on. [14] proposed a rapid method for image-text cross-modal hash retrieval, which utilized DenseNet and multi-head attention separately as ImgNet and TxtNet. [17–19] used thermal imaging to achieve nighttime semantic segmentation. Thermal images have some advantages over cameras, such as working in completely dark environments, being insensitive to changes in light and being robust to shadow effects. Motivated by the above observations, RTFNet [17] added thermal imaging features to the RGB encoder and learned supplementary information between different spectra through the well-trained encoder, thereby improving the segmentation quality. Besides, to consolidate the integration of information from both domains, multispectral information combining RGB and thermal imaging was fused to assist with segmentation [18]. Though the thermal image can work in low-light environments, its stability is affected by several factors such as ambient temperature and humidity. Additionally, the resolution of thermal imaging is relatively low, which limits its application in pixel-level tasks. For the case of nighttime image processing, a straightforward solution is to enhance the input of nighttime images and update existing daytime methods for semantic segmentation. For example, NightLab [20] proposed a regularized light adaptation module that corrects the light in both image-level and region-level segmentation recently. In the aspect of image enhancement, researchers have made many attempts, such as training neural networks for restoring image quality from coarse to fine [21–24], such as [23] leveraged CNN to enhance the feature extraction capability and post-processed the output with YUV color space to further correct the whole image. [24] optimized the enhancement process by simulating the conditional distribution of normally exposed images.

Challenges. Currently, night-time scene parsing (NTSS) employs various approaches to enhance its performance. Some models adjust or incorporate convolution and attention on the basis of daytime semantic segmentation (DTSS) to augment recognition in dark image areas. However, deep learning in the spatial domain limits the potential of convolution and attention. Other models attempt domain adaptation on top of daytime semantic segmentation systems to narrow the gap between nighttime and daytime illumination. However, these models are not guaranteed to be effective in real-world applications due to the different illumination distributions in various nighttime scenarios.

The fundamental reason is that exposure levels impact the quality of RGB images. The excessive high-frequency information in the image has led to poor segmentation performance in previous models. As shown in Fig. 1, we provide an illustration of spectrum graph on Cityscapes [25] and Nightcity [9] in 2d (Fig.1b) and 3d (Fig. 1c and d) views. We can see that the high-frequency regions (yellow regions) of the Nightcity-car are more abundant than the Cityscapes-car in Fig. 1b, forming a distinct circle. So much high-frequency component poses a challenge for handling nighttime semantic segmentation. Given these observations, a question arises: can we incorporate the frequency domain into the NTSS task to obtain multidimensional information?

In this paper, we propose dual-domain feature learning that relies on the RGB (spatial) and frequency domain to mimic human visual capabilities [26], enabling our model more sensitive to details and contours. Specifically, we incorporate attention mechanisms into the dual-domain feature-fusion learning process, and adjust the high-and-low frequency weight adaptively. Furthermore, we enhance the spatial domain features through frequency-to-spatial transformation and dual-domain cross-fusion operations. Our design utilizes spatial domain information to learn the capacity of human biological vision and goes beyond frequency domain information, which introduces an innovative and high-performing approach to the night semantic segmentation task. Our whole method is shown in Section 3. In summary, our contribution includes:

- We thoroughly analyzed existing nightly semantic segmentation frameworks, which are unable to handle complex exposures, and we analyzed the spatial and frequency domain features of daytime and nighttime road images to dissect the differences.
- We present a novel, efficient nightly semantic segmentation framework DDFL, which is composed of three key modules. The first module is the **Exposure Attention Correction Module** that repairs image exposure at coarse-and-fine levels with two sub-sections, mitigating complex exposure variations present within images. The second module is the **Frequency Domain Transformation Module** which employs attention mechanisms to adjust the weights of different frequencies adaptively and utilizes a residual module to accelerate convergence, thereby enhancing the coupling between high-and-low frequency domains. The third module is the **Dual-domain Cross-Fusion Module**, which reduces information loss during frequency domain conversion and facilitates cross-domain interaction. Our model is adaptable to varying levels of exposure, demonstrating robustness and flexibility.
- Comprehensive experiments were conducted on various network backbones. The results show that our DDFL achieves state-of-the-art performance in terms of three metrics, which are the prevalent metrics in semantic segmentation.

2. Related work

2.1. Semantic segmentation

Current research mainly focused on three approaches: expanding image resolution [27,28], utilizing contextual information [29–31] and introducing attention mechanisms [32–35]. Regarding expanding image resolution, HRNet [27] connected multiple subnets with different resolutions in parallel to maintain high-resolution representation, and

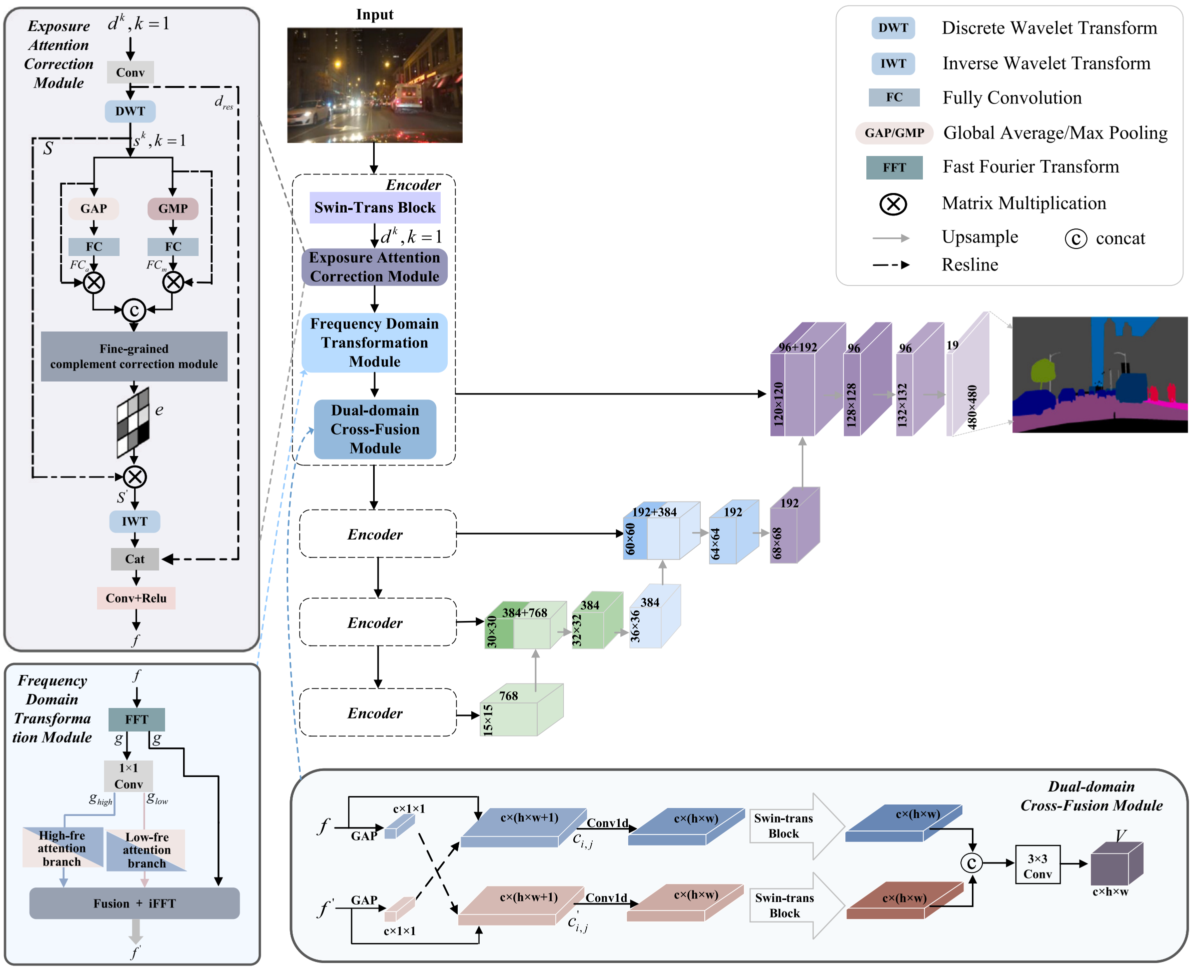

Fig. 2. The overall framework of the night-time semantic segmentation model with dual-domain feature learning. On the top left is the **E**xposure **A**ttention **C**orrection **M**odule (EACM), on the bottom left is the **F**requency **D**omain **T**ransformation **M**odule (FDTM) and on the bottom is the **D**ual-domain **C**ross-**F**usion **M**odule. (DCFM). Our model completes the segmentation task with four encodes and upsampling.

then repeatedly integrated each subnet with other parallel subnets to obtain dense high-level features. Of course, utilizing contextual information has a comparable effect in achieving the desired result. PSPNet [29] further aggregated more spatial features by expanding the receptive field. Recent work [31], such as combining the last layer of deep convolutional networks with a conditional random field, has been proposed to solve inaccurate localization in deep networks. More recently, there are also some people starting to focus on multi-modal semantic segmentation [36–39]. Zhou et al. [37] propose a cross-modal attention fusion module that explores the complementary information between RGB and thermal features. He et al. [38] introduce an SFAF-MA to enhance the ability to aggregate spatial features and fuse modality adaptation. Zhou et al. [39] introduce a DBCNet to extract high-level semantic features with a bilateral fusion of multiscale cross-modal data. However, these existing methods ignore the structural heterogeneity of different modalities, and they impose certain requirements on the dataset.

By contrast, attention mechanisms encoded spatial information over longer distances and decoded it to spatial resolution, forming a new approach to utilizing context. DANet [32] constructed a dual attention network from the perspective of space and channel to capture the dependency between two dimensions. Those methods maximize the information that the given pixel can obtain, capture global and local contextual information, and improve segmentation accuracy in a long-range dependency way. Thus, our work utilizes Swin-Transformer [35] as an auxiliary tool for extracting the superficial features, which has shown good performance in capturing global contextual information. However, a series of problems brought about by complex nighttime exposure still cannot be solved, which prompts us to design various modules to improve its performance.

2.2. Frequency domain in image processing

The application of frequency domain analysis in the field of image processing is commonly seen in fake image recognition [40], image restoration and synthesis [41], image compression [42] and image denoising [43,44]. For instance, fake images generated by various GANs can automatically be recognized through frequency domain analysis [40]. Jiang et al. [41] extensively investigated the frequency domain differences between real and fake images and improved the reconstructed and synthesized image quality by reducing those differences. Xu et al. [42] demonstrated that frequency-domain learning in image preprocessing preserved more effective image information than traditional spatial subsampling methods, and they proposed a dynamic channel selection method to identify trivial frequency components for removing unnecessary redundant information, thus improving segmentation accuracy and computational complexity. Around the same time, MBCNN [43] proposed a prior frequency of texture removal through a Learnable Bandpass Filter (LBF) to solve image restoration tasks involving texture and color recovery. However, these methods cannot compensate for the deficiency in capturing details using frequency domain information alone, so it is necessary to adopt additional spatial information to assist in pixel-level semantic segmentation. Therefore,

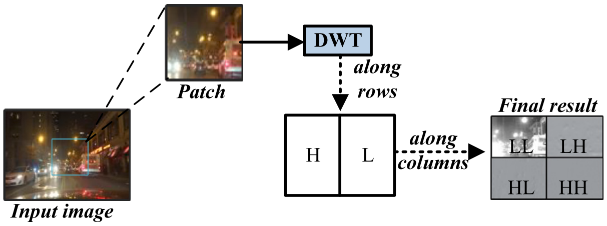

Fig. 3. Discrete wavelet transform single-level decomposition diagram. Top left: the input image; top right: DWT feature along row; lower right: DWT feature along column; lower left: DWT final result.

our model combines local features in the spatial domain and texture information in the frequency domain to learn effective representation in both domains.

2.3. Image enhancement

Image enhancement, as a method of improving image quality, removing image noise, increasing contrast, and enhancing image detail, has made numerous attempts in image processing. For example, Wang et al. [45] proposed a low-light image enhancement method using normalizing flows, a type of generative model. This method established a mapping model from low-light images to target images with higher brightness and contrast, and used the learned model to enhance the input image. In the same year, a semi-wavelet attention mechanism [46] was proposed for low-light image enhancement, which combined wavelet transform and attention mechanism to selectively enhance different frequency components of the input image. In addition to deep learning-based methods, some image enhancement algorithms based on traditional methods have also been researched and improved. For example, [47] used gray point calculation to transform the image histogram based on probability theory, thus improving the image effect and enhancing the overall brightness. Pan et al. [48] proposed a low-light image enhancement method based on the Retinex theory [49], which improved the lighting map to better differentiate between highlights and shadow areas in the image, and enhanced the image's details and contrast by adjusting the parameters of the lighting map. However, these methods were often limited to a single dominant light source and cannot handle images with complex lighting conditions, which prompts us to introduce various modules to correct tricky exposure issues.

3. Methods

In this section, we first describe the overall architecture of the proposed Dual-Domain Feature Learning model, followed by a careful presentation of each module.

3.1. Architecture overview

The overall structure of the model is shown in Fig. 2. We present an encoder and decoder architecture in an end-to-end manner. The encoding block in the encoder is comprised of four modules: Swin-Transformer (Swin-T) [35], Exposure Attention Correction Module (EACM) (Section 3.3), FDTM (Section 3.4) and DCFM (Section 3.5). The decoding block in the decoder utilizes U-net [30] to generate the segmentation map. The encoding block processes exposure, extracts and fuses spatial-and-frequency domain features from four different scales. Following it, the decoding block performs four upsampling operations and fuses the low-and high-level features to retain high-resolution details.

3.2. Discrete wavelet transform

The objective of Discrete Wavelet Transform (DWT) is to incorporate the attention mechanism to augment the spatial learning capacity of the model. We first employ Discrete Wavelet Transform DWT to enable our finely-grained analyses on various parts. The DWT decomposes the image into multiple scales and directions, allowing us to uncover subtle details in exposure images. The principle of its one-level decomposition is as follows:

Let $I \in \mathbb{R}^{3\times h\times w}$ be the input feature, $x_{i,j}$ denotes the pixel coordinates. Here, we use a two-dimensional Haar wavelet transform. The process can be divided into two steps. First, perform a one-dimensional Haar wavelet transform on the rows. The transformation formula is given by:

$$\begin{aligned} c_{m,n} &= \frac{1}{\sqrt{2}}\sum_{l=0}^{w-1}\left[x_{2m,l}\cdot\psi_{n,l}+x_{2m+1,l}\cdot\varphi_{n,l}\right], \\ d_{m,n} &= \frac{1}{\sqrt{2}}\sum_{l=0}^{w-1}\left[x_{2m,l}\cdot\varphi_{n,l}-x_{2m+1,l}\cdot\psi_{n,l}\right], \end{aligned} \tag{1}$$

where w is the width of the input feature I, c, d represent the low-and-high frequency coefficients, m, n represent the number of rows and columns respectively, φ and ψ are one-dimensional wavelet basis functions.

$$\begin{aligned} \varphi(x) &= \begin{cases} 1, & 0\le x<1, \\ 0, & \text{otherwise}\ . \end{cases} \\ \psi(x) &= \begin{cases} 1, & 0\le x<0.5, \\ -1, & 0.5\le x<1, \\ 0, & \text{otherwise}\ . \end{cases} \end{aligned} \tag{2}$$

Second, we perform a one-dimensional Haar wavelet transform on the columns. The transformation formula is given by:

$$\begin{aligned} c_{m,n} &= \frac{1}{\sqrt{2}}\sum_{l=0}^{h-1}\left[c_{l,n}\cdot\psi_{m,l}+d_{l,n}\cdot\varphi_{m,l}\right], \\ d_{m,n} &= \frac{1}{\sqrt{2}}\sum_{l=0}^{h-1}\left[c_{l,n}\cdot\varphi_{m,l}-d_{l,n}\cdot\psi_{m,l}\right], \end{aligned} \tag{3}$$

where the parameters c, d, m, n, φ and ψ are the same as above. Besides, h is the height of the input feature I. To gain a better understanding of the impact of DWT on extracting features from nighttime images, a randomly selected nighttime image was subjected to DWT operations, and the visualization of the feature result is shown in Fig. 3.

3.3. Exposure attention correction module

The build of DWT achieved fine-grained localization in the spatial and frequency domains and captured local features in the nighttime images. However, recovering the complex exposure patterns in nighttime images needs more than just local information. Thus, we propose the EACM to integrate and rectify image exposure at both the global and pixel levels, which sets itself apart from the conventional image enhancement network.

Formally, given an input image $I \in \mathbb{R}^{3\times h\times w}$, crop it by Patch Partition [35] to obtain $x_{i,j} \in \mathbb{R}^{48\times\frac{h}{4}\times\frac{w}{4}}$, then we downsample it to obtain features $d^k \in \mathbb{R}^{c'\times h'\times w'}$, k is adapted from [35]. We provide an instance of d^k, where k equals 1, in the first encoder of Fig. 2. The d^k is defined as:

$$d^k = STB^k\left(\left(PatchMerge\left(x_{i,j}\right)\right), k\in\{1,2,3,4\}, \tag{4}$$

where STB represents the SwinTransBlock, k denotes the downsampling level, and c', h', w' change as the value of k. Specifically, when $k=1$, $PatchMerge$ in the above formula is Linear Embedding. Here, STB, $PatchMerge$ and Linear Embedding are all adopted from [35].

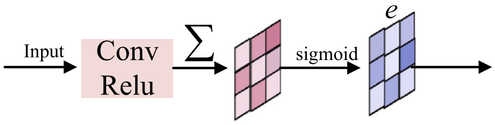

Fig. 4. Fine-grained complement correction module consists of three components, including convolution, summation, and sigmoid. e is our exposure attention map.

Next is the preliminary adaptive correction algorithm we have designed, detailed at the top-left corner of Fig. 2. We first process the downscaled features d^k by channel-wise halving and DWT to yield the features s^k, as:

$$s^k = DWT(Conv_{3\times3}(d^k)), k \in \{1,2,3,4\}, \tag{5}$$

then we proceed to rectify the exposure of the feature s^k. To enhance pixel-level nighttime scene, we flexibly apply the fully convolutional weight FC_a and FC_m to dynamically change the frequency spectrum through exposure-aware coefficients.

$$\begin{aligned} FC_a &= FC\left(GAP\left(s^k_{i,j}\right)\right)_{weight}, \\ FC_m &= FC\left(GMP\left(s^k_{i,j}\right)\right)_{weight}, \end{aligned} \tag{6}$$

where GAP and GMP stand for the global average and maximum pooling operation respectively, and FC stands for the fully convolutional operation. After re-scaling the feature map with the exposure-aware coefficients FC_a and FC_m, we perform a concatenation operation to combine the advantages of both average pooling and max pooling, as $D(\cdot)$:

$$D\left(s^k_{i,j}\right) = Cat\left(s^k_{i,j} \cdot FC_a, s^k_{i,j} \cdot FC_m\right), \tag{7}$$

where Cat means the concat operation.

Subsequently, we further propose a Fine-grained complement correction module (FCC) in the Fig. 4 to correct exposure at the pixel level. The module reduces the effect of high frequencies, leading to more even illumination. Besides, we conducted ablation experiments about the FCC to prove its validity in Fig. 6 of Section 4.2.1. After obtaining the preliminary results of adaptive correction $D\left(s^k_{i,j}\right)$, we reduce the dimensionality and learn the correlations between pixels in the same dimension, so that each pixel can learn more information when summed along the channel. The fine-grained correction process is defined as $C(\cdot)$:

$$C\left(s^k_{i,j}\right) = Softmax\left(\sum Relu\left(Conv_{3\times3}\left(D\left(s^k_{i,j}\right)\right)\right)\right), \tag{8}$$

where $C(x^k_{i,j})$ forms our exposure attention map e in Fig. 4. It implements fine-grained exposure correction at the pixel level. Here, we do not simply apply the softmax function, but rather progressively correct the exposure through two operations: global pooling to correct global exposure and exposure attention to correct the remaining uncorrected pixels at a fine-grained level.

We multiply $S(S = s^1, s^2, s^3, s^4$ obtained from Eq. (5)) and e yields S' ($S' = s^{1'}, s^{2'}, s^{3'}, s^{4'}$), getting the corrected map S' to the inverse wavelet transformation (IWT) (top-left corner of Fig. 2).

$$\begin{aligned} X(i,j) = &\sum_{k_1}\sum_{k_2} c_{j,k_1,k_2} \cdot \psi_{j,k_1,k_2}(i,j) \\ &+ \sum_{k_1}\sum_{k_2} d_{j,k_1,k_2} \cdot \phi_{j,k_1,k_2}(i,j), \end{aligned} \tag{9}$$

where $X(i,j)$ is the pixel value of the image after IWT, c_{j,k_1,k_2} is the wavelet coefficient of size j and position (k_1,k_2), $\psi_{j,k_1,k_2}(i,j)$ is scale function coefficients of size j and position (k_1,k_2), $\phi_{j,k_1,k_2}(i,j)$ is a two-dimensional scale function. After this, we map S' back to the normal spatial domain in order to compensate for the loss of information in the sub-bands, which is caused by DWT. Besides, we design a residual block to reduce the loss of low-light information during network propagation. Specifically, the inversed feature is then concatenated with the residual block to reduce the loss of low-light information during network propagation.

Finally, the convolutional, ReLU, and residual operations are performed to speed up model convergence and obtain the final corrected output feature f, as:

$$f = Relu\left(Conv_{3\times3}\left(Cat\left(IWT\left(S'\right), d_{res}\right)\right)\right), \tag{10}$$

and d_{res} represents the downsample of the feature d^k.

3.4. Frequency domain transformation module

To enhance the coupling between high-and-low frequency domains, we propose the FDTM to partition the frequency domain of the image, and employ attention mechanisms to enable its adaptive adjustment. We first employ Fast Fourier Transform (FFT) to map the spatial domain features to the frequency domain. Then two adaptive convolutions are utilized to learn both high and low frequency features. This ensures that the model avoids ignoring the high frequency information, which was a common issue with previous methods.

Formally, let $f \in \mathbb{R}^{c\times h\times w}$ be the input spatial feature to the FFT, where c, h, w vary with four downsamples. The FFT, defined as $F(u,v)$, is expressed mathematically as follows:

$$F(u,v) = \iint_{-\infty}^{+\infty} f(x,y)e^{-i2\pi(ux+vy)}dxdy, \tag{11}$$

where $f(x,y)$ is the pixel value of the feature f at coordinates (x,y), with u and v as frequency variables ranging from negative to positive infinity. The exponent $e^{-i2\pi(ux+vy)}$ represents a complex rotation factor. After transforming the feature f into the frequency domain, we get g_{high} and g_{low}, which is then divided into high and low frequency components by 1×1 convolution in the channel, as:

$$\left[g_{high}, g_{low}\right] = Conv_{1\times1}(FFT(f)). \tag{12}$$

To further enhance the corresponding feature in each frequency band, we pass the feature maps g_{high} and g_{low} into Dot-Product Attention (DPA) [50], which capture rich correlation information between each item in the input feature. DPA determines the degree of correlation between each frequency domain sequence. It also scales the results, accelerating the model's attention-based learning of high and low-frequency information.

However, we note that many detail-informants are compressed when the feature f is transformed into the frequency domain, leading to inaccurate extraction of deep information. Thus, we introduce the Inverse Fast Fourier Transform (iFFT), concat and DPA to get $f' \in \mathbb{R}^{c\times h\times w}$, as Eq. (13), which doubles the model's ability to acquire detailed information by gaining ancillary information from two domains (lower-left corner of Fig. 2).

$$f' = iFFT\left(Cat\left(DPA\left(g_{high}\right), DPA\left(g_{low}\right), g\right)\right), \tag{13}$$

here, the output of the DPA is fused with the residual branch and then transformed back into the spatial-domain feature using $iFFT$, as:

$$f'(x,y) = \frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} F(u,v)e^{i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}, \tag{14}$$

where $F(u,v)$ represents a two-dimensional frequency domain feature, while $f(x,y)$ represents its corresponding spatial domain feature. M and N are the lengths of the feature in the x and y directions, respectively.

We use the Eq. (14) to decompose frequency domain feature $F(u,v)$ into a weighted sum of sine waves with frequencies (u,v). Each sine wave's weight is multiplied by $e^{i2\pi(\frac{ux}{M}+\frac{vy}{N})}$, and then all the sine waves are added together to obtain the two-dimensional spatial domain feature $f(x,y)$. The inverse transformation produces the feature f', which is used in Section 3.5, so we can model the interaction information between the frequency-and-spatial domain at a deeper level. The specific details of the FDTM are as Algorithm 1. Please note, in Algorithm 1, 'freq' represents 'frequency'.

Algorithm 1 Frequency Domain Transformation Module

```
Input: The feature matrix from EACM $X \in \mathbb{R}^{c' \times h' \times w'}$;
Output: The feature based on high–low freq processing;
1: Calculate the freq features $g_{low}$ and $g_{high}$ according to Eq. (12) ;
2: Calculate attention scores separately for $g_{low}$ and $g_{high}$;
3: $g_{low}' = DPA(g_{low}),\ \ g_{high}' = DPA(g_{high})$;
4: Concate $g_{low}'$, $g_{high}'$ and $g$ to achieve enriched feature;
5: Calculate freq-to-spatial features $h$ according to Eq. (14).
```

3.5. Dual-domain cross-fusion module

Separate learning of high and low frequency information helps our network to analyze nighttime light conditions. However, information loss may occur due to truncation and quantization in FFT [51]. Skip-integration allows the network to effectively grasp various levels of feature granularity, while attention mechanisms focus on specific aspects by swiftly filtering out irrelevant features from the extensive feature set. Therefore, Zhu et al. [52] proposed a cross-channel attention framework, which utilized the structure to achieve limited attention, and identified as many crucial local object regions as possible in both channel and spatial dimensions. Considering these aspects, we propose the Dual-domain Cross-Fusion Module (DCFM) to skip-integrate the features f' processed by FFT with the spatial domain features f, as shown in the lower part of Fig. 2. Our DCFM consists of two components to achieve deep integration. In the first part, GAP prevents the useful information from being overwhelmed by a large amount of irrelevant information. And the cross-concate helps the module link dual-domain information at different scales. Considering the flattening feature is complex, computationally efficient one-dimensional convolutions are more suitable. We propose the one-dimensional convolution to learn cross-domain interactions on the same dimension with fewer parameters. The above processing is as follows:

$$
\begin{aligned}
c_{i,j} &= Conv1d(Cat(GAP(f), f')), \\
c'_{i,j} &= Conv1d(Cat(GAP(f'), f)),
\end{aligned}
\tag{15}
$$

where the features f and $f' \in \mathbb{R}^{c \times h \times w}$ are flattened along their dimensions and cross-concated to obtain the feature maps $c_{i,j}$, $c'_{i,j} \in \mathbb{R}^{c \times (h+w+1)}$.

In the second part, attention mechanism is used to enhance the different features and fuse them wisely, as:

$$
V = Conv_{3\times3}\left(Cat\left(STB\left(c_{i,j}\right), STB\left(c'_{i,j}\right)\right)\right),
\tag{16}
$$

here, the model obtains a wider range of pixel connectivity information by feeding the preprocessed fusion features $c_{i,j}$, $c'_{i,j}$ into the STB. Finally, the two branches are connected to fuse the features in different proportions to consolidate the feature $v \in V$. Our method is capable of exploring feature correlations between spatial and frequency domain information and learning similar features among objects of the same category.

3.6. Loss function

To address our model falling into local optimization, we incorporate a punishment mechanism to provide real-time supervision and correction to the pixel loss. Specifically, we add the mIoU loss as a punishment factor to the loss function. The mIoU is defined as follows:

$$
L_{mIoU} = 1 - \frac{1}{M+1}\sum_{i=0}^{M}\frac{TP}{FN+FP+TP},
\tag{17}
$$

where M represents the number of categories, TP refers to the regions where the model predicts positive and is actually positive, FN refers to the regions where the model predicts negative but is actually positive, and FP refers to the regions where the model predicts positive but is actually negative.

Besides, we also utilize Online Hard Example Mining loss (Ohem loss) [53] for computation segmentation loss. This Ohem loss named L_{ohem} allows our model focus on the difficult examples with larger losses during the training process and assigns them higher weights to aid the model in achieving better convergence. The L_{ohem} is based on cross-entropy loss, and the expression is as follows:

$$
L_{ohem} = -\frac{1}{N_{ohem}}\frac{1}{N}\sum_{i=1}^{N_{ohem}}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(P_{ic}),
\tag{18}
$$

here, N and M represent the number of total samples and categories separately, and N_{ohem} is the number of samples with the cross-entropy loss greater than the threshold. Besides, p_{ic} indicates the predicted probability that observation i belongs to category c, and the $y_{i,c}$ is defined as:

$$
y_{ic} = \begin{cases} 1, \text{label}_i = c, \\ 0,\ \text{otherwise}, \end{cases}
\tag{19}
$$

where $\text{label}_i = c$ means the true label of sample i is c.

For our image segmentation tasks, each pixel in a batch is converted into a long vector, sorted according to loss, and those pixels with losses above the threshold are selected as hard samples. The two hyperparameters improve experimental training results to some extent. Through the joint optimization of the mIoU loss and Ohem loss penalty factors, our model training can be more efficient in learning feature information and improving the model performance. The total loss function is formulated as:

$$
L = \alpha \cdot L_{mIoU} + L_{ohem},
\tag{20}
$$

where the parameter α is an empirically determined coefficient in this context.

4. Experiments

4.1. Experimental setups

Datasets: To verify the performance of the model, we compare our proposed model with the state-of-the-art approaches on the **Nightcity** and **BDD0OK-night** [54] datasets and we test our result on the **Nightcity**+ [20]:

Nightcity: It is a large-scale dataset of real nighttime images, containing 4297 finely annotated images, with 2998 images used for training and 1299 images for validation. The dataset's images have a resolution of 512 × 1024 and are compatible with cityscape labeling, consisting of 19 classes.

NightCity+: It is an update of NightCity where some labeling errors of the validation set are corrected and it is resized to the resolution of 1024 × 2048.

BDD0OK-night: It comprises 320 images in the training set and 34 images in the validation set, with a resolution of 720 × 1280 and 19 categories.

Evaluation metrics: We evaluate segmentation accuracy with four typical evaluation metrics, which are Mean Pixel Accuracy (MPA), mean Intersection over Union (mIoU), and Frequency Weighted Intersection over Union (FWIoU). Specifically, MPA measures the proportion of correctly classified pixels in model predictions, aiding in assessing the overall performance of the model at the pixel level. mIoU is the intersection of the inferred segmentation and the ground truth, divided by the union. FWIoU computes a weighted sum of IoU for each class to balance the impact of class frequencies on evaluation metrics. The metrics are computed as

$$
MPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}},
\tag{21}
$$

Table 1
Experimental results of different models on the NightCity dataset. DDFL is our introduced method. **Bold** represents the best results. ‘_’ means the second results. ‘–’ indicates that no source code has been provided.

Methods	Venue&Year	Original task	Resolution	Backbone	MPA	FWIoU (%)	mIoU (%)	
							NightCity	NightCity+
UNet [30]	MICCAI2015	–	512 × 1024	–	76.07	70.61	48.21	49.17
PSPNet [29]	CVPR2017	DTSS	512 × 1024	ResNet-101	88.37	81.56	52.67	53.09
PSPNet [29]	CVPR2017	DTSS	1024 × 2048	ResNet-101	89.65	82.23	53.02	53.95
Deeplabv3+ [31]	ECCV2018	DTSS	512 × 1024	ResNet-101	88.23	81.05	52.17	53.42
Deeplabv3+ [31]	ECCV2018	DTSS	1024 × 2048	ResNet-101	90.02	82.95	53.41	54.13
DANet [32]	CVPR2019	DTSS	1024 × 2048	ResNet-101	77.56	72.01	50.76	51.58
CCNet [33]	ICCV2019	DTSS	512 × 1024	ResNet-101	76.98	69.78	49.81	51.39
UperNet [35]	ICCV2021	DTSS	512 × 1024	Swin-T	90.83	83.76	54.93	56.32
EGNet [9]	TIP2021	NTSS	512 × 1024	ResNet-101	–	–	51.8	–
NightLab(Deeplabv3+) [20]	CVPR2022	NTSS	1024 × 2048	ResNet-101	–	–	–	56.21
FDLNet(UperNet) [55]	TIP2023	NTSS	512 × 1024	Swin-T	91.46	85.39	55.54	57.04
FDLNet(UperNet) [55]	TIP2023	NTSS	1024 × 2048	Swin-T	91.72	85.84	–	57.39
DDFL(UNet)	–	NTSS	512 × 1024	ResNet-101	77.67	72.10	50.87	51.95
DDFL(PSPNet)	–	NTSS	512 × 1024	ResNet-101	90.51	84.31	54.55	55.47
DDFL(UperNet)	–	NTSS	512 × 1024	ResNet-101	91.05	85.21	55.13	55.97
DDFL(UNet)	–	NTSS	512 × 1024	Swin-T	77.81	72.44	50.95	52.33
DDFL(PSPNet)	–	NTSS	512 × 1024	Swin-T	91.17	85.25	55.16	56.07
DDFL(UperNet)	–	**NTSS**	**512 × 1024**	**Swin-T**	**92.65**	**87.14**	**56.73**	**57.41**

$$mIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}}, \tag{22}$$

$$FWIoU = \frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}}, \tag{23}$$

where k is the number of classes, p_{ij} indicates the predicted probability that observation i belongs to category j, p_{ii} is similar to p_{ij}.

Implementation Details: All experiments were carried out using the PyTorch deep learning framework. During the training phase, we employ the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and a multi-learning strategy. We set the initial learning rate and weight decay coefficient to 0.01 and 5e−4, respectively. In addition, we set the batch size to 16 and performed data augmentation on the input images as suggested in [35], including randomly scaling and horizontally flipping as well as Gaussian blur.

Our experimental environment is Ubuntu 16.04.4, equipped with a single 24 GB NVIDIA GeForce GTX 3090 GPU. During the training process, we initialize the backbone of our model with the pre-trained weights of the Swin-T [35] model, which was trained on the ADE20K dataset. The two hyperparameters of the OHEM loss function, namely the threshold T and the minimum sample number Min, are set to 0.7 and $1e5$ respectively. The parameter α of the total loss function is set to 0.4. The initial learning rate for the pre-trained Swin-T model is 0.001, and for the EACM and other modules, it is set to 0.01. The training is performed with 1500 epochs to update the model, and we obey the default training pipeline as described in [35].

4.2. Experimental results

Comparison on the NightCity. To validate the effectiveness of our proposed method, we compared our model with several state-of-the-art models for DTSS, NTSS and other tasks, including PSPNet [29], DeeplabV3+ [31], DANet [32], CCNet [33], Uper-Swin-T [35] for DTSS, and EGNet [9], NightLab [20], FDLNet [55] for NTSS. Simultaneously, for a comprehensive demonstration of the effectiveness of our method, ResNet101 [56] and Swin-T are utilized as backbones for DDFL respectively. From Table 1, we see that our model performs best across the metrics of MPA, FWIOU, and MIOU. Fig. 7 shows comparative annotation results across four different scenarios. As evident, our model achieves a very close similarity to the actual results in detail. In terms of metrics, the day-time methods do not perform satisfactorily due to their lack of consideration for night-time exposure conditions, with the highest mIoU reaching only 56.32%. Compared to EGNet, NightLab, and FDLNet, our method improves accuracy, with the highest increase in mIoU reaching 4.93% and the MPA reaching 92.65. Although our model has a relatively smaller improvement compared to FDLNet(UperNet) with the resolution of 1024 × 2048, our method requires a lower resolution (512 × 1024) and relaxes the strict requirements for dataset resolution. Overall, our proposed method shows better performance than the state-of-the-art models.

Comparison on the BDD100k. On this dataset, we compared our model with seven methods: UNet [30], PSPNet [29], DeeplabV3+ [31], CCNet [33], Uper-Swin-T [35], EGNet [9] and FDLNet [55]. ResNet101 and Swin-T are still used as backbone networks to form the control group. The comparison results (Table 2) show that DDFL (UPerNet) significantly outperforms the other methods. In all three evaluation metrics, our model achieves both first and second place. The findings indicate that our model with the UperNet decoder head achieves the best performance of 28.97% at the same resolution. We also can see that using UperNet as the decoder head and ResNet-101 as the backbone our model still achieves second place, 3.68% higher than the FDLNet. In general, these experiments verify the superiority and generality of our method, and our method repairs exposure areas and uses dual-domain information.

4.2.1. Ablation experiment

To further validate the effectiveness of different modules in our proposed model, a series of ablation experiments were conducted on the NightCity dataset.

I. Exposure Attention Correction Module (EACM). To investigate the reasonableness of the EACM module, we perform ablation experiments that encompass both the complete module (Table 3) and its internal structure (Fig. 6).

First, for the ablation experiment of the whole module, we remove the exposure attention module or replace the EACM with the DAU module [46]. The down-sampling was based on the Swin-Trans-block, and the up-sampling was based on the UperNet. From these ablation experiments, we observed significant fluctuations in the results after removing the EACM.

(1) In Exp1, we use a 1 × 1 convolution to learn the feature information of the discrete wavelet transformation. We find Exp1 has limited capacity in utilizing the high and low-frequency information. As a result, the performance is not observed to be satisfactory.

Table 2
Experimental results of different models on the BDD100k dataset. DDFL is our introduced method. **Bold** represents the best results. '_' means the second results. '–' indicates that no source code has been provided.

Model	Task	Backbone	Resolution	MPA	FWIoU	mIoU%
UNet [30]	–	–	512 × 1024	32.59	30.76	15.34
PSPNet [29]	DTSS	ResNet-101	512 × 1024	48.28	47.76	19.75
Deeplabv3+ [31]	DTSS	ResNet-101	512 × 1024	50.49	49.25	22.79
CCNet [33]	DTSS	ResNet-101	512 × 1024	34.15	33.93	17.06
Uper-Swin-T [35]	DTSS	Swin-T	512 × 1024	51.08	50.90	24.73
EGNet [9]	NTSS	ResNet-101	512 × 1024	–	–	15.68
FDLNet(DeepLabv3+) [55]	NTSS	ResNet-101	512 × 1024	52.87	53.05	25.13
DDFL(UNet)	NTSS	ResNet-101	512 × 1024	33.28	32.54	16.45
DDFL(PSPNet)	NTSS	ResNet-101	512 × 1024	54.13	54.41	25.89
DDFL(UperNet)	NTSS	ResNet-101	512 × 1024	59.82	59.93	28.81
DDFL(UNet)	NTSS	Swin-T	512 × 1024	33.65	33.27	16.87
DDFL(PSPNet)	NTSS	Swin-T	512 × 1024	56.77	56.94	26.03
DDFL(UperNet)	**NTSS**	**Swin-T**	**512 × 1024**	**61.24**	**61.51**	**28.97**

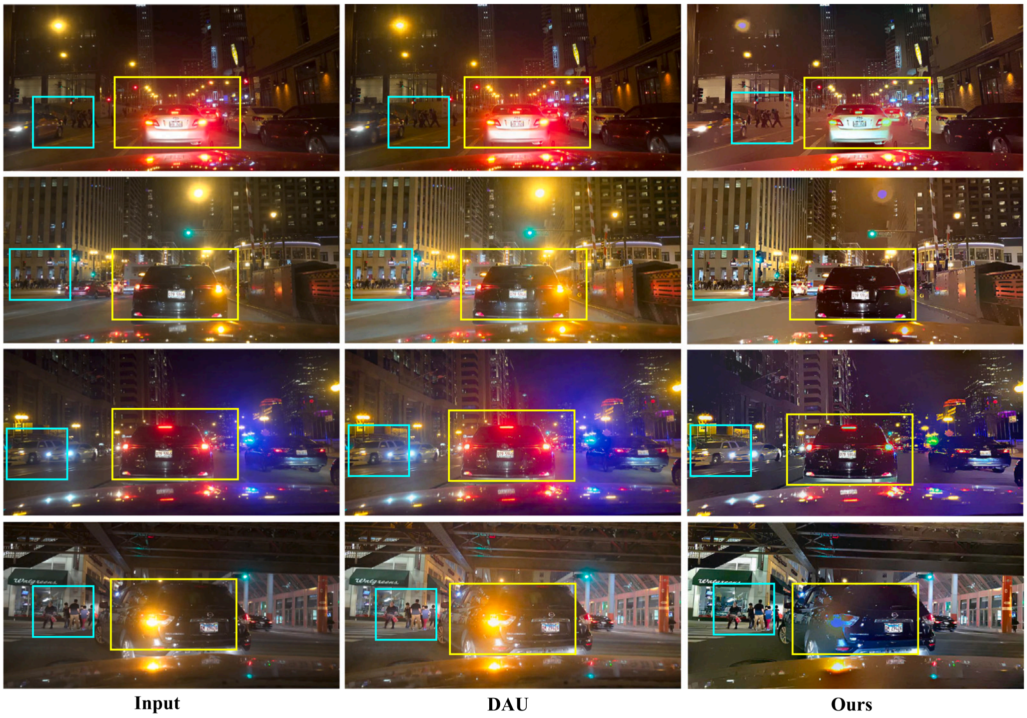

Fig. 5. Comparative ablation experiments between the DAU module and our EACM. Among them, the blue frame indicates underexposed areas and yellow indicates overexposed areas. The columns from left to right are input, DAU and ours.

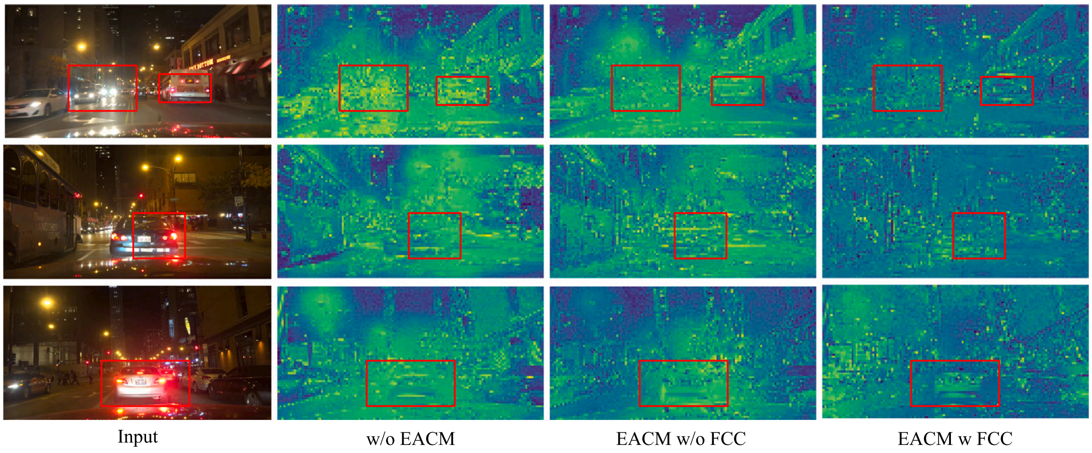

Fig. 6. Exposure features at different stages of the EACM module. "w/o" means without, and "w" means with.

Table 3
Impact of our EACM on model performance. DDFL is our introduced method. Bold represents the best results.

Exp	Model	mIoU%
1	DDFL(w/o EACM)	54.31
2	DDFL(w/o EACM)+DAU	55.85
3	**DDFL(w EACM)**	**57.41**

(2) In Exp2, the DAU module is used to replace our EACM, providing more detail and clarity on what is being referred to. We improve the performance by 1.54% owing to the dual-channel and spatial attention of the DAU. However, the diversity of nighttime image exposure conditions enables only a slight improvement in model performance. Even with attention to both channel and space, the influence of multiple areas with different exposure levels on segmentation performance cannot be ignored.

(3) In Exp3, the EACM is retained to form a control with other experiments. We can see the model's performance improve from 54.31% to **57.41%**. This indicates that our EACM, based on attention-guided exposure guidance in a region-based manner using convolution and so on, plays a significant role in moderately weakening. Meanwhile, our model enhances task regions while retaining complete feature information, ultimately enhancing model performance.

To further illustrate the issue at hand, we conduct a qualitative analysis of the module by loading the weights trained based on the DAU module and the EACM into the model. We then visualize the images after exposure correction and conduct a comparison, as shown in Fig. 5, where the green box indicates areas of underexposure and the yellow box indicates areas of overexposure. As can be observed, the DAU mechanism is inclined to share feature information across different dimensions. In the second column of Fig. 5, the entire image region is enhanced, with the area inside the green box being properly adjusted. However, the expansion of the light effect area inside the yellow box has had a counterproductive effect, which is unfavorable for the segmentation of vehicles and surrounding pedestrians and vehicles. Our model, on the other hand, has effectively addressed this issue. In the third column of Fig. 5, our model has enhanced the underexposed areas in shadow, such as pedestrians, traffic light poles, and distant car fronts, inside the green box. Inside the yellow box, our model has eliminated the light effect area of the taillights to achieve precise recognition of the contours of the vehicles and surrounding pedestrians. In summary, our model has achieved both low-light enhancement and overexposure correction tasks in a single image.

Second, we design an ablation study to demonstrate the superiority of the proposed FCC. Fig. 6 shows the exposure features at three stages of the EACM module: "w/o EACM" indicates the removal of EACM, "EACM w/o FCC" (Fig. 4) indicates the removal of FCC from the EACM module and "EACM w FCC" indicates the complete EACM (Fig. 2), respectively. When EACM is removed, the exposure is concentrated around vehicles, streetlights, and indoor lights, leading to significant exposure differences across different areas in the entire image. After adding the EACM w/o FCC, the exposure features exhibit an overall improvement, resulting in a more uniform overall illumination. However, the severe overexposure caused by complex lighting remains intense, which can easily impact the subsequent object segmentation. In comparison, "EACM w FCC" demonstrates an extremely uniform distribution of light, with no apparent strong lighting or shadows throughout the entire scene, creating a very comfortable overall picture. Through the aforementioned experiments, we have found that the two modules of EACM are mutually compatible, enabling exposure correction at both coarse and fine granularities.

II. Frequency Domain Transformation Module (FDTM). We investigate the impact of different approaches to learning frequency domain information on model performance. In particular, we conduct experiments using Swin-Trans-block-based downsampling and UperNet-based upsampling to explore the effects of these methods (Exp4∼Exp6), as shown in Table 4.

Table 4
The Impact of FDTM on model performance. DDFL is our introduced method. Bold represents the best results.

Exp	Model	mIoU%
4	DDFL(w/o FDTM)	54.36
5	DDFL(w/o FDTM)+FU	54.97
6	**DDFL(w FDTM)**	**57.41**

Table 5
Comparison experiments with Swin-Trans-T+UperNet baseline. Bold represents the best results.

Index	Model	mIoU%
7	Swin-Trans-T+UperNet	54.47
8	Swin-Trans-T+EACM+UperNet	54.98
9	Swin-Trans-T+EACM+FDTM+UperNet	55.85
10	**Swin-Trans-T+EACM+FDTM+DCFM+UperNet**	**57.41**

(4) In Exp4, we remove the FDTM and fuse spatial domain information double to perform ablation experiments on the FDTM. With the assistance of the Swin-Trans-UperNet network and other modules proposed in this paper, we achieve a performance of 54.36%. However, compared with the control group of Exp 6, performance declines by 3.05%, indicating that frequency domain information plays a certain auxiliary role in night-time semantic segmentation.

(5) In Exp5, to demonstrate that our FDTM can learn frequency domain information better, we introduce the FU module from FFC [57]. The FU module converts the spatial domain to the frequency domain, updates frequency domain information globally and then converts back to the spatial domain. We improve the performance by 0.61% compared to Exp 4, indicating that global learning of frequency domain features enables the model to capture details and differentiate between different classes. However, due to the reason that high and low frequency information in the frequency domain has different effects on image analysis, the performance improvement is not significant. High frequency makes the image sharper and clearer, facilitating detailed segmentation, while low frequency is more conducive to differentiating between different classes. Our introduction of self-attention-based high and low frequency band learning tasks assists the model in segmentation.

(6) In Exp6, we keep the model unchanged in order to form a control with other experiments. Here, we reach the highest mIoU, **57.41%**.

III. Our Model with baseline. As shown in Table 5, we have conducted four experiments.

(7) In Exp7, we utilize Swin-Trans-Tiny+UperNet as a baseline and get the mIoU 54.47%.

(8) In Exp8, we design an exposure correction module using exposure attention to enhance performance from 54.47% to 54.98%.

(9) In Exp9, we design an FDTM based on the characteristics of high and low frequency information using both attention mechanisms, enabling a further improvement in the performance of 0.87%.

(10) In Exp10, in order to fully exploit the benefits of both domains, we design a cross-domain fusion module and get the mIoU **57.41%**, which differs from previous approaches that only focus on either frequency or time domain features.

Table 6
Comparison of the complexities of various models and their accuracy on Nightcity+. **Bold** represents the best results. '_' means the second results. '–' indicates that no source code has been provided.

Methods	Backbone	Params/M ↓	Flops/G↓	MPA	mIoU (%)
UNet [30]	–	13	254	76.07	49.17
PSPNet [29]	ResNet-101	66	526	88.37	53.09
Deeplabv3+ [31]	ResNet-101	59	499	88.23	53.42
CCNet [33]	ResNet-101	67	557	76.98	51.39
Uper-Swin-T [35]	Swin-T	60	945	90.83	56.32
NightLab(Deeplabv3+) [20]	ResNet-101	–	–	–	56.21
FDLNet(UperNet) [55]	Swin-T	83	1206	<u>91.46</u>	<u>57.04</u>
DDFL(Unet)	Swin-T	86	1280	77.81	52.33
DDFL(PSPNet)	Swin-T	88	1319	91.17	56.07
DDFL(UperNet)	Swin-T	101	1473	**92.65**	**57.41**

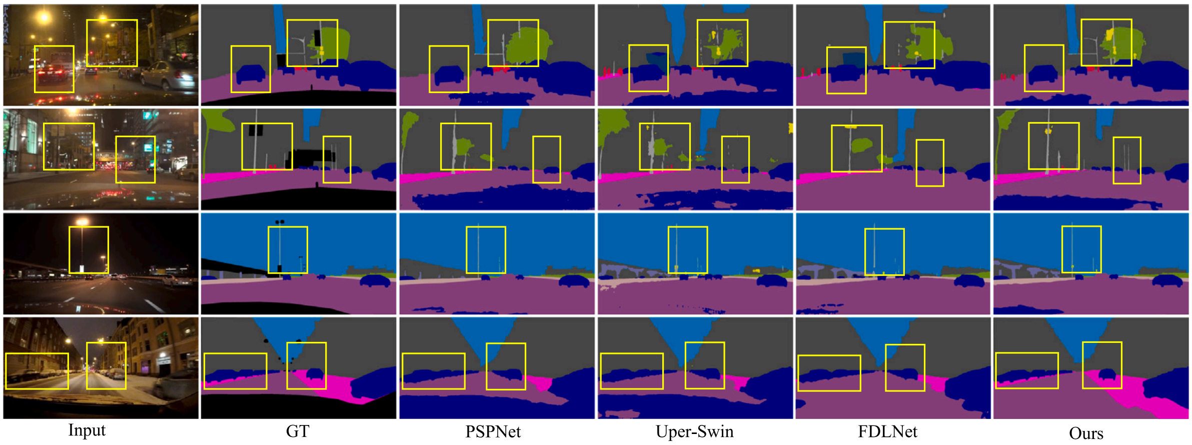

Fig. 7. Qualitative comparison on the NightCity dataset. We use yellow boxes to frame the areas where our model represents well.

We can see that our three modules improve model performance in various ways such as exposure correction, details extraction and feature fusion. They make the whole model capable of fusing dual-domain information across multiple scales, learning dual-domain contextual features.

4.2.2. Computational complexity

We provide a comparison between FLOPs for parameter quantities in Table 6. We adopt the results in seven methods and evaluate our method on a single 24 GB NVIDIA GeForce GTX 3090 GPU. It can be seen that our parameters do not increase significantly, yet achieving the highest scores across multiple evaluation metrics.

4.3. Visual analysis

To provide a more intuitive demonstration of the performance of our model in this paper, we have selected some challenging nighttime images from the NightCity dataset and conducted a comparative analysis with Uper-swin and PSPNet methods. The results are shown in Fig. 7, in which we have highlighted some regions with bounding boxes to facilitate clearer comparisons.

In the first line, our model exhibits superior performance compared to PSPNet and Uper-Swin in recognizing street lamps with more complete and accurate features, as well as vehicles with clearer contours. Moving on to the second line, our model achieves more precise segmentation for nearby street lamps compared to the other two models. For distant street lamps that suffered from underexposure and hence, appeared blurry, our model demonstrates an advanced ability to accurately recognize and segment them, producing more comprehensive results. In the third line, despite good lighting conditions, all three models achieve decent segmentation of details. However, our model outperforms the other two in segmenting the contours of street lamps. In the fourth line, the presence of numerous vehicles causes underexposure in the lower part of the vehicles. While PSPNet and Uper-Swin could identify the location of the vehicles, their contours are rough, and they are unable to pinpoint the position of the vehicle tires. In contrast, our model produces high-quality exposure correction, making the underexposed areas have minimal impact on our segmentation results. These results demonstrate that our model can focus on more details, such as street lamps, vehicle wheels, and vehicle contours, enabling superior segmentation outcomes. Since nighttime lighting conditions are complex and differ significantly from daytime lighting, the aforementioned results suggest that our model can achieve good segmentation performance by using exposure attention for image correction and simultaneously learning spatial and frequency domain information. Therefore, our model demonstrates superior performance in nighttime semantic segmentation tasks.

5. Conclusion

In this paper, we propose a model utilizing spatial-and-frequency domain features to handle the semantic segmentation task of complex night images. Specifically, the Exposure Attention Correction Module (EACM) helps the model correct the exposure. Since image enhancement is a difficult task, the EACM dynamically adjusts its ability to correct different-level exposure. Moreover, it gradually corrects the exposure from both coarse and fine grain to consolidate the effect. Furthermore, our Frequency Domain Transformation Module (FDTM) distributes the weights of high-and-low frequency information transformed from FFT. It adjusts the learning effort through two different branch sub-functions. Besides, the Dual-domain Cross Fusion Module (DCFM) integrates two different domains in an efficient and economical way which not only considers the respective information from the two but also the cross-fertilization information. When training, our model is optimized through OHEM loss and mIoU loss based on a penalty mechanism. Finally, our proposed model achieves a mIoU of **57.41%** and

28.97% on NightCity+ and BDD100k datasets, respectively, demonstrating good accuracy at low image resolutions compared to other daytime and nighttime semantic segmentation methods. We will further improve the segmentation speed and accuracy in more complex scenarios, including improving and exploring semantic segmentation for other types of multi-modal data.

CRediT authorship contribution statement

Xiao Lin: Conceptualization, Methodology, Investigation, Funding acquisition, Writing – original draft. **Peiwen Tan:** Methodology, Investigation, Writing – original draft, Writing – review & editing. **Zhengkai Wang:** Visualization, Investigation. **Lizhuang Ma:** Resources, Supervision. **Yan Li:** Resources, Supervision, Writing – review & editing.

Declaration of competing interest

We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Data availability

Data will be made available on request.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62072126 and supported by grants from the National Natural Science Foundation of China (No. 62302167), Shanghai Sailing Program (23YF1410500). This work was also supported by the 2022-2023 Shanghai Municipal Special Project for Promoting High-Quality Development of Industries (No. 2211106).

References

[1] Y. Li, J. Cai, Q. Zhou, H. Lu, Joint semantic-instance segmentation method for intelligent transportation system, IEEE Trans. Intell. Transp. Syst. (2022).

[2] J. Zhang, T. Fukuda, N. Yabuki, Automatic generation of synthetic datasets from a city digital twin for use in the instance segmentation of building facades, J. Comput. Des. Eng. 9 (5) (2022) 1737–1755.

[3] Y. Hou, R. Volk, L. Soibelman, A novel building temperature simulation approach driven by expanding semantic segmentation training datasets with synthetic aerial thermal images, Energies 14 (2) (2021) 353.

[4] L. Pulvirenti, M. Chini, N. Pierdicca, L. Guerriero, P. Ferrazzoli, Flood monitoring using multi-temporal COSMO-SkyMed data: Image segmentation and signature interpretation, Remote Sens. Environ. 115 (4) (2011) 990–1002.

[5] W. Wu, J. Yan, Y. Zhao, Q. Sun, H. Zhang, J. Cheng, D. Liang, Y. Chen, Z. Zhang, Z.-C. Li, Multi-task learning for concurrent survival prediction and semi-supervised segmentation of gliomas in brain MRI, Displays 78 (2023) 102402.

[6] W. Cai, B. Zhai, Y. Liu, R. Liu, X. Ning, Quadratic polynomial guided fuzzy C-means and dual attention mechanism for medical image segmentation, Displays 70 (2021) 102106.

[7] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, K. Dietmayer, Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges, IEEE Trans. Intell. Transp. Syst. 22 (3) (2020) 1341–1360.

[8] H. Wang, Y. Chen, Y. Cai, L. Chen, Y. Li, M.A. Sotelo, Z. Li, SFNet-N: An improved SFNet algorithm for semantic segmentation of low-light autonomous driving road scenes, IEEE Trans. Intell. Transp. Syst. 23 (11) (2022) 21405–21417.

[9] X. Tan, K. Xu, Y. Cao, Y. Zhang, L. Ma, R.W. Lau, Night-time scene parsing with a large real dataset, IEEE Trans. Image Process. 30 (2021) 9085–9098.

[10] X. Wu, Z. Wu, H. Guo, L. Ju, S. 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, 2021, pp. 15769–15778.

[11] H. Gao, J. Guo, G. Wang, Q. 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, 2022, pp. 9913–9923.

[12] Q. Xu, Y. Ma, J. Wu, C. Long, X. Huang, Cdada: A curriculum domain adaptation for nighttime semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2962–2971.

[13] C. Li, W. Xia, Y. Yan, B. Luo, J. Tang, Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation, IEEE Trans. Neural Netw. Learn. Syst. 32 (7) (2020) 3069–3082.

[14] B. Li, D. Yao, Z. Li, RICH: A rapid method for image-text cross-modal hash retrieval, Displays 79 (2023) 102489.

[15] B. Li, Z. Li, Large-scale cross-modal hashing with unified learning and multi-object regional correlation reasoning, Neural Netw. 171 (2024) 276–292.

[16] W. Zhao, X. Wu, Boosting entity-aware image captioning with multi-modal knowledge graph, IEEE Trans. Multimed. 26 (2024) 2659–2670.

[17] Y. Sun, W. Zuo, M. Liu, Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes, IEEE Robot. Autom. Lett. 4 (3) (2019) 2576–2583.

[18] H. Zhou, C. Tian, Z. Zhang, Q. Huo, Y. Xie, Z. Li, Multispectral fusion transformer network for RGB-thermal urban scene semantic segmentation, IEEE Geosci. Remote Sens. Lett. 19 (2022) 1–5.

[19] H. Xiong, W. Cai, Q. Liu, MCNet: Multi-level correction network for thermal image semantic segmentation of nighttime driving scene, Infrared Phys. Technol. 113 (2021) 103628.

[20] X. Deng, P. Wang, X. Lian, S. 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, 2022, pp. 16938–16948.

[21] M. Afifi, K.G. Derpanis, B. Ommer, M.S. Brown, Learning multi-scale photo exposure correction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9157–9167.

[22] F. Lv, F. Lu, J. Wu, C. Lim, MBLLEN: Low-light image/video enhancement using CNNs, in: BMVC, vol. 1, 2018, p. 4.

[23] Z. Lyu, A. Peng, Q. Wang, D. Ding, An efficient learning-based method for underwater image enhancement, Displays 74 (2022) 102174.

[24] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, A. Kot, Low-light image enhancement with normalizing flow, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 3, 2022, pp. 2604–2612.

[25] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3213–3223.

[26] F.W. Campbell, J.G. Robson, Application of Fourier analysis to the visibility of gratings, J. Phys. 197 (3) (1968) 551.

[27] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, et al., Deep high-resolution representation learning for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 43 (10) (2020) 3349–3364.

[28] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 801–818.

[29] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2881–2890.

[30] O. Ronneberger, P. Fischer, T. 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, Springer, 2015, pp. 234–241.

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.

[32] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.

[33] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, W. Liu, Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 603–612.

[34] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P.H. Torr, et al., Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6881–6890.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[36] S. Dong, W. Zhou, X. Qian, L. Yu, GEBNet: Graph-enhancement branch network for RGB-T scene parsing, IEEE Signal Process. Lett. 29 (2022) 2273–2277.

[37] W. Zhou, S. Dong, M. Fang, L. Yu, CACFNet: Cross-modal attention cascaded fusion network for RGB-T urban scene parsing, IEEE Trans. Intell. Veh. (2023) 1–10.

[38] X. He, M. Wang, T. Liu, L. Zhao, Y. Yue, SFAF-MA: Spatial feature aggregation and fusion with modality adaptation for RGB-thermal semantic segmentation, IEEE Trans. Instrum. Meas. 72 (2023) 1–10.

[39] W. Zhou, T. Gong, J. Lei, L. Yu, DBCNet: Dynamic bilateral cross-fusion network for RGB-t urban scene understanding in intelligent vehicles, IEEE Trans. Syst. Man, Cybern.: Syst. 53 (12) (2023) 7631–7641.

[40] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, T. Holz, Leveraging frequency analysis for deep fake image recognition, in: International Conference on Machine Learning, PMLR, 2020, pp. 3247–3258.

[41] L. Jiang, B. Dai, W. Wu, C.C. Loy, Focal frequency loss for image reconstruction and synthesis, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 13919–13929.

[42] K. Xu, M. Qin, F. Sun, Y. Wang, Y.-K. Chen, F. Ren, Learning in the frequency domain, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1740–1749.

[43] B. Zheng, S. Yuan, G. Slabaugh, A. Leonardis, Image demoireing with learnable bandpass filters, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3636–3645.

[44] J. Wang, J. Wu, Z. Wu, J. Jeong, G. Jeon, Wiener filter-based wavelet domain denoising, Displays 46 (2017) 37–41.

[45] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, A. Kot, Low-light image enhancement with normalizing flow, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 2604–2612.

[46] C.-M. Fan, T.-J. Liu, K.-H. Liu, Half wavelet attention on M-net+ for low-light image enhancement, in: 2022 IEEE International Conference on Image Processing, ICIP, IEEE, 2022, pp. 3878–3882.

[47] Y. Xie, L. Ning, M. Wang, C. Li, Image enhancement based on histogram equalization, in: Journal of Physics: Conference Series, vol. 1314, IOP Publishing, 2019, 012161.

[48] X. Pan, C. Li, Z. Pan, J. Yan, S. Tang, X. Yin, Low-light image enhancement method based on retinex theory by improving illumination map, Appl. Sci. 12 (10) (2022) 5257.

[49] E.H. Land, The retinex theory of color vision, Sci. Am. 237 (6) (1977) 108–129.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[51] D. James, Quantization errors in the fast Fourier transform, IEEE Trans. Acoust. Speech Signal Process. 23 (3) (1975) 277–283.

[52] Q. Zhu, W. Kuang, Z. Li, A collaborative gated attention network for fine-grained visual classification, Displays (2023) 102468.

[53] A. Shrivastava, A. Gupta, R. Girshick, Training region-based object detectors with online hard example mining, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 761–769.

[54] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, T. Darrell, Bdd100k: A diverse driving dataset for heterogeneous multitask learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2636–2645.

[55] Z. Xie, S. Wang, K. Xu, Z. Zhang, X. Tan, Y. Xie, L. Ma, Boosting night-time scene parsing with learnable frequency, IEEE Trans. Image Process. (2023).

[56] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[57] L. Chi, B. Jiang, Y. Mu, Fast fourier convolution, Adv. Neural Inf. Process. Syst. 33 (2020) 4479–4488.