

TLIS: Two-stage low-light image instance segmentation

Wei Li, Ya Huang*, Xinyuan Zhang, Guijin Han

School of Automation, Xi'an University of Posts and Telecommunications, 618 West Chang'an Avenue, Xi'an, 710121, Shaanxi, China.

Contributing authors: wli@xupt.edu.cn; xiyouhy@stu.xupt.edu.cn;
xiyouzxy@stu.xupt.edu.cn; hanguijin@xupt.edu.cn;

Abstract

Abstract: Current instance segmentation models perform adequately under normal light conditions, but encounter significant challenges in low-light scenarios, where the feature information of objects becomes less discernible, thus impeding accurate segmentation. The present study proposes a two-stage methodology for instance segmentation in low-light images. In stage-I, we enhance the low-light images while denoising and enhancing the detail information. To mitigate degradation phenomena such as noise and loss of details caused by low-light image enhancement (LLIE), we propose a post-processing module for enhancing details while reducing noise. This module uses a diffusion approach with an attention mechanism to model the conditional distribution between under-enhanced and naturally illuminated images. The enhancement results from the first stage are used as input to the segmentation network in the second stage, and we observe that incorporating fine-grained features into the mask branch of the segmentation network positively affects the accurate delineation of finer masks. To achieve this objective, we devise a wavelet feature fusion module to effectively integrate the high-resolution and low-resolution features within the feature pyramid, thereby preserving more intricate details. We achieve great segmentation results on LIS, an extremely low-light instance segmentation dataset. Detailed Comparative experiments and ablation studies show the advantages and excellent generalization ability of our model.

Keywords: Instance segmentation, Low light image, Post-processing detail enhancement denoising module, Wavelet feature fusion module

1 Introduction

Instance segmentation is an important direction in computer vision, which plays a huge role in medical images, automated driving, robotics, and other applications. Instance segmentation has achieved good results with the help of deep learning, but the segmentation accuracy of these methods can be significantly reduced by images in low-light scenes.

Existing state-of-the-art methods for low-light image enhancement are based on deep learning. Zhang et al. [1] proposes that the KinD network decomposes an image into two components, enabling adjustment of lighting and removal of degradation. There are uneven illumination artifacts in KinD, and Zhang et al. [2] develops KinD++ using

multi-scale attention to solve this problem. Zero-DCE [3] designs a set of non-reference loss functions to formulate low-light enhancement as an image-specific curve estimation task with a deep network. In view of the difficulty of obtaining pairs of low-light/normal light images, EnlightenGAN [4] proposes an unsupervised generative adversarial network that can be trained without image pairs and has strong generalization. Wang et al. [5] develops LLFlow, which utilizes conditional normalization flow to map naturally exposed images into Gaussian distributions to capture manifolds.

Recently, diffusion models (DMs) have shown significant advantages in image generation and denoising. Some typical DMs, such as DDPM [6] and fractional-based generation modeling methods [7], follow the principle of first adding

noise through forward propagation and then removing noise through propagation. In addition, DMs perform well in several image-wise tasks, such as image segmentation [8], object detection [9], unconditional image generation [10], etc. The learning capacity of DMs during training surpasses that of other generative models. Therefore, some works on low-light image enhancement have also introduced DMs. Yuan et al. [11] uses the denoising diffusion probability model (DDPM) with random damage conditions to enhance the performance of star field images. The LLDiffusion [12] algorithm seamlessly integrates degradation and image priors into the diffusion process, effectively enhancing low-light images. While DMs have achieved excellent results in low-light enhancement work, their drawbacks are also evident, such as large computational consumption of resources, high configuration requirements, and lengthy training times. To solve these problems, Jiang et al. [13] proposes a wavelet-based conditional diffusion model, which uses the advantages of wavelet transform to considerably improve the inference speed without loss of accuracy. PyDiff [14] uses a novel propagation method to gradually increase the resolution of the image in the form of a pyramid, which effectively accelerates the inference speed of the diffusion model. Moreover, low-light image enhancement may introduce different forms of image degradation during the enhancement process. Savvas et al. [15] proposes a diffusion model-based post-processing denoiser LPDM, and adopts a different denoising process to avoid the expensive iterative back-propagation process. We note that frequently removing noise comes at the cost of removing details, while downstream tasks such as image segmentation require more high-frequency details. The focus of our approach, unlike LPDM, is to simultaneously address noise and preserve image details. This ensures that the enhanced image closely resembles reality and is well-suited for instance segmentation tasks.

The instance segmentation model under normal illumination achieves great segmentation results. Mask R-CNN [16] is the basis of many two-stage model improvements. It adopts a strategy of first detection and then segmentation. CondInst [17] employs a fully convolutional network for single-stage instance segmentation, thereby eliminating the need for feature alignment during RoI cropping and mask generation. BCNet [18] is a two-layer instance segmentation network that effectively addresses the challenge of severe occlusion between instances by separating the modeling of occlusions and occluded objects. The study of instance segmentation in low-light conditions is still in its early stages compared to that of instances in normal light conditions, and the current research is relatively limited. Previous studies employ low-light enhancement techniques or image denoising algorithms as a preprocessing step. The processing criterion remains unchanged in our model. However, we place greater emphasis on mitigating noise generation and preserving detailed information during image enhancement. This ensures that the resulting image is more conducive to downstream tasks such as segmentation, with a heightened focus on achieving fine segmentation.

Our work aims to build a two-stage framework for instance segmentation of low-light images, which can effectively improve the instance segmentation accuracy of low-light images even in dark and noisy conditions. To this

end, we first investigate whether the enhanced low-light images can be directly used for instance segmentation, and find that noise and detail loss will be introduced in the process of low-light image enhancement. For instance segmentation of low-light images, the image needs to not only increase its brightness but also preserve more high-frequency details with little noise. Second, fine instance segmentation requires more detailed features to generate accurate masks. We have demonstrated that instance segmentation of low-light images can be efficiently achieved using a two-stage processing strategy of first enhancement and then segmentation. In summary, our contributions are as follows:

1) We propose a diffusion model-based detail enhancement denoising module (DEDM), which primarily addresses the degradation phenomenon caused by the LLIE network in low-light image enhancement. This module not only facilitates straightforward image denoising but also enhances the generation of detailed information. At the same time, our module effectively improves the performance of dark images in downstream tasks.

2) We utilize the potential of wavelet decomposition in feature fusion to design a wavelet feature fusion module (WFFM) for improving the feature extraction stage in the segmentation network, which better integrates multiscale feature information in the feature pyramid as input for the segmentation branch, which is conducive to generating refined masks. Remarkably, it is also beneficial for instance segmentation in normal light.

3) We compare our method with various pipeline methods on the LIS [19] very low light image dataset and demonstrate its excellent low light image segmentation performance. We conduct multiple ablation experiments on each module of the algorithm to demonstrate the effectiveness and generalization of our method.

2 Proposed Method

In this chapter, we first give a brief review of the diffusion model (Section 2.1), and then, in (Section 2.2), we detail the application of the detail enhanced denoising module in the first stage. In order to generate more detailed masks, we describe in (Section 2.3) the segmentation process by adding the instance segmentation network of the wavelet feature fusion module.

2.1 Diffusion Model

DMs are a form of generative models that gradually transform Gaussian noise distributions into training data by learning Markov chains. DMs add noise sequentially in the forward process and denoise in the back propagation process. c is used to represent low-light images, and x_0 is the image under normal light. Given a conditional data distribution $x_0 \sim q(x_0|c)$, the forward noise process is defined as [6]:

$$q(x_{1:T}|x_0) := \prod_{t=1}^T q(x_t|x_{t-1}), \quad (1)$$

$$q(x_t|x_{t-1}) := N(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t\mathbf{I}), \quad (2)$$

Where $\beta_t \in (0, 1)$ is the variance, N is the Gaussian distribution, $t \in \{1, \dots, T\}$ is the time step, and T is set to a large

number ($T = 1000$) so that the backward process can better approximate the Gaussian distribution. As t increases, the noise I gradually increases, and the amount of noise added each time is positively correlated with t . Subject to input x_0 , when $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{s=1}^t \alpha_s$, can be expressed as:

$$q(x_t|x_0) := N(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)\mathbf{I}). \quad (3)$$

Thus, in the case of a given x_0 , x_t can be expressed as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon, \quad (4)$$

Where $\varepsilon \sim N(0, 1)$ is a random source. The inverse process of the conditional diffusion model is to model $q(x_{t-1}|x_t, c)$ through x_t and c , which learns the inverse process of dynamic perception degradation:

$$p_\theta(x_{0:T}|c) := p(x_T) \prod_{t=1}^T p_\theta(x_{t-1}|x_t, c), \quad (5)$$

$$p_\theta(x_{t-1}|x_t, c) := N(x_{t-1}; z_\theta(x_t, t), \sum_\theta(x_t, t)). \quad (6)$$

To avoid learning variance, let $\sum_\theta(x_t, t) = \sigma_t^2 \mathbf{I}$, where $\sigma_t^2 = \frac{1-\alpha_t-1}{1-\alpha_t} \beta_t$ be a time-varying constant. Thus, the only learnable component is z . In the process of each training, the DM is parameterized by the denoising autoencoder $\varepsilon_\theta(x_t, t, c)$ to predict x_0 , and the final image $\hat{x}_0$ is obtained through the repeated denoising process. Therefore, the corresponding model loss function is defined as follows:

$$L_{DM} = \mathbb{E}_{x_0, c, \varepsilon, t} [\|\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\varepsilon, t, c)\|^2], \quad (7)$$

Given x_t , the model training is constrained using the difference between the predicted value ε_θ and the true value ε . Finally, through the diffusion model, we can calculate the estimated $\hat{x}_0$ of x_0 :

$$x_0 \approx \hat{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}x_t - (\sqrt{\frac{1}{\bar{\alpha}_t}} - 1)\varepsilon_\theta(x_t, t, c). \quad (8)$$

2.2 Detail Enhancement Denoising Module

The task of instance segmentation requires normal light images with intricate details and minimal noise. While addressing low-light scenarios, it seems plausible to employ LLIE for image enhancement. However, this may introduce diverse degradation artifacts in the processed images. LPDM adopts a diffusion model for post-processing denoising and reduces the computational overhead of the reverse generation process of typical diffusion models. LPDM estimates the noise in the enhanced image $\hat{x}_0^\eta$:

$$n_\phi = \varepsilon_\theta(\hat{x}_0^\eta, \phi, c), \quad (9)$$

Where ϕ takes the value of 300, and when an estimate of noise n_ϕ is obtained, the noise is subtracted from $\hat{x}_0^\eta$ using the modification of formula (8):

$$\hat{x}_0^{DM} = \xi(\hat{x}_0^\eta, n_\phi, s) := \frac{1}{\sqrt{\bar{\alpha}_s}}\hat{x}_0^\eta - (\sqrt{\frac{1}{\bar{\alpha}_s}} - 1)n_\phi, \quad (10)$$

Among them, s is the time step and should be significantly less than ϕ . Most importantly, the value of s will directly affect the effect of correction.

Following the work of LPDM, we aim to design a detail enhanced denoising model that enhances image detail while denoising and avoiding artifacts. We propose DEDM by modifying the structure of the U-Net network part in the diffusion model, as shown in Fig. 1, the image I_1 is obtained by fusing the low-light image c and the noise image x_t . I_1 and the encoder feature I_2 are used as the input of the detail attention (DA) module, and the output result is input into the symmetry layer of the decoder. The detail attention module can assist the model in incorporating appropriate detailed feature information during the decoding process, thereby enhancing its ability to recover image details while simultaneously eliminating noise. The operation of DA can be expressed as follows:

$$I_a = \text{soft max}(I_1' \odot I_2'), \quad (11)$$

$$I_b = (I_a \odot I_1') \oplus I_1', \quad (12)$$

$$I_c = (I_a \odot I_2') \oplus I_2', \quad (13)$$

$$I_{out} = \text{Conv}[I_b, I_c], \quad (14)$$

Where, $I_1' \in \mathbb{R}^{C' \times H' \times W'}$ is obtained from $I_1 \in \mathbb{R}^{C \times H \times W}$ after a 3×3 convolution operation, $I_2' \in \mathbb{R}^{C' \times H' \times W'}$ is obtained from $I_2 \in \mathbb{R}^{C \times H \times W}$ after a 1×1 convolution operation, $[\cdot, \cdot]$ means that features are serialized in the channel dimension, $\odot$ means the dot product operation, $\oplus$ means the addition of two features. Conv is a 1×1 convolution operation, and I_{out} as the final output, its size and symmetry decoding layer of the same size.

The application of DA makes our DEDM not only an enhanced image denoising device, but also effectively preserves more texture details, which effectively lays the foundation for the next instance segmentation task.

2.3 Wavelet Feature Fusion Module

The enhancement effect of low-light images is crucial for the quality of segmentation, and the utilization of feature information in segmentation networks also plays a decisive role. In low-light scenes, there are also objects that block and overlap with each other. Therefore, BCNet is adopted as the baseline of the second-stage instance segmentation network, which decouples the boundary of the occluding objects and the occluded instance, effectively improving the instance segmentation effect of the occluded object. In general, the model is divided into three parts: feature extraction network, detection head, and mask head. The feature extraction network adopts ResNet-FPN [20], the region of interest features are extracted from the FPN multi-scale feature $\{P_3, P_4, P_5\}$ for mask prediction. We use WFFM to enhance the feature extraction network and designate the improved segmentation model as W-BCNet, as depicted in Fig. 2.

The feature map with more detailed information is more conducive to generating fine masks as the input of mask branches. To this end, we hope to add feature figure P_2 with the highest resolution of the feature pyramid. An intuitive method is the feature fusion of P_2 and P_3, P_4, P_5 by simple addition through up-sampling or down-sampling, but it cannot efficiently use multi-scale features.

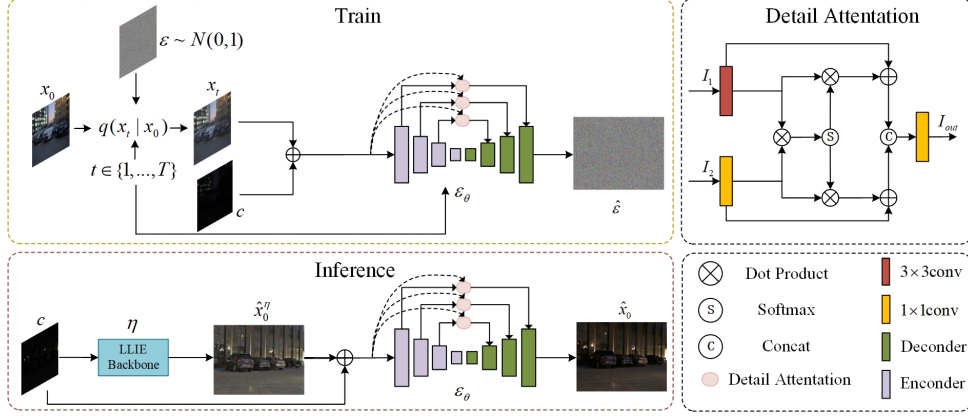


Fig. 1 Flowchart of the training and inference phases of DEDM.

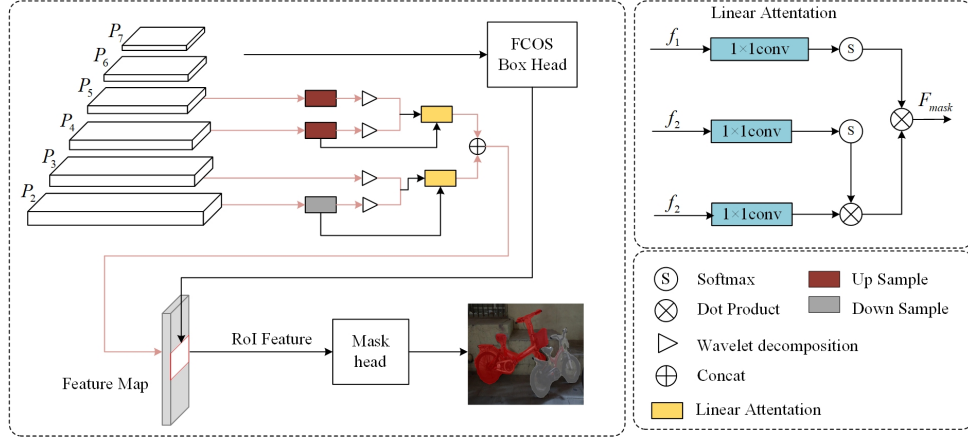


Fig. 2 W-BCNet instance segmentation framework. (We omit the image input as well as ResNet in the feature extraction network for ease of presentation. The pink arrow is WFFM.)

We find that the wavelet transform is able to extract different frequency components from the image, with high-frequency components representing the details in horizontal, vertical and diagonal directions, so that introducing high-frequency components is more beneficial to preserve the detailed features. We choose the Haar wavelet transform to use four convolution kernels to extract components of different frequencies for features of different scales in the feature pyramid, fully considering that high-frequency details and low-frequency components are indispensable for the composition of features. We fuse the decomposed components through linear attention and re-use them as input to the mask branch. We refer to this process as wavelet feature fusion, which can effectively combine high-frequency detail components and low-frequency components in features of different scales. The whole fusion process is shown in Fig. 2. Specifically, we first up-sample P_4 , P_5 to get P'_4 and P'_5 , and down-sample P_2 to get P'_2 , so that their dimensions are the same as P_3 , and then use Haar wavelet transform to decompose them into four components respectively, including a principal component and three horizontal, vertical and diagonal detail high-frequency components. We

denote this process as follows:

$$H(P_2') = LL(P_2'), LH(P_2'), HL(P_2'), HH(P_2'), \quad (15)$$

$$H(P_3) = LL(P_3), LH(P_3), HL(P_3), HH(P_3), \quad (16)$$

$$H(P_4') = LL(P_4'), LH(P_4'), HL(P_4'), HH(P_4'), \quad (17)$$

$$H(P_5') = LL(P_5'), LH(P_5'), HL(P_5'), HH(P_5'), \quad (18)$$

Here, $LL \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$ represents the low-frequency principal component, $LH \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$ represents the detail component in the horizontal direction, $HL \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$ represents the detail component in the vertical direction, and represents the detail component in the diagonal direction. $HH \in \mathbb{R}^{C \times \frac{H}{2} \times \frac{W}{2}}$ denotes the Haar wavelet transform.

The high-frequency detail features contribute to more precise segmentation, while the low-frequency components are also indispensable. We combine the low-frequency principal component of P_3 and the high-frequency detail component of P'_4 and the low-frequency principal component of P'_5 and the high-frequency detail component of P'_4 respectively through linear attention. We add the obtained $A_{2,3}$

and $A_{4,5}$ to obtain the fused feature $F_{mask} \in \mathbb{R}^{C \times H \times W}$. The entire process can be expressed as:

$$q_i = W_{q_i} \otimes f_i + b_i, \quad (19)$$

$$[k_i, v_i] = W_{k_i, v_i} \otimes [LL, LH, HL, HH] + b_i, \quad (20)$$

$$A_{n,m} = \text{soft max}(q_i) \odot (\text{soft max}(k_i) \odot v_i), \quad (21)$$

$$F_{mask} = A_{4,5} \oplus A_{2,3}, \quad (22)$$

Here, $i \in \{1, 2\}$, f_1 and f_2 represent P_4 and P_2 , respectively, while q_i denotes the corresponding output of feature f_i . k_1, v_1 are the tensor obtained by concatenating $LL(P_5'), LH(P_4'), HL(P_4'), HH(P_4')$ along the channel. The procedure for obtaining k_2, v_2 when $i = 2$ is consistent with the above. $\otimes$ is the convolution operation, W_{q_i} is the weight, b_i is the bias. The result F_{mask} contains both high frequency and low frequency information, enabling better preservation of details for fine segmentation.

Table 1 Comparison of segmentation performance between DEDM and LPDM on different LLIE networks.

Method	AP	AP ₅₀	AP ₇₅	AP ^{box}	AP ₅₀ ^{box}	AP ₇₅ ^{box}
RetinexNet	18.6	31.4	19.1	21.7	35.2	23.5
+LPDM	23.6	38.9	23.7	26.8	43.4	28.4
+DEDM	24	39.6	24.2	27.4	44.1	28.9
EnlightenGAN	20.1	34.8	19.7	22.5	38.2	26.9
+LPDM	27.5	49.8	27.6	31.3	52.4	37.1
+DEDM	28.1	50.7	28.2	32.2	53.5	37.9
ZeroDCE++	21.3	37.5	20.8	23.7	40.1	27.8
+LPDM	33.2	54.8	32.4	38.4	60.1	39.0
+DEDM	33.8	55.7	33.0	39.3	61.2	39.8
LLFormer	25.6	41.4	23.6	26.1	43.3	30.8
+LPDM	32.1	54.2	32.1	37.2	60.0	38.1
+DEDM	33.4	55.6	33.9	38.8	61.2	40.1

Table 2 Comparison of segmentation performance of different LLIE networks with BCNet and W-BCNet. (where the method without * means that the segmentation network uses BCNet and the method with * means that the segmentation network uses W-BCNet.)

Method	AP	AP ₅₀	AP ₇₅	AP ^{box}	AP ₅₀ ^{box}	AP ₇₅ ^{box}
RetinexNet	18.6	31.4	19.1	21.7	35.2	23.5
RetinexNet*	19.2	32.2	20.6	22.3	36.1	24.6
EnlightenGAN	20.1	34.8	19.7	22.5	38.2	26.9
EnlightenGAN*	20.5	35.5	20.8	23.0	38.8	27.8
ZeroDCE++	21.3	37.5	20.8	23.7	40.1	27.8
ZeroDCE++*	22.0	38.5	22.2	24.5	41.2	29.2
LLFormer	25.6	41.4	23.6	26.1	43.3	30.8
LLFormer*	26.1	42.0	24.6	26.6	43.9	31.9

2.4 Framework

As shown in Fig. 3, we implement a two-stage processing method for low light image enhancement and instance segmentation, effectively solving the degradation phenomenon

in image enhancement and the problem of rough segmentation masks.

Table 3 WFFM affect network normal light image segmentation.

Method	backbone	AP	AP ₅₀	AP ₇₅
mask-R-CNN		35.6	57.6	38.1
+WFFM	R-50-FPN	36.4	58.6	39.4
CondInst		35.9	56.9	38.3
+WFFM	R-50-FPN	36.5	57.7	39.1
BCNet		38.4	59.6	41.5
+WFFM	R-50-FPN	39.4	60.6	42.7

Method	backbone	AP _S	AP _M	AP _L
mask-R-CNN		18.7	38.3	46.6
+WFFM	R-50-FPN	19.4	38.9	47.2
CondInst		19.1	38.6	46.8
+WFFM	R-50-FPN	19.5	39.1	46.4
BCNet		21.9	40.9	49.3
+WFFM	R-50-FPN	22.8	41.7	50.1

In stage-I, we adopt ZeroDCE++ [21] as LLIE network to enhance low-light image. Since the enhanced image still has problems such as noise and artifacts, we add DA to the U-Net structure of the diffusion model to construct a DPDM to enhance the learning of detailed information in the diffusion process. After the LLIE enhanced image is processed by DPDM, as shown in Fig. 3, the obtained image $\hat{x}_0$ is more clear and rich in texture information.

We take the output image $\hat{x}_0$ of the first stage as the input of the second stage, and construct the W-BCNet model to segment the input image to get the segmentation result x_{mask} . The combination of ResNet and FPN cannot extract detailed features efficiently. We design a new fusion method WFFM to take the multi-scale feature fusion in FPN as the input of the segmentation branch. WFFM makes full use of high-frequency detail information, so that the input feature map of the mask branch contains more detail features, and the generated mask is more refined.

3 Experiment

In this section, we introduce the dataset used in this paper and its related evaluation metrics, and define the relevant training details. In addition, we design DEDM and WFFM ablation studies for quantitative comparison of the proposed approaches to ours.

3.1 Dataset and Evaluation Index

Our study focuses on training using the paired non-synthetic low-light dataset LIS, which comprises 2230 image pairs randomly divided into 1561 training pairs and 669 test pairs. The dataset encompasses diverse scenes, including both indoor and outdoor environments, with each image pair meticulously annotated at the pixel level. For every scene in the LIS dataset, four types of images were captured: long-exposure normal-light images and short-exposure low-light images in both RAW and sRGB formats.

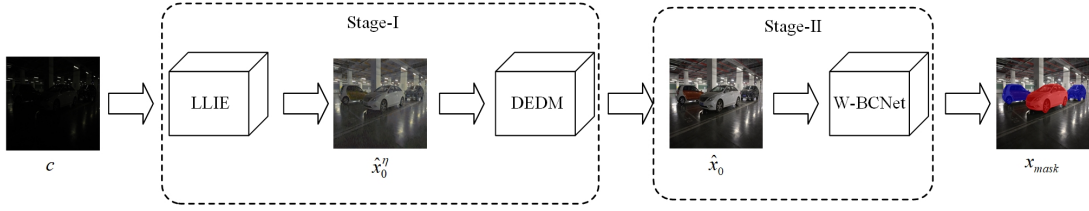


Fig. 3 Overall framework.

Table 4 Quantitative comparisons of low-light instance segmentation.

Preprocessing method	Segmentation Method	AP	AP ₅₀	AP ₇₅	AP ^{box}	AP ₅₀ ^{box}	AP ₇₅ ^{box}
-	BCNet	20.9	37.1	19.7	22.9	40.2	27.5
RetinexNet	BCNet	19.4	31.7	19.6	21.4	35.2	25.7
EnlightenGAN	BCNet	20.1	34.8	19.7	22.5	38.2	26.9
ZeroDCE++	BCNet	21.3	37.5	20.8	23.7	40.1	27.8
KinD++	BCNet	23.9	38.2	21.3	24.5	40.9	29.1
URetinexNet	BCNet	24.9	38.6	21.5	25.2	41.4	30.3
LLFormer	BCNet	25.6	41.4	23.6	26.1	43.3	30.8
LPDM	BCNet	33.2	54.8	32.4	38.4	60.1	39
LIS	BCNet	32.8	53.7	32.6	38.4	59.9	43.2
DPDM	BCNet	33.8	55.7	33	39.3	61.2	39.8
DPDM	W-BCNet	34.5	56.7	34.4	40.1	62.3	41.2

In this work, we exclusively utilize the normal and low-light images in the sRGB format. We use the COCO[22] dataset evaluation metrics AP (average threshold from 0.5 to 0.95, interval 0.05), AP₅₀ (IoU 0.5), AP₇₅ (IoU 0.75) to measure the performance of segmentation, including the AP_S, AP_M, AP_L for multiscale segmentation performance. We also provide detection results including AP^{box}, AP₅₀^{box}, and AP₇₅^{box}.

3.2 Training Details

We adopt a two-stage training approach, using Adam optimizer and PyTorch framework for training and testing on NVIDIA RTX 3080Ti Gpus. In stage-I, we train DPDM, where the time step T is set to 1000, the learning rate is set to, and the loss function is defined as in Equation (6). We employ random flipping for data augmentation and train 8000 epochs on the LIS training set with a batch size of 4 and gradient accumulation of 8 batches to simulate a batch size of 32. In stage-II, our instance segmentation framework is trained on the COCO2017 dataset. To speed up the convergence of the model, we adopt the pre-trained model of BCNet for model initiation, which also has the same loss function defined by BCNet. The model is trained for 10k iterations with batch size set to 8 and initial vector set to 0.01.

3.3 Ablation Study

In this section, we conduct ablation studies to validate the efficacy of our proposed modules and techniques. Firstly, we perform an ablation study on DEDM and confirm the significance of detail enhancement and denoising in dark-light image enhancement, as illustrated in Fig. 2, which effectively enhances segmentation performance. Subsequently, we propose integrating wavelet feature modules

to construct a novel instance segmentation network called W-BCNet and compare its segmentation performance with different LLIE networks. Additionally, we also evaluate the performance of incorporating WFFM into a conventional optical segmentation network.

DEDM: We compare DEDM with LPDM and discuss their disparities. While both networks employ DMs for post-processing operations on enhanced images, our DEDM primarily focuses on detailed information recovery while also addressing denoising. To validate the superior performance of DEDM, we incorporate four LLIE networks into both LPDM and DEDM for comparative analysis. In order to ensure fairness, We set the segmentation network as BCNet. As depicted in Table 1, it is evident that the addition of post-processing operations significantly enhances the segmentation accuracy of the four LLIE networks. Most notably, ZeroDCE++ with only LPDM operation sees an AP₅₀ improvement from 37.5 to 54.8, while the AP₅₀^{box} improvement for EnlightenGAN was 14.2(from 38.2 to 52.4). When our DEDM is also added to ZeroDCE++, AP₅₀ is directly increased from 37.5 to 55.7, which is 0.9 higher than LPDM, and the AP₅₀^{box} of the EnlightenGAN is directly increased by 15.3(from 38.2 to 53.5). Therefore, we can conclude that the enhanced images generated by DEDM can effectively improve the instance segmentation performance of low-light images.

WFFM: As evident from Table 2, the segmentation accuracy of D-BCNet is significantly improved compared to the baseline, thereby highlighting the effectiveness of the wavelet feature fusion module. In contrast to traditional approaches such as multiscale feature fusion, which involve sampling features at different resolutions or employing direct sum fusion, our proposed method enables the generation of fine masks while preserving global information for the mask branch. WFFM introduces a feature map P_2 rich in texture detail information and then fuses the high-frequency detail information of each scale feature

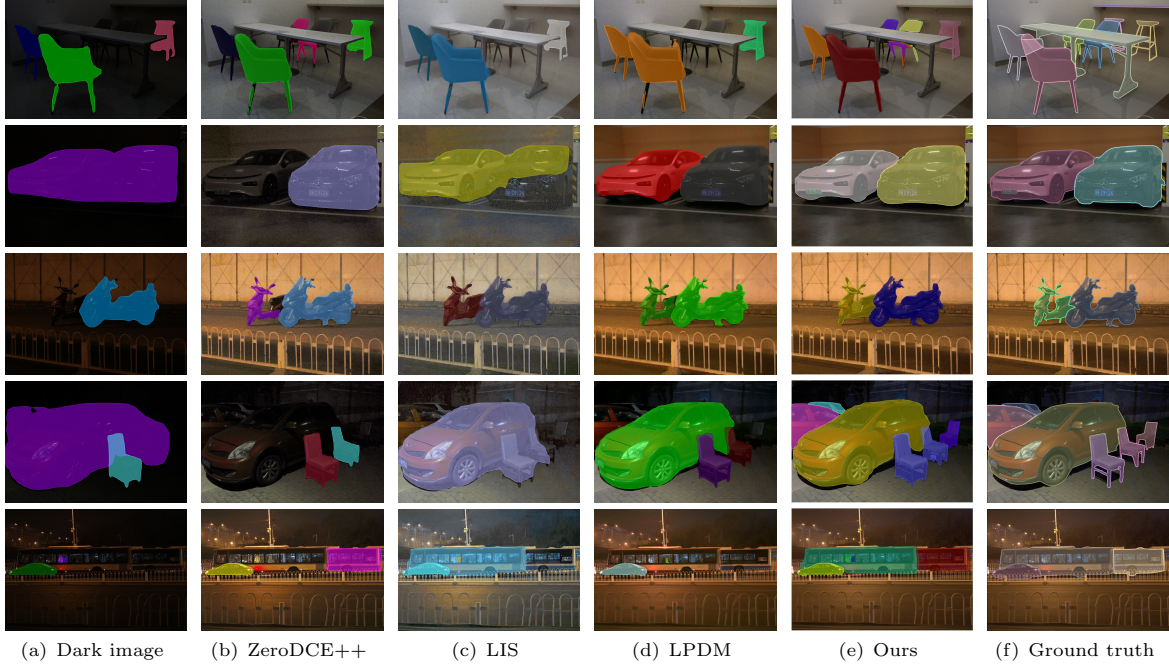


Fig. 4 Comparison of various pipeline segmentation results on LIS dataset.

obtained by wavelet decomposition to help us solve the coarse mask problem. The LLIE network adopt here does not employ any post-processing methods. When BCNet is used in the segmentation network, LLFormer achieves the highest accuracy with AP_{50} of 41.4. However, our W-BCNet can make its segmentation accuracy reach 42.0, which is beyond 0.6 of the BCNet network. We also have clear advantages over different methods.

In this study, we conduct experiments to assess the impact of WFFM on a normal light image instance segmentation network. Specifically, WFFM is incorporated into the backbone networks of CondInst (a one-stage instance segmentation network), Mask R-CNN (a two-stage instance segmentation network), and BCNet. The models are trained on the COCO dataset using ResNet-50-FPN as the backbone network. As shown in Table 3, CondInst achieves a boost of 0.6 AP and the two-stage network BCNet achieves a boost of up to 1.0 AP. The segmentation effect of WFFM is more clearly improved in normal light, which also achieves excellent results for improving the AP value of multi-scale images. It also helps the network to focus on detailed features during the feature extraction stage. Fortunately, although we use it for instance segmentation in low-light images, it can also improve the segmentation task in normal lighting. We can argue that WFFM can be applied in the domain of additional attention to detail features.

3.4 Comparative Experiments

In addition, we compare the proposed method with four pipelines, namely direct prediction of low-light images, prediction of augmented images, prediction of augmented images with post-processing denoising, and an end-to-end

method for instance segmentation of low-light images. We choose some deep learning based low-light image enhancement methods such as RetinexNet [23], EnlightenGAN, ZeroDCE++, KinD++, URetinex-Net [24], LLFormer [25] and uses LPDM for denoising. We also consider the LIS end-to-end low-light instance segmentation method, which takes RAW images as input, and the remaining methods take sRGB images as input. Table 4 shows all the pipelines.

In addition to our method, the baseline network BCNet is used for the remaining segmentation methods. To reflect the actual application situation more accurately, we consider the LIS dataset as a test set. As shown in Table 4, the segmentation accuracy of BCNet instance network in low-light images is only 20.9 AP directly compared to the segmentation performance in normal-light environment. We use to add the augmentation network pipeline, which we think would highly boost the segmentation accuracy, but the results are not very favorable, the base accuracy remains the same or even decreased. We conjecture that the degradation phenomenon is caused by the enhancement process, which leads to increased noise and loss of detail. To test our conjecture, we introduce LPDM into the pipeline, and the results prove that the segmentation results are significantly improved, such as Zero-DCE add LPDM. Demonstrating the importance of denoising for enhanced images. We then experimented with LIS and although it is an end-to-end network, the augmented images and segmentation results did not live up to our hopes.

Although these enhancement and denoising steps have significantly improved the segmentation accuracy, we find that, as shown in Fig. 4, the pipeline partition graphics division of heavy leakage of segmentation, error and mask rough phenomenon. To this end, we hope to enhance image

denoising and at the same time pay attention to the detail of the image processing, and the image feature extraction part make improvements. The proposed method uses the enhanced network with ZeroDCE++ plus DEDM post-processing, and the segmentation method uses W-BCNet, and our accuracy is 1.7 AP higher than the most competitive LIS. It turns out that our method achieves good results and can achieve advanced segmentation results even in complex scenes.

4 Conclusion

In this paper, we introduce a two-stage approach for instance segmentation of low-light images by first enhancing and then segmenting. In stage-I, we propose a DEDM post-processing method for the degradation of the basic enhancement network, which enhances the image and solves the problem of noise and detail loss. In stage-II, wavelet feature fusion is used to efficiently use high-frequency detail features in the feature extraction stage to achieve fine segmentation. It is worth noting that our method highly fills the gap of low-light image instance segmentation methods. We hope that our experimental results and ideas can inspire more work in low light environments.

Declarations

- Funding
This work was supported by the Key Research and Development Plan General Project of Shaanxi Provincial Science and Technology Department under Grants (No.2023-YBGY-032).
- Conflict of interest/Competing interests (check journal-specific guidelines for which heading to use)
The authors declare that they have no competing interests.
- Ethics approval
Not applicable.
- Consent to participate
Not applicable.
- Consent for publication
Not applicable.
- Availability of data and materials
The datasets used or analyzed during the current study are available from the corresponding author on reasonable request.
- Code availability
Not applicable.
- Authors' contributions
Wei Li wrote the manuscript, Ya Huang revised the article and designed the experiment, Xinyuan Zhang visualized the results and sorted out the data, and Guijin Han provided financial and basic support.

References

- [1] Zhang, Y., Zhang, J., Guo, X.: Kindling the Darkness: A Practical Low-light Image Enhancer (2019)

- [2] Zhang, Y., Guo, X., Ma, J., Liu, W., Zhang, J.: Beyond brightening low-light images. *International Journal of Computer Vision* **129**(2) (2021)
- [3] Guo, C., Li, C., Guo, J., Loy, C.C., Hou, J., Kwong, S., Cong, R.: Zero-reference deep curve estimation for low-light image enhancement. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1777–1786 (2020). <https://doi.org/10.1109/CVPR42600.2020.00185>
- [4] Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P., Wang, Z.: Enlighten-gan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing* **30**, 2340–2349 (2021) <https://doi.org/10.1109/TIP.2021.3051462>
- [5] Wang, Y., Wan, R., Yang, W., Li, H., Chau, L.-P., Kot, A.C.: Low-Light Image Enhancement with Normalizing Flow (2021)
- [6] Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models (2020)
- [7] Song, Y., Ermon, S.: Generative Modeling by Estimating Gradients of the Data Distribution (2020)
- [8] Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: DifuMask: Synthesizing Images with Pixel-level Annotations for Semantic Segmentation Using Diffusion Models (2023)
- [9] Chen, Z., Gao, R., Xiang, T.-Z., Lin, F.: Diffusion Model for Camouflaged Object Detection (2023)
- [10] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10674–10685 (2022). <https://doi.org/10.1109/CVPR52688.2022.01042>
- [11] Yuan, Y., Wu, J., Wang, L., Jing, Z., Leung, H., Zhu, S., Pan, H.: Learning to Kindle the Starlight (2022)
- [12] Wang, T., Zhang, K., Shao, Z., Luo, W., Stenger, B., Kim, T.-K., Liu, W., Li, H.: LLDiffusion: Learning Degradation Representations in Diffusion Models for Low-Light Image Enhancement (2023)
- [13] Jiang, H., Luo, A., Han, S., Fan, H., Liu, S.: Low-Light Image Enhancement with Wavelet-based Diffusion Models (2023)
- [14] Zhou, D., Yang, Z., Yang, Y.: Pyramid Diffusion Models For Low-light Image Enhancement (2023)
- [15] Panagiotou, S., Bosman, A.S.: Denoising Diffusion Post-Processing for Low-Light Image Enhancement (2023)

- [16] He, K., Gkioxari, G., Dollr, P., Girshick, R.: Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017). <https://doi.org/10.1109/ICCV.2017.322>
- [17] Tian, Z., Shen, C., Chen, H.: Conditional Convolutions for Instance Segmentation (2020)
- [18] Ke, L., Tai, Y.-W., Tang, C.-K.: Deep Occlusion-Aware Instance Segmentation with Overlapping BiLayers (2021)
- [19] Chen, L., Fu, Y., Wei, K., Zheng, D., Heide, F.: Instance Segmentation in the Dark (2023)
- [20] Lin, T.-Y., Dollr, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature Pyramid Networks for Object Detection (2017)
- [21] Li, C., Guo, C., Loy, C.C.: Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(8), 4225–4238 (2022) <https://doi.org/10.1109/TPAMI.2021.3063604>
- [22] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollr, P.: Microsoft COCO: Common Objects in Context (2015)
- [23] Wei, C., Wang, W., Yang, W., Liu, J.: Deep Retinex Decomposition for Low-Light Enhancement (2018)
- [24] Wu, W., Weng, J., Zhang, P., Wang, X., Yang, W., Jiang, J.: Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5891–5900 (2022). <https://doi.org/10.1109/CVPR52688.2022.00581>
- [25] Wang, T., Zhang, K., Shen, T., Luo, W., Stenger, B., Lu, T.: Ultra-High-Definition Low-Light Image Enhancement: A Benchmark and Transformer-Based Method (2022)