2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Latent Disentanglement for Low Light Image Enhancement

Zhihao Zheng[1] , Mooi Choo Chuah[1]

Abstract— **Many learning-based low light image enhancement (LLIE) algorithms are based on the Retinex theory. However, the Retinex-based decomposition models introduce corruptions which limit their enhancement performance. In this paper, we propose a Latent Disentangle-based Enhancement Network (LDE-Net) for low light vision tasks. The latent disentanglement module disentangles the input image in latent space such that no corruption remains in the disentangled Content and Illumination components. For LLIE task, we design a Content-Aware Embedding (CAE) module that utilizes Content features to direct the enhancement of the Illumination component. For downstream tasks (e.g. nighttime UAV tracking and low light object detection), we develop an effective light-weight enhancer based on the latent disentanglement framework. Comprehensive quantitative and qualitative experiments demonstrate that our LDE-Net significantly outperforms state-of-the-art methods on various LLIE benchmarks. In addition, the great results obtained by applying our framework on the downstream tasks also demonstrate the usefulness of our latent disentanglement design.**

I. Introduction

The emergence of low-cost embedded devices with great computing power and small powerful cameras have empowered many robotic applications in recent years. Deep learning based models have been designed for such robotic applications, however, such models are typically trained using images captured during day time. In real world deployments, we typically encounter scenarios with varying light conditions e.g. night-time operations. Images captured under low-light conditions typically contain noise, low contrasts and result in performance degradation using models trained on normal light images. To overcome such limitations, much research has been spent on designing low-light image enhancement techniques.

In the past, many deep learning-based methods have been proposed to address the LLIE problems. Some general methods regard low-light image enhancement as a restoration task and construct models to learn an overall mapping between low-light and normal-light images. Although such general methods achieve better results than traditional approaches, most of them still exhibit unsatisfactory enhancement results such as uneven illumination, low efficiency, lack of robustness to noise and structure details etc.

Decomposition-based LLIE methods first decompose images into two components and then design specific enhancement technique for each component. Among all these decomposition methods, LLIE models based on Retinex theory have attracted the most attention. Retinex LLIE methods usually decompose images into a illumination map and a reflectance image, then learn to bright up the illumination map and subsequently fuse the improved illumination map with the denoised reflectance image. They impose a structure-aware loss function to constrain the enhancement of the illumination component during training but such design results in worse decomposition results.

Based on the above observation, we propose a latent disentanglement framework for low-light image enhancement and other downstream vision tasks (e.g. object detection and tracking). Our design disentangles images into light-specific component (i.e. Illumination) and light-invariant component (i.e. Content) in latent feature space for better disentangle results. Then, only the Illumination component requires enhancement during the enhancement stage since the Content component is invariant under varying light conditions. Furthermore, we propose a Content-Aware Embedding module (CAE) to explicitly learn the correlation between the Content and Illumination features, thereby improving the performance of the Illumination enhancement network. Last but not least, taking advantage of our latent disentanglement design, we can easily develop light-weight enhancement networks to improve performance for downstream low light vision tasks.

In summary, the contributions are summarized as:

- Unlike existing Retinex-based decomposition methods, we propose a transformer-based latent disentanglement framework to get better disentanglement results for low light vision tasks.
- We design a Content-Aware Embedding module (CAE), which learns the correlation between the Content and Illumination features, thereby improving the LLIE performance of our LDE-Net. Extensive experimental results on LLIE benchmark datasets demonstrate that our LDE-Net outperforms SOTA methods.
- Benefiting from the proposed latent disentanglement design, we implement extremely light weight enhancement network for downstream low light vision tasks (e.g. nighttime UAV tracking and low light object detection) that boost the performance of existing methods. Due to the page limits The full version of our LDE-Net with the light-weight design for downstream tasks can be found at the following link: https://arxiv.org/pdf/2408.06245.

*This work was partially supported by National Science Foundation Grant CPS 1931867, and a gift from Qualcomm Technologies, Inc.

[1]Computer Science and Engineering department, P.C. Rossin College of Engineering and Applied Science, Lehigh University, Bethlehem, PA 18015, USA. {zhzc21@lehigh.edu, chuah@cse.lehigh.edu}

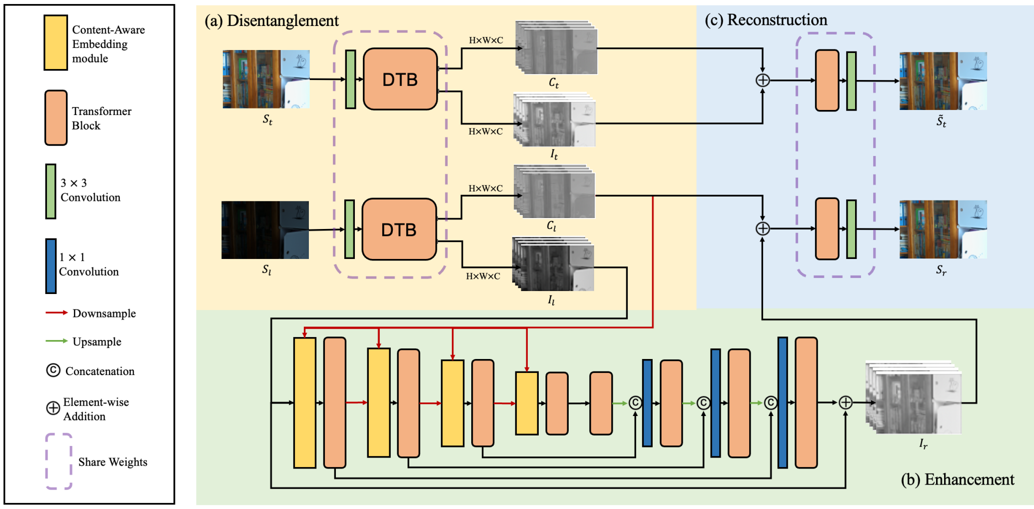

Fig. 1. Overview of the proposed LDE-Net. (a)The disentanglement module disentangle an input image to the Content and Illumination components, both of which are latent space features. (b)Content-Aware Illumination Enhancement module enhances the Illumination with the guidance of Content feature. (c)The reconstruction module reconstructs a new image with the Content and restored Illumination component.

II. Related Work

Recent deep learning based methods show promising results [1], [2], [3], [4]. We can further divide existing designs into Retinex-based methods and end-to-end methods.

Retinex-based methods. Retinex-based methods use deep network to decompose and enhance an image. Wei et al. proposed a two-stage Retinex-based method called Retinex-Net [4]. Inspired by Retinex-Net, Zhang et al. proposed two refined methods, called KinD [5] and KinD++ [6]. Recently, Wu et al. proposed a novel deep unfolding Retinex-based network to further integrate the strengths of model-based and learning-based methods [7].

End-to-End methods. In comparison to Retinex-based method, recent end-to-end methods directly learn an overall mapping to generate an enhanced image. Lore et al. [8] made the first attempt by proposing a deep autoencoder model named Low-Light Net (LLNet). Later on, various end-to-end methods are proposed. Physics-based concepts, e.g. Laplacian pyramid [9], normalization flow [10] and wavelet transform [11], are proposed to improve model interpretability and lead to more satisfactory results. In [12], [13], adversarial learning is introduced to capture the visual properties. In [14], [15], the light enhancement is creatively formulated as a task of image specific curve estimation using zero-shot learning. Recently, Transformer-based methods have been applied to LLIE tasks. Wang et al. [16] adopts a modified swin-transformer block to build a U-shaped network; Zamir et al. [17] introduces a transposed transformer block to improve feature aggregation for image restoration; Wang et al. [18] design an Axis-based transformer block to further improve efficiency of the attention mechanism on LLIE tasks.

III. Methodologies

A. Motivation

It is extremely challenging to extract the illumination component from a single image. Most of the previous decomposition models are based on the well-known Retinex theory. According to the Retinex theory, an input image $S \in R^{h\times w\times 3}$ can be decomposed into a 3-channel reflectance image $R \in R^{h\times w\times 3}$and a one-channel illumination map $I \in R^{h\times w\times 1}$ as

$$S = R \odot I \tag{1}$$

where, $\odot$ denotes the element-wise multiplication. Modern Retinex methods use deep learning models to estimate the illumination map I and corresponding reflectance image R using the above equation. Small errors randomly generated by such models will be amplified by this dot-product related operation and lead to inaccurate estimation. Thus, previous Retinex-based decomposition models [4], [19], [5], [7], [20] cannot obtain satisfactory decomposition results and hence require further enhancement to the decomposed reflectance images during the adjustment step. To overcome this problem, we use a disentanglement strategy inspired by Causal theory to extract the illumination component from the input image by operating in the latent feature space instead of the RGB image pixel space as in the traditional Retinex-based approaches.

B. Overview

In this work, we propose a Latent Disentangle-based Enhancement Network (LDE-Net) as illustrated in Fig 1 which consists of three modules: the disentanglement module, the enhancement module and the reconstruction module. The overall model operations can be represented as follows:

$$S_r = M_{recon}(M_{enh}(I, C), C) = M_{recon}(M_{enh}(M_{dis}(S))) \tag{2}$$

where S is an input image and S_r is the restored image. The disentanglement module M_{dis} disentangle images into the light-invariant Content C and the light-specific Illumination component I , both of which are latent space features, as shown in Fig 1 (a). Then, the enhancement module M_{enh} incorporates both Content and Illumination component to

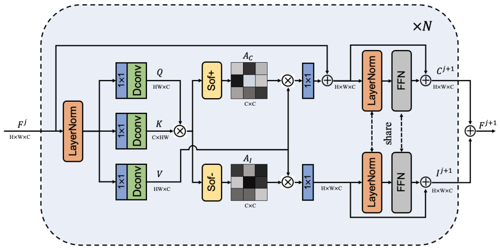

Fig. 2. Illustration of the implementation details of the Disentanglement Transformer Block. (DTB)

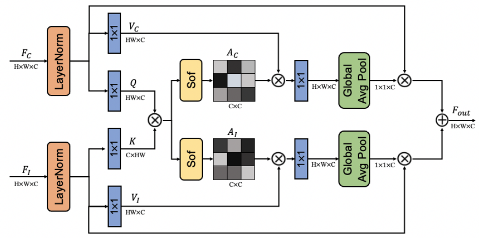

Fig. 3. Illustration of the implementation details of the Content-Aware Embedding Module (CAE).

bright up the illumination, as illustrated in Fig 1 (b). Finally, the reconstruction module M_{recon} reconstructs a new image with the Content and restored Illumination component , as shown in Fig 1 (c).

C. Latent Disentanglement

Different from previous decomposition models, our design is a disentangle-then-reconstruct architecture. To make disentanglement light weight and preserve high-resolution information, we avoid using hierarchical structure for both disentanglement and reconstruction modules. Given an input image $\mathbf{S} \in R^{h\times w\times 3}$, the disentanglement module M_{dis} first employs a 3×3 convolution as a embedding layer to extract shallow feature $\mathbf{F}_0 \in R^{h\times w\times c}$. Next, the Disentangle Transformer Blocks (DTB) disentangle $\mathbf{F}_0$ into the Content $\mathbf{C} \in R^{h\times w\times c}$ and Illumination $\mathbf{I} \in R^{h\times w\times c}$ in latent space. The disentanglement module M_{dis} can be formulated as:

$$\mathbf{C}, \mathbf{I} = M_{dis}(\mathbf{S}) \tag{3}$$

Then, the Content $\mathbf{C}$ and Illumination $\mathbf{I}$ are combined together as the input to the following transformer blocks [17] in the reconstruction module M_{recon}. Finally, we apply a 3×3 convolution to yield a new image $\tilde{\mathbf{S}} \in R^{h\times w\times 3}$. The reconstruction module M_{recon} can be described as:

$$\tilde{\mathbf{S}} = M_{recon}(\mathbf{C} + \mathbf{I}) \tag{4}$$

Disentanglement Transformer Block. We build Disentanglement Transformer Block (DTB) based on the transposed transformer block in Restormer [17].

As shown in Fig 2, given a feature $\mathbf{Y} \in R^{h\times w\times c}$ output from the normalization layer, we first generate query ($\mathbf{Q}$), key ($\mathbf{K}$) and value ($\mathbf{V}$) projections by applying 1×1 convolutions followed by 3×3 depth-wise convolutions to encode channel-wise spatial context, as $\mathbf{Q} = W_d^Q W_p^Q \mathbf{Y}$, $\mathbf{K} = W_d^K W_p^K \mathbf{Y}$, $\mathbf{V} = W_d^V W_p^V \mathbf{Y}$, where $W_p^{(\cdot)}$ is the 1×1 point-wise convolution and $W_d^{(\cdot)}$ is the 3×3 depth-wise convolution. Next, we reshape query ($\mathbf{Q}$) and key ($\mathbf{K}$) so that their dot-product generates a transposed-attention map $\mathbf{A} \in R^{c\times c}$. Then, we calculate Content attention $\mathbf{A}_C$ and Illumination attention $\mathbf{A}_I$ through the Softmax functions. Overall, the disentanglement attentions are formulated as:

$$\begin{aligned} &\mathbf{Q}, \mathbf{K}, \mathbf{V} = W_d^Q W_p^Q \mathbf{Y}, W_d^K W_p^K \mathbf{Y}, W_d^V W_p^V \mathbf{Y} \\ &\mathbf{A}_C = Softmax(\hat{\mathbf{Q}} \cdot \hat{\mathbf{K}^T})\mathbf{V} \\ &\mathbf{A}_I = Softmax(-\hat{\mathbf{Q}} \cdot \hat{\mathbf{K}^T})\mathbf{V} \end{aligned} \tag{5}$$

where, $\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \hat{\mathbf{V}} \in R^{hw\times c}$ are obtained after reshaping tensors from the original size $R^{h\times w\times c}$.

Similar to the conventional vision transformer [22], $\mathbf{A}_C$ and $\mathbf{A}_I$ are then fed into a normalization layer and Feed-Forward Network(FFN) To get better disentanglement results, we implement N Disentanglement Transformer Blocks(DTB) in the disentanglement module (In this work, we set $N = 2$). Given the input feature $\mathbf{F}^j$, the $j+1^{th}$ DTB can be formulated as follows:

$$\begin{aligned} &\mathbf{A}_C^{j+1}, \mathbf{A}_I^{j+1} = DAM(LN(\mathbf{F}^j)) \\ &\mathbf{C}^{j+1} = FFN(LN(\mathbf{A}_C^{j+1} + \mathbf{F}^j)) + \mathbf{A}_C^{j+1} + \mathbf{F}^j \\ &\mathbf{I}^{j+1} = FFN(LN(\mathbf{A}_I^{j+1})) + \mathbf{A}_I^{j+1} \\ &\hat{\mathbf{F}}^{j+1} = \mathbf{C}^{j+1} + \mathbf{I}^{j+1} \end{aligned} \tag{6}$$

where, $\mathbf{F}^{j+1}$ is the output feature, $LN(\cdot)$ represents the normalization layer. Notice that the Content feature $\mathbf{C}^N$ and Illumination feature $\mathbf{I}^N$ in the last block N are the final output from the distanglement moduleM_{dis}.

Disentanglement Loss Functions. Since we do not have the ground-truth disentanglement of real images, we follow the training approach in [4]. The disentanglement module M_{dis} takes the paired low-light image $\mathbf{S}_l$ and the normal-light image $\mathbf{S}_t$ as inputs, then estimates the Content $\mathbf{C}_l$ and the Illumination $\mathbf{I}_l$ for $\mathbf{S}_l$, $\mathbf{C}_t$ and $\mathbf{I}_t$ for $\mathbf{S}_t$, respectively.

$$\begin{aligned} &\mathbf{C}_l, \mathbf{I}_l = M_{dis}(\mathbf{S}_l) \\ &\mathbf{C}_t, \mathbf{I}_t = M_{dis}(\mathbf{S}_t) \end{aligned} \tag{7}$$

Recall that the Content component is corruption-free and invariant under different light condition, we impose strong constraints between $\mathbf{C}_l$ and $\mathbf{C}_t$. The Content consistency loss L_{cc} is introduced to constrain the consistency of Content components:

$$L_{cc} = \|\mathbf{C}_l - \mathbf{C}_t\|_1 \tag{8}$$

Based on our proposed latent disentanglement design, we employ no constraints on Illumination component. Then, to ensure that both $\mathbf{C}_l$ and $\mathbf{C}_t$ can reconstruct a new image that is similar to the input image $\mathbf{S}_l/\mathbf{S}_t$ with the corresponding Illumination $\mathbf{I}_l/\mathbf{I}_t$, the reconstruction loss L_{recon} can be formally described as:

$$L_{recon} = \sum_{i=l}^{t}\sum_{j=l}^{t} \|M_{recon}(\mathbf{C}_i + \mathbf{I}_j) - \mathbf{S}_j\|_1 \tag{9}$$

The disentanglement loss L_{dis} can be formulated as:

$$L_{dis} = L_{cc} + \lambda_{recon} L_{recon} \tag{10}$$

where λ_{recon} denote the coefficient to balance two losses.

Methods	LOL-v1 Dataset PSNR(↑)	LOL-v1 Dataset SSIM(↑)	LOL-v2 Dataset PSNR(↑)	LOL-v2 Dataset SSIM(↑)	LDIS Dataset PSNR(↑)	LDIS Dataset SSIM(↑)
Retinex-Net [4]	16.77	0.560	15.47	0.567	16.12	0.573
KinD++ [5]	18.45	0.779	18.63	0.799	16.52	0.582
LIME [21]	16.05	0.486	17.16	0.480	16.25	0.539
IAT [15]	23.38	0.809	23.50	0.824	17.34	0.692
LLFlow [10]	21.15	0.852	22.37	0.865	18.01	0.789
Restormer [17]	22.37	0.816	19.94	0.827	17.13	0.727
Uformer [16]	19.61	0.755	19.41	0.657	17.01	0.694
LLFormer [18]	23.65	0.816	21.46	0.821	18.67	0.685
LDE-Net(our)	25.02	0.910	25.49	0.950	19.38	0.848

TABLE I

QUANTITATIVE COMPARISONS ON LOL-V1, LOL-V2 AND LDIS DATASETS. THE HIGHEST RESULT IS IN RED COLOR WHILE THE SECOND HIGHEST RESULT IS IN BLUE COLOR. OUR LDE-NET SIGNIFICANTLY OUTPERFORMS SOTA ALGORITHMS.

D. Content-Aware Illumination Enhancement

The disentanglement module M_{dis} yields the Content and Illumination features. We only need to restore the degraded low-light Illumination $\mathbf{I}_l$ to normal-light $\mathbf{I}_t$.

Our illumination enhancement module is a Transformer-based 4-level hierarchical encoder-decoder network, as shown in Fig. 1 (b). From top to bottom levels, the encoder-decoder hierarchically reduces spatial dimension and doubles the channel dimension. We use our proposed Content-Aware Embedding (CAE) module to adaptively fuse the Illumination feature with the Content feature at the input of each encoder block. For feature downsampling and upsampling, we apply pixel-unshuffle and pixel-shuffle operations [23]. To help recover high-resolution representations, the encoder features are concatenated with the decoder features via skip connections [24].

Content-Aware Embedding Module. When restoring Illumination with the guidance of Content, the main challenge is to efficiently exchange information between the two components and adaptively fuse them together. To tackle this issue, we propose Content-Aware Embedding (CAE) module to refine the Illumination features as shown in Fig.3.

Our CAE module uses the Cross-Attention mechanism to fuse Content feature $\mathbf{F}_C$ and Illumination feature $\mathbf{F}_I$. In order to fuse high-resolution features with a low computational cost, we use transposed-attention mechanism [17] to calculate channel-wised attentions with the input features. Moreover, we use the Global Average Pooling function to learn adaptive fuse weights instead of directly fuse dense features from the cross-attention layer.

E. Loss functions for LLIE

The training objective for the LLIE task is to minimize the difference between the normal-light image and enhanced image. Thus, our loss functions for LLIE as follows:

$$L_{enh} = \|\mathbf{S}_r - \mathbf{S}_t\|_1 + \lambda_s(1 - SSIM(\mathbf{S}_r, \mathbf{S}_t)) + \lambda_p\|\theta(\mathbf{S}_r) - \theta(\mathbf{S}_t)\|_2 \tag{11}$$

where, $\mathbf{S}_r$ and $\mathbf{S}_t$ represent enhanced image and corresponding normal-light image. $SSIM(\cdot,\cdot)$ is structural similarity [25] and $\theta(\cdot)$ denotes the process to extract deep features form a pre-trained network. λ_s and λ_p are the weighting factors for balancing each item.

IV. EXPERIMENTS

A. Experimental Settings

Datasets. We evaluate the proposed LDENet on three datasets involving both indoor and outdoor scenes, namely LOL-v1[4], LOL-v2[26] and LDIS [27]. The LOL-v1 dataset [4] contains 485 low/normal light image pairs for training and 15 pairs for testing. The LOL-v2 dataset contains 689 image pairs for training and 100 pairs for testing. The low-light images of LOL-v1/v2 are collected by changing exposure time and ISO of normal light images. The Light-Dark Indoor Scenes (LDIS) dataset [27] consists of 389 low/normal-light image pairs for training and 98 pairs for testing. Although LDIS is collected for low light indoor semantic segmentation tasks, it is collected using a mobile device (e.g. an Apple iPad Pro 4^{th} Gen) under real world condition (e.g. by turning on/off all the lights). Compared with previous datasets, such realistic low light dataset offers more diverse lighting conditions for the enhancement task.

Metrics. We adopt the widely used peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [25]as the evaluation measures.

Implementation Details. We implement the proposed framework using PyTorch. The model is trained with the Adam [28] optimizer (β_1 = 0.9 and β_2 = 0.999). The learning rate is initially set to 2×10^{-4} and then steadily decreased to 1×10^{-6} by the cosine annealing scheme [29] during the training process. Patches at the size of 128×128 are randomly cropped from the low/normal-light image pairs as training samples. The batch size is 8. The training data is augmented with random HSV, rotation and flipping. Our LDE-Net is trained with the loss function as shown in Eq. 10 and Eq. 11, and we empirically set the parameter $\lambda_{recon} = 0.2$, $\lambda_p = \lambda_s = 1$.

B. Low-light Image Enhancement

Quantitative Results. We quantitatively compare the proposed method with SOTA enhancement algorithms in Table I. The evaluation code for KinD++ [5] and LLFlow [10] models involve an unreasonable adjustment to the enhanced images using ground-truth images so we evaluated their models using the corrected evaluation code. As shown in Table I, our method outperforms the best SOTA model on three datasets in all metrics. Compared with the LLFormer

Fig. 4. Qualitative results of LOL-v1 (top) and LDIS (bottom). Previous methods either collapse by noise, or distort color, or produce blurry and under-/over-exposed images. While our algorithm can effectively reconstruct well-exposed image details.

	LOL-v1		LDIS	
Metrics	Retinex-Net	Ours	Retinex-Net	Ours
$PSNR(M_r(\mathbf{C}_l, \mathbf{I}_l), \mathbf{S}_l)(\uparrow)$	44.81	48.32	51.81	48.34
$PSNR(M_r(\mathbf{C}_t, \mathbf{I}_t), \mathbf{S}_t)(\uparrow)$	35.84	48.46	43.94	50.38
$PSNR(M_r(\mathbf{C}_t, \mathbf{I}_l), \mathbf{S}_l)(\uparrow)$	37.58	36.85	30.25	37.22
$PSNR(M_r(\mathbf{C}_l, \mathbf{I}_t), \mathbf{S}_t)(\uparrow)$	22.80	34.92	17.19	36.02

TABLE II

CROSS-DATASET ABLATION STUDY. WE PRE-TRAIN BOTH METHODS ON LOL-V1 DATASET AND TEST THEM ON LOL-V1 AND LDIS DATASET.

CAE	T-MSA	W-MSA	A-MSA	PSNR(↑)	SSIM(↑)
✓	✓			25.02	0.910
	✓			23.82	0.901
		✓		22.83	0.883
			✓	23.75	0.896

TABLE III

WE CONDUCT ABLATION STUDY ON LOL-V1 DATASET. PSNR AND SSIM ARE REPORTED

[18] and LLFlow [10] , our method improves PSNR by 1.73 dB and significantly increases SSIM by 5.8 % on LOL-v1 datasets. Our LDE-Net also outperforms IAT [15] and LLFlow [10] on the LOL-v2 dataset, achieving a significant gain of PSNR by 1.99 dB and SSIM by 8.5 %. This trend is also observed on the more challenging LDIS dataset as shown in Table I. Our approach consistently outperforms other methods in all metrics. All these results clearly suggest the effectiveness of our LDE-Net, especially on more challenging and diverse conditions.

Qualitative Results. The visual comparisons of the proposed LDE-Net and SOTA algorithms are shown in Fig. 4. Please zoom in for a better review. Among the enhanced images on LOL-v1 dataset, one can see that Retinex-Net causes serious color distortion while LLFlow generates a blurry image. Restormer and LLFormer fail to suppress the noise and contain over-/under-exposed regions. Among visual quality comparisons from the LDIS dataset, none of previous methods can achieve acceptable results on the image from LDIS. SOTA methods such as LLFlow and LLFormer fail to suppress the noise while Restormer introduces black spots and unnatural artifacts. Although our LDE-Net losses some texture details around shadow areas (e.g. under the chairs), it is capable of achieving satisfactory visual quality on lighting consistency, noise suppression, and structure detail. In summary, our proposed LDE-Net produces enhancement results with sharper details, comfortable contrast, and vivid colors, making it more satisfying and superior to other SOTA methods.

C. Ablation Study

In this section, we report two ablation studies we perform to demonstrate (i) the usefulness of our latent-space disentanglement approach (in terms of its performance and transferability), (ii) the effectiveness of various attention mechanisms and the proposed CAE module.

Cross-Dataset Disentanglement. We conduct a cross-dataset ablation study to compare our latent disentanglement framework with the SOTA decomposition method Retinex-Net [4]. We evaluate our latent disentanglement using paired low-/normal-light images $\mathbf{S}_l$ and $\mathbf{S}_t$. Based on the disentanglement objective, both $\mathbf{C}_l$ and $\mathbf{C}_t$ can be used to reconstruct a new image that is similar to the input image $\mathbf{S}_l/\mathbf{S}_t$ with the corresponding Illumination component $\mathbf{I}_l/\mathbf{I}_t$. Thus, we adopt PSNR metrics to calculate the difference between the reconstructed images and input images as shown in Table II (e.g. $PSNR(M_r(\mathbf{C}_l, \mathbf{I}_l), \mathbf{S}_l)$ represents the PSNR between $\mathbf{S}_l$ and the reconstructed image from $\mathbf{C}_l + \mathbf{I}_l$). Since the decomposition in Retinex-Net [4] share similar concept with our disentanglement approach, we use the same PSNR metrics to report their performance.

We pre-train our disentanglement framework and the Decom-Net from Retinex-Net on LOL-v1 dataset and test on both LOL-v1 and LDIS datasets as shown in Table II. The results on LOL-v1 dataset show that our method estimates much more consistent light-invariant components than Retinex-Net. Unlike Retinex-based decomposition, our disentanglement method doesn't introduce corruptions that need to be restored in the subsequent enhancement step. Further, the results on the LDIS dataset demonstrates that our method also outperforms the Retinex-Net in terms of the cross-dataset transferability and robustness under variant light conditions.

Self-Attention Scheme. We conduct an ablation to study the effectiveness of various self-attention schemes as shown in Table III. For fair comparison, we remove the CAE module and only implement 4-level hierarchical encoder-decoder network with different multi-head self-attention (MSA) mechanisms. We compared the transposed MSA (T-MSA) used in Restormer [17] with window-based MSA (W-MSA) proposed by Swin-Transformer and Axis-based MSA (A-MSA) proposed by LLFormer [18]. The results show that T-MSA achieves the best performance.

Adaptive Fusion. The first row of Table III shows the impact of adding the CAE module. The enhancement performance improves PSNR by 1.2 dB with the CAE module which shows the usefulness of our CAE design.

V. Conclusions

In this paper, we propose a novel transformer-based latent disentanglement framework to get better disentanglement results for low-light image enhancement and other downstream low-light vision tasks. Then, we design a Content-Aware Embedding module (CAE), which explicitly improves the performance of Illumination enhancement by learning the correlation between the disentangled Content and Illumination features. Thirdly, we take the advantage of the proposed latent disentanglement framework to develop extremely light-weight enhancement network for downstream low-light vision tasks. Finally, extensive experimental results using two public LLIE datasets and one nighttime UAV tracking dataset demonstrate that our proposed Latent Disentangle-based Enhancement Network (LDE-Net) outperforms SOTA methods.

References

[1] K. Jiang, Z. Wang, Z. Wang, C. Chen, P. Yi, T. Lu, and C.-W. Lin, "Degrade is upgrade: Learning degradation for low-light image enhancement," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, 2022, pp. 1078–1086.

[2] 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 transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[3] M. Afifi and M. S. Brown, "Deep white-balance editing," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2020, pp. 1397–1406.

[4] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," *arXiv preprint arXiv:1808.04560*, 2018.

[5] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 1632–1640.

[6] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *International Journal of Computer Vision*, vol. 129, pp. 1013–1037, 2021.

[7] W. Wu, J. Weng, P. Zhang, X. Wang, W. Yang, and J. Jiang, "Uretinex-net: Retinex-based deep unfolding network for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5901–5910.

[8] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[9] S. Lim and W. Kim, "Dslr: Deep stacked laplacian restorer for low-light image enhancement," *IEEE Transactions on Multimedia*, vol. 23, pp. 4272–4284, 2020.

[10] Y. Wang, R. Wan, W. Yang, H. Li, L.-P. Chau, and A. Kot, "Low-light image enhancement with normalizing flow," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2604–2612.

[11] C.-M. Fan, T.-J. Liu, and 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.

[12] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang, "Enlightengan: Deep light enhancement without paired supervision," *IEEE transactions on image processing*, vol. 30, pp. 2340–2349, 2021.

[13] Y. Jin, W. Yang, and R. T. Tan, "Unsupervised night image enhancement: When layer decomposition meets light-effects suppression," in *European Conference on Computer Vision*. Springer, 2022, pp. 404–421.

[14] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 1780–1789.

[15] Z. Cui, K. Li, L. Gu, S. Su, P. Gao, Z. Jiang, Y. Qiao, and T. Harada, "You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction." in *BMVC*, 2022, p. 238.

[16] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general u-shaped transformer for image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 683–17 693.

[17] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5728–5739.

[18] T. Wang, K. Zhang, T. Shen, W. Luo, B. Stenger, and T. Lu, "Ultra-high-definition low-light image enhancement: A benchmark and transformer-based method," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, 2023, pp. 2654–2662.

[19] Y. Zhang, X. Di, B. Zhang, and C. Wang, "Self-supervised image enhancement network: Training with low light images only," *arXiv preprint arXiv:2002.11300*, 2020.

[20] X. Yi, H. Xu, H. Zhang, L. Tang, and J. Ma, "Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12 302–12 311.

[21] X. Guo, Y. Li, and H. Ling, "Lime: Low-light image enhancement via illumination map estimation," *IEEE Transactions on image processing*, vol. 26, no. 2, pp. 982–993, 2016.

[22] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-token vit: Training vision transformers from scratch on imagenet," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 558–567.

[23] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[24] O. Ronneberger, P. Fischer, and 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.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[26] W. Yang, W. Wang, H. Huang, S. Wang, and J. Liu, "Sparse gradient regularized deep retinex network for robust low-light image enhancement," *IEEE Transactions on Image Processing*, vol. 30, pp. 2072–2086, 2021.

[27] X. Ying, B. Lang, Z. Zheng, and M. C. Chuah, "Delving into light-dark semantic segmentation for indoor scenes understanding," in *Proceedings of the 1st Workshop on Photorealistic Image and Environment Synthesis for Multimedia Experiments*, 2022, pp. 3–9.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[29] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.