Unsupervised Low Light Image Enhancement Transformer Based on Dual Contrastive Learning

Fengji Ma
mafengji@buaa.edu.cn
Jinping Sun(✉)
sunjinping@buaa.edu.cn

School of Electronic and
Information Enginnering,
Beihang University, Beijing, China

Abstract

Low-light image enhancement aims to recover normal-light images from the images captured under very dim environments. While deep learning-based methods have achieved substantial success in this field, most of them require paired training data, which is difficult to be collected. We propose an Unsupervised Dual Contrastive Learning Transformer (UDCL-Transformer) where the unsupervised contrastive learning is for the first time introduced to the low light image enhancement task. From a different yet new perspective, we explore contrastive learning with an adversarial training effort to leverage unpaired low-light images and normal-light images. Our proposed method leveraged dual contrastive learning and generative adversarial networks to restore low light image. Patch-wise contrastive learning maximizes the mutual information between raw and restored images. Pixel-wise contrastive learning encourages the restored images to approach the positive samples and keep away from the negative samples in the embedding space. Generator based on Parallel Convolution Transformer (PC-Former) is proposed to capture the rich features of global and local context for better aggregate information. Extensive experiments with comparisons to recent approaches further demonstrate the superiority of our proposed method.

1 Introduction

Limited by weather factors and equipment reasons, the existence of low-light images will not only affect the visual effect, but also negatively affect the downstream visual tasks. It affects the reliability of the model in advanced vision tasks, further misleading machine systems, such as autonomous driving. This makes image augmentation a meaningful low-level vision task. Low light image enhancement is a typical ill-posed problem, and traditional Low-Light Image Enhancement (LLIE) algorithms tend to limit the solution space with priors [2, 7, 17]. However, these images are often significantly different from normal light images and may introduce artifacts in regions that do not satisfy the prior.

Deep learning based methods have achieved great success in the field of computer vision, and researchers have proposed a large number of LLIE methods [6, 13, 18, 20, 21, 22, 30, 33, 35, 38, 39] based on deep convolutional neural network (CNN). These methods

can generally be divided into two categories, one is direct end-to-end image enhancement models, which utilize network models to learn the mapping from low-light images to corresponding normal-light images [6, 13, 20, 35, 38], ignoring the underlying physics. The second is the CNN method inspired by Retinex [18, 21, 22, 30, 33, 39], where CNN is used as a first attempt to decompose the low-light image into two components, i.e., reflectance and illuminanation, which are then post-processed [39] or directly a reflectance layer is used as the final result [30], inheriting a similar idea compared to traditional Retinex-based methods. Due to its strong interpretability and general prior knowledge based on physical models, methods to introduce Retinex theory into network design have received more attention, and Retinex-based CNN methods generally outperform end-to-end methods[15]. Besides these CNN-based methods, Zhang et al.[40] proposed the first Transformer-based LLIE method which naturally and implicitly captures the structural relationships of different regions in an image. However, there exists several issues: (1) Most existing methods adopt ground-truth as positive samples to guide the training of low-light enhancement network based reconstruction loss. The way to maximize the mutual information between input and output data is to be found. How to use the negative and positive samples is the key to maximize the mutual information. (2) CNN-based methods are limited by the receptive field, which are weak in capturing the long-distance dependence relationship and have disadvantages in extracting global context information. (3) How to train Transformer method of LLIE task with unpaired dataset is also a to-be-solved problem.

To address these issues, we propose a novel unsupervised dual contrastive learning paradigm. To effectively train the network in an unsupervised manner, in addition to the patch-wise contrastive learning loss, we formulate a pixel-wise contrastive learning loss to encourage the restored images and the normal light images (positive samples) to pull together in the representation space while pushing them away from the low light ones (negative samples). Inspired by [19, 29], We also propose a parallel convolution Transformer (PC-Former) to capture the rich features of global and local context. Our main contributions are as follows:

- We propose a novel unsupervised dual contrastive learning Transformer-based generative adversarial network. To the best of our knowledge, we are the first to combine a Transformer-based generator with contrastive learning for LLIE task. We also propose a novel multi-head self-attention with parallel convolution for information aggregation.

- We formulate an effective dual contrastive learning method to train our proposed UDCL-Transformer. Specifically, we employ pixel-wise contrastive learning to learn a representation that pulls the restored images and normal light images (positives) together while pushing them away from the low light ones (negatives). We also leverage patch-wise contrastive learning to maximize the mutual information between corresponding patches of the raw image and the restored image to capture the content and detail correspondences between two image domains.

2 Related Work

2.1 Unsupervised Learning Low Light Image Enhancement

Training a deep model on paired data may result in overfitting and limited generalization capability. To solve this issue, an unsupervised learning method named EnligthenGAN [13]

is proposed. The EnlightenGAN adopts an attention-guided U-Net [27] as the generator and uses the global-local discriminators to ensure the enhanced results look like realistic normal-light images. Zhu et al. [41] propose a three-branch CNN, called RRDNet, for underexposed images restoration. The RRDNet decomposes an input image into illumination, reflectance, and noise via iteratively minimizing specially designed loss functions. To drive the zero-shot learning, a combination of Retinex reconstruction loss, texture enhancement loss, and illumination-guided noise estimation loss is proposed. Liu et al. [18] propose a Retinex inspired unrolling method for LLIE, in which the cooperative architecture search is used to discover lightweight prior architectures of basic blocks and non-reference losses are used to train the network. Different from the image reconstruction-based methods [13, 18, 41], a deep curve estimation network, Zero-DCE [6], is proposed. Zero-DCE formulates the light enhancement as a task of image-specific curve estimation, which takes a lowlight image as input and produces high-order curves as its output. These curves are used for pixel-wise adjustment on the dynamic range of the input to obtain an enhanced image.

2.2 Contrastive Learning

Contrastive learning are widely used in self-supervised representation learning, where the contrastive losses are inspired by noise contrastive estimation [8], triplet loss [11] or N-pair loss [28]. These approaches aim to learn an embedding that brings the associated features close to each other, while the irrelevant samples are pushed away. Existing efforts mainly apply contrastive learning into high level vision tasks, since these tasks inherently suit for modeling the contrast between positive and negative samples/features. Recently, several studies have attempted to apply contrastive learning to low-level vision tasks. The design choices of the InfoNCE loss [24], which aims to learn an embedding or an encoder that associates corresponding patches to each other, was first introduced into image translation by [26]. Han et al. [10] proposed an unsupervised contrastive learning method using InfoNCE loss for underwater Image restoration. Wu et al. [34] develop a contrastive regularization term to leverage the information of both hazy and clean images for image dehazing. Taking into account two kinds of comparative learning methods, Wang et. al. [31] proposed a dual contrastive learning method for real-world image dehazing. Different from these works, we explore unsupervised contrastive learning from an adversarial training perspective to leverage unpaired normal-light and low-light images. Our proposed network does not require paired data during training. By training the network in an unsupervised yet adversarial manner, we can better utilize unpaired positive/ negative data to promise the enhancement performance.

3 Method

We employ a U-shape Transformer network as the generator module. We aim to learn a mapping to enable low light image enhancement. Unsupervised Dual Contrastive Learning Transformer (UDCL-Transformer) has a generator G as well as a discriminator D. G enables the mapping from low light domain to normal light domain and D ensures that the translated images belong to the correct image domain. The first half of the generator is defined as an encoder, while the second half is a decoder and denoted G_{enc} and G_{dec} respectively. The framework of our proposed UDCL-Former is shown in Fig. 1.

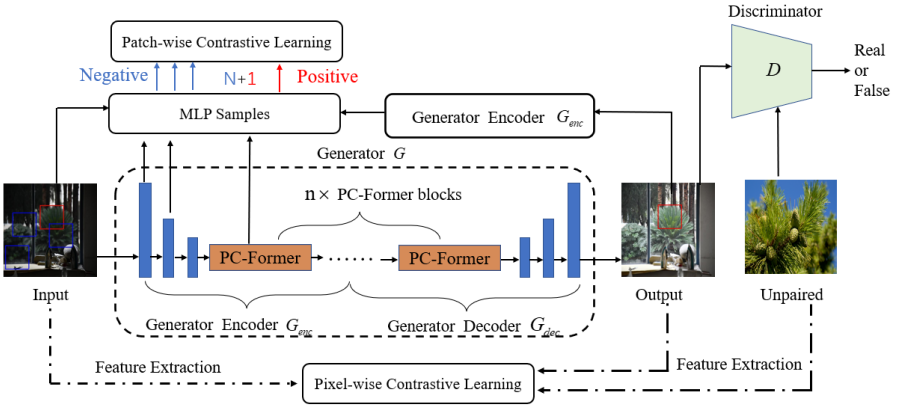


Figure 1: The network structure of UDCL-Transformer. UDCL-Transformer targets to learn a mapping of low light image to restored image. We use a Transformer-based generator with Parallel Convolution Transformer (PC-Former) blocks and define the first half of the Generator G to be Generator Encoder G_{enc}. The last half of the Generator G is Generator Decoder G_{dec}. In pixel-wise contrastive learning, we denote the group of unpaired label (normal-light image) and the restored image as the positive pair. Similarly, the negative pair is generated by the group of low light image and the restored image. In patch-wise contrastive learning, given the red "query" from the generated restored image, we set up an (N + 1)-way classification problem and denote the two corresponding patches (red "positive") as the positive sample, while the other N patches (blue "negative") are the negative samples.

3.1 Generator Based on Parallel Convolution Transformer

As demonstrated in Fig. 2, we propose a Transformer with parallel convolution (PC-Former) block of generator G in our proposed method. Long-short range multi-head self-attention (LSR-MHSA) extracts high frequency information and low frequency information in spatial-wise with parallel convolution. Inspired by [5, 29], which introduced convolution layer to act as a positional embedding, we use the convolution to extract high frequency information. In contrast to [5], we use parallel convolution for more spatial information aggregation rather than encoding position information implicitly [12]. In contrast to [29], we use zero padding and shifted-window to capture local relationships within the local window and enable connections across windows. Recent works [25, 29] note that multi-head self attention (MHSA) is a low-pass filtering. Although the spatial information aggregation weight of MHSA is dynamic, the weight is always positive, making it work like smoothing. As a counterpart to MHSA's dynamic spatial information aggregation style, we perform static additional convolution on value V. Thus the spatial information aggregation scheme is

$$\mathrm{MHSA}_{\mathrm{PC}} = \mathrm{Softmax}(QK^T/\sqrt{d}+B)V + \mathrm{Conv}(V) \tag{1}$$

where $\mathrm{MHSA}_{\mathrm{PC}}$ represents long-short range multi-head self-attention with parallel convolution module, $\mathrm{Conv}(\cdot)$ can be either DWConv [4] or a ConvBlock (convolution layer with activation function). We still use the attention mechanism to aggregate information within the window, but also use convolution to aggregate information in the neighborhood without

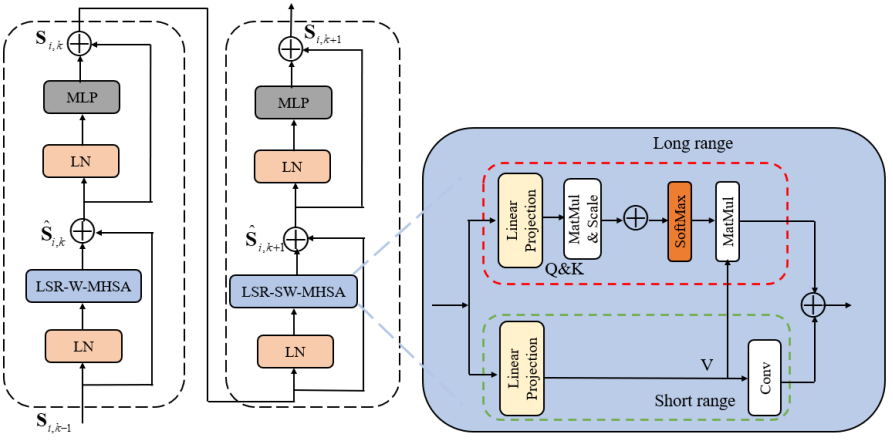


Figure 2: Detailed structure of PC-Former.

considering window partitioning. Most importantly, PC-Former's convolution layer is performed on Value V before window partitioning, thus it provides the capability to aggregate information between windows.

$$\begin{aligned}
\hat{\mathbf{S}}_{i,k} &= \mathrm{W-MHSA}_{\mathrm{PC}}(\mathrm{LN}(\mathbf{S}_{i,k-1})) + \mathbf{S}_{i,k-1} \\
\mathbf{S}_{i,k} &= \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{S}}_{i,k-1})) + \hat{\mathbf{S}}_{i,k-1} \\
\hat{\mathbf{S}}_{i,k+1} &= \mathrm{SW-MHSA}_{\mathrm{PC}}(\mathrm{LN}(\mathbf{S}_{i,k})) + \mathbf{S}_{i,k} \\
\mathbf{S}_{i,k+1} &= \mathrm{MLP}(\mathrm{LN}(\hat{\mathbf{S}}_{i,k})) + \hat{\mathbf{S}}_{i,k}
\end{aligned} \tag{2}$$

where $\hat{\mathbf{S}}_{i,k}$ and $\mathbf{S}_{i,k}$ denote the output features of the LSR-(S)W-MSA and the MLP module for k − th layer depth of i − th PC-Former block , respectively.

3.2 Discriminator

The function of the discriminator D is to judge whether a given image is a real clean image or a fake image produced by the generator, thus guiding the generator to produce more realistic images. We use the same Patch-GAN discriminator architecture, which passes domain translation through five downsampling Convolutional-Normalization-LeakeyReLU layers. Least-Square GAN (LSGAN) loss [23] has been proved to be more effective than the vanilla GAN loss, as it can ensure that the training process to be more stable. We adopt the LSGAN loss to train our network. The definition of adversarial loss can be expressed as:

$$L_{adv}(G) = \mathrm{E}_{G(x)\sim P_{fake}}[(D(G(x)) - 1)^2] \tag{3}$$

$$L_{adv}(D) = \mathrm{E}_{y\sim P_{real}}[(D(y) - 1)^2] + \mathrm{E}_{G(x)\sim P_{fake}}[(D(G(x)))^2] \tag{4}$$

where x refers to low-light image, y refers to the unpaired normal-light image and $G(x)$ represents the enhanced image.

3.3 Pixel-Wise Contrastive Learning

Inspired by [34], we develop a novel pixel-wise contrastive learning loss to encourage the restored images to be close to the positive samples while keeping away from the negative ones in the embedding space. All these samples are randomly chosen from the images and unpaired label from each other. In addition to constructing the positive and negative pairs, we need to find a latent feature space of these pairs for contrast. Here, we employ a pre-trained VGG-19 network to extract the feature maps of different samples. Therefore, the pixel-wise contrastive loss can be expressed as:

$$L_{PiC} = \sum_{i=1}^{n} \omega_i \frac{\|\psi_i(\tilde{r}) - \psi_i(G(x))\|_1}{\|\psi_i(\tilde{x}) - \psi_i(G(x))\|_1} \tag{5}$$

where $\tilde{r}$ and $\tilde{x}$ represent the group of unpaired labels (normal light images) and low light images. ψ_i, $i = 1,2,...,n$, refer to extracting the i-th hidden features from the VGG-19 network pre-trained on ImageNet. We choose the 2-nd, 7-th, 12-th, 21-th, 30-th layer of VGG-19 network. In Pixel-wise contrastive learning, we set the weights $\omega_I$corresponding to the extracted features of the five layers as $[1/32, 1/16, 1/8, 1/4, 1]$. Different from perceptual loss [12], which measures the visual difference between the prediction and the ground truth by leveraging multi-layer features extracted from a pre-trained deep neural network, Pixel-wise Contrastive learning L_{PiC} adopt low light image (input of model) as negatives to constrain the solution space.

3.4 Patch-Wise Contrastive Learning

Following the setting of [10, 26], we use a noisy contrastive estimation (NCE) framework to maximize the mutual information between inputs and outputs. The idea behind contrastive learning is to correlate two signals, i.e., the "query" and its "positive" example, in contrast to other examples in the dataset (referred to as "negatives").

$$\ell(\mathbf{v}, \mathbf{v}^+, \mathbf{v}^-) = -log(\frac{\exp(\text{sim}(\mathbf{v}, \mathbf{v}^+)/\tau)}{\exp(\text{sim}(\mathbf{v}, \mathbf{v}^+)/\tau) + \sum_{i=1}^{N} \exp(\text{sim}(\mathbf{v}, \mathbf{v}_i^-)/\tau)}) \tag{6}$$

where $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v}/\|\mathbf{u}\|\|\mathbf{v}\|$ denotes the cosine similarity between $\mathbf{u}$ and $\mathbf{v}$. τ denotes a temperature parameter to scale the distance between the query and other examples, we use 0.07 as default. We set the numbers of negatives N as 255. We use G_{enc} (consisted of G_{enc} and a 2-layer MLP) to extract features $g_l = G^l_{enc}(x)$. l represents l-th selected layer in G_{enc}. For the patches, after having a stack of features, each feature actually represents one patch from the image. We denote the spatial locations in each selected layer as $s \in \{1,2,...,S_l\}$, where S_l is the number of spatial locations in each layer. The Patch-wised contrastive learning loss can be described as,

$$L_{PaC} = \mathrm{E}_x \sum_{l=1}^{L} \sum_{s=1}^{S_l} \ell(\hat{g}_l^s, g_l^s, g_l^{S\setminus s}) \tag{7}$$

In order to prevent generators from unnecessary changes and keep the structure identical, we add an identity loss as follows.

$$L_{idt} = \mathrm{E}_y[\|G(y) - y\|_1] \tag{8}$$

	Datasets	LOL[53]		MIT[1]		LSRW[9]	
	Metrics	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
T	LIME[7]	15.7586	0.4439	17.5976	0.8179	15.4775	0.4627
	JIEP[2]	16.7856	0.5664	19.5241	0.8690	14.9076	0.5039
	SRLL[17]	15.9872	0.5109	17.6464	0.7793	14.6694	0.5061
S	RetinexNet[53]	16.7741	0.4287	13.7474	0.7394	15.9062	0.4765
	KinD[39]	18.7913	0.7086	17.0935	0.8307	14.8176	0.5691
	STAR[40]	19.9301	0.7896	21.3597	0.8405	15.9629	0.5881
U	EnGAN[13]	15.6314	0.5781	16.4371	0.7966	16.0677	0.4755
	RUAS[18]	19.1076	0.7168	20.0945	0.8734	16.3186	0.6814
	RRD[41]	14.2261	0.5316	18.5372	0.8642	15.8906	0.5276
	Ours	19.6394	0.6901	20.8741	0.8721	16.5984	0.6903

Table 1: Quantitative results (PSNR and SSIM) of state-of-the-art methods and ours on the MIT-Adobe FiveK[1], LOL[53] and LSRW[9] datasets. The best results is in red whereas the second best one is in blue. T, S and U are traditional methods, supervised learning methods and unsupervised learning methods, respectively.

Such an identity loss also encourages the mappings to preserve structure and detail between the input and output. The total loss function can be formulated as

$$L = \lambda_{adv} L_{adv} + \lambda_{PiC} L_{PiC} + \lambda_{PaC} L_{PaC} + \lambda_{idt} L_{idt} \tag{9}$$

where $\lambda_{adv}, \lambda_{PiC}, \lambda_{PaC}, \lambda_{idt}$ are set as $1, 0.5, 1, 10$.

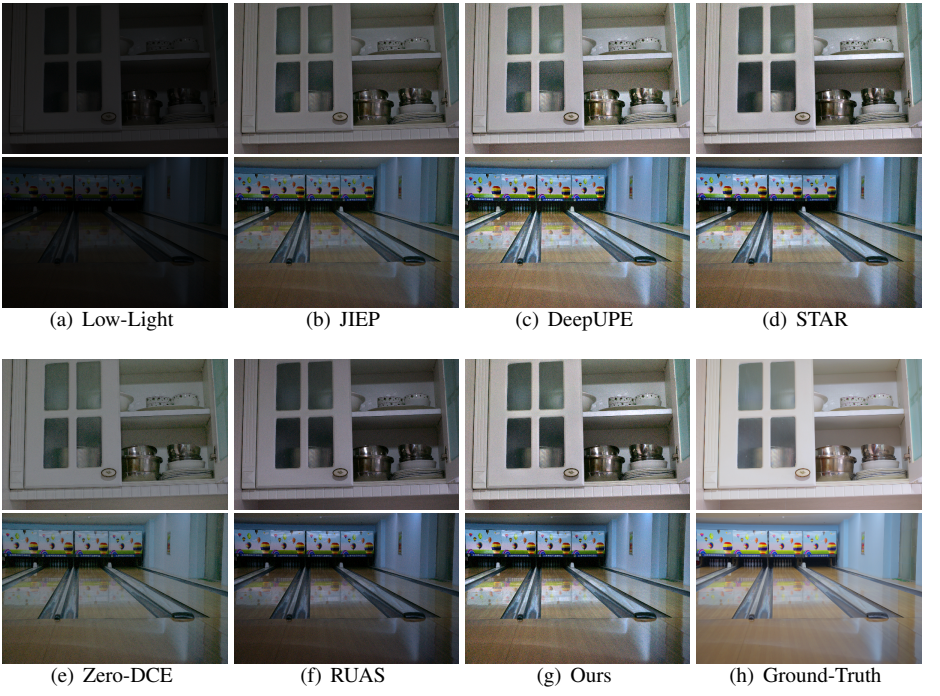
4 Experiments

4.1 Dataset and Implementation Details

Inspired by [13], we collect a larger-scale unpaired training set, that covers diverse image qualities and contents. We assemble a mixture 1645 low light and 1828 normal light images from several datasets released in [1, 3, 9, 53], without the need to keep any pair Manual inspection and selection are performed to remove images of medium brightness. We load all images in 512×512 resolution during training. For UDCL-Transformer, the batch size is set to 1. The ADAM optimizer is employed for optimization with learning rate 0.002. We train our method and other baselines using a Nvidia 3090 Ti GPU. In this paper, the number of PC-Former is 4. Embedding dimensions in PC-Former are [256, 256, 256, 256]. MLP ratios are [2, 4, 4, 2,]. The numbers of blocks are [2, 4, 4, 2]. The number of multi heads are [4, 6, 6, 4]. The shfited window size in PC-Former block is 8. We set the total number of epochs to 200 and adopt a linear decay strategy to adjust learning rate after 100 epochs.

4.2 Quantitative and Qualitative Comparisons

We compare our method with a rich collection of state-of-the-art (SOTA) methods for low-light image enhancement, including LIME[7], JIEP [2], SRLL [17], Retinex-Net [53], KinD [39], EnGAN [13], RRD [41], RUAS [18]. Also, we compared our framework with a recent transformer structures for low-light image enhancement tasks, STAR [40]. We adopt Peak



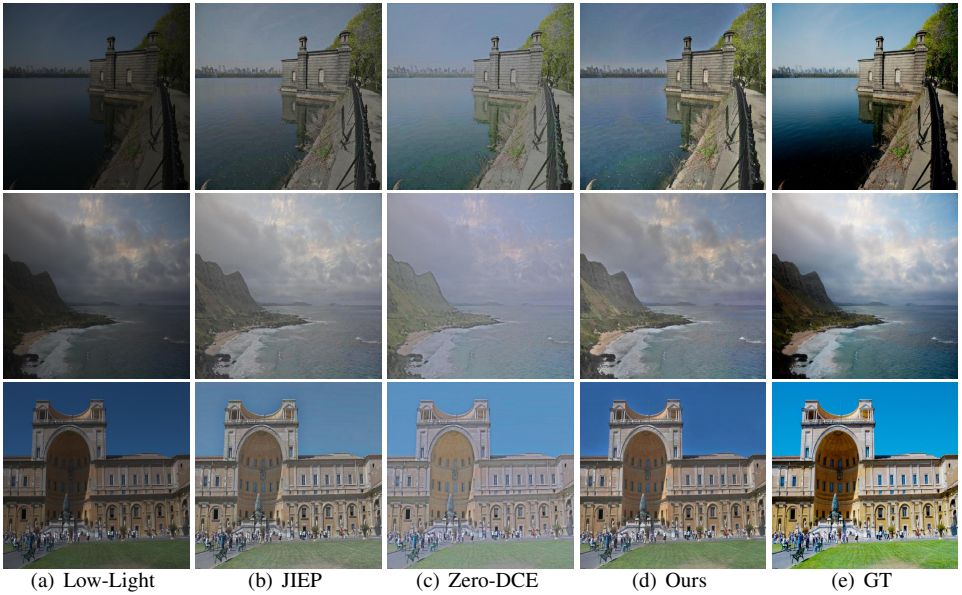
(a) Low-Light (b) JIEP (c) DeepUPE (d) STAR

(e) Zero-DCE (f) RUAS (g) Ours (h) Ground-Truth

Figure 3: Visual comparisons on low-light image from LOL dataset [33]

Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [32] for evaluation. In general, a higher SSIM means more high-frequency details and structures in results. Table 1 shows the comparisons on LOL [33], MIT-Adobe FiveK (MIT) [1], LSRW [9]. Our method surpasses all the baselines. Note that we obtain these numbers either from the respective papers or by running the respective public code. Ours (UDCL-Transformer) yields the best performance.

We present visual samples on LOL [33] dataset in Fig. 3 for comparing our method with several SOTA methods. Our result shows better visual quality with higher contrast, more precise details, color consistency, and better brightness. While the original images in LOL dataset have apparent noise and weak illumination, our method can still produce more realistic results. Our result shows better visual quality with higher contrast, more precise details, color consistency, and better brightness. In Fig. 3, Zero-DCE [6] produces color deviation. STAR [40] over-smoothes the details while DeepUPE [30] suffers over-exposed and noise. JIEP [2] and RUAS [18] fail to enhance the brightness. Meanwhile, we present visual samples on MIT-Adobe FiveK [1] dataset in Fig. 4 for comparing our method with several SOTA methods. Zero-DCE [6] produces color deviation and JIEP [2] has insufficient brightness enhancement. From these visual comparison experiments, Our proposed method produces a visually pleasing result while avoiding over-exposure artifacts during the processing of enhancement. Others either do not enhance dark details enough or generate over-exposure artifacts. Our proposed method achieves a great performance in color and brightness.



(a) Low-Light (b) JIEP (c) Zero-DCE (d) Ours (e) GT

Figure 4: Visual comparisons on low-light image from MIT-Adobe FiveK dataset [33]

(a) Raw Detection (b) Enhanced Detection

Figure 5: Two examples of face detection before (Raw Detection) and after (Enhance Detection) enhanced by proposed method UDCL-Transformer. DSFD [16] is our face detector.

4.3 Ablation Study

In Table 2, we analyze the effect of different components in UDCL-Transformer and the weights in loss functions. We also analyze the effect of dual contrastive learning (pixel-wise and patch-wise contrastive learning).

4.4 Dark Face Detection

We investigate the performance of low-light image enhancement methods on the face detection task under low-light conditions. Specifically, we use the latest DARK FACE dataset [37]

Generator (PC-Former)		Dual Contrastive Learning			Baseline(Ours)
U-shape skip connection	residual block	w/o L_{PiC}	w/o L_{PaC}	w/o $L_{PiC}+L_{PaC}$	
18.9043	19.3917	18.8561	18.7611	18.0816	
hyperparameters($\lambda_{adv},\lambda_{PiC},\lambda_{PaC},\lambda_{idt}$)					**19.6394**
(0.1,0.5,1,10)	(1,0.05,1,10)	(1,1,1,10)	(1,0.5,0.1,1)	(1,0.5,1,1)	
18.1762	18.5182	18.9091	18.7716	18.8162	

Table 2: Ablation Analysis (PSNR) on UDCL-Transformer, loss functions and hyperparameters on LOLdataset [33]. w/o represents without.

that composes of 10,000 images taken in the dark. Since the bounding boxes of test set are not publicly available, we perform evaluation on 1000 images of the training and validation sets, which are chosen form 6000 images. A deep face detector [16], trained on WIDER FACE dataset [36], is used as the baseline model. Observing the examples in Fig. 5, our method lightens up the faces in the extremely dark regions and preserves the well-lit regions, thus improves the performance of face detector in the dark.

5 Conclusion

We propose a novel unsupervised dual contrastive learning paradigm. To the best of our knowledge, we are the first to combine a Transformer-based generator with contrastive learning for LLIE task. We formulate an effective dual contrastive learning method to train our proposed UDCL-Transformer. Specifically, we employ pixel-wise contrastive learning to learn a representation that pulls the restored images and normal light images (positives) together while pushing them away from the low light ones (negatives). We also leverage patch-wise contrastive learning to maximize the mutual information between corresponding patches of the raw image and the restored image to capture the content and detail correspondences between two image domains. We also propose a parallel convolution Transformer (PC-Former) to capture the rich features of global and local context.

References

[1] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR 2011*, pages 97–104. IEEE, 2011.

[2] Bolun Cai, Xianming Xu, Kailing Guo, Kui Jia, Bin Hu, and Dacheng Tao. A joint intrinsic-extrinsic prior model for retinex. In *Proceedings of the IEEE international conference on computer vision*, pages 4000–4009, 2017.

[3] Jianrui Cai, Shuhang Gu, and Lei Zhang. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4): 2049–2062, 2018.

[4] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[5] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Weiming Zhang, Nenghai Yu, Lu Yuan, Dong Chen, and Baining Guo. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12124–12134, 2022.

[6] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020.

[7] Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016.

[8] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

[9] Jiang Hai, Zhu Xuan, Ren Yang, Yutong Hao, Fengzhu Zou, Fang Lin, and Songchen Han. R2rnet: Low-light image enhancement via real-low to real-normal network. *arXiv preprint arXiv:2106.14501*, 2021.

[10] Junlin Han, Mehrdad Shoeiby, Tim Malthus, Elizabeth Botha, Janet Anstee, Saeed Anwar, Ran Wei, Lars Petersson, and Mohammad Ali Armin. Single underwater image restoration by contrastive learning. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 2385–2388. IEEE, 2021.

[11] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[12] Md Amirul Islam, Sen Jia, and Neil DB Bruce. How much position information do convolutional neural networks encode? *arXiv preprint arXiv:2001.08248*, 2020.

[13] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 30:2340–2349, 2021.

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[15] Chongyi Li, Chunle Guo, Linghao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Lighting the darkness in the deep learning era. *arXiv preprint arXiv: 2104.10729*, 2021.

[16] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. Dsfd: dual shot face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5060–5069, 2019.

[17] Mading Li, Jiaying Liu, Wenhan Yang, Xiaoyan Sun, and Zongming Guo. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, 27(6):2828–2841, 2018.

[18] Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10561–10570, 2021.

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

[20] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 61:650–662, 2017.

[21] Fengji Ma and Haitao Li. Underexposed image enhancement via unsupervised feature attention network. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[22] Long Ma, Tengyu Ma, Risheng Liu, Xin Fan, and Zhongxuan Luo. Toward fast, flexible, and robust low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5637–5646, 2022.

[23] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[25] Namuk Park and Songkuk Kim. How do vision transformers work? *arXiv preprint arXiv:2202.06709*, 2022.

[26] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020.

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[28] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[29] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *arXiv preprint arXiv:2204.03883*, 2022.

[30] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6849–6857, 2019.

[31] Yongzhen Wang, Xuefeng Yan, Fu Lee Wang, Haoran Xie, Wenhan Yang, Mingqiang Wei, and Jing Qin. Ucl-dehaze: Towards real-world image dehazing via unsupervised contrastive learning. *arXiv preprint arXiv:2205.01871*, 2022.

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[33] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[34] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10551–10560, 2021.

[35] Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2281–2290, 2020.

[36] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016.

[37] Wenhan Yang, Ye Yuan, Wenqi Ren, Jiaying Liu, Walter J Scheirer, Zhangyang Wang, Taiheng Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, et al. Advancing image understanding in poor visibility environments: A collective benchmark study. *IEEE Transactions on Image Processing*, 29:5737–5752, 2020.

[38] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *European Conference on Computer Vision*, pages 492–511. Springer, 2020.

[39] Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1632–1640, 2019.

[40] Zhaoyang Zhang, Yitong Jiang, Jun Jiang, Xiaogang Wang, Ping Luo, and Jinwei Gu. Star: A structure-aware lightweight transformer for real-time image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4106–4115, 2021.

[41] Anqi Zhu, Lin Zhang, Ying Shen, Yong Ma, Shengjie Zhao, and Yicong Zhou. Zero-shot restoration of underexposed images via robust retinex decomposition. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.