CLIP-RestoreX: Restore Image Structure and Perception in Exposure Correction

Xiang Huang[1], **Qing Zhang**[1,2*], **Jian-Fang Hu**[1,2], **Wei-Shi Zheng**[1,2]

[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
huangx398@mail2.sysu.edu.cn, zhangq93@mail.sysu.edu.cn, hujf5@mail.sysu.edu.cn, wszheng@ieee.org

Abstract

Exposure correction aims to adjust the exposure of an under- and over-exposed image to enhance its overall visual quality. The core challenge of this task lies in that it requires to faithfully restore both the structure and perception information. In this work, we present a novel exposure correction method, referred to as CLIP-RestoreX, that leverages structural and perceptual priors from CLIP to tackle exposure correction. Specifically, we in CLIP-RestoreX propose to perform exposure correction by aligning CLIP-based structural and perceptual feature of the impaired image with its ground-truth image. To better restore the damaged structural information and perceptual information, we further design a frequency-domain based feature enhancement diffusion model, where we utilize the globality of Fourier transform to help reveal potential the relationship within the features. We conduct extensive experiments on several benchmark datasets. The results demonstrate that the proposed CLIP-RestoreX outperforms state-of-the-art exposure correction methods.

Code —
https://github.com/HXDreamChaser/CLIP-RestoreX

Introduction

When the dynamic range of the camera fails to accommodate the complex lighting conditions of the environment, the captured image often exhibits unwanted under- or over-exposed areas. In this work, we study the exposure correction problem, which aims to adjust the exposure of an under- or over-exposed image to enhance its visual quality by restoring both the structure and perception information degraded by the undesired exposure conditions. Specifically, the structural information refers to the geometric elements of an image, such as edges, contours and lines, while the perceptual information relates to how human visual system interprets images, including aspects like color, tone and contrast. With the rapid advancement of deep learning, existing exposure correction methods that focus on color correction (Baek et al. 2023), exposure consistency (Huang et al. 2023) and detail enhancement (Wang et al. 2023) have made significant progress. However, for images that are severely under- and over-exposed, these methods still struggle to effectively restore the impaired structural and perceptual information simultaneously.

Figure 1: Comparison with previous methods. In the first and second rows, our CLIP-RestoreX can better preserve the original structural information of the image. For example, our method recovers more cloud details. In the third and fourth rows, our CLIP-RestoreX can better preserve the original perceptual information of the image, for example, it restores the blue sky and the purple background of the woman, which are closer to GT. These results demonstrate that our CLIP-RestoreX can effectively restore structural and perceptual information in severely under- and over-exposed images.

To restore the structure and perceptual information damaged by severe under- and over-exposure, we introduce a method that leverages the structural and perceptual priors from CLIP (Radford et al. 2021) to improve exposure correction (named CLIP-RestoreX). Our method takes inspiration from (Guo et al. 2022; Wei et al. 2021; Wang, Chan, and Loy 2023) that CLIP encodes structural and perception information in shallow and deep layers, respectively. Specifically, We use CLIP to extract complete structural and perceptual information from the ground-truth image, and also the structural and perceptual feature degraded by severe under- or over-exposure. In order to restore the dam-

*Corresponding author.

aged structural information and perceptual information, we design a frequency-domain based feature enhancement diffusion model (FFEDM) to take full advantage of the globality of Fourier transform for better structural and perceptual information restoration. Then, we integrate the resulting structure and perception features of FFEDM into the shallow and deep blocks of the Restormer (Zamir et al. 2022) for exposure correction.

It is worth noting that our method is different from previous methods that apply CLIP to low-level vision tasks, such as CLIP-lit (Liang et al. 2023), DA-CLIP (Luo et al. 2024) and (Morawski et al. 2024). These methods all assist model training by calculating the similarity between text encode and image encode, but do not explore how to utilize the structural and perceptual information in CLIP image encoder. As illustrated in Figure 1 and Figure 3, our method successfully recovers structural and perceptual information that existing methods fail to capture. The contributions of this paper are as follows:

- We propose to leverage structural and perceptual priors from CLIP for exposure correction of severely under- and over-exposed image.
- We develop a frequency domain based feature enhancement diffusion model to restore the CLIP features degraded by severe underexposure and overexposure.
- Extensive experiments on several benchmark datasets show that our method performs favorably against previous state-of-the-art methods.

Related Work

Exposure Correction

Some traditional methods used histogram-based techniques to improve brightness and contrast (Abdullah-Al-Wadud et al. 2007; Reza 2004; Tian and Cohen 2017; Ying et al. 2017), while other methods used Retinex theory (Land 1977) to improve the brightness of the illumination component (Guo, Li, and Ling 2016; Cai et al. 2017; Li et al. 2018; Ren et al. 2020; Zhang et al. 2018, 2020). In addition, some methods use dual illumination estimation for exposure correction (Zhang, Nie, and Zheng 2019; Wang et al. 2019). With the continuous advancement of deep learning, it has achieved remarkable success in numerous fields (Li et al. 2024b,d,c; Lin et al. 2024, 2022, 2023; Lin, Zhou, and Zheng 2024; Xu et al. 2023), including exposure correction(Wang et al. 2019). MSEC (Afifi et al. 2021) provided a dataset containing 24000 images with different exposures and designed a coarse to fine network to complete the two sub tasks of color and detail enhancement. CMEC (Nsampi, Hu, and Wang 2021) introduces a deep feature matching loss to address the issue of inconsistent correction, enabling the network to learn exposure invariant representations in the feature space. ECLNET (Huang et al. 2022c) proposes an Exposure Consistency Processing (ECP) module to address the issue of different correction procedures for underexposure and overexposure in exposure correction. This module uniformly learns representations of underexposure and overexposure in the feature space, and develops an Exposure Consistency Constraint (ECC) strategy to further assist in exposure consistency learning. Eyiokur et al. (Eyiokur et al. 2022) designed an image encoder, continuous residual blocks, and image decoder to synthesize corrected images to address the issues of reduced contrast and low visibility of content caused by exposure errors in images. ENC (Huang et al. 2022a) has designed a multi exposure correction framework based on exposure standardization and compensation (ENC) module to address the issue of different correction programs and connect different exposure representations. FECNet (Huang et al. 2022b) utilizes spatial frequency information interaction in both spatial and frequency domains, recovering amplitude to enhance brightness through amplitude sub networks and reconstructing phase recovery details through phase sub networks. DAC (Wang et al. 2023) proposes decoupling contrast enhancement and detail restoration in each convolution process to address the issues of contrast reduction and detail distortion. Its designed contrast aware (CA) and detail aware (DA) units can replace traditional convolution (TConv) to improve performance. ERL (Huang et al. 2023) explored sample relationships within small batches to address the optimization inconsistency issue in exposure correction. LACT (Baek et al. 2023) proposed a brightness aware color transformation method that enables complex color transformations in both under- and over-exposed images.

The above methods are typically designed to address either the under- or over-exposure problem in the input image. LCDPNet (Wang, Xu, and Lau 2022) proposed a new dataset in which images exhibit both underexposure and overexposure. RECNet (Liu et al. 2024) proposes a region aware exposure correction network for this type of mixed exposure problem, which processes mixed exposure by adaptively learning and bridging different region exposure representations. CESC (Li et al. 2024a) enhances under- and over-exposed images by learning to estimate and correct tone shifts. COTF (Li et al. 2024e) integrates global transformation with pixel by pixel transformation in an efficient manner, using image adaptive 3D LUT to adjust the overall appearance and pixel transformation to compensate for local context. Although these existing exposure correction methods have achieved certain results, they are unable to effectively restore the damaged structures and perceptual information in the image when faced with severe underexposure and overexposure.

CLIP in Low-Level Vision

Currently, CLIP (Radford et al. 2021) has also been explored and applied in low-level vision. CLIP-LIT (Liang et al. 2023) is the first to use CLIP for low-level restoration tasks. It utilizes the rich priors embedded in the CLIP model and employs an iterative prompt learning strategy to generate more accurate prompts, thereby better characterizing backlit and well-lit images. DA-CLIP (Luo et al. 2024) trained an additional controller that can accurately predict the degraded embeddings of LQ images. Integrating embeddings into image restoration networks through cross attention to guide model learning for high fidelity image reconstruction. Morawski et al. (Morawski et al. 2024) exploit the

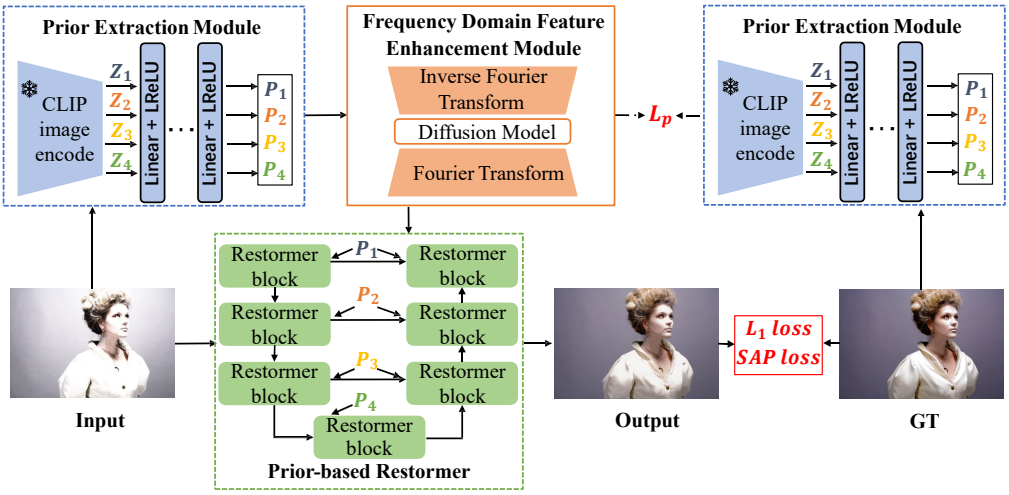

Figure 2: The overview of the proposed CLIP-RestoreX. In the first stage, we use the prior extraction module to obtain the structural and perceptual information (Z_1, Z_2, Z_3, Z_4) from the GT and transform it into a structural perceptual prior (P_1, P_2, P_3, P_4) through fully connected layers. The second stage is similar to the first stage. We input the under- and over-exposed images into the prior extraction module to obtain the damaged structural and perceptual features. The difference is that we input P_1, P_2, P_3, P_4 into the frequency domain based feature enhancement diffusion model we designed. Through FFEDM, the feature obtained from under- and over-exposed images through CLIP can be made similar to the prior obtained from GT. Finally, we optimized the model using L_p, L_1 loss, and our designed structural and perceptual loss.

rich CLIP prior and the zero-shot capability of CLIP to improve zero-reference low-light image enhancement during the training phase. They first pre-train a pair of prompts to capture enhanced low-light images in advance through prompt learning and a simple data augmentation strategy without the need for paired or unpaired normal-light data.

Previous works show that CLIP can benefit various low-level vision tasks, but there is no work exploring how to utilize CLIP for exposure correction severe under- and over-exposure. We take the structural and perceptual information damaged by severe under- and over-exposure as the starting point, and combine the structural and perceptual priors in CLIP to improve exposure correction.

Our Method

Given a severely overexposed or underexposed image, the goal of exposure correction is to produce an normal exposure image I_{out}. The main challenge of exposure correction is to restore the structural and perceptual information in severely under- and over-exposed images.

Overview

To address the main challenges, we propose to leverage structural and perceptual priors from CLIP to improve exposure correction. As shown in Figure 2, our method consists a prior extraction module and a feature enhancement module. Specifically, given an under- and over-exposure image as input, we first extract the damaged structure and perceptual feature through the prior extraction module, and then use the feature enhancement module to transform the damaged structure and perceptual feature into complete structure and perceptual priors. Finally, we inject the prior alone with the input image into Restormer to produce the image with normal exposure.

Prior Extraction Module

Given an image as input, PEM first leverages the CLIP image encoder to obtain image features from various blocks. For convenience, we note the features obtained from shallow to deep blocks as $Z = \{Z_1, ..., Z_n\}$, where n is the number of blocks. After that, PEM adopts a conversion module, which includes an MLP alone with non-linear activation functions, to convert the features into structural and perceptual priors, i.e., $P = \{P_1, ..., P_n\}$.

Frequency Domain Based Feature Enhancement Diffusion Model

CLIP can extract the complete structural and perceptual information from the normal-exposed images (we note them as GT), but for severely under- and over-exposed images, the structural and perceptual feature extracted by CLIP are damaged. Hence, it is necessary to enhance these feature and convert it into the feature that GT can provide. To this end, we propose to enhance the feature with a frequency domain based feature enhancement module (FFEDM). Specifically, given the feature P extracted by PEM, a Fast Fourier Transform (FFT) is adopted to model the relationship with the feature so that the frequency domain features $P^f = \{P_1^f, ..., P_n^f\}$ can be obtained. After that, a frequency domain based diffusion model is utilized to enhance the features to the complete prior that GT can provide. Subsequently, the enhanced features are transformed back to the original domain via inverse Fourier transform.

Discussion. The goal of FFEDM is to enhance the structural and perceptual feature of the improperly exposed images so that they could be aligned with the structural and perceptual information of the normal-exposed images. Note that the L_1 loss could be a naive way to this end. However, the L_1 loss just makes the average distances of the two features closest while ignoring the relationship within the features. On the contrary, our FFEDM can focus on the relationship within the feature. In frequency domain, the exposure correction problem can be simplified to adjustment of the amplitude and phase component, with the former related to brightness condition while the latter for the rest, which helps allow more efficient and effective exposure correction learning. This has been validated by various previous methods, such as FECNet (Huang et al. 2022b).

$$P_i^f = \mathrm{FFT}\left(P_i\right), P_i^{f'} = \mathrm{diff}\left(P_i^f\right), P_i' = \mathrm{IFFT}\left(P_i^{f'}\right), \tag{1}$$

where P_i is the i-th prior from P. FFT is the Fourier transform, P_i^f is the frequency domain of the i-th prior, diff refers to the diffusion model, and IFFT indicates the Inverse Fourier Transform. Our frequency-domain based diffusion model is based on the conditional denoising diffusion probability model (Ho, Jain, and Abbeel 2020; Rombach et al. 2022), and the optimization method is similar to previous work (Xia et al. 2023). Assuming the prior is x_0 and the noise vector at time step t is x_t, the diffusion process of our diffusion model is as follows:

$$q\left(x_t \mid x_{t-1}\right) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right), \tag{2}$$

where β_t is the predefined scale factor, and $\mathcal{N}$ represents the Gaussian distribution. The above equation can be further simplified as:

$$q\left(\mathbf{x}_t \mid \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right), \tag{3}$$

where $\alpha_t = 1-\beta_t, \bar{\alpha}_t = \prod_{i=0}^{t}\alpha_i$.

In the reverse process, the DM method samples the Gaussian random noise vector $\mathbf{x}_t$ and gradually denoises $\mathbf{x}_t$ until it reaches our first stage prior $\mathbf{x}_0$:

$$q\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_t\left(\mathbf{x}_t, \mathbf{x}_0\right), \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\mathbf{I}\right), \tag{4}$$

where $\boldsymbol{\mu}_t\left(\mathbf{z}_t, \mathbf{x}_0\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}\right)$, $\boldsymbol{\epsilon}$ represents the noise in $\mathbf{x}_t$. In this paper, we use the prior output by CLIP as condition(c), so our reverse process is as follows:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{y}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\mathbf{x}_t, \mathbf{c}, t\right)\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_t, \tag{5}$$

where $\boldsymbol{\epsilon}_\theta$ is the denoising network used to estimate $\boldsymbol{\epsilon}$. Our training objective is:

$$\nabla_{\boldsymbol{\theta}}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \mathbf{c}, t\right)\right\|_2^2. \tag{6}$$

Prior Injection into Restormer

Restormer is an efficient image restoration model. In this paper, we inject priors into Restormer to restore the structural and perceptual information in severely under- and over-exposed images. We use P_1', P_2', P_3', P_4' in the order they appear in the CLIP as shallow and deep blocks priors for the Restormer, respectively. Suppose the input of the attention layer and FFN layer of a Restormer block is $x \in \mathbb{R}^{H\times W\times C}$, and the prior to be injected is $P_i' \in \mathbb{R}^D$. Our injection module is:

$$x' = W_1 P_i' \odot \mathrm{Norm}(x) + W_2 P_i', \tag{7}$$

where $\odot$ indicates element-wise multiplication and Norm denotes layer normalization, W_1, W_2 represents linear layer and $W_1 P_i', W_2 P_i' \in \mathbb{R}^C$.

Training and Inference

During training, our loss function mainly consists of two parts. The first part is the loss between the prior extracted by CLIP from INPUT and the prior extracted from GT. The second part is the loss between the final output image of the model and GT.

The first part is the L_1loss between the prior vectors extracted by CLIP from Input and GT.

$$L_p = \left\|IFFT\left(p_{gt}\right) - IFFT\left(p_{\text{input}}\right)\right\|_1, \tag{8}$$

In order to fully utilize the structure and perceptual priors in the CLIP, we also use the vectors output by different blocks in the CLIP as losses to constrain the model's reconstruction of structural and perceptual information. The calculation method for the loss of structural and perceptual we designed is as follows:

$$L_{SAP} = \left\|\psi_i\left(I_{gt}\right) - \psi_i\left(I_{out}\right)\right\|_1, \tag{9}$$

where $\psi(\cdot)$is the feature vector output by CLIP image encode, and i indicates that the feature vector output by the $i-th$ block is extracted. Finally, our loss is:

$$L_{\text{total}} = L_p + \left\|I_{gt} - I_{\text{out}}\right\|_1 + L_{SAP}. \tag{10}$$

In the inference phase, we first extract the damaged structural and perceptual feature from the improperly exposed image through the prior extraction module, then use the frequency-domain based feature enhancement diffusion to transform the damaged structural and perceptual feature into structural and perceptual priors. Finally, we inject the priors into Restormer and output the corrected image.

Experiment

Experimental Settings

Datasets We conduct experiments on three public datasets, i.e., LCDP (Afifi et al. 2021), MSEC (Wang, Xu, and Lau 2022), and SICE (Cai, Gu, and Zhang 2018). LCDP is a non-uniform exposure dataset with both overexposure and underexposure in a single image. It contains 1415 training sample pairs, 100 validation sample pairs, and 218 test sample pairs. A set of images in the MSEC dataset has 5 different brightnesses, and its exposure values (EVs) are -1.5,

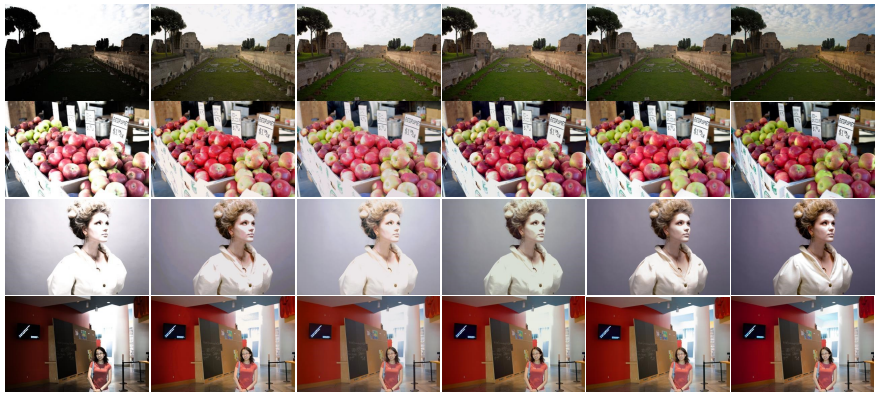

Figure 3: Visual comparisons between our method and state-of-the-art methods on the LCDP dataset. In the first row, we better restore the details of the clouds in the sky (structural information). In the second row, we better restore the color of the apple (perceptual information). In the third row, we better restore the purple background of the woman (perceptual information). In the fourth row we better restore the structural information of the roof.

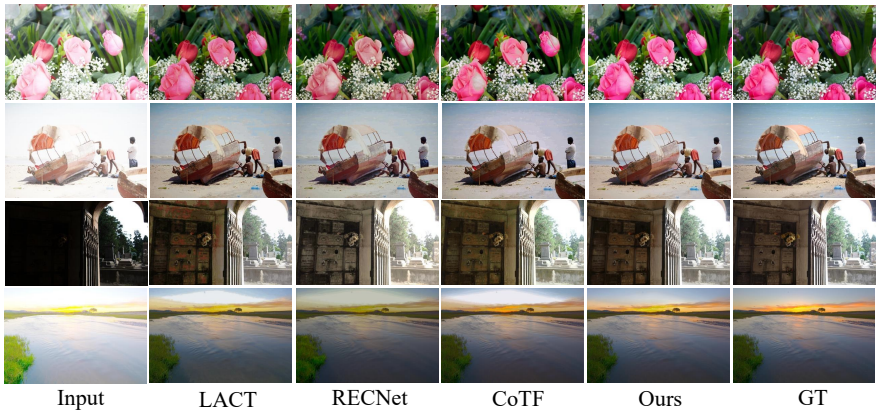

Figure 4: Visual comparisons between our method and state-of-the-art methods on the MSEC and SICE datasets. In the first row, we better restore the color of the flower (perceptual information). In the second row, we better restore the color of the background (perceptual information). In the third row, we better restore the detail information of the wall (structural information). In the fourth row, we better restore the detail information of the sky (structural information).

-1, 0, 1, and 1.5 respectively. Its test set is adjusted by 5 human experts. MSEC includes 17,675 training sample pairs, 750 validation sample pairs, and 5,905 test sample pairs. As in the previous method (Huang et al. 2022a), for the SICE dataset, we take the middle exposure subset as the ground truth, while the second and last second exposure subsets are set to the under- and over-exposed images, respectively.

Evaluation Metrics Following previous works on exposure correction, we use the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) (Wang et al. 2004) as the evaluation metrics. PSNR evaluates the degree of distortion between two images. The larger the value, the smaller the distortion. A higher SSIM indicates a higher structural similarity between the two images.

Implementation Details The architecture of the pre-trained CLIP image encoder used in our experiment is VIT-B-32. We separate the image encoder into 4 blocks ($n = 4$). The number of blocks of Restormer is [2, 2, 2, 2, 4]. In the first stage of training, the batch-size is 4, the image is cropped to 224 $\times$ 224, the initial learning rate is $2e^{-4}$, and the cosine annealing strategy is used for updating. In order to enable CLIP to obtain more structural and perceptual information in the image, in the second stage, we use a larger patch size. In the second stage of training, our batch-size is 1, the image is cropped to 512 $\times$ 512, and the learning rate is $2e^{-4}$. We train models with Adam (Kingma and Ba 2014) optimizer ($\beta_1 = 0.9, \beta_2 = 0.99$). Our models are implemented using Pytorch and run on NVIDIA RTX8000 GPUs.

Comparisons with State-of-the-art Methods

We compare our model with several state-of-the-art exposure correction methods, including MSEC (Afifi et al. 2021), ENC (Huang et al. 2022a), LCDPNet (Wang, Xu, and Lau 2022), FECNet (Huang et al. 2022b), ECLNet (Huang et al.

2022c), LACT (Baek et al. 2023), RECNet (Liu et al. 2024), and CoTF (Li et al. 2024e). As shown in Table 1, the proposed CLIP-RestoreX significantly outperforms existing methods among all metrics on three datasets. Moreover, we perform visualization comparisons to further evaluate the effectiveness of our model. Figure 3 and Figure 4 show comparisons of the visualization results between SOTA methods and ours. It can be observed that our method recovers the structural and perceptual information lost more effectively.

Methods	LCDP		MSEC		SICE	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MSEC	20.12	0.6462	20.21	0.8194	19.13	0.5675
ENC-SID	22.27	0.7824	22.45	0.8499	20.37	0.6795
ENC-DRBN	23.55	0.8528	22.35	0.8531	20.52	0.7171
LCDPNet	23.24	0.8420	22.30	0.8566	18.80	0.6452
FECNet	23.31	0.8310	22.84	0.8663	20.96	0.6733
ECLNet	23.08	0.7931	22.58	0.8686	20.65	0.6943
LACT	24.27	0.8621	23.53	0.8728	22.11	0.7102
RECNet	22.77	0.8183	22.12	0.8636	20.45	0.6926
CoTF	23.89	0.8581	23.46	0.8733	21.51	0.7151
Ours	**24.95**	**0.8864**	**23.57**	**0.8764**	**23.21**	**0.7302**

Table 1: Quantitative comparison between the proposed method and SOTA methods on the LCDP, MSEC, and SICE datasets.

Ablation Study

In this section, we perform ablation studies on the LCDP dataset to demonstrate the effectiveness of each component in our proposed method, including structure and perception priors, FFEDM and SAP loss.

Structure and Perception Priors We first verify the shallow and deep blocks of CLIP can serve as the structure and perceptual priors through quantitative and qualitative experimental results. For better comparison, we ablate the SAP loss and FFEDM module to reflect the role of different blocks in CLIP for exposure correction. As shown in Table 2, compared with the model using features from both shallow and deep layer (i.e., CLIP-3, 6, 9, 12), variants with only shallow or deep layer (i.e., CLIP-1, 2, 3, 4 or CLIP-9, 10, 11, 12) suffer from performance deterioration. Such results suggest that the help of features from shallow and deep layers are complementary for exposure correction. Besides, we analyze that shallow and deep layers of the VIT-B-32 model pre-trained on ImageNet(Deng et al. 2009) cannot provide different assistance for exposure correction, which can be found in the supplementary materials.

As can be seen from Figure 5, using 1-2-3-4 blocks as a prior allows the model to restore structural information of areas damaged by under- and over-exposures exposure (for example, 9-10-11-12 blocks failed to restore structural information on the little girl's knees that was damaged by overexposure), and using 9-10-11-12 blocks as a prior restores the perceptual information better (for example, the 9-10-11-12 blocks restores the woman's purple background deeper). In contrast, using 3-6-9-12 blocks can combine the advantages of both, restoring both structural information and perceptual information such as color.

Variants	Shallow layer	Deep layer	PSNR	SSIM
CLIP-1,2,3,4	✓	×	24.42	0.8824
CLIP-9,10,11,12	×	✓	24.63	0.8815
CLIP-3,6,9,12	✓	✓	**24.75**	**0.8835**

Table 2: Comparison of results of different blocks of CLIP as priors. The number after CLIP indicates the i-th block. The settings in our default model are colored in gray. Note that we ablate SAP loss and FFEDM module here to evaluate the role of CLIP's different blocks for better comparison.

FFEDM and SAP Loss Next, we verify the effectiveness of FFEDM and SAP loss. From the results in Table 3, it can be seen that using CLIP prior, FFEDM, and SAP loss can bring better results. Comparing the second and third rows, the sixth and seventh rows, we can see that using SAP loss can effectively help the model training. Comparing the fourth and fifth rows, we can see that CLIP can provide effective prior information for exposure correction. Comparing the fifth and sixth rows, we can see the effectiveness of the FFDEM we proposed.

CLIP-P	DM	FFEDM	L_{SAP}	PSNR	SSIM
×	×	×	×	23.67	0.8671
×	×	×	✓	24.10	0.8711
×	✓	×	×	24.14	0.8808
✓	✓	×	×	24.75	0.8835
✓	×	✓	×	24.86	0.8858
✓	×	✓	✓	24.95	0.8864

Table 3: Ablation experiments of FFEDM and SAP loss on the LCDP dataset. CLIP-P represents the outputs of CLIP's 3-6-9-12 blocks as Priors. DM represents the diffusion model in DIFFIR. FFEDM is our frequency domain based feature enhancement diffusion model. SAP loss is the loss function we designed with the outputs of CLIP's 3-6-9-12 blocks as constraints. The settings in our default model are colored in gray.

Analysis of Structure and Perception Priors

Here, we first analyze why the shallow and deep blocks of CLIP can serve as the structural and perceptual priors for the exposure correction task. To this end, we extracted 700 unpaired data from the LCDP training set and output them to CLIP, and then visualized the output of 1, 3, 7, 11 blocks of CLIP. As shown in Figure 6, the orange points are under/over-exposed images, and the blue points are normal images. As the block goes deeper, the distribution of orange points gradually spreads from the middle area where they are initially gathered to the entire area. This is because the shallow blocks focus on the structural information of the

Figure 5: Comparison of exposure correction results using different CLIP layers. It can be seen that the shallow layers help recover better structural information , while the deep layers benefits restore more vivid color.

images, and severely under/over-exposed images destroy the structure of some regions to a certain extent. Therefore, the shallow blocks can easily distinguish between normal images and images with damaged structures. The deeper blocks pay more attention to perceptual information, so the orange points with different semantics will gradually spread out.

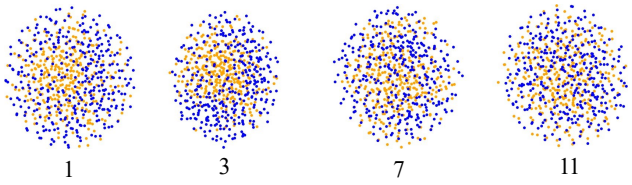

Figure 6: T-SNE visualizations of the features of the incorrect-exposed and normal-exposed images extracted by various blocks (i.e., the 1st, 3rd, 7th, and 11th blocks) in CLIP. The orange points indicate abnormally exposed images, while the blue points indicate normally exposed images. As the block goes deeper, the two types of features become increasingly indistinguishable.

We conducted further analysis on the MSEC (Afifi et al. 2021) training set. In the MSEC training set, a set of images includes five images ranging from dark to bright and one normal image (GT), where the middle brightness is closest to the brightness of GT. The semantic information of these five images is almost the same, and the main difference is the brightness. In other words, the high-level perceptual information that humans pay more attention to in these five images is very similar. We still follow the above method to obtain the vectors output by 1, 3, 7, 11 blocks in the pre-trained model (VIT-B-32) on ImageNet and CLIP image encoder (VIT-B-32), and then calculate the distance between the vector pairs (INPUT and its corresponding GT). We use the distance between the middle brightness image and GT as the benchmark, and count the changes in the distance between other brightness images and GT relative to benchmark. From Figure 7, we can see that the distance changes of shallow blocks in CLIP image encoder are larger when facing brightness changes, while the distance changes of deep blocks are smaller. However, the distance changes of shallow blocks and deep blocks of VIT-B-32 pre-trained on ImageNet are almost the same when facing brightness changes.

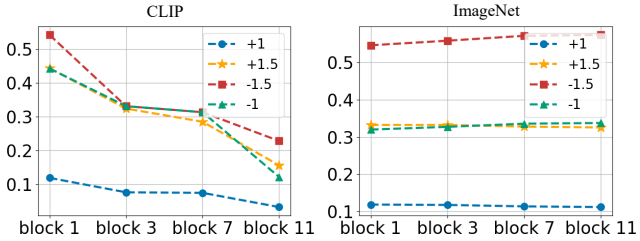

Figure 7: Comparison of the changes in the distances between exposure values (EVs) -1.5, -1, +1, +1.5 and GT relative to the distance between exposure value 0 and GT on the MSEC training set. We define the distance between the image with exposure value 0 and the feature vector output by GT after passing through 1-3-7-11 blocks as 1 (benchmark), and then count the change in the distance between the image with exposure values -1.5, -1, +1, +1.5 and the feature vector of GT relative to benchmark in 1-3-7-11 blocks.

Conclusion

We have presented a novel exposure correction method that can effectively deal with previously challenging severely overexposed or underexposed images. Our method is built upon the observation that CLIP encodes the structural and perceptual information required by exposure correction. Hence, we design our method to leverage CLIP priors for exposure correction. Besides, we develop a frequency domain-based feature enhancement diffusion model to restore the structural and perceptual information damaged by under- and over-exposure. Experiments on various benchmark datasets show that the proposed method outperforms state-of-the-art exposure correction methods, both qualitatively and quantitatively.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2023YFA1008503), NSFC(92470202, U21A20471, 62471499), Guangdong NSF Project (No. 2023B1515040025), Guangdong Basic and Applied Basic Research Foundation (2023A1515030002). The authors would like to thank Kun-Yu Lin and Yuan-Ming Li for their valuable suggestions on model design and writing.

References

Abdullah-Al-Wadud, M.; Kabir, M. H.; Dewan, M. A. A.; and Chae, O. 2007. A dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, 53(2): 593–600.

Afifi, M.; Derpanis, K. G.; Ommer, B.; and Brown, M. S. 2021. Learning multi-scale photo exposure correction. In *CVPR*, 9157–9167.

Baek, J.-H.; Kim, D.; Choi, S.-M.; Lee, H.-j.; Kim, H.; and Koh, Y. J. 2023. Luminance-aware color transform for multiple exposure correction. In *ICCV*, 6156–6165.

Cai, B.; Xu, X.; Guo, K.; Jia, K.; Hu, B.; and Tao, D. 2017. A joint intrinsic-extrinsic prior model for retinex. In *ICCV*, 4000–4009.

Cai, J.; Gu, S.; and Zhang, L. 2018. Learning a deep single image contrast enhancer from multi-exposure images. *IEEE Transactions on Image Processing*, 27(4): 2049–2062.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Eyiokur, F.; Yaman, D.; Ekenel, H. K.; and Waibel, A. 2022. Exposure correction model to enhance image quality. In *CVPR*, 676–686.

Guo, S.; Liu, L.; Gan, Z.; Wang, Y.; Zhang, W.; Wang, C.; Jiang, G.; Zhang, W.; Yi, R.; Ma, L.; et al. 2022. Isdnet: Integrating shallow and deep networks for efficient ultra-high resolution segmentation. In *CVPR*, 4361–4370.

Guo, X.; Li, Y.; and Ling, H. 2016. LIME: Low-light image enhancement via illumination map estimation. *IEEE Transactions on Image Processing*, 26(2): 982–993.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *NeurIPS*, volume 33, 6840–6851. Curran Associates, Inc.

Huang, J.; Liu, Y.; Fu, X.; Zhou, M.; Wang, Y.; Zhao, F.; and Xiong, Z. 2022a. Exposure normalization and compensation for multiple-exposure correction. In *CVPR*, 6043–6052.

Huang, J.; Liu, Y.; Zhao, F.; Yan, K.; Zhang, J.; Huang, Y.; Zhou, M.; and Xiong, Z. 2022b. Deep fourier-based exposure correction network with spatial-frequency interaction. In *ECCV*, 163–180. Springer.

Huang, J.; Zhao, F.; Zhou, M.; Xiao, J.; Zheng, N.; Zheng, K.; and Xiong, Z. 2023. Learning sample relationship for exposure correction. In *CVPR*, 9904–9913.

Huang, J.; Zhou, M.; Liu, Y.; Yao, M.; Zhao, F.; and Xiong, Z. 2022c. Exposure-consistency representation learning for exposure correction. In *ACMM*, 6309–6317.

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

Land, E. H. 1977. The retinex theory of color vision. *Scientific American*, 237(6): 108–129.

Li, M.; Liu, J.; Yang, W.; Sun, X.; and Guo, Z. 2018. Structure-revealing low-light image enhancement via robust retinex model. *IEEE Transactions on Image Processing*, 27(6): 2828–2841.

Li, Y.; Xu, K.; Hancke, G. P.; and Lau, R. W. 2024a. Color Shift Estimation-and-Correction for Image Enhancement. In *CVPR*, 25389–25398.

Li, Y.-M.; Huang, W.-J.; Wang, A.-L.; Zeng, L.-A.; Meng, J.-K.; and Zheng, W.-S. 2024b. EgoExo-Fitness: Towards Egocentric and Exocentric Full-Body Action Understanding. *arXiv preprint arXiv:2406.08877*.

Li, Y.-M.; Wang, A.-L.; Lin, K.-Y.; Tang, Y.-M.; Zeng, L.-A.; Hu, J.-F.; and Zheng, W.-S. 2024c. TechCoach: Towards Technical Keypoint-Aware Descriptive Action Coaching. *arXiv preprint arXiv:2411.17130*.

Li, Y.-M.; Zeng, L.-A.; Meng, J.-K.; and Zheng, W.-S. 2024d. Continual Action Assessment via Task-Consistent Score-Discriminative Feature Distribution Modeling. *IEEE Transactions on Circuits and Systems for Video Technology*.

Li, Z.; Zhang, F.; Cao, M.; Zhang, J.; Shao, Y.; Wang, Y.; and Sang, N. 2024e. Real-Time Exposure Correction via Collaborative Transformations and Adaptive Sampling. In *CVPR*, 2984–2994.

Liang, Z.; Li, C.; Zhou, S.; Feng, R.; and Loy, C. C. 2023. Iterative prompt learning for unsupervised backlit image enhancement. In *ICCV*, 8094–8103.

Lin, K.-Y.; Ding, H.; Zhou, J.; Tang, Y.-M.; Peng, Y.-X.; Zhao, Z.; Loy, C. C.; and Zheng, W.-S. 2024. Rethinking clip-based video learners in cross-domain open-vocabulary action recognition. *arXiv preprint arXiv:2403.01560*.

Lin, K.-Y.; Du, J.-R.; Gao, Y.; Zhou, J.; and Zheng, W.-S. 2023. Diversifying Spatial-Temporal Perception for Video Domain Generalization. In *NeurIPS*, volume 36, 56012–56026. Curran Associates, Inc.

Lin, K.-Y.; Zhou, J.; Qiu, Y.; and Zheng, W.-S. 2022. Adversarial partial domain adaptation by cycle inconsistency. In *ECCV*, 530–548. Springer.

Lin, K.-Y.; Zhou, J.; and Zheng, W.-S. 2024. Human-Centric Transformer for Domain Adaptive Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, J.; Fu, H.; Wang, C.; and Ma, H. 2024. Region-Aware Exposure Consistency Network for Mixed Exposure Correction. In *AAAI*, volume 38, 3648–3656.

Luo, Z.; Gustafsson, F. K.; Zhao, Z.; Sjölund, J.; and Schön, T. B. 2024. Controlling Vision-Language Models for Multi-Task Image Restoration. In *ICLR*.

Morawski, I.; He, K.; Dangi, S.; and Hsu, W. H. 2024. Unsupervised Image Prior via Prompt Learning and CLIP Semantic Guidance for Low-Light Image Enhancement. In *CVPR*, 5971–5981.

Nsampi, N. E.; Hu, Z.; and Wang, Q. 2021. Learning Exposure Correction Via Consistency Modeling. In *BMVC*, 12.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Ren, X.; Yang, W.; Cheng, W.-H.; and Liu, J. 2020. LR3M: Robust low-light enhancement via low-rank regularized retinex model. *IEEE Transactions on Image Processing*, 29: 5862–5876.

Reza, A. M. 2004. Realization of the contrast limited adaptive histogram equalization (CLAHE) for real-time image enhancement. *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, 38: 35–44.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Tian, Q.-C.; and Cohen, L. D. 2017. Global and local contrast adaptive enhancement for non-uniform illumination color images. In *ICCV*, 3023–3030.

Wang, H.; Xu, K.; and Lau, R. W. 2022. Local color distributions prior for image enhancement. In *ECCV*, 343–359. Springer.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring clip for assessing the look and feel of images. In *AAAI*, volume 37, 2555–2563.

Wang, R.; Zhang, Q.; Fu, C.-W.; Shen, X.; Zheng, W.-S.; and Jia, J. 2019. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, 6849–6857.

Wang, Y.; Peng, L.; Li, L.; Cao, Y.; and Zha, Z.-J. 2023. Decoupling-and-aggregating for image exposure correction. In *CVPR*, 18115–18124.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Wei, J.; Wang, Q.; Li, Z.; Wang, S.; Zhou, S. K.; and Cui, S. 2021. Shallow feature matters for weakly supervised object localization. In *CVPR*, 5993–6001.

Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; and Van Gool, L. 2023. Diffir: Efficient diffusion model for image restoration. In *ICCV*, 13095–13105.

Xu, Z.; Shang, H.; Yang, S.; Xu, R.; Yan, Y.; Li, Y.; Huang, J.; Yang, H. C.; and Zhou, J. 2023. Hierarchical painter: Chinese landscape painting restoration with fine-grained styles. *Visual Intelligence*, 1(1): 19.

Ying, Z.; Li, G.; Ren, Y.; Wang, R.; and Wang, W. 2017. A New Image Contrast Enhancement Algorithm Using Exposure Fusion Framework. In *Computer Analysis of Images and Patterns*, 36–46. Springer International Publishing.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 5728–5739.

Zhang, Q.; Nie, Y.; and Zheng, W.-S. 2019. Dual illumination estimation for robust exposure correction. In *Computer Graphics Forum*, volume 38, 243–252. Wiley Online Library.

Zhang, Q.; Nie, Y.; Zhu, L.; Xiao, C.; and Zheng, W.-S. 2020. Enhancing underexposed photos using perceptually bidirectional similarity. *IEEE Transactions on Multimedia*, 23: 189–202.

Zhang, Q.; Yuan, G.; Xiao, C.; Zhu, L.; and Zheng, W.-S. 2018. High-quality exposure correction of underexposed photos. In *ACMM*, 582–590.