Semantically Contrastive Learning for Low-Light Image Enhancement

Dong Liang[1], Ling Li[1], Mingqiang Wei[1*], Shuo Yang[2], Liyan Zhang[1], Wenhan Yang[3], Yun Du[1], Huiyu Zhou[4]

[1]Nanjing University of Aeronautics and Astronautics, MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Collaborative Innovation Center of Novel Software Technology and Industrialization
[2]vivo Mobile Communication   [3]Nanyang Technological University   [4]University of Leicester
{liangdong, liling, zhangliyan, muyun}@nuaa.edu.cn, mingqiang.wei@gmail.com, shuo.yang@vivo.com, wenhan.yang@ntu.edu.sg, hz143@leicester.ac.uk

Abstract

Low-light image enhancement (LLE) remains challenging due to the unfavorable prevailing low-contrast and weak-visibility problems of single RGB images. In this paper, we respond to the intriguing learning-related question – if leveraging both accessible unpaired over/underexposed images and high-level semantic guidance, can improve the performance of cutting-edge LLE models? Here, we propose an effective semantically contrastive learning paradigm for LLE (namely SCL-LLE). Beyond the existing LLE wisdom, it casts the image enhancement task as multi-task joint learning, where LLE is converted into three constraints of contrastive learning, semantic brightness consistency, and feature preservation for simultaneously ensuring the exposure, texture, and color consistency. SCL-LLE allows the LLE model to learn from *unpaired* positives (normal-light)/negatives (over/underexposed), and enables it to interact with the scene semantics to regularize the image enhancement network, yet the interaction of high-level semantic knowledge and the low-level signal prior is seldom investigated in previous methods. Training on readily available open data, extensive experiments demonstrate that our method surpasses the state-of-the-arts LLE models over six independent cross-scenes datasets. Moreover, SCL-LLE's potential to benefit the downstream semantic segmentation under extremely dark conditions is discussed. *Source Code:* https://github.com/LingLIx/SCL-LLE.

Introduction

Capturing images with underexposures remains a significant source of errors in camera imaging. Such degenerated images severely hinder some downstream tasks from operating smoothly, such as semantic segmentation or object detection in vision-based driving assistance systems.

Data-driven methods have been proposed to formulate the LLE problem by learning strict pixel correspondence of image pairs via strong supervisions, *e.g.*, (Lore, Akintayo, and Sarkar 2017; Wei et al. 2018; Zhang, Zhang, and Guo 2019; Xu et al. 2020; Ren et al. 2019). Although having achieved remarkable successes in LLE, they only adopt the normal-illumination images as positive samples to guide the training while neglecting to exploit the over/underexposed images

(a) Input   (b) w/o Neg. samples   (c) Ours
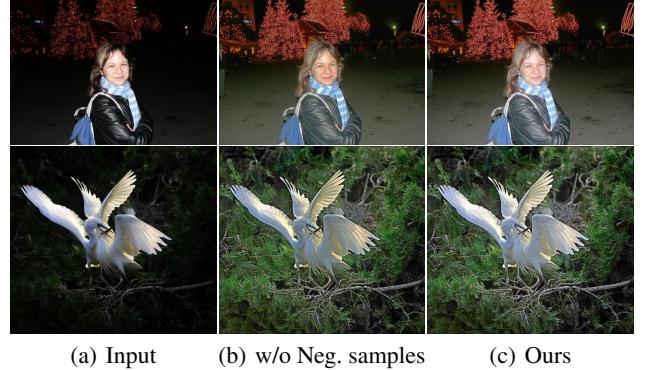
Figure 1: Impact of training data. (b) no negative samples are involved in training.

as negative samples. On the one hand, high-quality positive samples without brightness and color defects are challenging to acquire in practice. For example, Ignatov *et al.* propose a scheme to acquire them from a DSLR camera to refine the imaging of a mobile phone camera (Ignatov et al. 2017). It brings complicated registration and pixel-by-pixel calibration to image pairs. On the other hand, negative samples with error exposures are readily available. As shown in Fig. 1, our method yields visually better results, as the negative samples also provide supervisory information to generate more discriminative visual mapping.

Another drawback of current methods is the ignorance of the important semantic information. As shown in Fig. 3, the enhancement strategies for the background and foreground should be different. Scene semantics can help to distinguish different areas for further enhancement. In addition, there is natural brightness consistency inside a semantic category. Employing such consistency can help to avoid local uneven exposure. Although methods (Fan et al. 2020; Xie et al. 2020) apply semantic information as guidance to improve image enhancement, they ignore the consistency between pixels of the same semantic category.

The motivation of this paper can be summarized as follows: 1) to release from the strict pixel correspondence of training image pairs and to leverage over/underexposed images as negative samples for building a more flexible and

robust visual mapping method; 2) to effectively leverage the semantic information to distinguish enhancement areas and to keep brightness consistency inside the same semantic category. To this end, we develop a novel paradigm – semantically contrastive enhancement learning for low-light image enhancement (SCL-LLE for short). SCL-LLE casts the image enhancement task as multi-task joint learning, where LLE is converted into three constraints of contrastive learning, semantic brightness consistency, and feature preservation for simultaneously ensuring the exposure, texture, and color consistency. From the respective of ease to use, we use readily accessible training data – the Cityscapes (Cordts et al. 2016) dataset, to provide input images with semantic ground truths, and the Part1 of SICE dataset (Cai, Gu, and Zhang 2018), to provide unpaired negative/positive samples. The contributions of this paper are three folds:

- SCL-LLE removes pixel-correspond paired training data, and provides a more flexible way: 1) training with unpaired images in different real-world domains, 2) training with unpaired negative samples, allowing us to leverage readily available open data to build a more generalized and discriminative LLE network.

- Low and high-level vision tasks (i.e., LLE and semantic segmentation) promote each other. A semantic brightness consistency loss is introduced to ensure smooth and natural brightness recovery of the same semantic category. The enhanced images lead to better performance on the downstream semantic segmentation.

- SCL-LLE is compared with SOTAs via comprehensive experiments on six independent datasets, in terms of visual quality, no and full-referenced image quality assessment, and human subjective survey. All results consistently endorse the superiority of the proposed approach.

Related Work

Low-light Image Enhancement

Early efforts are commonly made towards the use of handcrafted priors with empirical observations (Pizer et al. 1990; Land 1977; Xu et al. 2014; Guo, Li, and Ling 2016) to deal with the LLE problem. Histogram equalization (Pizer et al. 1990) used a cumulative distribution function to regularize the image's pixel values to distribute overall intensity levels evenly. Retinex model (Land 1977) and its multi-scale version (Jobson et al. 1997) decomposed the brightness into illumination and reflectance and then process them separately. Guo $et\ al.$ (Guo, Li, and Ling 2016) introduced a structural prior to refine the initial illumination map and finally synthesized the enhanced image according to the Retinex theory. However, these handcrafted constraints/priors are not self-adaptive enough to recover image details and color, resulting in washing out details, local under/over-saturation, uneven exposure, or halo artifacts around objects.

Lore $et\ al.$ (Lore, Akintayo, and Sarkar 2017) proposed a variant of the stacked sparse denoising autoencoder to enhance the degraded images. RetinexNet (Wei et al. 2018) leveraged a deep architecture based on Retinex to enhance low-light images. Zhang $et\ al.$ (Zhang, Zhang, and Guo

2019) developed three subnetworks for layer decomposition, reflectance restoration, and illumination adjustment based on Retinex. RUAS (Liu et al. 2021) constructed the overall LLE network architecture by unfolding its optimization process. The above methods are trained based on image pairs with strict pixel correspondence.

Recently, (Jiang et al. 2021) reported an unsupervised method using normal-light images which do not have low-light images as correspondences. Zero-DCE (Guo et al. 2020) reformulated the LLE task as an image-specific curve estimation problem with a fixed default brightness value without any reference image. Different from (Jiang et al. 2021) and (Guo et al. 2020), our method learns the appropriate brightness from the unpaired negative samples.

In another recent work, Fan $et\ al.$ (Fan et al. 2020) used semantic information to guide the reconstruction of the reflection of Retinex to eliminate noise. In contrast, we directly introduce semantic information to the brightness reconstruction and pay more attention to the dependency among the semantic elements via the interaction of high-level semantic knowledge and low-level signal priors.

Contrastive Learning

Contrastive learning (He et al. 2020; Chen et al. 2020; Sermanet et al. 2018; Tian, Krishnan, and Isola 2019; Henaff 2020) is from the self-supervised learning paradigm, which is characterized by using pretext tasks to mine its own supervisory information from original data for downstream tasks. For a given input, contrastive learning aims to pull it together with the positives and push it apart from negatives in a representation space. Previous works have applied contrastive learning to high-level vision tasks because these tasks are inherently suited for modeling the contrast between positive and negative samples. A recent work (Wu et al. 2021) has demonstrated that it can improve image dehaze.

In this paper, we present contrastive learning to remove the dependency of training data and allows us to use both positive and negative samples, helping to collect training samples from available open data. In addition, most of the existing contrastive learning methods rely heavily on a large number of negative samples and thus require either large batches or memory banks (Li et al. 2021). In our approach, we employ only a couple of negative samples for one positive sample and introduce a random mapping strategy to avoid the risk of overfitting.

Methodology

Problem Formulation and Overall Architecture

Fundamentally, low-light image enhancement can be regarded as seeking a mapping function F, such that $I_H = F(I_L)$ is the desired image, which is enhanced from the input image I_L. In our design, we introduce two different priors: one is the contrastive samples including the negatives I_N, i.e., the under/overexposed images, and the positives I_P, i.e., the normal-light images; the other is the semantic-related priors including the semantic segmentation ground truth y and the semantic prediction $\tilde{y}$. Therefore, we formu-
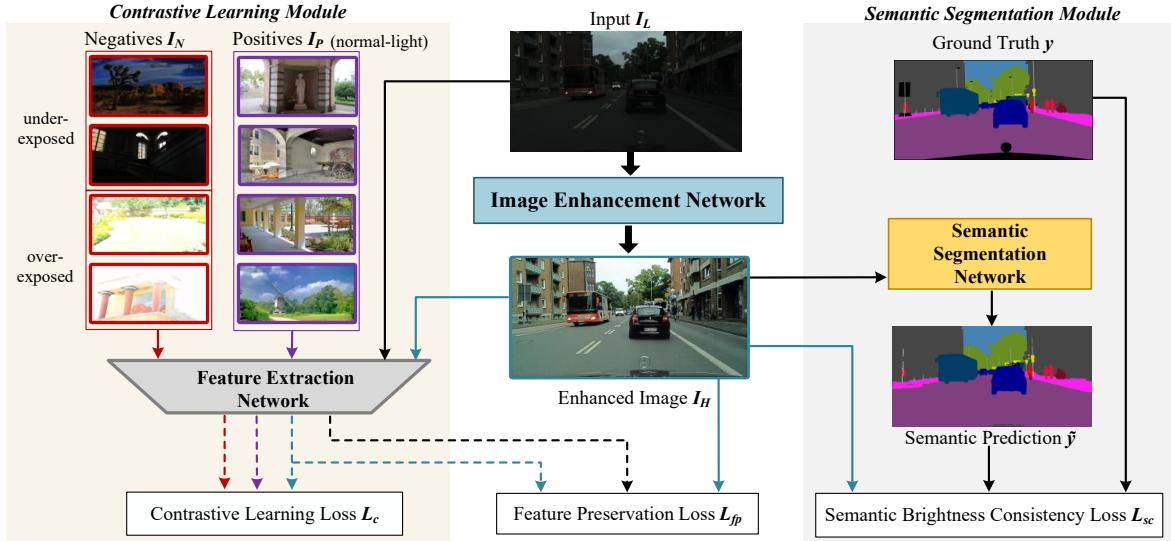
Figure 2: Overall architecture of our proposed SCL-LLE. It includes a low-light image enhancement network, a contrastive learning module and a semantic segmentation module.

late a new mapping function as follows:

$$I_H = F(I_L, I_N, I_P, y, \tilde{y}) \tag{1}$$

At the top level, we design a novel semantically contrastive learning framework for LLE (call SCL-LLE) as shown in Fig. 2 to obtain a better mapping function F than the existing methods do.

SCL-LLE consists of an image enhancement network, a semantic segmentation network, and a feature extraction network. Specifically, given an I_L input, the image enhancement network is first applied, and the enhanced result is then fed into the following semantic segmentation network. For the task modules, we leverage three mainstream networks: The low-light image enhancement network is a U-Net like backbone (Guo et al. 2020), which remaps every pixel by generating the pixel correction curve; the semantic segmentation network that we use here is the popular DeepLabv3+ (Chen et al. 2018); the feature extraction network is the VGG-16 (Simonyan and Zisserman 2014). We use three losses to correspond to the three aspects – Contrastive learning, semantic brightness consistency, and feature preservation.

Contrastive Brightness Restoration

We use the normal-light and over/underexposed images as the positive samples and negative samples to restore the brightness of the low-light images. Note that, to ensure the flexibility of the method, the positive and negative samples can be select in different scenes with each other and with the input image, i.e., they are unpaired with the input image and also unpaired with each other.

The goal is to learn a representation to pull together "positive" pairs in the latent feature space and push apart the representation between "negative" pairs. We need to consider two aspects: to construct the "positive" pairs and "negative"

pairs, and to find the latent feature space of these pairs for contrasting. In our method, the positive pair is generated by a normal-light image I_P and an enhanced image I_H by the low-light image enhancement network. Similarly, the negative pair is generated by an over/underexposure image I_N and an enhanced image I_H. For the latent feature space, besides using the appearance of the image, we choose a pre-trained VGG-16 to extract the feature map $f \in \mathbb{R}^{C \times H \times W}$, where G_{ij}^l is the inner product between the feature maps i and j in the layer l:

$$G_{ij}^l = \sum_k f_{ik}^l f_{jk}^l \tag{2}$$

where k represents the vector length. We then get a set of Gram matrices $\{G^1, G^2, \ldots, G^L\}$ from layers $1, \ldots, L$ in the feature extraction network. The Gram matrix G is a quantitative description of latent image features. Similar to the triplet loss (Schroff, Kalenichenko, and Philbin 2015; Hermans, Beyer, and Leibe 2017), our goal is:

$$d(G(I_H), G(I_P)) \ll d(G(I_H), G(I_N)) \tag{3}$$

$$d(E(I_H), E(I_P)) \ll d(E(I_H), E(I_N)) \tag{4}$$

where E represents the expectation. We wish that the distance d between features I_H and I_P is smaller than the distance between features I_H and I_N. The contrastive learning loss L_c can be expressed as:

$$L_c = max \{d(G(I_H), G(I_P)) - d(G(I_H), G(I_N)) + \alpha, 0\}$$
$$+ max \{d(E(I_H), E(I_P)) - d(E(I_H), E(I_N)) + \beta, 0\} \tag{5}$$

where α and β are hyperparameters (i.e., the margin in the triplet loss), we set them to 0.04 and 0.3 respectively in our experiments. In our implementation, the triplet loss is used to formulate the above loss function, which is a particular case of the contrastive loss when the number of the positive

and negative samples is one. The details of the loss derivation can be found in supplementary material.

In the latest theoretical work (Li et al. 2021), the authors argued that negative pairs are necessary, but one is sufficient for a triplet loss. They also observe that contrastive learning for visual representation can gain significantly from randomness. Inspired by this work, to collect positive and negative samples, we leverage SICE (Cai, Gu, and Zhang 2018) dataset, which contains low-contrast images with different exposure levels and their corresponding high-quality images. The SICE dataset includes indoor/outdoor 589 scenes (Part1 with 360 scenes and Part2 with 229 scenes) with a total number of 4413 multi-exposure images. All 360 standard images in all the scenes of Part1 are used as positive samples. We choose one underexposed and one overexposed sample as negatives for each scene, a total of 720 images from Part1 of SICE. To improve the robustness of the model, during training, positive and negative samples are randomly selected in each iteration.

Semantic Brightness Consistency Constraint

To better maintain the details of the image and make full use of its semantic information, we propose a semantic consistency loss to restrict brightness consistency and smoothness. This constraint can ensure that the same semantic parts in the enhanced image are consistent, which proves to be critical in avoiding local over/underexposures.

In a real scene, the elements belonging to the same semantic category have a clustered or adjacent location and should be similar with consistent brightness (*e.g.*, the sky and the road). The existing enhancement networks cannot make the parts of inconsistent brightness to be smooth. Based on this observation, we define an average value B of the brightness level of the overall pixels in each semantic category as follows:

$$B_s = \frac{1}{n} \sum_{i \in \theta_s} (B^i_{I_H}) \tag{6}$$

where s represents the s-th category, and we can attain multiple averages representing individual categories separately $\{B_1, B_2, \ldots\}$. n represents the number of the semantic pixels in this category with the ground truth y. We denote θ_s as the pixel index collection belonging to category s, $B^i_{I_H}$ as the brightness level in the enhanced image I_H at the category s. The semantic brightness consistency loss L_{sc} is defined as:

$$L_{sc} = \sum_{s=1}^{S} \sum_{i \in \theta_s} (B^i_{I_H} - B_s)^2 - \sum_{s=1}^{S} (p_s * \log q_s) \tag{7}$$

where S is the number of the categories of the semantic prediction $\tilde{y}$, p_s represents the ground truth value, and q_s represents the predicted value in the semantic prediction $\tilde{y}$.

To collect input images with semantic ground truths, we use the Cityscapes (Cordts et al. 2016) dataset, which is a street scene dataset and contains 5000 images with pixel-level fine/coarse annotations. There are 2975 images for training, 500 for validation, and 1525 for testing. We train our network on the training images of the Cityscapes dataset

for two reasons. First, it contains the fine semantic annotation information needed in our experiments. Second, Cityscapes captured street scenes with a low-end sensor (1/3 in CMOS 2MP), having the prevailing low-quality imaging samples. We collect all the samples with fine annotations from the low dynamic range (LDR) training set. We further adjust the brightness of these images to augment more severe no normal-light imaging conditions.

Feature Preservation

Many low-level visual tasks (Ledig et al. 2017; Kupyn et al. 2018; Johnson, Alahi, and Fei-Fei 2016) use the perceptual loss to make desired images and their ground truth perceptually consistent. To be distinct from the typical usage of the perceptual loss, we reformulate it as a feature retention loss to preserve the image features before and after enhancement. The feature retention loss L_{fr} is defined as:

$$L_{fr} = \frac{1}{C_l W_l H_l} (f^l(I_L) - f^l(I_H))^2 \tag{8}$$

where $f^l(I_L)$ denotes the feature map $f \in \mathbb{R}^{C \times H \times W}$ of the input image I_L in the layer l, and $f^l(L_H)$ is the feature map of the enhanced image I_H in the layer l.

Since the color naturalness is one of the significant concerns of LLE, we add a color constancy term L_{cc} incorporating with the feature retention term, following the way reported in (Guo et al. 2020). It is based on the gray-world color constancy hypothesis (Buchsbaum 1980) that the pixel averages of the three channels tend to be of the same value. L_{cc} constrains the ratio of three channels to prevent potential color deviations in the enhanced image. In addition, to avoid aggressive and sharp changes between neighboring pixels, an illumination smoothness penalty term is also embedded in L_{cc}. The formulation of L_{cc} can be expressed as:

$$L_{cc} = \sum_{\forall (p,q) \in \xi} (J^p - J^q)^2$$
$$+ \lambda \frac{1}{M} \sum_{m=1}^{M} \sum_{p \in \xi} (|\nabla_x A^p_m| + |\nabla_y A^p_m|), \xi = \{R, G, B\} \tag{9}$$

where J^p denotes the average intensity value of p channel in the enhanced image, (p, q) represents a pair of channels, M is the number of the iterations, and ∇_x and ∇_y denote the horizontal and vertical gradient operations, respectively. The curve parameter map A (Guo et al. 2020) is the output at each iteration. We set λ to 200 in our experiments for the best outcome. The feature preservation loss L_{fp} is the sum of L_{fr} and L_{cc}.

Efficient Training Details

In our implementation, the semantic segmentation network is pre-trained on the Cityscapes dataset, and the feature extraction network is pre-trained on ImageNet (Russakovsky et al. 2015). We train SCL-LLE end-to-end while fixing the weights of the semantic segmentation network and the feature extraction network. The back-propagated operation

| (a) Input | (b) LIME | (c) RetinexNet | (d) ISSR |
| (e) Zero-DCE | (f) EnlightenGAN | (g) RUAS | (h) Ours |

Figure 3: Comparison of SCL-LLE and the state-of-the-art methods over VV and DICM datasets with zoom-in regions. Our method enables the enhanced images to look more realistic, and recovers better details in both foreground and background.

updates only the weights in the image enhancement network. Hence, most network computation is done in the image enhancement network, which efficiently learns I_H from $(I_L, I_N, I_P, y, \tilde{y})$ to recover the enhanced image with a wide variety of scenes and semantics. We resize the training images to the size of 384×384. As for the numerical parameters, we set the maximum epoch as 50 and the batch size as 2. The model is optimized using the Adam optimizer with a fixed learning rate of $1e^{-4}$.

Experiments

Cross-dataset Peer Comparison

For testing images, we use six publicly available independent cross-scenes low-light image datasets from other reported works, i.e., DICM (Lee, Lee, and Kim 2012), MEF (Ma, Zeng, and Wang 2015), LIME (Guo, Li, and Ling 2016), NPE (Wang et al. 2013), VV* and the Part2 of SICE (Cai, Gu, and Zhang 2018)).

We compare the proposed method with six representative heterogeneous state-of-the-art methods, including a latest conventional method LIME (Guo, Li, and Ling 2016); a GAN-based method EnlightenGAN (Jiang et al. 2021); three Retinex-based methods RetinexNet (Wei et al. 2018), RUAS (Liu et al. 2021), and ISSR (Fan et al. 2020), where ISSR also leverages semantic knowledge; and Zero-DCE (Guo et al. 2020) which leverages the same backbone enhancement network with ours.

---

*https://sites.google.com/site/vonikakis/datasets

Visual quality comparison. We first examine whether the proposed methods can achieve visually pleasing results in brightness, color, contrast, and naturalness. We observe from Fig. 3 that all the SOTAs sacrifice over/under/uneven exposure in global or local areas. Specifically, LIME leads to color artifacts in strong local edges; RetinexNet and EnlightenGAN cause global color distortions with details missing; ISSR and RUAS generate severe global and local over/underexposure. In contrast, our method recovers more details and better contrast in both foreground and background, thus enabling the enhanced images to look more realistic with vivid and natural color mapping.

No-referenced image quality assessment. We adopt Natural Image Quality Evaluator (NIQE) (Mittal, Soundararajan, and Bovik 2013) and UNIQUE (Zhang et al. 2021) as no-reference image quality assessment to evaluate image restoration. Smaller NIQE and larger UNIQUE indicate more naturalistic and perceptually favored quality. The NIQE and UNIQUE results on five datasets are reported in Table 1. For the NIQE and UNIQUE, SCL-LLE wins on two of the five datasets and has the best overall averaged results.

Full-referenced image quality assessment. For full-reference image quality assessment, we employ the Peak Signal-to-Noise Ratio (PSNR,dB) and Structural Similarity (SSIM) metrics to compare the performance of different methods quantitatively. Since the five datasets used in the previous test contain no standard images, we use the Part2 of the SICE dataset (Cai, Gu, and Zhang 2018) here. As shown in Table 2, our method obtained the best PSNR and SSIM

Methods	DICM		LIME		MEF		VV		NPE		Average	
	NIQE	UN.	NIQE	UN.	NIQE	UN.	NIQE	UN.	NIQE	UN.	NIQE	UN.
Input	4.26	0.72	4.36	0.70	4.26	0.72	3.52	**0.74**	4.32	**1.17**	4.13	0.75
(TIP'17) LIME	3.75	0.78	3.85	0.53	3.65	0.65	**2.54**	0.44	4.44	0.93	3.55	0.69
(BMVC'18) R.-Net	4.47	0.75	4.60	0.52	4.41	0.97	2.70	0.36	4.60	0.81	4.13	0.69
(ACMMM'20) ISSR	4.14	0.59	4.17	**0.83**	4.22	0.87	3.57	0.62	4.02	0.99	4.03	0.68
(CVPR'20) Z.-DCE	3.56	0.82	3.77	0.73	3.28	1.22	3.21	0.48	3.93	1.07	3.50	0.81
(TIP'21) E.GAN	3.55	0.63	**3.70**	0.49	**3.16**	1.03	3.25	0.58	3.95	1.07	3.47	0.69
(CVPR'21) RUAS	5.21	-0.17	4.26	0.34	3.83	0.73	4.29	-0.04	5.53	0.13	4.78	0.04
Ours	**3.51**	**0.87**	3.78	0.76	3.31	**1.25**	3.16	0.49	**3.88**	1.08	**3.46**	**0.85**

Table 1: NIQE ↓ and UNIQUE (UN.) ↑ scores on DICM, LIME, MEF, VV, and NPE datasets.

Methods	LIME	R.Net	ISSR	Z.-DCE	E.GAN	RUAS	Ours
PSNR ↑	13.67	16.98	15.01	14.78	17.82	10.62	**17.95**
SSIM ↑	0.62	0.66	0.65	0.62	0.66	0.44	**0.68**

Table 2: PSNR and SSIM on the Part2 of the SICE dataset.

Methods	DICM		LIME		MEF		VV		NPE		Average	
	NIQE	UN.	NIQE	UN.	NIQE	UN.	NIQE	UN.	NIQE	UN.	NIQE	UN.
Input	4.26	0.72	4.36	0.70	4.26	0.72	3.52	**0.74**	4.32	**1.17**	4.13	0.75
w/o L_c	4.31	0.64	4.36	0.57	4.25	0.56	4.10	0.70	4.28	1.02	4.27	0.66
w/o L_{sc}	3.53	0.83	3.85	**0.76**	3.32	1.18	3.21	0.50	3.98	1.02	3.49	0.82
w/o L_{fr}	3.54	0.80	3.88	0.71	3.32	1.22	3.18	0.47	3.97	1.03	3.50	0.80
w/o Neg. samples	3.55	0.81	3.84	0.72	3.36	1.14	**3.14**	0.38	3.95	1.01	3.49	0.78
w/o overexp. Neg.	3.59	0.75	3.91	0.59	3.36	1.24	3.15	0.37	4.12	0.87	3.54	0.74
w/o underexp. Neg.	4.58	0.57	4.52	0.48	4.69	0.46	3.58	0.65	4.36	0.86	4.38	0.58
Ours	**3.51**	**0.87**	**3.78**	**0.76**	**3.31**	**1.25**	3.16	0.49	**3.88**	1.08	**3.46**	**0.85**

Table 3: Ablation study. NIQE ↓ and UNIQUE (UN.) ↑ scores on the testing sets.

among all the peer methods.

Human subjective survey. We conduct a human subjective survey for comparisons. For each image in the five test datasets enhanced by seven methods , we ask 11 human subjects to rank the enhanced images. These subjects are instructed to consider: 1) Whether or not the images contain visible noise; 2) Whether the images have over or underexposure artifacts; 3) Whether the images show non-realistic color or texture distortion. We stipulate the score of each image from 1 to 5, and the lower the value is, the better the image quality will be. The final results are shown in Fig. 4 , and our method is recognized as the best imaging quality.

Ablation Study

Contribution of each loss. We perform ablation studies to demonstrate the effectiveness of each loss component. Note that since the color consistency item L_{cc} is initially proposed and tested in Zero-DCE (Guo et al. 2020), we consider it as a baseline item without doing ablation study. Thus the feature preservation loss L_{fp} is tested using the first item L_{fr}. The visualized samples with their corresponding histograms of the effects of L_c, L_{sc} and L_{fr} functions are shown in Fig. 5 (a-e). Table 3 shows the average NIQE and UNIQUE scores of each loss on five test sets. The contrastive learning loss L_c plays a significant role in controlling the exposure level. The results without the semantic brightness consistency constraint loss L_{sc} and without the feature retention
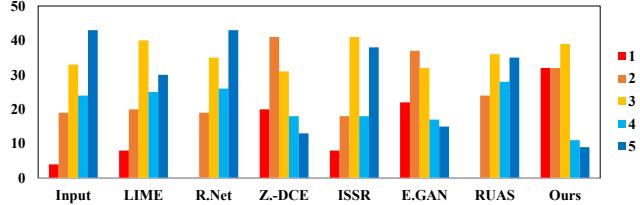

Figure 4: The result in the human subjective survey. The color-changing from hot to cool means the quality transition from the best to worst; y-axis denotes the number of images in each ranking index.

loss L_{fr} have relatively lower contrast than the final result.

Impact of negative samples. To verify the rationality of the negative sample selection in our experiment, we retrain SCL-LLE with: 1) 360 positive samples; 2) 360 positive and 360 underexposed negative samples; 3) 360 positive and 360 overexposed negative samples; and 4) 360 positive and 360 underexposed and 360 overexposed negative samples. As shown in Fig. 1, Fig. 5 (g) and Table 3, after removing the negative samples, SCL-LLE tends to use a biased white balance (color temperature) to enhance the well-lit regions. Meanwhile, as shown in Fig. 5 (h-i), removing the overexposed or underexposed negative samples leads to the white balance offset in the opposite direction. The above degraded

(a) Input (b) w/o L_c (c) w/o L_{sc} (d) w/o L_{fr} (e) Ours

(f) Input (g) w/o neg. samples (h) w/o overexposed (i) w/o underexposed (j) Ours
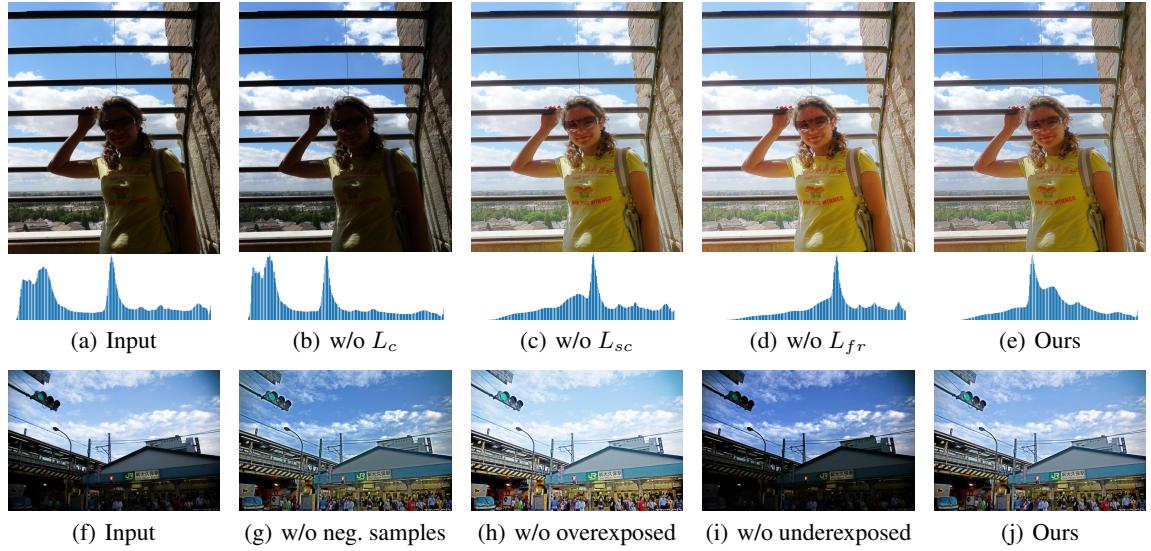
Figure 5: Ablation study on the contribution of each loss and the impact of negative samples.

Methods	Frankfurt				Lindau				Munster			
	γ=1	γ=2	γ=6	γ=10	γ=1	γ=2	γ=6	γ=10	γ=1	γ=2	γ=6	γ=10
Input	**73.831**	**72.453**	49.879	33.577	**62.295**	**61.498**	30.536	17.894	**71.839**	70.118	45.519	24.658
LIME	69.507	70.361	50.547	36.928	54.708	59.259	34.639	16.877	68.888	69.191	53.132	29.482
R.-Net	58.586	56.471	45.620	29.463	43.938	44.943	24.045	13.725	59.888	56.118	45.559	22.785
ISSR	68.310	70.297	44.433	30.866	56.137	59.919	25.414	16.973	68.638	69.907	39.667	23.176
Z.-DCE	64.297	69.127	58.046	41.385	49.190	56.599	41.522	22.151	63.403	70.153	55.300	33.109
E.GAN	70.669	70.624	50.757	34.435	57.242	59.117	32.599	15.982	68.671	66.948	48.965	29.182
RUAS	41.374	66.277	51.890	35.535	41.482	55.621	36.501	19.713	53.152	69.300	52.015	29.763
Ours	67.437	70.100	**58.488**	**42.047**	51.909	60.200	**41.558**	**22.853**	65.698	**70.710**	**56.379**	**33.442**

Table 4: The segmentation results (mIoU ↑) of the input low-light image after enhancement on Cityscapes.

results indicate the rationality and necessity of the usage of negative samples in our training framework.

Semantic Segmentation with LLE

Since there is no semantic annotation in current low-light image datasets, we use subsets Frankfurt, Lindau, and Munster in the validation set of Cityscapes to test the semantic segmentation performance before and after enhancement. In addition, we use the standard positive gamma transformation with a series of gamma values to simulate images with lower brightness. As shown in Table 4, when using original input (γ=1), the semantic segmentation with all the enhancement models could not surpass the initial input. The reason may be in two aspects: 1) the current LLE methods, including ours, pay more attention to preserving visual-pleasing results, and more or less destroy the topological information on local areas; and 2) the normal-light scenes dominate the Cityscapes dataset. When γ becomes larger, mIoU of segmentation after using the LLE methods has been significantly better than those using the original image. Among all the methods, the segmentation performance with our method tends to be the best when the scene tends to be dark. These results encourage us to bridge the gap between the current LLE methods and the downstream tasks.

Conclusion

We have proposed an effective semantically contrastive learning paradigm (SCL-LLE) to solve the LLE problem. Based on a feature extraction network and a semantic segmentation network, we cast image enhancement as multi-task joint learning, where SCL-LLE is converted into three constraints of contrastive learning, semantic brightness consistency, and feature preservation for simultaneously ensuring the color, texture, and exposure consistency. Experiments exhibit clear improvements of our method over existing state-of-the-arts LLE models on six cross-domain datasets. Meanwhile, experiments also reveal that our SCL-LLE has the potential to guide the downstream semantic segmentation task to gain better performance.

Acknowledgments

References

Buchsbaum, G. 1980. A spatial processor model for object colour perception. *Journal of the Franklin institute*.

Cai, J.; Gu, S.; and Zhang, L. 2018. Learning a Deep Single Image Contrast Enhancer from Multi-Exposure Images. *IEEE Transactions on Image Processing*, 2049–2062.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European Conference on Computer Vision*, 801–818.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607.

Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3223.

Fan, M.; Wang, W.; Yang, W.; and Liu, J. 2020. Integrating Semantic Segmentation and Retinex Model for Low-Light Image Enhancement. In *ACM International Conference on Multimedia, Virtual Event*, 2317–2325.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1780–1789.

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

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 9729–9738.

Henaff, O. 2020. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, 4182–4192.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In Defense of the Triplet Loss for Person Re-Identification. *arXiv preprint arXiv:1703.07737*.

Ignatov, A.; Kobyshev, N.; Timofte, R.; Vanhoey, K.; and Van Gool, L. 2017. DSLR-quality photos on mobile devices with deep convolutional networks. In *IEEE International Conference on Computer Vision*, 3277–3285.

Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; and Wang, Z. 2021. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 2340–2349.

Jobson; Daniel; J.; Rahman; Zia-ur; Woodell; Glenn; and A. 1997. A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image Processing*.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 694–711.

Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; and Matas, J. 2018. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 8183–8192.

Land, E. H. 1977. The retinex theory of color vision. *Scientific american*, 108–129.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4681–4690.

Lee, C.; Lee, C.; and Kim, C.-S. 2012. Contrast enhancement based on layered difference representation. In *IEEE International Conference on Image Processing*, 965–968.

Li, W.; Yang, X.; Kong, M.; Wang, L.; Huo, J.; Gao, Y.; and Luo, J. 2021. Triplet is All You Need with Random Mappings for Unsupervised Visual Representation Learning. *arXiv:2107.10419*.

Liu, R.; Ma, L.; Zhang, J.; Fan, X.; and Luo, Z. 2021. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition*, 10561–10570.

Lore, K. G.; Akintayo, A.; and Sarkar, S. 2017. LLNet: A deep autoencoder approach to natural low-light image enhancement. *Pattern Recognition*, 650–662.

Ma, K.; Zeng, K.; and Wang, Z. 2015. Perceptual quality assessment for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 3345–3356.

Mittal, A.; Soundararajan, R.; and Bovik, A. 2013. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters*, 209–212.

Pizer, S.; Johnston, R.; Ericksen, J. P.; Yankaskas, B.; and Muller, K. 1990. Contrast-limited adaptive histogram equalization: speed and effectiveness. *Conference on Visualization in Biomedical Computing*, 337–345.

Ren, W.; Liu, S.; Ma, L.; Xu, Q.; Xu, X.; Cao, X.; Du, J.; and Yang, M.-H. 2019. Low-Light Image Enhancement via a Deep Hybrid Network. *IEEE Transactions on Image Processing*, 4364–4375.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 211–252.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 815–823.

Sermanet, P.; Lynch, C.; Chebotar, Y.; Hsu, J.; Jang, E.; Schaal, S.; Levine, S.; and Brain, G. 2018. Time-contrastive networks: Self-supervised learning from video. In *IEEE International Conference on Robotics and Automation*, 1134–1141.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive multiview coding. *arXiv:1906.05849*.

Wang, S.; Zheng, J.; Hu, H.-M.; and Li, B. 2013. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, 3538–3548.

Wei, C.; Wang, W.; Yang, W.; and Liu, J. 2018. Deep Retinex Decomposition for Low-Light Enhancement. In *British Machine Vision Conference*.

Wu, H.; Qu, Y.; Lin, S.; Zhou, J.; Qiao, R.; Zhang, Z.; Xie, Y.; and Ma, L. 2021. Contrastive Learning for Compact Single Image Dehazing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 10551–10560.

Xie, J.; Bian, H.; Wu, Y.; Zhao, Y.; Shan, L.; and Hao, S. 2020. Semantically-guided low-light image enhancement. *Pattern Recognition Letters*, 308–314.

Xu, K.; Yang, X.; Yin, B.; and Lau, R. W. 2020. Learning to Restore Low-Light Images via Decomposition-and-Enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Xu, Q.; Jiang, H.; Scopigno, R.; and Sbert, M. 2014. A novel approach for enhancing very dark image sequences. *Signal processing*, 309–330.

Zhang, W.; Ma, K.; Zhai, G.; and Yang, X. 2021. Uncertainty-Aware Blind Image Quality Assessment in the Laboratory and Wild. *IEEE Transactions on Image Processing*, 3474–3486.

Zhang, Y.; Zhang, J.; and Guo, X. 2019. Kindling the Darkness: A Practical Low-light Image Enhancer. In *ACM International Conference on Multimedia*, 1632–1640.