

# Improving Visual and Downstream Performance of Low-Light Enhancer with Vision Foundation Models Collaboration

Yuxuan Gu<sup>1</sup>, Haoxuan Wang<sup>1</sup>, Pengyang Ling<sup>1,2</sup>, Zhixiang Wei<sup>1</sup>,  
 Huaian Chen<sup>1†</sup>, Yi Jin<sup>1†</sup>, and Enhong Chen<sup>1</sup>

<sup>1</sup>University of Science and Technology of China

<sup>2</sup>Shanghai AI Laboratory

## Abstract

In this paper, we observe that the **collaboration** of various foundation models can perceive semantic and degraded information within images, thereby guiding the low-light enhancement process. Specifically, we propose a self-supervised low-light enhancement framework based on the multiple foundation models **collaboration** (dubbed **FoCo**), aimed at improving both the visual quality of enhanced images and the performance in high-level applications. **At the feature level**, FoCo leverages the rich features from various foundation models to enhance the model’s semantic perception during training, thereby reducing the gap between enhanced results and high-quality images from a high-level perspective. **At the task level**, we exploit the robustness-gap between strong foundation models and weak models, applying high-level task guidance to the low-light enhancement training process. Through the collaboration of multiple foundation models, the proposed framework **shows better enhancement performance and adapts better to high-level tasks**. Extensive experiments across various enhancement and application benchmarks demonstrate the qualitative and quantitative superiority of the proposed method over numerous state-of-the-art techniques.

## 1. Introduction

Images taken under short exposure or low-light conditions frequently suffer from significant degradation, including loss of color information and the introduction of substantial noise. Such degradation severely impacts the performance of downstream tasks [19, 36]. Low-light image enhancement (LIE) aims to restore input images captured in low-light environments to high-quality, well-lit images. With the advancement of deep learning technology in computer vision, recent research [2, 25, 31, 33, 34, 39] has achieved low-light image enhancement through supervised training on paired datasets. However, collecting

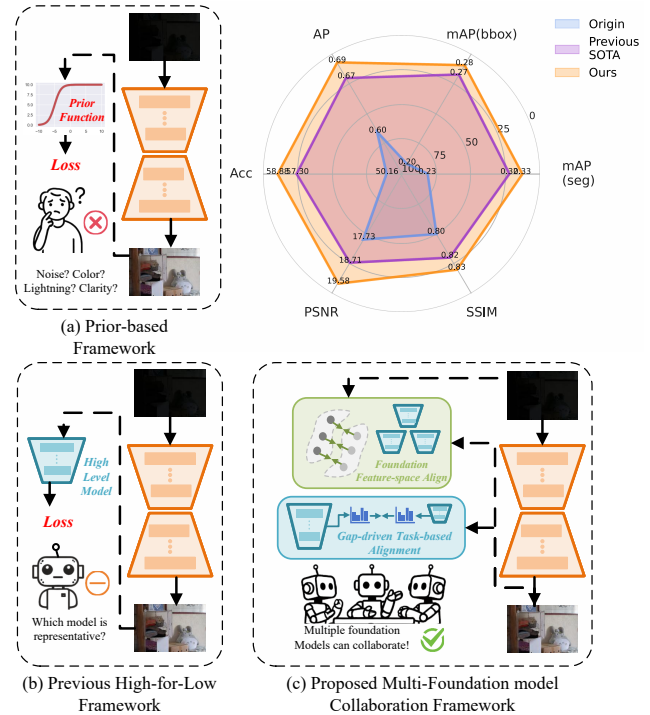


Figure 1. The proposed approach demonstrates substantial improvements over existing unsupervised methods in terms of enhancement tasks (PSNR, SSIM) and application tasks (classification/Acc, face detection/AP, and instance segmentation/mAP), attributed to the proposed Multi-VFM Collaboration Framework. Compared with previous unsupervised LIE learning paradigms (a) and (b), the proposed framework (c) can make full use of the foundation models from a collaborative perspective to improve visual quality and downstream task performance.

the substantial amount of paired images required for these methods is both time-consuming and costly, which highlights the broad potential applications of self-supervised approaches [6, 7, 13, 18]. As shown in Fig. 1 (a), early self-supervised training paradigms mainly relied on manually defined priors, such as smoothness priors, brightness priors, or histogram equalization. The drawback of such prior-based frameworks is their lack of connection with semantic

<sup>†</sup> Corresponding authors.

information, and the efficacy of these prior functions can diminish under complex natural conditions. For instance, the smoothness prior may sometimes degrade semantic information, limiting these methods’ performance in practical applications and impacting the visual quality of results.

Recognizing these limitations, as illustrated in Figure 1 (b), some approaches [15, 16, 27, 29, 37] have introduced high-level models into the training of Low-light Image Enhancement (LIE) models, significantly improving the model’s ability to retain semantic information. However, this line of research typically relies on a specific high-level model and does not fully explore the potential of different vision models, resulting in only limited visual improvements. In recent years, the emergence of various foundation vision models widely applied in real-world scenarios has led us to consider integrating multiple foundational models from a collaborative perspective, which aims to improve the visual quality and practical performance of LIE methods more effectively.

In this work, we explore the use of multiple foundation models to strengthen the self-supervised enhancement process, as shown in Figure 1 (c). *To this end, We address two key challenges: (1) how to effectively utilize pre-trained knowledge? (2) how to handle differences between models?* To deal with these challenges, we propose **FoCo**, a Low-light Image Enhancement (LIE) framework based on the multiple foundation models ensemble, aligning the enhancement results with high-quality images from both feature-level and task-level perspectives. As the core strategies to implement FoCo, we introduce two optimization strategies: **Foundation Feature-space Alignment (FoFA)** and **Gap-driven task-based Alignment (GaTA)**. FoFA focuses on creating a unified feature space for foundation models, using adversarial learning to align enhanced results with high-quality image sets. GaTA leverages robustness differences across models to model degradation scenarios, using robust pseudo-labels to directly incorporate high-level task training into the enhancement model’s optimization process.

In summary, the contributions of this paper are as follows:

- We are the first to explore the collaborative effects of multiple vision foundation models on the LIE task, and then propose **FoCo**, a framework designed to enhance both the human- and machine-understanding performance of unsupervised LIE methods by leveraging complementary interactions among multiple foundational vision models.
- We introduce two key strategies **FoFA** and **GaTA** to fully exploit the pre-trained knowledge of foundation models, which align enhancement results and high-quality images from the perspective of foundation models.
- We conduct extensive experiments on benchmarks for low-light enhancement, object detection, instance seg-

mentation, and classification. The experimental results show significant performance improvements, demonstrating the proposed method’s effectiveness in enhancing visual quality and application performance.

## 2. Related works

### 2.1. Low-light Image Enhancement

Supervised methods, such as RetinexNet [25], KinD [33], and more recent approaches like Retinexformer [2], Diff-Retinex [31], and GLARE [39], utilize paired datasets and advanced model architectures to enhance image quality. While successful, these methods face challenges due to their dependence on large-scale paired datasets, which are difficult to obtain in real-world scenarios. Unsupervised methods, such as EnLightenGAN [11], and CoLIE [4], aim to overcome this limitation by eliminating the need for paired data, leveraging adversarial learning and multimodal approaches. However, these methods often neglect integration with high-level models, limiting their ability to preserve semantic information and adapt to downstream tasks.

### 2.2. High-Level guided Enhancement

Recent works [15, 16, 27, 29, 37] have increasingly focused on leveraging high-level models to supervise low-light enhancement models, aiming to improve the fidelity of the enhancement results. For example, SGZ [37] incorporates the concept of focal loss to maximize the logits of a pre-trained semantic segmentation model, enhancing the model’s ability to preserve semantic information. SKF [27] integrates features from semantic segmentation models into the enhancement model and proposes a semantic-based histogram equalization loss and adversarial loss based on segmentation results. Methods like NeRCO [29] and CLIP-LIT [16] employ pre-trained CLIP models to align enhancement results with high-quality images during adversarial training. The methods above have significant success, but a common limitation of these approaches is their reliance on a specific high-level model, and they often do not fully exploit the model’s potential.

## 3. Methods

### 3.1. FoCo Framework

To implement the low-light enhancement process, we first define a simple enhancer  $G$  for enhancing a low-light input image  $I$ , and then use some priors to constrain the behavior of  $G$ . By incorporating the smoothness and consistency losses  $\mathcal{L}_C$  borrowed from [21] and [6], we can construct a baseline low-light enhancement model that fully relies on simply defined hand-crafted prior. However, this dependency on these hand-crafted prior compromises

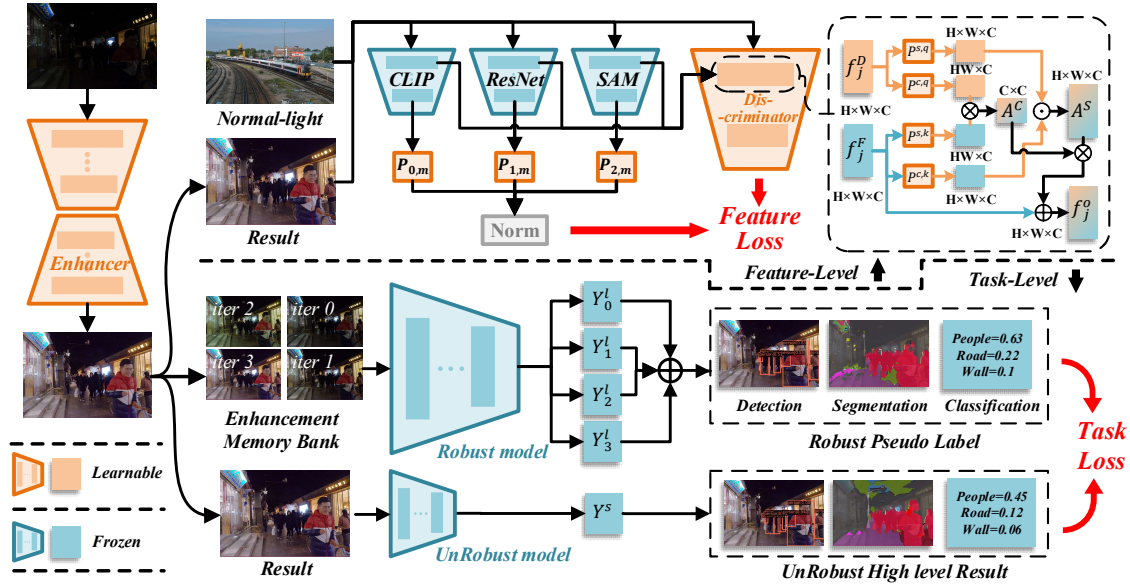


Figure 2. Overview of the proposed low-light enhancement method, which enhances a basic LIE model using multi-foundation model ensemble. Given the set of pre-trained foundation models, the enhancement model first performs basic enhancement, followed by supervision from the ensemble through two strategies: (1) **Feature-Level**: adversarial training aligns enhanced results with high-quality images in the unified feature space, and (2) **Task-Level**: robustness differences between models are used to align the enhancement results with high-quality images on high-level tasks.

the model’s visual quality and adaptability to downstream tasks. Mathematically, such a straightforward image-level enhancement approach can be formalized as:

$$\arg \min_{\theta} d(G(I), \hat{I}), \quad (1)$$

where  $\hat{I}$  represents the ideal normal-light image,  $\theta$  denotes the parameters of the enhancement model,  $d(\cdot)$  is the distance measurement function. To address the limitations of the prior-based approach, we introduce FoCo framework designed to enhance this model from the perspective of multiple **foundation** models **collaboration**. Specifically, we propose (1) Foundation Feature-space Alignment (FoFA), at the feature level (indicated in green), and (2) Gap-driven Task-based Alignment (GaTA), at the task level (indicated in blue). These strategies leverage the feature extraction process (denoted as  $F^f$ ) and task decision process (denoted as  $F^t$ ) of foundation models. Thus, FoCo’s optimization objective can be expressed as:

$$\arg \min_{\theta} \underbrace{d(G(I), \hat{I})}_{\mathcal{L}_C} + \underbrace{d(F^f(G(I)), F^f(\hat{I}))}_{\text{FoFA}} + \underbrace{d(F^t(G(I)), F^t(\hat{I}))}_{\text{GaTA}}. \quad (2)$$

The corresponding framework is illustrated in Figure 2.

### 3.2. Foundation Feature-space Alignment

**Observation.** Figure 3 (a) displays a GAN-enhanced low-light image from the LIS [3] dataset, alongside its Mask-RCNN detect result. The GAN was trained in an unpaired

manner on the LIS dataset. It can be observed that the GAN-enhanced result suffers from significant semantic degradation, likely due to the discriminator’s limited semantic awareness, which fails to guide the generator effectively. This observation leads us to a potential solution: could we enhance the discriminator by incorporating a feature extractor with strong semantic awareness? As shown in Figure 3 (b), we use CLIP, SAM and ResNet separately to extract feature maps from both images. Incorporating these base models into the discriminator training process, as shown in Figure 3 (a), led to significant improvements in detection performance. However, SAM produced darker results, CLIP introduced color bias, and ResNet maintained limited semantic integrity. We hypothesize that these differences stem from the varying focus of each foundational model: CLIP prioritizes the primary objects in the image and is less sensitive to color bias and subtle degradations, SAM captures fine-grained semantics but lacks sensitivity to lighting conditions, while ResNet offers weaker semantic understanding compared to the other models but is more susceptible to degradations and lighting variations, which still contributes to the discriminator’s learning. Interestingly, ResNet’s lack of robustness to degradations may become a key factor in complementing CLIP and SAM, while the discriminator needs to identify degradation and illumination. As shown in Figure 3 (c), where we observed that ResNet’s features were more sensitive to lighting variations compared to the other models.

This inspires us that CLIP, SAM, and ResNet may complement each other at the feature level: CLIP highlights key objects, SAM provides fine-grained guidance, and

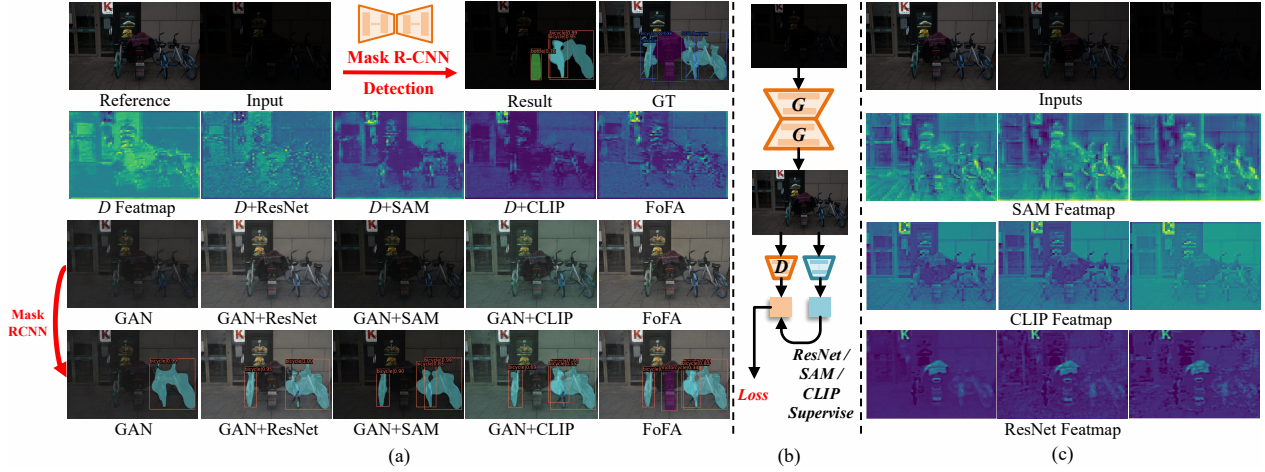


Figure 3. The perception capabilities of foundation models on unsupervised low-light enhancer training. (a) presents the discriminator feature maps, enhanced results, and instance segmentation outcomes using different foundation model-supervised discriminators. (b) illustrates the process of adversarial training supervised by the foundation model. (c) compares the robustness of CLIP and ResNet feature maps under varying lighting conditions. The results suggest that foundation models may collaborate at feature level.

ResNet adds degradation-related information to guide semantic learning in the discriminator. However, the feature spaces of CLIP, SAM, and ResNet are unaligned, preventing the discriminator from directly utilizing their outputs. The challenge now is how to effectively integrate CLIP, SAM, and ResNet into the discriminator training process.

**Method.** Based on this observation, Foundation Feature-space Alignment (FoFA) aims to leverage features from different foundation models to bring the enhanced results closer to a high-quality image set  $\hat{I}$ . Let  $F_{i,j}$  denote the feature extractor of the  $i$ -th pre-trained foundation models, extending up to the  $j$ -th layer, given the varied structures and training data of these models, our goal is to construct a unified feature space  $\hat{F}_j$  and find semantic and degradation information in it that is beneficial to the enhancement process. To this end, we use a series of convolutional projectors  $P_{i,j}$  to map the features from different spaces into a unified feature space. To align the numerical scale of these features with the subsequent modules, we normalize the output features, defined as:

$$\hat{F}_{i,j}^f(x) = \text{Norm}(P_{i,j}(F_{i,j}^f(x))), s.t. x \sim p(x), \quad (3)$$

$$\hat{F}_j^f(x) = \left\{ \hat{F}_{i,j}^f(x) \right\}_{i=0}^n, s.t. x \sim p(x), \quad (4)$$

where  $n$  denotes the count number of foundation models,  $P$  is the convolutional projectors and  $F_{i,j}^f$  denotes feature extraction with foundation models. Then we adopt an adversarial learning approach by introducing a discriminator  $D$ , which links the features from the foundation models with the enhancement model. The input to  $D$  consists of images along with their corresponding semantic information derived from the unified feature space. Specifically, inspired by [27], we employ cross-attention across both spatial and channel dimensions between the image features and the semantic features to generate attention maps  $A^c$  and  $A^s$ :

$$A^c = \text{Softmax} \left( P^{c,k} (f_j^F) \times P^{c,q} (f_j^D) / \sqrt{C} \right), \quad (5)$$

$$A^s = \text{Softmax} \left( P^{s,k} (f_j^F) \cdot P^{s,q} (f_j^D) / \sqrt{H \cdot W} \right), \quad (6)$$

where  $f_j^F = \hat{F}_j^f(x)$  and  $f_j^D$  is the feature map of  $j$ -th layer in  $D$ . These attention maps are then utilized to adjust the original image features to the output feature  $f_j^O$  as follows:

$$f_j^O = A^c \times A^s + f_j^D. \quad (7)$$

The optimization objective of  $D$  is to align its learned features with the outputs of the foundation model while simultaneously distinguishing between enhanced *low*-quality images and *high*-quality images. Throughout this process, it is crucial to ensure that the features input to the classification layer of  $D$  are closely related to the semantic and degradation information. Based on the observation, the loss function for the discriminator  $D$  can be expressed as follows:

$$\begin{aligned} \mathcal{L}_D = & -\mathbb{E}_{x \sim p_{high}(x)} [\log D(x) + \|D^f(x) - \hat{F}_m^f(x)\|] \\ & -\mathbb{E}_{x \sim p_{low}(x)} [\log(1 - D(G(x))) + \\ & \|D^f(G(x)) - \hat{F}_m^f(G(x))\|], \end{aligned} \quad (8)$$

$$\mathcal{L}_G = -\mathbb{E}_{x \sim p_{low}(x)} [\log D(G(x))], \quad (9)$$

where  $m$  denotes the layer count of the foundation model,  $D^f$  denotes feature extraction with discriminator. The discriminator  $D$  effectively projects the foundation models' features into a unified feature space, providing valuable guidance for the enhancement model. Unlike traditional image-level adversarial training, FoFA focuses on selected

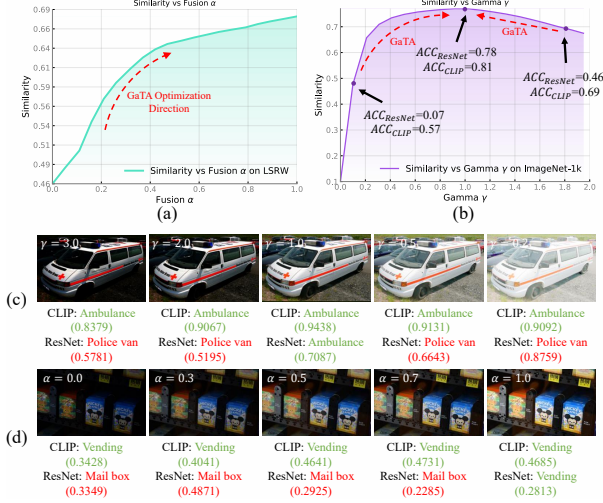


Figure 4. Robustness to illumination across models. (a) and (b) show the similarity of results between CLIP and ResNet under different lighting adjustments. (c) and (d) present specific image examples, along with the classification results and confidence scores. The goal of GaTA is to optimize the augmented results towards a direction that enhances correct recognition.

semantic features, with the goal of aligning enhanced results with the high-quality image set.

The strength of FoFA lies not only in its ability to preserve semantics but also in its optimization objective, which is designed to enhance compatibility with downstream tasks. In FoFA, the discriminator  $D$  is specifically designed to inherit the strong semantic perception capabilities of the multi-foundation models. *From another perspective, if the semantic information can no longer distinguish the enhanced result from the high-quality image, the high-level model’s perception of the enhanced result will be indistinguishable from that of the high-quality image.* Therefore, for downstream tasks, FoFA effectively transforms degraded images into the high-level models’ familiar domain.

### 3.3. Gap-driven Task-based Alignment

**Observation.** FoFA improves the similarity between augmented results and high-quality images at the feature level; however, a constraint is still needed to assess whether the augmented results are easy to recognize. In Section 3.2, we observed that ResNet is more sensitive to illumination compared to CLIP. This raises the question: can we also observe robustness differences between these models at the task level? To explore this, we generated images under different lighting conditions using two methods: (1) fusing low-light and normal-light image pairs from the LSRW dataset (Get the pixel value according to the fusion ratio  $\alpha$ ) (2) applying gamma transformations to images from ImageNet-1k. As shown in Figure 4, low-light conditions degrade the performance of both CLIP and ResNet, but ResNet is more significantly affected. Interestingly, their predictions are most similar under normal conditions, suggesting that CLIP’s ro-

business stems from its training on large-scale data, while ResNet is more sensitive to illumination and degradation. This motivates us to leverage these robustness differences to guide the enhancement model to produce clean and well-lit results, aiming at “detecting” the degradation and brightness, which can be seen as obstacles in recognition.

**Method.** *Gap-driven Task-based Alignment (GaTA) leverages the robustness differences between models to evaluate whether the augmented results are sufficiently good to be easily recognized.* From another view, GaTA employ the less robust model as a “degradation detector” while the robust model functions as a ground truth provider. We found that tasks such as segmentation, classification, and detection can be applied to GaTA, with CLIP and ResNet serving as representative robust and degradation-sensitive models, respectively. Let  $F^l$  represent the robust model and  $F^s$  represent the degradation-sensitive model. Following the observation and robustness study in [22], we can apply CLIP and ResNet for  $F^l$  and  $F^s$ , respectively. For each enhanced image, we first use the robust model to generate pseudo-labels and obtain the logits from the sensitive model:

$$Y^l = F^l(G(x)), s.t. x \sim p_{low}(x), \quad (10)$$

$$Y^s = F^s(G(x)), s.t. x \sim p_{low}(x). \quad (11)$$

A key issue to consider is that pseudo-labels are not always accurate. In extreme cases, if the generation quality of  $G(x)$  is very poor, even robust models may fail to provide fully accurate labels. Fortunately, during iterative training, we can collect multiple enhanced results for each training image, constructing a memory bank of size  $T$ . Leveraging this memory bank, we can apply a voting mechanism to correct the pseudo labels:

$$\hat{Y}^l = \arg \max_c \sum_{t=1}^T \mathbb{I} \left[ \arg \max_{c'} (p(Y_t^l = c')) = c \right], \quad (12)$$

where  $c'$  denotes the predicted class for each individual image while  $c$  represents the final pseudo-label. Finally, we apply a high-level task loss to supervise GaTA:

$$\mathcal{L}_{GaTA} = \mathcal{L}_h(Y^s, \hat{Y}^l), \quad (13)$$

where  $\mathcal{L}_h$  denotes the specific task loss for  $Y^s$  and  $\hat{Y}^l$ . Please note that both  $F_t^l$  and  $F_t^s$  are frozen during GaTA training. By exploiting the robustness differences between models in a straightforward manner, GaTA effectively tests for degradation at the task level, ensuring the enhanced results exhibit both visual quality and strong performance in downstream applications.

## 4. Experiments

### 4.1. Implementation Details

The proposed method is implemented using PyTorch, running on an NVIDIA TITAN RTX GPU alongside an In-

Table 1. Benchmarking results for low-light enhancement. Among unsupervised methods, we highlight the top-ranking scores in red and the second in blue.

| Datasets | Metrics | Supervised                   |                        |                     |                        | Unsupervised         |                     |                      |                        |                         |                     |                      |       |
|----------|---------|------------------------------|------------------------|---------------------|------------------------|----------------------|---------------------|----------------------|------------------------|-------------------------|---------------------|----------------------|-------|
|          |         | Retinexformer [2]<br>ICCV'23 | LLflow [24]<br>AAAI'22 | SKF [27]<br>CVPR'23 | Kind++ [34]<br>IJCV'21 | EnGAN [11]<br>TIP'21 | SCI [21]<br>CVPR'22 | RUAS [18]<br>CVPR'21 | ZeroDCE [6]<br>CVPR'20 | SCL-LLE [15]<br>AAAI'22 | SGZ [37]<br>WACV'22 | CoLIE [4]<br>ECCV'24 | Ours  |
| Huawei   | PSNR↑   | 17.98                        | 17.18                  | 19.31               | 16.52                  | 17.83                | 15.44               | 7.41                 | 16.38                  | 13.21                   | 16.91               | 13.96                | 18.61 |
|          | SSIM↑   | 0.61                         | 0.63                   | 0.61                | 0.50                   | 0.59                 | 0.55                | 0.60                 | 0.47                   | 0.49                    | 0.56                | 0.47                 | 0.72  |
|          | LPIPS↓  | 0.32                         | 0.31                   | 0.32                | 0.40                   | 0.35                 | 0.36                | 0.42                 | 0.41                   | 0.37                    | 0.35                | 0.39                 | 0.35  |
| Nikon    | PSNR↑   | 16.87                        | 18.06                  | 9.35                | 15.09                  | 15.86                | 17.52               | 7.75                 | 16.68                  | 15.11                   | 17.37               | 14.92                | 17.76 |
|          | SSIM↑   | 0.63                         | 0.54                   | 0.34                | 0.34                   | 0.58                 | 0.55                | 0.59                 | 0.58                   | 0.59                    | 0.61                | 0.55                 | 0.67  |
|          | LPIPS↓  | 0.33                         | 0.37                   | 0.53                | 0.53                   | 0.37                 | 0.38                | 0.46                 | 0.39                   | 0.38                    | 0.38                | 0.39                 | 0.37  |
| FiveK    | PSNR↑   | 23.28                        | 16.32                  | 16.59               | 14.11                  | 12.83                | 16.29               | 9.52                 | 13.50                  | 15.70                   | 10.62               | 18.71                | 19.58 |
|          | SSIM↑   | 0.84                         | 0.77                   | 0.77                | 0.70                   | 0.78                 | 0.83                | 0.66                 | 0.73                   | 0.77                    | 0.43                | 0.82                 | 0.83  |
|          | LPIPS↓  | 0.24                         | 0.19                   | 0.20                | 0.20                   | 0.14                 | 0.14                | 0.30                 | 0.23                   | 0.21                    | 0.33                | 0.16                 | 0.13  |

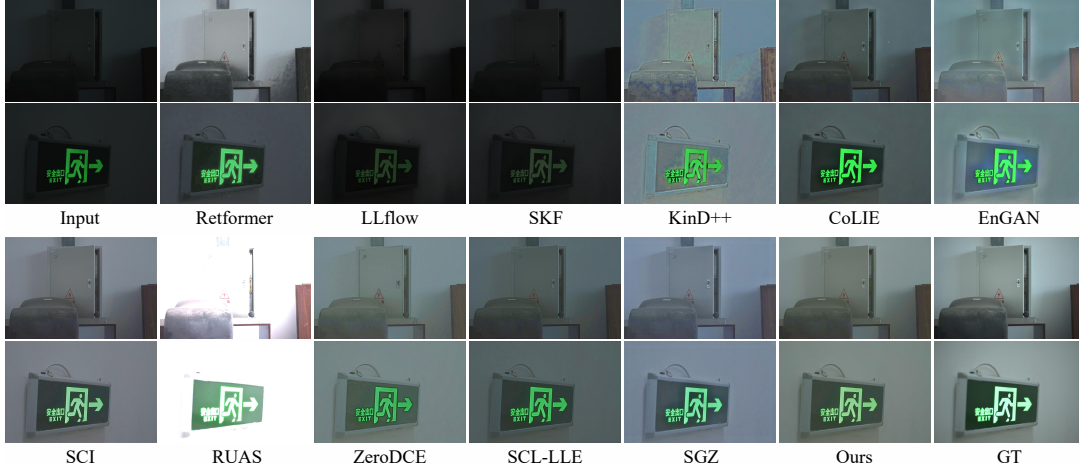


Figure 5. Visual Comparisons on Low-Light Enhancement Datasets.

tel(R) Xeon(R) Gold 6252 CPU @ 2.10GHz. Training is conducted with the Adam optimizer, using a learning rate of  $1e-4$ . All images are resized to  $448 \times 448$  pixels, and the model is trained with a batch size of 32 over 50 epochs. For ResNet, we utilize the pre-trained weights obtained from training on ImageNet, following the origin setting in [9]. For CLIP and SAM, we employ the official weights trained on their respective datasets. For the training set of high-level applications, we randomly sample low-light images from ExDark [20] and normal-light images from COCO [17] as training set and test the obtained weights on the test sets of all benchmarks. More information can be found in Supplementary.

## 4.2. Benchmarks

The performance of our method is evaluated across four metrics relevant to low-light enhancement: reference-based low-light enhancement, low-light image classification, low-light instance segmentation, and low-light face detection. Specifically, reference-based low-light enhancement is evaluated using the MIT-FiveK [1], Nikon, and Huawei [8] datasets, from which we select 500, 945, and 735 paired low-light and normal-light samples; Low-light image classification uses the CODaN [12] dataset, a 10-class dataset with a training set of 10,000 daytime images and a test set comprising 2,500 daytime and nighttime images. Low-light instance segmentation is performed using

the LIS [3] dataset with 2230 images. For low-light face detection, we use the darkface [30] dataset, containing 6,000 images captured under low-light conditions, from which we select 1800 images for testing.

To provide a comprehensive comparison, our method is benchmarked against 4 state-of-the-art supervised learning methods, including Retinexformer [2], LLFlow [24], SKF [27], and Kind++ [34], as well as 7 unsupervised learning methods, such as Zero-DCE[6], RUAS [18], EnlightGAN [11], SGZ [37], SCL-LLE [15] CoLIE [4] and SCI [21] across all tasks.

## 4.3. Enhancement Tasks

To validate the effectiveness of our method in terms of generalization performance, we evaluated the proposed framework on several benchmark datasets, i.e. MIT-FiveK [1], Nikon and Huawei [8] datasets, using PSNR, SSIM, and LPIPS as performance metrics. The measurement results are presented in Table 1. It is evident that our proposed method demonstrates advantages compared to other self-supervised approaches, exhibiting competitive performance in terms of PSNR, SSIM, and LPIPS.

Figure 5 illustrates a visual comparison between our method and other approaches. While alternative methods may struggle with incomplete degradation removal or loss of semantic information, our method consistently demonstrates its ability to effectively recover image details. This

Table 2. Performance comparison on different high-level benchmarks: CODaN for classification, Darkface for face detection, and LIS for instance segmentation. Highlighted in red are the top scores, and in blue are the second best methods.

| Datasets | Detectors    | Metrics            | Raw   | Supervised           |                   |                |                   | Unsupervised    |                |                 |                    |                    |                |                  |       |
|----------|--------------|--------------------|-------|----------------------|-------------------|----------------|-------------------|-----------------|----------------|-----------------|--------------------|--------------------|----------------|------------------|-------|
|          |              |                    |       | Retformer<br>ICCV'23 | LLflow<br>AAAI'22 | SKF<br>CVPR'23 | Kind++<br>IJCV'21 | EnGAN<br>TIP'21 | SCI<br>CVPR'22 | RUAS<br>CVPR'21 | ZeroDCE<br>CVPR'20 | SCL-LLE<br>AAAI'22 | SGZ<br>WACV'22 | CoLIE<br>ECCV'24 | Ours  |
| CoDAN    | ResNet18     | Acc                | 0.502 | 0.551                | 0.533             | 0.525          | 0.539             | 0.531           | 0.565          | 0.556           | 0.565              | 0.572              | 0.573          | 0.564            | 0.588 |
|          | EfficientNet | Acc                | 0.587 | 0.614                | 0.653             | 0.616          | 0.559             | 0.630           | 0.573          | 0.535           | 0.605              | 0.676              | 0.591          | 0.652            | 0.742 |
| Darkface | S3FD         | AP                 | 0.603 | 0.653                | 0.636             | 0.621          | 0.646             | 0.638           | 0.661          | 0.641           | 0.648              | 0.669              | 0.665          | 0.656            | 0.683 |
|          | DSFD         | AP                 | 0.274 | 0.337                | 0.348             | 0.351          | 0.367             | 0.331           | 0.349          | 0.326           | 0.367              | 0.334              | 0.358          | 0.334            | 0.381 |
| LIS      | MaskRCNN     | mAP <sup>box</sup> | 0.228 | 0.303                | 0.286             | 0.258          | 0.309             | 0.286           | 0.301          | 0.277           | 0.317              | 0.309              | 0.317          | 0.309            | 0.331 |
|          |              | mAP <sup>seg</sup> | 0.198 | 0.252                | 0.254             | 0.237          | 0.258             | 0.244           | 0.263          | 0.238           | 0.273              | 0.265              | 0.272          | 0.258            | 0.281 |

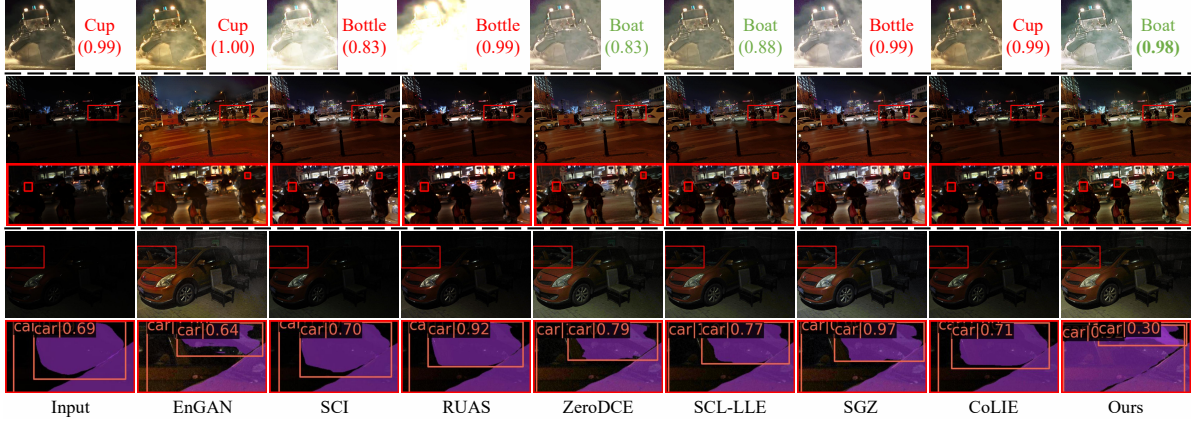


Figure 6. Visual Comparisons and High-level Results on CODaN, Darkface, and LIS Dataset. The first row shows CODaN and the classification results by ResNet (color denotes incorrect/correct), the second row shows the Darkface detection results by S3FD, and the third row shows the LIS and detection results and Mask-RCNN.

further substantiates the advanced capabilities of the proposed framework.

#### 4.4. High-level Tasks

**Low-Light Classification.** We evaluate the adaptability of various methods for low-light image classification tasks on the CODaN [12] dataset. For our experiments, we employ the official ResNet18 model, which was trained on the daytime portion of the dataset, as our primary detector. Additionally, to mitigate the impact of domain differences, we incorporate an EfficientNet [23] model trained on ImageNet, targeting classes corresponding to those in the CODaN dataset, as a supplementary detector. Using these two detectors, we assess Top-1 Accuracy as our evaluation metric. As shown in Table 2, our proposed method outperforms other low-light enhancement techniques, demonstrating exceptional performance. This superiority can be attributed to the more consistent illumination provided by the models enhanced with our framework, while the degraded results from other methods hinder classifier performance.

**Low-Light Face Detection.** We evaluate the performance of two detectors, S3FD [32] and DSFD [14], trained on the WiderFace [28] dataset, and on the DarkFace [30] dataset to assess the performance of various enhancement methods for low-light face detection. It is noteworthy that face detection is more granular and sensitive to semantic information compared to classification tasks; thus, local blurriness can significantly impair performance. Con-

sequently, the methods like EnGAN [11] and Retinexformer [25] demonstrate a marked decline relative to other tasks, whereas our proposed method consistently achieves excellent results across both detectors. This further underscores the advanced nature of our proposed framework.

**Low-Light Instance Segmentation.** We evaluated various low-light enhancement methods for their adaptability to low-light instance segmentation using the LIS [3] dataset. As one of the most fine-grained high-level tasks, instance segmentation imposes significant demands on enhancement techniques. We conducted experiments using Mask R-CNN [10] as provided by the official LIS framework, which was initially trained on the COCO dataset and subsequently tested on low-light images from the LIS dataset. Our results demonstrate that the proposed methods achieve superior performance, further confirming that the framework effectively preserves semantic information and enhances the quality of the results.

The above three tasks encompass three datasets, five detectors, and six evaluation metrics, providing a comprehensive assessment of the proposed method’s robustness across different low-light scenarios and its applicability to high-level tasks.

#### 4.5. Ablation Studies

We conduct extensive ablation studies to validate the efficacy of our proposed perspective and evaluate the contribution of each model component. All eval-

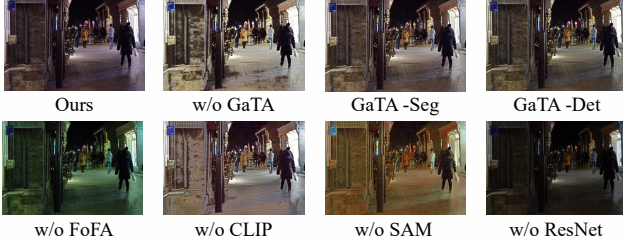


Figure 7. Visual Comparisons of ablation studies.

uations are carried out on FiveK for the enhancement task and on three high-level task datasets for other tasks. We select ResNet18/Top-1 Acc, S3FD/AP, and MaskRCNN/mAP segm as the high-level metrics.

**Ablation Studies on FoFA.** We employed a combination of different models to examine the supervisory capabilities of various foundation models at the feature level in the enhancement process. As shown in Table 3, removing any single foundation model from the FoFA negatively impacts the performance of the enhancement model across various tasks, and removing all of them results in even poorer outcomes. Overall, excluding CLIP and SAM reduces the model’s adaptability for high-level tasks, while the absence of ResNet leads to accuracy losses in image enhancement tasks. This is likely because, during FoFA training, CLIP and SAM primarily contribute robust semantic learning, which is critical for high-level tasks; in contrast, ResNet is more responsive to specific degradations and lighting variations, which helps the discriminator perceive visual quality in images. Our observations further validate this theory: as shown in Figure 7, removing ResNet results in significant lighting changes in the enhanced images, while omitting CLIP or SAM introduces blurred and distorted areas. If all foundation models are excluded, FoFA degrades to a conventional adversarial training process, where semantic losses in the enhancement results become more severe. In summary, each foundation model in FoFA contributes positively to the training of the enhancement model, and their integration yields the best enhancement outcomes.

Table 3. Quantitative analysis from ablation studies demonstrating the impact of foundation feature space alignment. The best results are marked in bold.

|          | PSNR         | SSIM        | LPIPS       | AP          | mAP~segm    | Top-1 Acc   |
|----------|--------------|-------------|-------------|-------------|-------------|-------------|
| w FoFA   | <b>19.58</b> | <b>0.83</b> | <b>0.13</b> | <b>0.68</b> | <b>0.28</b> | <b>0.59</b> |
| w/o FoFA | 18.07        | 0.81        | 0.20        | 0.64        | 0.24        | 0.54        |
| -CLIP    | 18.67        | 0.81        | 0.15        | 0.65        | 0.23        | 0.53        |
| -SAM     | 18.95        | 0.82        | 0.17        | 0.64        | 0.26        | 0.54        |
| -ResNet  | 18.24        | 0.81        | 0.19        | 0.67        | 0.27        | 0.57        |

**Ablation Studies on GaTA.** To assess the role of GaTA in training the enhancement model, we first examined the impact of removing GaTA, as shown in Table 4. Removing GaTA negatively affected model performance in both the enhancement and high-level tasks. This is because GaTA primarily functions to detect incomplete degradation

elimination in the enhanced results, thereby improving the model’s performance across various tasks. As illustrated in Figure 7, removing GaTA significantly increased the severity of degradation in the enhanced results, which not only impaired visual quality but also compromised the original semantic information.

Table 4. Quantitative analysis from ablation studies demonstrating the Gap-driven Task-based Alignment. The best results are marked in bold.

|                  | PSNR         | SSIM        | LPIPS       | AP          | mAP~segm    | Top-1 Acc   |
|------------------|--------------|-------------|-------------|-------------|-------------|-------------|
| w GaTA           | <b>19.58</b> | <b>0.83</b> | <b>0.13</b> | 0.68        | 0.28        | <b>0.59</b> |
| w/o GaTA         | 18.01        | 0.81        | 0.16        | 0.67        | 0.25        | 0.52        |
| +GaTA <i>Seg</i> | 18.50        | 0.79        | 0.19        | 0.67        | <b>0.30</b> | 0.49        |
| +GaTA <i>Det</i> | 18.75        | 0.82        | 0.19        | <b>0.70</b> | 0.27        | 0.54        |

Additionally, in Section 3.3, we implement GaTA using a classification task, with CLIP and ResNet employed to achieve  $F^l$  and  $F^s$ , respectively. Since GaTA leverages robustness differences between models to perceive degradation, various tasks—such as object detection or semantic segmentation—could potentially be used for GaTA as well. To test this, we trained  $F^l$  on a dataset with gamma transformations and noise augmentations, and  $F^s$  on the high-quality subset of the dataset, filtered using advanced image quality assessment metric Q-Instruct [26]. In practice, we used YOLOX [5], trained on COCO, and PSPNet [35], trained on ADE20k [38], to represent models for object detection and semantic segmentation, denote as *Det* and *Seg* respectively. Detailed experimental settings are provided in the supplementary materials. As shown in Table 4, both object detection and semantic segmentation tasks effectively implement GaTA, achieving performance comparable to ResNet and CLIP, further demonstrating the generalizability of the proposed method.

## 5. Conclusion

We propose a framework named FoCo designed to enhance the visual and applicational performance of low-light image enhancers by leveraging the collaborative potential of diverse foundation models. At the feature level, our approach uses Feature Semantic Alignment (FoFA) to integrate semantically relevant features into adversarial training, aligning the discriminator and the foundation model’s output features within a unified feature space. This alignment promotes feature-level consistency between enhanced and high-quality images. At the task level, we introduce Gap-driven task-based Alignment (GaTA), which utilizes the robustness disparities across models by using the less robust model as a degradation detector, thereby further improving enhancement quality. The proposed framework exhibits outstanding performance, providing a novel training paradigm that integrates self-supervised low-light enhancement methods with high-level applications.

**Acknowledgements:** This research was funded through Anhui Provincial Natural Science Foundation (Grant 2408085J024), National Natural Science Foundation of China (Grant 62401532), and Fundamental Research Funds for the Central Universities (Grant WK20900000065).

## 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. 6
- [2] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. *arXiv preprint arXiv:2303.06705*, 2023. 1, 2, 6
- [3] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 3, 6, 7
- [4] Tomáš Chobola, Yu Liu, Hanyi Zhang, Julia A Schnabel, and Tingying Peng. Fast context-based low-light image enhancement via neural implicit representations. In *European Conference on Computer Vision*, pages 413–430. Springer, 2025. 2, 6
- [5] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 8
- [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. 1, 2, 6
- [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. 1
- [8] 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. *Journal of Visual Communication and Image Representation*, 90: 103712, 2023. 6
- [9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 7
- [11] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlighten: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021. 2, 6, 7
- [12] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4399–4409, 2021. 6, 7
- [13] Chongyi Li, Chunle Guo, and Chen Change Loy. Learning to enhance low-light image via zero-reference deep curve estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4225–4238, 2021. 1
- [14] 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. 7
- [15] Dong Liang, Ling Li, Mingqiang Wei, Shuo Yang, Liyan Zhang, Wenhan Yang, Yun Du, and Huiyu Zhou. Semantically contrastive learning for low-light image enhancement. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1555–1563, 2022. 2, 6
- [16] Zhixin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, and Chen Change Loy. Iterative prompt learning for unsupervised backlit image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8094–8103, 2023. 2
- [17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6
- [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. 1, 6
- [19] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016. 1
- [20] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. 6
- [21] 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. 2, 6
- [22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5
- [23] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 7
- [24] Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with

- normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2604–2612, 2022. 6
- [25] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018. 1, 2, 7
- [26] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25490–25500, 2024. 8
- [27] Yuhui Wu, Chen Pan, Guoqing Wang, Yang Yang, Jiwei Wei, Chongyi Li, and Heng Tao Shen. Learning semantic-aware knowledge guidance for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1662–1671, 2023. 2, 4, 6
- [28] 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. 7
- [29] Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, and Jian Zhang. Implicit neural representation for cooperative low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12918–12927, 2023. 2
- [30] 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. 6, 7
- [31] Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, and Jiayi Ma. Diff-retinex: Rethinking low-light image enhancement with a generative diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12302–12311, 2023. 1, 2
- [32] Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017. 7
- [33] 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. 1, 2
- [34] Yonghua Zhang, Xiaojie Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *International Journal of Computer Vision*, 129:1013–1037, 2021. 1, 6
- [35] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 8
- [36] Chuanjun Zheng, Daming Shi, and Wentian Shi. Adaptive unfolding total variation network for low-light image enhancement. pages 4439–4448, 2021. 1
- [37] Shen Zheng and Gaurav Gupta. Semantic-guided zero-shot learning for low-light image/video enhancement. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, pages 581–590, 2022. 2, 6
- [38] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 8
- [39] Han Zhou, Wei Dong, Xiaohong Liu, Shuaicheng Liu, Xiongkuo Min, Guangtao Zhai, and Jun Chen. Glare: Low light image enhancement via generative latent feature based codebook retrieval. *arXiv preprint arXiv:2407.12431*, 2024. 1, 2