Distilling Semantic Priors from SAM to Efficient Image Restoration Models

Quan Zhang[1,3,*] Xiaoyu Liu[2,3,*] Wei Li[3] Hanting Chen[3] Junchao Liu[3] Jie Hu[3]
Zhiwei Xiong[2] Chun Yuan[1,†] Yunhe Wang[3,†]

[1]Tsinghua Shenzhen International Graduate School
[2]University of Science and Technology of China [3]Huawei Noah's Ark Lab

Abstract

In image restoration (IR), leveraging semantic priors from segmentation models has been a common approach to improve performance. The recent segment anything model (SAM) has emerged as a powerful tool for extracting advanced semantic priors to enhance IR tasks. However, the computational cost of SAM is prohibitive for IR, compared to existing smaller IR models. The incorporation of SAM for extracting semantic priors considerably hampers the model inference efficiency. To address this issue, we propose a general framework to distill SAM's semantic knowledge to boost exiting IR models without interfering with their inference process. Specifically, our proposed framework consists of the semantic priors fusion (SPF) scheme and the semantic priors distillation (SPD) scheme. SPF fuses two kinds of information between the restored image predicted by the original IR model and the semantic mask predicted by SAM for the refined restored image. SPD leverages a self-distillation manner to distill the fused semantic priors to boost the performance of original IR models. Additionally, we design a semantic-guided relation (SGR) module for SPD, which ensures semantic feature representation space consistency to fully distill the priors. We demonstrate the effectiveness of our framework across multiple IR models and tasks, including deraining, deblurring, and denoising.

1. Introduction

Image restoration (IR) [7, 12, 20] is an essential computer vision task that reconstructs high-quality (HQ) images from degraded low-quality (LQ) inputs. These inputs are impaired by distortions like noise [21, 61], blurring [4, 24, 59], and rain drops [9, 45]. To address this, IR methods incorporate explicit image priors and models of the distortion process. These constraints help narrow the solution space for feasible reconstruction. With the advent of deep learning [3, 57, 61, 64], data-driven techniques have achieved superior performance in IR tasks by learning strong statistical regularities. Previous studies [28, 47, 50, 51] have demonstrated the potential benefits of utilizing semantic priors obtained from segmentation for image restoration tasks. These priors contain valuable information about the texture and color characteristics of individual objects within an image. By incorporating these priors, a deeper understanding of the image content can be achieved, providing explicit instructions that guide the restoration process. This integration of semantic priors enhances the restoration performance by leveraging the rich knowledge encoded within the segmentation results.

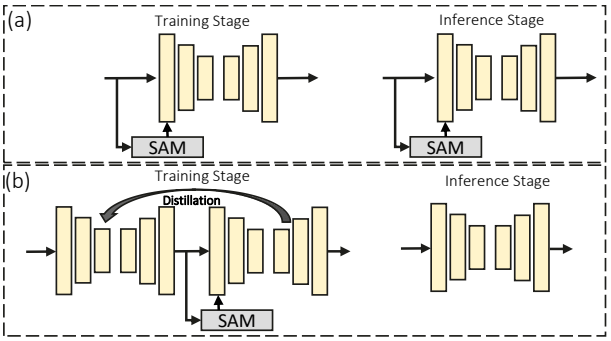

Figure 1. Comparison of training and inference pipelines between different manners of exploiting semantic priors from SAM. (a) Existing methods require the use of SAM at both the training and inference stages. (b) Our method only uses SAM at the training stage and preserves the same inference efficiency as the original image restoration model at the inference stage.

In recent years, the emergence of the Segment Anything Model (SAM) [19] has had a profound impact on various computer vision tasks [1, 8, 53, 55], including image restoration [17, 26, 34, 52]. By scaling up datasets and model capacity, this large foundation model exhibits capabilities beyond surface-level labeling and provides useful cues unavailable to mainstream networks. Specifically, SAM can extract advanced semantic priors through a holistic understanding of image content. However, the compu-

*Both authors contributed equally to this research, which was done during Quan Zhang and Xiaoyu Liu's internship at Huawei Noah's Ark Lab.
†Corresponding authors: yunhe.wang@huawei.com

tational cost associated with SAM poses a significant challenge when applied to IR, particularly when compared to existing smaller IR models. Incorporating SAM for extracting semantic priors during the inference stage considerably hampers the overall efficiency of the model. As shown in Fig. 1 (a), Existing methods require the use of SAM at both the training and inference stages. The SAM module is employed during training to learn spatial attention for capturing semantic priors. However, during inference, the SAM module is also utilized, which can potentially introduce additional computational overhead.

To address this issue, we propose a general framework that distills SAM's semantic knowledge to enhance existing IR models without interfering with their inference process. As shown in Fig. 1(b), the objective of our framework is to leverage the benefits of SAM's semantic priors while mitigating the computational burden and preserving the efficiency of the IR models. Our proposed framework consists of two key schemes: the semantic priors Fusion (SPF) scheme and the semantic priors Distillation (SPD) scheme with a semantic-guided relation (SGR) module. The SPF scheme focuses on fusing information from two sources: the restored image predicted by the original IR model and the semantic mask predicted by SAM. The SPD scheme aims to distill the fused semantic priors to boost the performance of the original IR model. In the SPF scheme, we combine these two sources as inputs of the IR network cascaded behind the original IR network, to refine the restored image and improve its quality. This fusion process enables the incorporation of SAM's semantic knowledge into the IR model without sacrificing efficiency. In the SPD scheme, we leverage the knowledge distillation manner [30, 32, 37] to distill the fused semantic priors obtained through the SPF scheme by enforcing the consistency between the original restored image and the refined original restored image. This scheme aims to enhance the performance of the original IR model by transferring and consolidating the valuable insights extracted by SAM. Additionally, we design a semantic-guided relation (SGR) module for SPD, which ensures consistency in the semantic feature representation space. This further enhances the distilled priors and promotes their effectiveness in improving the IR model's performance. Finally, we only utilize the distilled IR model to restore the degraded LQ inputs without segmentation masks from SAM.

To validate the effectiveness of our proposed framework, we conduct extensive experiments across multiple IR models and tasks, including deraining, deblurring, and denoising. The results demonstrate the potential of our approach in leveraging SAM's semantic knowledge to enhance the performance of existing IR models while addressing the computational challenges associated with SAM integration.

Overall, our contributions can be summarized as follows:

- We propose a general framework to distill semantic knowledge from SAM to boost existing IR models without interfering with their inference process.
- We propose an SPF scheme to fuse information between restored images from the IR model and semantic masks from SAM to refine the restoration.
- We propose an SPD scheme that uses a self-distillation manner with a designed semantic-guided relation module to transfer semantic priors from SPF into the original IR models.
- The effectiveness of our framework is demonstrated through experiments on various IR models and tasks.

2. Related Work

Semantic Priors for Image Restoration. Existing methods can deal with the degraded images by low-level and high-level vision interaction, which are categorized into two types [51]: loss-level methods and feature-level methods. Loss-level methods [29, 46, 66] focus on incorporating semantic priors by utilizing semantic-aware losses as additional objective functions during the training process of original vision tasks. However, these methods exploit semantic priors in an implicit manner, lacking sufficient interaction between semantic priors and IR tasks. Feature-level methods [28, 39, 42, 47] integrate semantic priors into the feature representation space by extracting intermediate features from semantic segmentation networks and combining them with image features. While these methods explicitly exploit semantic priors to significantly improve the performance of IR tasks, these methods modify the inference way of original IR models and still rely on the input of semantic inputs during the inference stage. Although some recent works have attempted to combine both loss-level and feature-level methods [51], existing methods still do not adequately exploit the semantic priors of the Segment Anything Model (SAM) to achieve sufficient interaction and inference efficiency. Our framework ensures effective interaction between semantic priors and IR tasks without interfering with the inference process of the IR models.

SAM VS. Other Segmentation Models. SAM is a foundation model for the image segmentation task with zero-shot generalization capacity, which can be used to solve a range of downstream segmentation problems on new data distributions using prompt engineering. Although existing segmentation models have achieved excellent performance in various segmentation tasks such as semantic segmentation [15, 33, 65], instance segmentation [5, 11, 31], panoptic segmentation [18, 54], and unified segmentation [25], these models heavily rely on annotated segmentation masks for training. However, in many image restoration tasks, such segmentation annotations are not available. This is where SAM stands out with its zero-shot generalization capability. SAM can effectively handle image restoration tasks

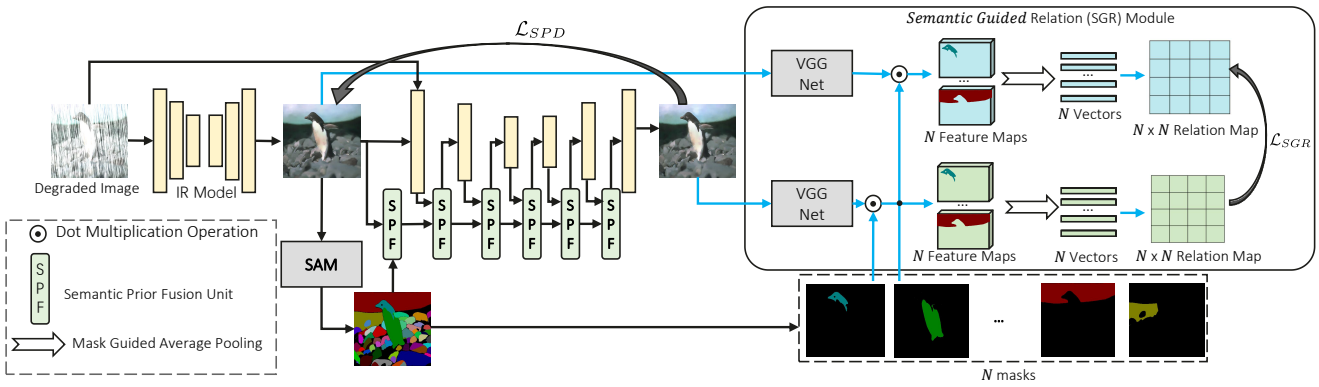

Figure 2. The workflow of our proposed framework to distill semantic knowledge from SAM to boost existing IR models without interfering with their inference process.

without the need for specific segmentation annotations. In addition, SAM provides more fine-grained semantic information compared to instance and category labels through segmentation at different granularity. This finer level of semantic information is crucial for image restoration tasks, as it enables the utilization of richer semantic priors in the restoration process.

3. Method

In this section, we introduce our proposed framework, as illustrated in Fig 2. This framework comprises two essential schemes: the semantic priors Fusion (SPF) scheme (detailed in Sec. 3.2) and the semantic priors Distillation (SPD) scheme with a semantic-guided relation (SGR) module ((detailed in Sec. 3.3)). The SPF scheme focuses on fusing information from two distinct sources: the restored image predicted by the original IR model and the semantic mask predicted by SAM. On the other hand, the SPD scheme aims to distill the fused semantic priors to enhance the performance of the original IR model.

3.1. Problem Analysis and Definition

Semantic priors offer invaluable guidance for restoring colors, contrasts, and texture consistency in degraded images. As a large-scale foundation model, SAM contains extensive parameters trained on diverse distributions of image data. This broad exposure equips SAMs with rich semantic knowledge to inform restoration processes. By leveraging SAMs' understanding of semantic concepts and contexts, we can provide restoration models with informative cues to improve fidelity. Moreover, the scale and generalizability of the SAM allow for capturing higher-order semantics beyond the scope of restoration datasets. These holistic scene-level cues can further boost coherence in restored images.

Given a degraded low-quality image $I_{LQ} \in \mathbb{R}^{3\times H\times W}$, we can obtain its segmentation mask by utilizing SAM's automatic mode:

$$M_{LQ} = f_{SAM}(I_{LQ}), \tag{1}$$

where, M_{LQ} represents the semantic priors extracted from the degraded image I_{LQ} and is the segmentation masks of each objects, and f_{SAM} refers to the SAM model.

The objective of existing solutions [17, 26, 52] is to utilize both the semantic priors M_{LQ} and the degraded input image I_{LQ} to restore a high-quality image I_{HQ}, represented as $(M_{LQ}, I_{LQ}) \rightarrow I_{HQ}$. In our framework, we propose two distinct designs to overcome two drawbacks in the conventional pipeline:

1) Directly combining the semantic priors M_{LQ} from SAM and the input image I_{LQ} at the feature level is impractical since it hampers the inference efficiency of the original IR model and requires the input of SAM's prior during the inference stage. To address this, we cascade two IR models, f^{IR1} and f^{IR2}, together. The second IR model, f^{IR2}, is responsible for fusing the semantic priors, and then the fused model's capabilities are distilled to the first IR model, f^{IR1}. This allows us to utilize f^{IR1} during the inference process, thereby maintaining the efficiency of the original IR model.

2) Since the semantic priors M_{LQ} is extracted from the severely degraded image I_{LQ}, it may contain segmentation errors. To address this, we propose to extract the semantic priors from the restored image obtained from f^{IR1}:

$$I^1_{HQ} = f_{IR1}(I_{LQ}),\ M = f_{SAM}(I^1_{HQ}), \tag{2}$$

where, I^1_{HQ} represents the output of f^{IR1}, and $M \in \mathbb{R}^{N\times H\times W}$ is the segmentation with N mask channels.

3.2. Semantic Priors Fusion

The SPF scheme is used to fuse the semantic priors M and preliminary restored image I^1_{HQ} from the IR model f^{IR1} into the second IR model f^{IR2} for obtaining the enhanced restored image I^2_{HQ}. To provide a concrete example of how the SPF (Semantic Prior Fusion) scheme works, let's consider the feature maps F_i from the i^{th} building block of the image restoration model f^{IR2}. The SPF scheme consists of multiple SPF units, where the number of SPF units matches the number of building blocks in f^{IR2}. Each SPF unit com-

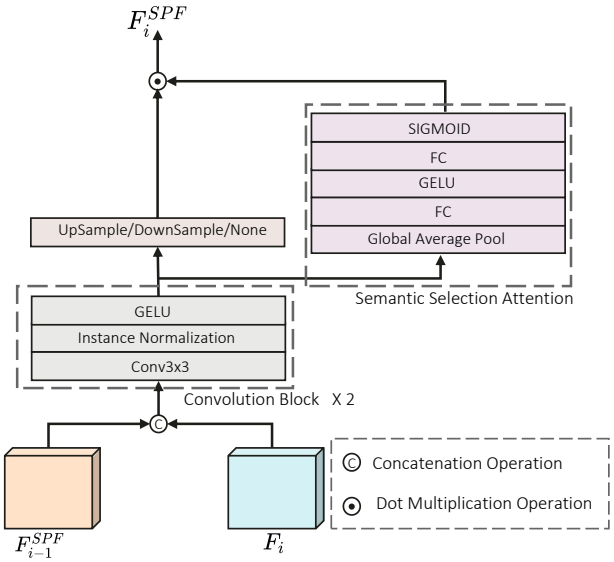

Figure 3. Architecture of the Semantic Prior Fusion (SPF) unit.

bines the semantic prior M extracted from the restored image and the corresponding feature maps F_i.

Without loss of generality, this formulation exemplifies how SPF incorporates semantic guidance into intermediate feature representations within the IR model f^{IR2}:

$$F_{i+1} = f(F_i^{SPF}),$$
$$\begin{cases} F_i^{SPF} = f_i^{SPF}([F_{i-1}^{SPF}, F_i]) \text{ if } i > 1 \\ F_i^{SPF} = f_i^{SPF}([I_{HQ}^1, M]) \text{ if } i = 1 \end{cases}, \quad (3)$$

where $f(\cdot)$ represents the next building block to generate the next feature maps, f_i^{SPF} represents the i^{th} SPF unit in the designed SPF scheme to fuse the feature map F_i and the semantic prior M from SAM. $[\cdot]$ represents the concatenation operation.

The architecture of the proposed SPF unit f_i^{SPF} is detailed in Fig. 3. We first concatenate the input feature maps F_{i-1}^{SPF} and F_i. This combined input is fed into two convolutional blocks to extract an intermediate fused representation. To align with the designs of different IR network building blocks, we introduce resize operations to match dimensions as needed. Additionally, a Semantic Selection Attention module is proposed to focus the SPF fusion on effective semantic priors. It implements a dot product to selectively highlight salient semantics.

By repeating this SPF unit across network stages, semantic guidance is systematically integrated into the IR model.

3.3. Semantic Priors Distillation

During the initial training stage, the output of f^{IR2} is the high-quality restored image I_{HQ}^2, which aims to incorporate semantic priors and has better image quality compared to the restored image I_{HQ}^1 from the output of f^{IR1}. To transfer the capabilities of f^{IR2} to f^{IR1} and improve the performance of f^{IR1} to match or approach the performance of f^{IR2}, we propose a semantic priors distillation scheme. This scheme facilitates the convergence of both networks during training.

In the semantic priors distillation scheme, we leverage the high-quality restored image I_{HQ}^2 obtained from f^{IR2} as a teacher signal to guide the training of f^{IR1}. The key points of the scheme are as follows:

Firstly, we distill the semantic priors from f^{IR2} to f^{IR1} by minimizing the smooth L-1 loss between I_{HQ}^1 and I_{HQ}^2. In image restoration tasks, there may be cases where the restored images contain artifacts. By using the smooth L1 loss, we can reduce the influence of these artifacts and provide more robust training:

$$\mathcal{L}_{SPD} = \begin{cases} \left\| I_{HQ}^1, -I_{HQ}^2 \right\|_1 - 0.5 & \text{if } \left\| I_{HQ}^1, -I_{HQ}^2 \right\|_1 > 1, \\ 0.5 \times \left\| I_{HQ}^1, -I_{HQ}^2 \right\|_1^2 & \text{if } \left\| I_{HQ}^1, -I_{HQ}^2 \right\|_1 < 1, \end{cases} \quad (4)$$

where $\|\cdot\|_1$ represents the L1 loss. Through minimizing the distillation loss, we encourage the semantic priors extracted by f^{IR1} to converge to the semantic priors from f^{IR2} as the two networks are jointly optimized.

Secondly, we introduce a semantic-guided relation (SGR) module to facilitate the transfer of semantic priors from f^{IR2} to f^{IR1} while ensuring consistency between I_{HQ}^1 and I_{HQ}^2 in the semantic feature representation space. To achieve this, we utilize a pre-trained VGG model to extract semantic-aware feature maps $F_{VGG}^1 \in \mathbb{R}^{512 \times \frac{H}{8} \times \frac{W}{8}}$ and $F_{VGG}^2 \in \mathbb{R}^{512 \times \frac{H}{8} \times \frac{W}{8}}$ from I_{HQ}^1 and I_{HQ}^2, respectively. We then employ the semantic priors M to generate N object masks $m^0, m^1, ..., m^N$, which are resized to the dimensions of $\frac{H}{8} \times \frac{W}{8}$. This allows us to obtain the mask-guided semantic features as follows:

$$F_{VGG}^1(n) = F_{VGG}^1 \odot m^n, \; F_{VGG}^2(n) = F_{VGG}^2 \odot m^n, \quad (5)$$

where $\odot$ denotes the dot product. $F_{VGG}^1(n)$ and $F_{VGG}^2(n)$ represent the semantic features guided by the n^{th} mask.

Next, we calculate the semantic relationship knowledge between the mask-guided semantic features and distill this knowledge from f^{IR2} to assist f^{IR1} in obtaining semantic priors for improved image restoration performance. The semantic relationship is formulated as:

$$R_{VGG}^1(n_1, n_2) = \frac{F_{VGG}^1(n_1) F_{VGG}^1(n_2)}{\|F_{VGG}^1(n_1)\|_2 \, \|F_{VGG}^1(n_2)\|_2}, \quad (6)$$

where $\|\cdot\|_2$ represents the L2 loss, and $R_{VGG}^1(n_1, n_2)$ represents the semantic relationship between $F_{VGG}^1(n_1)$ and $F_{VGG}^1(n_2)$. The calculation of $R_{VGG}^2(n_1, n_2)$ follows a similar approach.

To align the semantic relationships obtained from f^{IR2}

and f^{IR1}, we use the following formulation:

$$\mathcal{L}_{SGR} = \sum_{n_1,n_2=1,n_1\neq n_2}^{N} \frac{\left\| R^1_{VGG}(n_1,n_2) - R^2_{VGG}(n_1,n_2) \right\|_2}{N^2 - N}, \tag{7}$$

where F_{SGR} represents the SGR loss calculated using the L2 loss function.

3.4. Overall Optimization

In the training stage, the IR models f^{IR1} and f^{IR2} undergo independent supervision using Groundtruth, with their respective reconstruction loss functions denoted as $\mathcal{L}^1_{recon}$ and $\mathcal{L}^2_{recon}$.

The IR model f^{IR2} is only supervised by $\mathcal{L}^2_{recon}$ and the overall objective function of the IR model f^{IR1}, denoted as $\mathcal{L}$, combines these aforementioned losses as follows:

$$\mathcal{L} = \mathcal{L}^1_{recon} + \lambda_1 \mathcal{L}_{SPD} + \lambda_2 \mathcal{L}_{SGR}, \tag{8}$$

where the weights λ_1 and λ_2 are used to balance the contribution of the corresponding loss terms. Although all loss functions are optimized simultaneously, $\mathcal{L}_{SPD}$ and $\mathcal{L}_{SGR}$ only affect the gradient backpropagation of f^{IR1} and are not propagated through f^{IR2} with the stop-gradient mechanism. The parameters of SAM and VGG models are also frozen during the training stage.

4. Experiments

4.1. Datasets

Deraining Task. We evaluate our proposed framework on multiple publicly available datasets for the rain removal task, including Rain200L/H [56], DID [60], and DDN [6]. The Rain200L/H dataset comprises 1,800 synthetic rainy images for training and 200 images for testing. The DID and DDN datasets consist of 12,000 and 12,600 synthetic images, respectively, with varying rain directions and density levels. Both the DID and DDN datasets provide 1,200 rainy images for testing. The results obtained from these datasets demonstrate the effectiveness of our method in handling diverse types of spatially varying rain streaks, and they indicate a successful reduction of rain artifacts.

Furthermore, we incorporate two additional synthetic deraining datasets, Cityscape-syn100/200 [50], to assess the quality of the restored images in relation to their impact on downstream segmentation tasks. These datasets are synthesized based on the Cityscape dataset [2], allowing us to evaluate image quality using downstream segmentation metrics on the validation set.

Deblurring Task. We conduct an evaluation of our framework on the GoPro dataset [35] for the image deblurring task. The GoPro dataset contains a total of 2,103 training images and 1,111 test images. To generate the blurs of different strengths, a varying number of successive latent frames are averaged together. The images in this dataset are captured using a GoPro camera at a frame rate of 240 fps.

Denoising Task. We conduct an evaluation of our framework on the SenseNoise dataset [63] for the image denoising task. The SenseNoise dataset comprises 500 diverse scenes, each consisting of high-resolution images. The dataset includes both indoor and outdoor scenes, and it provides high-quality ground truth images for reference.

4.2. Evaluation Metrics

We employ two commonly used metrics, Peak Signal-to-Noise Ratio (PSNR) [14] and Structural Similarity Index (SSIM) [48], to evaluate the performance of our image restoration tasks. Additionally, we introduce the Fréchet Inception Distance (FID) [13] as a measure of the subjective visual quality perceived by humans. In the context of the cityscape-syn datasets, we utilize pixel accuracy (PA), intersection over union (IOU), and DICE [41] as segmentation metrics to assess the performance of the downstream segmentation tasks.

4.3. Implementation Details

All the components of our framework are trained simultaneously. We utilize the PyTorch platform within the Python environment and employ NVIDIA Tesla V100 GPUs with 32 GB memory for training. The framework is trained using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We set the learning rate to $1e-4$ and utilize a batch size of 8 for a total of 200 epochs.

4.4. Selected Baseline Methods

In order to demonstrate the effectiveness of our general framework, we conduct experiments using several well-established image restoration models. For the rain removal task, we select two representative models: RCDNet [45] and Efficientderain [9]. For the deblurring and denoising tasks, we chose the widely recognized Uformer model [49] as our representative model. These models are carefully selected to cover a range of restoration tasks and showcase the versatility of our framework.

4.5. Quantitative and Qualitative Results

Deraining Task. In Table 1 and Table 2, we provide the validation results of our framework on the deraining task. The tables illustrate that our framework consistently outperforms the original RCDNet and Efficientderain models, yielding significant performance gains of both objective and subjective visual metrics across multiple datasets. Specifically, our framework achieves an average improvement of 0.91 dB PSNR over the RCDNet model and 1.02 dB PSNR over the Efficientderain model. These improvements are primarily attributed to the introduction of advanced capabilities for suppressing noise and artifacts while preserv-

Method	R200L			R200H			DID			DDN		
	PSNR ↑	SSIM ↑	FID ↓	PSNR ↑	SSIM ↑	FID ↓	PSNR ↑	SSIM ↑	FID ↓	PSNR ↑	SSIM ↑	FID ↓
RESCAN [27]	36.09	0.9697	-	26.75	0.8353	-	33.38	0.9417	-	31.94	0.9345	-
PReNet [38]	37.80	0.9814	-	29.04	0.8991	-	33.17	0.9481	-	32.60	0.9459	-
MSPFN [16]	38.85	0.9827	-	29.36	0.9034	-	33.72	0.9550	-	32.99	0.9333	-
MPRNet [58]	39.47	0.9825	-	30.67	0.9110	-	33.99	0.9590	-	33.10	0.9347	-
RCDNet [45]	39.04	0.9846	5.42	30.27	0.9063	31.75	34.12	0.9561	26.02	33.07	0.9483	20.09
RCDNet+	40.26	0.9874	4.19	30.85	0.9142	30.05	34.67	0.9609	24.00	33.84	0.9547	19.16
	+1.22	+0.0028	-1.23	+0.58	+0.0079	-1.70	+0.55	+0.0048	-2.02	+0.77	+0.0064	-0.93
Efficientderain [9]	34.42	0.9641	14.89	24.20	0.8100	86.23	31.99	0.9120	25.09	31.75	0.9234	24.46
Efficientderain+	35.70	0.9718	9.67	25.31	0.8479	56.49	32.72	0.9181	21.74	32.48	0.9323	20.77
	+1.28	+0.0077	-5.22	+1.11	+0.0379	-29.74	+0.73	+0.0061	-3.35	+0.73	+0.0008	-3.69

Table 1. Quantitative comparison on multiple deraining datasets to evaluate our framework for the draining task. '+' represents the IR models enhanced by our proposed framework.

Figure 4. The qualitative comparison of IR models with and without our framework on various deraining datasets.

ing texture and color consistency. Furthermore, we evaluate the segmentation performance of different restored images using an HRNet [43] model pre-trained on the Cityscape dataset, as shown in Table 2. The results demonstrate that our framework consistently achieves downstream segmentation improvements in terms of IoU, PA, and DICE metrics and significantly better segmentation results compared to the segmentation results of degraded images in 'Rain'.

In Fig. 4, we visualize the deraining results on the benchmark datasets. It can be seen that our framework assists the existing IR models not only in removing rain streaks more effectively but also in the preservation of high-fidelity object boundaries.

We also visually depict the deraining and corresponding segmentation results on the Cityscape-syn datasets in Figure 5. From the deraining perspective, our framework effectively aids the IR models in removing a larger number of rain streaks and restoring the intricate structure and content of rainy images. Additionally, our framework provides enhanced semantic priors to the image restoration models, enabling the generation of restored images with richer semantic information. This, in turn, facilitates better segmentation

Method	Cityscape-syn 100 mm						Cityscape-syn 200 mm					
	PSNR ↑	SSIM ↑	FID ↓	IoU ↑	PA ↑	DICE ↑	PSNR ↑	SSIM ↑	FID ↓	IoU ↑	PA ↑	DICE ↑
No Rain	INF	1	0	0.8020	0.9650	0.9323	INF	1	0	0.8020	.9650	0.9323
Rain	22.06	0.7513	164.66	0.5023	0.7452	0.6138	16.77	0.6076	253.03	0.2128	0.4481	0.3076
RCDNet [45]	33.72	0.9852	9.09	0.7869	0.9618	0.9263	31.85	0.9761	13.69	0.7701	0.9585	0.9204
RCDNet+	34.88	0.9869	8.37	0.7881	0.9620	0.9268	33.04	0.9787	12.90	0.7741	0.9590	0.9214
	+1.16	+0.0017	-0.72	+0.0013	+0.0002	+0.0005	+1.19	+0.0026	-0.79	+0.0040	+0.0005	+0.0010
Efficientderain [9]	35.06	0.9886	8.56	0.7807	0.9610	0.9250	33.40	0.9827	10.71	0.7692	0.9592	0.9216
Efficientderain+	36.29	0.9912	6.17	0.7862	0.9622	0.9272	34.65	0.9858	8.19	0.7770	0.9605	0.9240
	+1.23	+0.0026	-2.39	+0.0065	+0.0012	+0.0022	+1.25	+0.0031	-2.52	+0.0078	+0.0013	+0.0024

Table 2. Quantitative comparison on the synthesized deraining datasets to evaluate our framework for the downstream segmentation task.

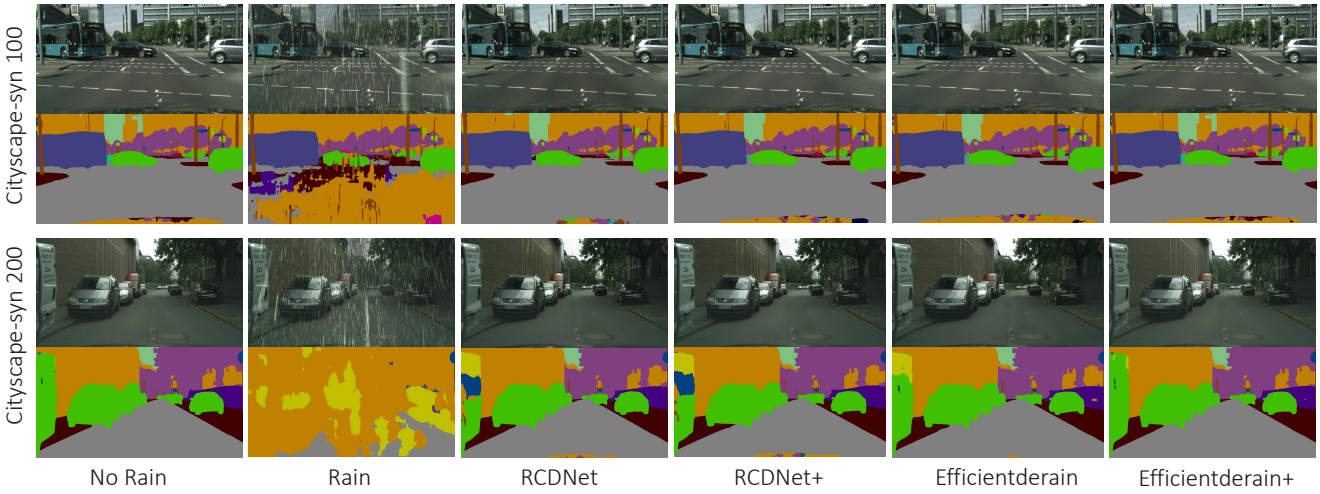

Figure 5. The qualitative comparison of IR models with and without our framework on the cityscape datasets.

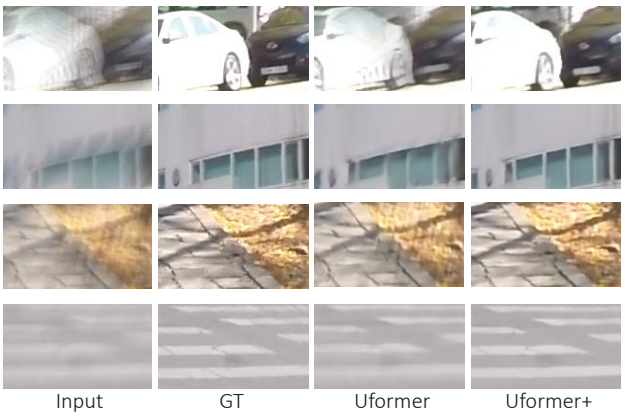

Figure 6. The qualitative comparison of IR models with and without our framework on the GoPro dataset.

results for the downstream segmentation task. The incorporation of semantic priors through our framework enhances both deraining and segmentation performance, demonstrating its effectiveness in addressing the challenges posed by rainy images.

Deblurring and Denoising Tasks. We validate our framework on the deblurring and denoising tasks In Table 3 and Table 4 with the representative model Uformer, respectively. It can be observed that our framework achieves over 0.1 dB PSNR performance improvement of Uformer without inferring with its inference processing on both of these two tasks.

In Figure 6, we present visual results on the GoPro dataset for the deblurring task. These visualizations clearly demonstrate the effectiveness of our framework in enhancing the performance of the Uformer model. Our framework enables the Uformer model to effectively handle dynamic blurring and significantly improve the quality of the deblurred images. Moving on to Figure 7, we showcase visual results on the SenseNoise dataset for the denoising task. These results further illustrate the effectiveness of our framework in enhancing the performance of the Uformer model. This indicates that our framework enhances the denoising capabilities of the Uformer model, resulting in improved image quality and more accurate noise reduction.

4.6. Ablation Study

Components of the Proposed Framework. In Table 5, we present the results of an ablation study conducted on

Figure 7. The qualitative comparison of IR models with and without our framework on the SenseNoise dataset.

Method	PSNR ↑	SSIM ↑	FID ↓
DeblurGAN [22]	28.70	0.858	-
DeblurGANv2 [23]	29.55	0.934	-
SRN [44]	30.26	0.934	-
DBGAN [62]	31.10	0.942	-
MT-RNN [36]	31.15	0.945	-
DMPHN [59]	31.20	0.940	-
Uformer [49]	32.10	0.949	11.36
Uformer+	32.21	0.950	11.10
	+0.11	+0.001	-0.26

Table 3. Quantitative comparison on the GoPro dataset to evaluate our framework for the deblurring task.

Method	PSNR ↑	SSIM ↑	FID ↓
UNet [40]	34.92	0.9130	20.61
CBDNet [10]	35.00	0.9140	17.73
Uformer [49]	35.14	0.9139	17.37
Uformer+	35.25	0.9151	17.09
	+0.11	+0.0012	-0.28

Table 4. Quantitative comparison on the SenseNoise dataset to evaluate our framework for the denoising task.

the main components of our proposed framework. The cascaded networks f^{IR1} and f^{IR2} in our framework demonstrate improved performance, as the cascading architecture enhances the output of f^{IR2}. Furthermore, by incorporating the semantic priors from SAM using the SPF scheme, we observe further performance improvement. Both the SPD scheme and the SGR module consistently enhance the performance of the output of f^{IR1} while maintaining the performance of f^{IR2} itself. Notably, we achieve comparable performance between the outputs of f^{IR1} and f^{IR2}.

In addition, we evaluate the impact of different semantic priors obtained from SAM and the commonly used segmentation model, PSPNet. As SAM provides more detailed and comprehensive semantic priors through segmentation at various levels of granularity, our framework, which integrates SAM, achieves superior performance compared to incorporating semantic priors from instance and category labels alone.

Hyperparameters. We conduct ablation experiments to investigate the impact of hyperparameters λ_1 and λ_2 in balancing the losses $\mathcal{L}_{PSD}$ and $\mathcal{L}_{SGR}$, respectively. The results of these experiments are summarized in Table 6. Various values are tested to identify an optimal weight for each hyperparameter. Based on the experimental findings, we empirically set $\lambda_1 = 0.005$ and $\lambda_2 = 200$ as they yield the best performance. These values are chosen to strike a balance between the two losses and achieve optimal results for our framework.

Method	f^{IR1}			f^{IR2}		
	PSNR ↑	SSIM ↑	FID ↓	PSNR ↑	SSIM ↑	FID ↓
f^{IR1}	34.09	0.9847	5.02	-	-	-
+ f^{IR2}	34.01	0.9831	5.13	34.21	0.9851	4.68
+ SPF, (w. SAM)	34.03	0.9836	5.11	34.89	0.9859	4.67
+ SPD, (w. SAM)	34.84	0.9853	4.71	34.91	0.9865	4.52
+ SGR, (w. SAM)	**35.29**	**0.9864**	**4.49**	**35.30**	**0.9864**	**4.51**
+ SGR, (w. PSPNet)	35.08	0.9858	4.62	35.09	0.9858	4.65

Table 5. Albaltion studies on components of our framework.

λ_1	0.0005	0.005	0.05
PSNR / SSIM	35.12 / 0.9859	**35.29 / 0.9864**	35.16 / 0.9861
λ_2	20	200	2000
PSNR / SSIM	35.26 / 0.9863	**35.29 / 0.9864**	35.21 / 0.9863

Table 6. Ablation study on loss weights of our framework.

5. Conclusion

We propose a general framework to distill the semantic knowledge of the segment anything model (SAM) and boost existing image restoration (IR) models. By incorporating the semantic priors fusion (SPF) and semantic priors distillation (SPD) schemes, we successfully enhance the performance of multiple IR models across tasks such as deraining, deblurring, and denoising. Our framework addresses the computational cost limitations of SAM while effectively leveraging its semantic priors.

Acknowledgement

This work was supported by the National Key R&D Program of China (2022YFB4701400/4701402), in part by the SSTIC Grant (KJZD20230923115106012), in part by Shenzhen Key Laboratory (ZDSYS20210623092001004), and in part by the Beijing Key Lab of Networked Multimedia. We gratefully acknowledge the support of MindSpore, CANN (Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References

[1] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, pages 1316–1326, 2023. 1

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 5

[3] Shuting Dong, Feng Lu, Zhe Wu, and Chun Yuan. Dfvsr: directional frequency video super-resolution via asymmetric and enhancement alignment network. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 681–689, 2023. 1

[4] Shuting Dong, Zhe Wu, Feng Lu, and Chun Yuan. Enhanced image deblurring: An efficient frequency exploitation and preservation network. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7184–7193, 2023. 1

[5] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, pages 6910–6919, 2021. 2

[6] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *CVPR*, pages 3855–3863, 2017. 5

[7] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984. 1

[8] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable medical image segmentation. *arXiv preprint arXiv:2306.13465*, 2023. 1

[9] Qing Guo, Jingyang Sun, Felix Juefei-Xu, Lei Ma, Xiaofei Xie, Wei Feng, Yang Liu, and Jianjun Zhao. Efficientderain: Learning pixel-wise dilation filtering for high-efficiency single-image deraining. In *AAAI*, volume 35, pages 1487–1495, 2021. 1, 5, 6, 7

[10] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *CVPR*, pages 1712–1722, 2019. 8

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 2

[12] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 1

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 5

[14] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 5

[15] Jitesh Jain, Anukriti Singh, Nikita Orlov, Zilong Huang, Jiachen Li, Steven Walton, and Humphrey Shi. Semask: Semantically masked transformers for semantic segmentation. In *ICCV*, pages 752–761, 2023. 2

[16] Kui Jiang, Zhongyuan Wang, Peng Yi, Chen Chen, Baojin Huang, Yimin Luo, Jiayi Ma, and Junjun Jiang. Multi-scale progressive fusion network for single image deraining. In *CVPR*, pages 8346–8355, 2020. 6

[17] Zheyan Jin, Shiqi Chen, Yueting Chen, Zhihai Xu, and Huajun Feng. Let segment anything help image dehaze. *arXiv preprint arXiv:2306.15870*, 2023. 1, 3

[18] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *CVPR*, pages 9404–9413, 2019. 2

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 1

[20] Johannes Kopf, Boris Neubert, Billy Chen, Michael Cohen, Daniel Cohen-Or, Oliver Deussen, Matt Uyttendaele, and Dani Lischinski. Deep photo: Model-based photograph enhancement and viewing. *ACM transactions on graphics (TOG)*, 27(5):1–10, 2008. 1

[21] Vladimir Kulikov, Shahar Yadin, Matan Kleiner, and Tomer Michaeli. Sinddm: A single image denoising diffusion model. In *ICML*, pages 17920–17930. PMLR, 2023. 1

[22] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *CVPR*, pages 8183–8192, 2018. 8

[23] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *ICCV*, pages 8878–8887, 2019. 8

[24] Dasong Li, Yi Zhang, Ka Chun Cheung, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Learning degradation representations for image deblurring. In *ECCV*, pages 736–753. Springer, 2022. 1

[25] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *CVPR*, pages 3041–3050, 2023. 2

[26] Siwei Li, Mingxuan Liu, Yating Zhang, Shu Chen, Haoxiang Li, Hong Chen, and Zifei Dou. Sam-deblur: Let segment anything boost image deblurring. *arXiv preprint arXiv:2309.02270*, 2023. 1, 3

[27] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *ECCV*, pages 254–269, 2018. 6

[28] Yi Li, Yi Chang, Changfeng Yu, and Luxin Yan. Close the loop: a unified bottom-up and top-down paradigm for joint image deraining and segmentation. In *AAAI*, volume 36, pages 1438–1446, 2022. 1, 2

[29] Ding Liu, Bihan Wen, Xianming Liu, Zhangyang Wang, and Thomas S Huang. When image denoising meets high-level vision tasks: A deep learning approach. *arXiv preprint arXiv:1706.04284*, 2017. 2

[30] Xiaoyu Liu, Bo Hu, Wei Huang, Yueyi Zhang, and Zhiwei Xiong. Efficient biomedical instance segmentation via knowledge distillation. In *MICCAI*. Springer, 2022. 2

[31] Xiaoyu Liu, Wei Huang, Zhiwei Xiong, Shenglong Zhou, Yueyi Zhang, Xuejin Chen, Zheng-Jun Zha, and Feng Wu. Learning cross-representation affinity consistency for sparsely supervised biomedical instance segmentation. In *ICCV*, 2023. 2

[32] Xiaoyu Liu, Yueyi Zhang, Zhiwei Xiong, Wei Huang, Bo Hu, Xiaoyan Sun, and Feng Wu. Graph relation distillation for efficient biomedical instance segmentation. *arXiv preprint arXiv:2401.06370*, 2024. 2

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2

[34] Zhihe Lu, Zeyu Xiao, Jiawang Bai, Zhiwei Xiong, and Xinchao Wang. Can sam boost video super-resolution? *arXiv preprint arXiv:2305.06524*, 2023. 1

[35] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, pages 3883–3891, 2017. 5

[36] Dongwon Park, Dong Un Kang, Jisoo Kim, and Se Young Chun. Multi-temporal recurrent neural networks for progressive non-uniform single image deblurring with incremental temporal training. In *ECCV*, pages 327–343. Springer, 2020. 8

[37] Minh Pham, Minsu Cho, Ameya Joshi, and Chinmay Hegde. Revisiting self-distillation. *arXiv preprint arXiv:2206.08491*, 2022. 2

[38] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *CVPR*, pages 3937–3946, 2019. 6

[39] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu. Deep video dehazing with semantic segmentation. *IEEE transactions on image processing*, 28(4):1895–1908, 2018. 2

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 8

[41] Reuben R Shamir, Yuval Duchin, Jinyoung Kim, Guillermo Sapiro, and Noam Harel. Continuous dice coefficient: a method for evaluating probabilistic segmentations. *arXiv preprint arXiv:1906.11031*, 2019. 5

[42] Ziyi Shen, Wei-Sheng Lai, Tingfa Xu, Jan Kautz, and Ming-Hsuan Yang. Exploiting semantics for face image deblurring. *International Journal of Computer Vision*, 128:1829–1846, 2020. 2

[43] Ke Sun, Yang Zhao, Borui Jiang, Tianheng Cheng, Bin Xiao, Dong Liu, Yadong Mu, Xinggang Wang, Wenyu Liu, and Jingdong Wang. High-resolution representations for labeling pixels and regions. *arXiv preprint arXiv:1904.04514*, 2019. 6

[44] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *CVPR*, pages 8174–8182, 2018. 8

[45] Hong Wang, Qi Xie, Qian Zhao, and Deyu Meng. A model-driven deep neural network for single image rain removal. In *CVPR*, pages 3103–3112, 2020. 1, 5, 6, 7

[46] Li Wang, Dong Li, Yousong Zhu, Lu Tian, and Yi Shan. Dual super-resolution learning for semantic segmentation. In *CVPR*, pages 3774–3783, 2020. 2

[47] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *CVPR*, pages 606–615, 2018. 1, 2

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[49] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022. 5, 8

[50] Yanyan Wei, Zhao Zhang, Huan Zheng, Richang Hong, Yi Yang, and Meng Wang. Sginet: Toward sufficient interaction between single image deraining and semantic segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6202–6210, 2022. 1, 5

[51] 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 *CVPR*, pages 1662–1671, 2023. 1, 2

[52] Zeyu Xiao, Jiawang Bai, Zhihe Lu, and Zhiwei Xiong. A dive into sam prior in image restoration. *arXiv preprint arXiv:2305.13620*, 2023. 1, 3

[53] Defeng Xie, Ruichen Wang, Jian Ma, Chen Chen, Haonan Lu, Dong Yang, Fobo Shi, and Xiaodong Lin. Edit everything: A text-guided generative system for images editing. *arXiv preprint arXiv:2304.14006*, 2023. 1

[54] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *CVPR*, pages 8818–8826, 2019. 2

[55] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023. 1

[56] Wenhan Yang, Robby T Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *CVPR*, pages 1357–1366, 2017. 5

[57] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5728–5739, 2022. 1

[58] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *CVPR*, pages 14821–14831, 2021. 6

[59] Hongguang Zhang, Yuchao Dai, Hongdong Li, and Piotr Koniusz. Deep stacked hierarchical multi-patch network for image deblurring. In *CVPR*, pages 5978–5986, 2019. 1, 8

[60] He Zhang and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *CVPR*, pages 695–704, 2018. 5

[61] Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6360–6376, 2021. 1

[62] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn

Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *CVPR*, pages 2737–2746, 2020. 8

[63] Yi Zhang, Dasong Li, Ka Lung Law, Xiaogang Wang, Hongwei Qin, and Hongsheng Li. Idr: Self-supervised image denoising via iterative data refinement. In *CVPR*, pages 2098–2107, 2022. 5

[64] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 43(7):2480–2495, 2020. 1

[65] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, pages 2881–2890, 2017. 2

[66] Shen Zheng and Gaurav Gupta. Semantic-guided zero-shot learning for low-light image/video enhancement. In *WACV*, pages 581–590, 2022. 2