Transfer CLIP for Generalizable Image Denoising

Jun Cheng, Dong Liang, Shan Tan*
Huazhong University of Science and Technology, Wuhan, China
{jcheng24, liangdong, shantan}@hust.edu.cn

Abstract

Image denoising is a fundamental task in computer vision. While prevailing deep learning-based supervised and self-supervised methods have excelled in eliminating in-distribution noise, their susceptibility to out-of-distribution (OOD) noise remains a significant challenge. The recent emergence of contrastive language-image pre-training (CLIP) model has showcased exceptional capabilities in open-world image recognition and segmentation. Yet, the potential for leveraging CLIP to enhance the robustness of low-level tasks remains largely unexplored. This paper uncovers that certain dense features extracted from the frozen ResNet image encoder of CLIP exhibit distortion-invariant and content-related properties, which are highly desirable for generalizable denoising. Leveraging these properties, we devise an asymmetrical encoder-decoder denoising network, which incorporates dense features including the noisy image and its multi-scale features from the frozen ResNet encoder of CLIP into a learnable image decoder to achieve generalizable denoising. The progressive feature augmentation strategy is further proposed to mitigate feature over-fitting and improve the robustness of the learnable decoder. Extensive experiments and comparisons conducted across diverse OOD noises, including synthetic noise, real-world sRGB noise, and low-dose CT image noise, demonstrate the superior generalization ability of our method.

1. Introduction

Image denoising is a significant task in computer vision and image processing. Current supervised denoising methods leveraging powerful deep neural networks and large-scale datasets have achieved exceptional performance in both synthetic and real-world noise removal [33, 60]. However, these supervised denoisers tend to overfit the noise present in the training datasets, resulting in poor generalization to out-of-distribution (OOD) noise [6]. On the other hand, unsupervised and self-supervised denoising methods

*Corresponding author.

[12, 18, 26, 28, 44, 54] directly focus on the target domain in which the target noisy images reside and hence bypass OOD generalization. Nevertheless, these methods are inherently vulnerable to unseen noise [7] and the collection of target noisy datasets is not always available. Therefore, it is critical to enhance the generalization of deep denoisers.

OOD generalization has been popular research in high-level vision tasks like image recognition and segmentation [51, 70]. In contrast, attention to OOD generalization within image denoising is limited. Existing research in this area primarily consists of two aspects: generalization across degradation levels and generalization across degradation types. Regarding the former, some works trained blind denoisers [18, 60, 62] or bias-free networks [41, 63] to handle noise with varying levels. However, these methods are confined to specific noise and cannot generalize to unseen noise types. For the latter, several works aimed to fortify models against general OOD noise. Particularly, MaskDenoising [6] incorporated dropout units into the model training to enforce the denoiser to learn the reconstruction of image contents. DIL [32] built upon causality and meta-learning and encouraged the model to learn distortion-invariant representations. HAT [58] designed an adversarial attack for deep denoisers and then conducted adversarial training.

Recently, through solving the image-text alignment problem based on hyper-scale datasets, the contrastive language-image pre-training (CLIP) model [48] has demonstrated remarkable generalization capacity in downstream open-world image recognition tasks. A series of extensions on CLIP through frozen models [65, 69], model fine-tuning [49], visual prompts [71], distillations [19, 31], and so on [34] have been proposed to transfer the generalization ability of CLIP from classification to dense prediction tasks, including open-vocabulary segmentation [34] and zero-shot depth estimation [65]. However, the feasibility of CLIP for robust restoration in low-level tasks remains unexplored. We therefore ask, is CLIP robust to image noise and can we transfer it for generalizable image denoising?

In this paper, we find that the dense feature maps from the frozen ResNet image encoder of CLIP within specific scales exhibit remarkable resilience to noise, a property that

is not easy to obtain via supervised learning. These features of clean images and their noisy counterparts show significant similarities in terms of cosine and CKA [25] similarity measures. In addition, these features maintain a clear distinction for images with different contents and semantics. Such distortion-invariant and content-related properties are desirable for generalizable denoising as the robust and distinctive features commendably represent the latent image regardless of the corruption in the noisy measurement. As a result, we propose an asymmetrical encoder-decoder denoising network by integrating the frozen ResNet image encoder of CLIP and a learnable image decoder. The multi-scale features of noisy images from the frozen encoder, as well as an extra dense feature represented by the noisy image, are progressively incorporated into the decoder to recover high-quality images. Through supervised training on a single noise type and noise level, the proposed concise denoiser, termed CLIPDenoising, exhibits good generalization capacity to various OOD noises.

By employing the frozen image encoder, the image denoising task turns into recovering clean images from fixed features. During training, the inherent similarity of training images [5] as well as their respective dense features will inevitably affect feature diversity, leading to potential feature overfitting. Therefore, we propose progressive feature augmentation to randomly perturb these features from the frozen CLIP with increasing randomness at deeper scales. In total, our contributions are summarized as follows:

- We identify that dense features from the frozen ResNet encoder of CLIP possess distortion-invariant and content-related properties. Leveraging this finding, we incorporate these features along with the noisy image into a learnable image decoder to construct a generalizable denoiser.
- We propose the progressive feature augmentation strategy to further improve the robustness of our approach.
- To the best of our knowledge, we are the first to utilize CLIP for generalizable denoising. Extensive experiments and comparisons on various OOD noises, including synthetic noise, real-world sRGB noise, and low-dose CT noise, demonstrate superior generalization of our method.

2. Related works

2.1. Deep Learning-based Image Denoising

Supervised denoising methods generally build upon powerful deep architectures (e.g., CNNs [62, 63], non-local networks [66], and Transformer [33, 35, 60]), large-scale paired datasets (e.g., SIDD [1]), and diverse optimization targets (e.g., L1/L2 losses [67], adversarial loss [27] or diffusion loss [50]) and have achieved state-of-the-art performance. However, the strong reliance on paired datasets and *i.i.d.* assumption makes them vulnerable to unseen and OOD noise [6]. To circumvent this limitation, many unsu-

pervised and self-supervised denoising methods [4, 12, 18, 23, 26, 28, 44, 46, 54] have been introduced to directly handle the target noisy images. While effective, these methods bypass tackling the generalization problem, leading to insufficient improvement in the OOD robustness of deep denoisers. Consequently, there remains a significant gap in research concerning generalizable denoising.

2.2. OOD Generalization in Image Denoising

Existing research on generalizable image denoising primarily consists of generalization across degradation levels and generalization across degradation types. The former addresses known noise types at unknown levels during inference, while the latter strives for general OOD robustness. Regarding the former, DnCNN [62] proposed to train blind denoisers, which are capable of handling specific noise types with varying levels. Mohan et al. [41] discovered that bias-free (BF) denoisers trained on limited noise ranges exhibited robustness to unseen noise levels. Consequently, the BF architecture was adopted in subsequent models like DRUNet [63] and Restormer [60]. Regarding the latter, GainTuning [42] employed a test-time training strategy to optimize the denoiser for each noisy input. Chen et al. [7] disentangled latent clean features from multiple corrupted versions of the same image to achieve OOD generalization. MaskDenoising [6] revisited the commonly employed dropout operation in high-level vision tasks and integrated these units into model training. DIL [32] combined counterfactual distortion augmentation and meta learning-based optimization to develop a generalizable restoration network. HAT [58] incorporated the adversarial attack and adversarial training to improve the OOD generalization of deep denoisers. Despite these advancements, there remains room for further improvement in generalizability.

2.3. CLIP-based Generalization

CLIP has demonstrated remarkable generalization abilities in open-world image recognition [48]. Subsequent works, such as MaskCLIP [69], DenseCLIP [49], ZegCLIP [71], and others [34, 38], extended CLIP to dense prediction tasks, enabling zero-shot or open-vocabulary image segmentation. There are also studies explicitly distilling CLIP while maintaining its zero-shot performance [19, 31, 55]. Nevertheless, due to a distinct domain gap, there remains a lack of research on harnessing CLIP's exceptional generalization capacity for low-level vision tasks. This gap in exploration serves as a primary motivation for our work.

2.4. Foundation Models for Image Restoration

Utilizing foundation models to solve domain-specific tasks has become prevalent in computer vision. Similar to high-level vision tasks, many works have integrated large pre-trained models for image restoration. Diffusion model

stands as a state-of-the-art generation method and many papers have leveraged the pre-learned diffusion priors to address various image restoration tasks [9, 13, 21, 22, 72]. Regarding Segment Anything model [24], some works have integrated it into image deblurring [30], image dehazing [20], and super-resolution [37, 56]. Yu et al. [59] introduced the semantic pyramid auto-encoder to enable large language models to perform image deblurring and inpainting. Additionally, Luo et al. [39] leveraged CLIP to predict high-quality image features and degradation features, subsequently integrating them into the image restoration model for universal image restoration. While existing works predominantly concentrate on leveraging foundation models to enhance image restoration performance, our paper aims to emphasize the enhancement of OOD generalization ability.

3. Method

In this section, we first check whether CLIP that has been trained on hyper-scale image-text datasets enjoys some good properties for generalizable denoising. Based on the analysis in Section 3.1, we propose the simple and generalizable denoiser in Section 3.2, followed by the strategy of progressive feature augmentation in Section 3.3.

3.1. Analyzing Features of CLIP Image Encoder

CLIP offers two variants of image encoders, i.e., ResNet [16] and ViT [11]. The ResNet version extracts multi-scale feature maps through sequential Conv-blocks and Pooling operations, while the ViT version breaks down images into smaller 16×16 patches and then employs standard Transformer operations. For deep learning-based image denoising, low-level image details and textures are critical for reconstructing high-quality images [61]. Existing methods utilize either the *single-scale* pipeline [33, 35] or the *encoder-decoder* architecture with skip connections [60, 63] to preserve spatial details. As the ViT architecture directly processes overly downsampled image features, it abandons spatial image details and hence is not suitable for image denoising. Consequently, we focus on analyzing and utilizing the ResNet for our further analysis and method.
Distortion-invariant property. We examine the dense feature maps before each (average or attention)-pooling operation in the CLIP ResNet image encoder. (Refer Alg. 1 in the Supplementary Material for details). This yields a total of five multi-scale features, denoted as $\mathbf{F}^1 \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$ and $\mathbf{F}^i \in \mathbb{R}^{\frac{H}{2^i} \times \frac{W}{2^i} \times 2^i C}, i \in \{2, \cdots 5\}$, where $H \times W$ are the spatial dimensions of the input image, and C is the base channel number. To assess the robustness of these features, we start with a clean image $\mathcal{I}_c$ and introduce diverse *i.i.d.* Gaussian noises to create corresponding noisy images $\mathcal{I}_n$. Note that the image intensity range used in this section is $[0, 1]$. The clean image features $\mathbf{F}_c^i$ and the noisy image features $\mathbf{F}_n^i$ are obtained by passing both the high-quality

(a) *Lena* corrupted by different levels of *i.i.d.* Gaussian noise

(b) Left plot uses RN50 model with various noise levels while the right uses ResNets with different model sizes under std=0.1
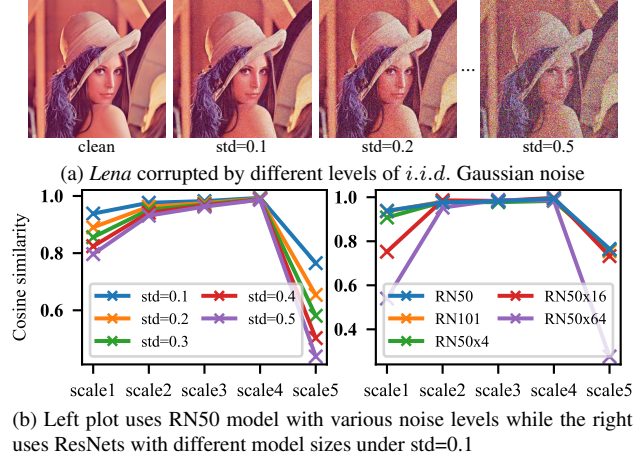
Figure 1. Feature similarity analysis of the CLIP ResNet image encoder for image *Lena*. Cosine similarity between $\mathbf{F}_c^i$ and $\mathbf{F}_n^i$ with regard to different noise levels and model sizes is displayed

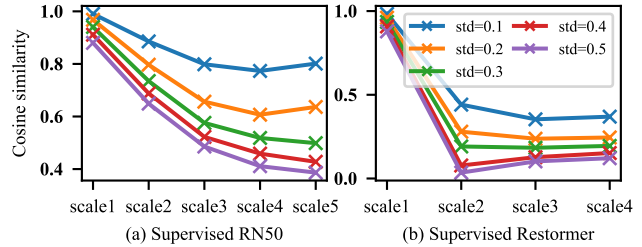(a) Supervised RN50   (b) Supervised Restormer

Figure 2. Feature similarity analysis of ResNet50 (supervised training for image classification, not from CLIP) and Restormer (supervised training for blind Gaussian noise removal)

and degraded images through the frozen ResNet. Subsequently, the cosine similarity between $\mathbf{F}_c^i$ and $\mathbf{F}_n^i$ is computed at each scale i. We show the result in Fig. 1, where five distinct noise levels and five pre-trained ResNets with increasing sizes (i.e., more residual blocks within each scale of ResNet) provided by CLIP are considered.

From Fig. 1, we observe that the initial *four* features $\mathbf{F}_n^i$ from the RN50, under various corruption levels, show significant similarities to their corresponding $\mathbf{F}_c^i$, with higher similarity at deeper scales i. On the other hand, as the model size increases, the resemblance between $\mathbf{F}_c^1$ and $\mathbf{F}_n^1$ dramatically decreases even under smaller std=0.1. Additional results using *CKA similarity* metric [25], *Poisson degradation*, and *other image* are given in Figs. 9, 10, 11 in the Supplementary Material, which all suggest similar observations. From these findings, we conclude that features $\mathbf{F}_n^i, i \in \{1, \cdots 4\}$ from CLIP frozen RN50 are robust and distortion-invariant, which is essential for building generalizable denoisers. Additionally, we contrast these findings with feature analyses of ResNet50 trained on ImageNet for supervised image classification and Restormer trained on *i.i.d.* Gaussian noise with $\sigma \in [0, 0.2]$ for blind denoising,
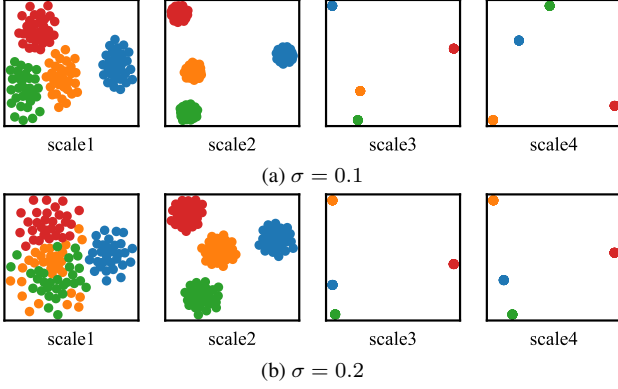
Figure 3. The t-SNE plots of $\mathbf{F}^i, i \in \{1, \cdots 4\}$ from four corrupted images with diverse contents, i.e., *Lena*, *Baboon*, *F16* and *Peppers* from set9 [10] under $i.i.d.$ Gaussian noise with two noise levels. Different colors denote features of different images
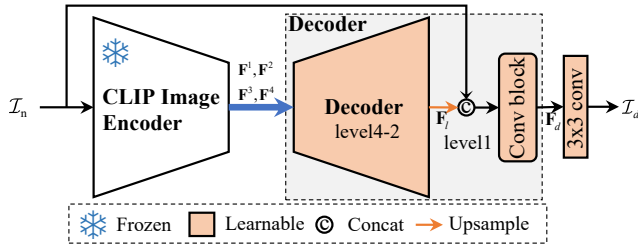

Figure 4. The CLIPDenoising for generalizable image denoising, which comprises the frozen RN50 encoder from CLIP, a learnable image decoder, and 3×3 convolution

and report the results in Fig. 2. The distinction between Fig. 1 and Fig. 2 underlines that such distortion-invariant property is not universal and originates from CLIP. We present a brief discussion in Section 5 about why the RN50 image encoder of CLIP holds this appealing property.

Conetent-related property. We then check whether the above features from CLIP frozen RN50 are content-related, that is, if features of two noisy images with distinctive contents are different. Given M distictive clean images $\mathcal{I}_c^m, m \in \{1, \cdots M\}$, we generate multiple noisy images $\mathcal{I}_n^m$ from $\mathcal{N}(\mathcal{I}_c^m, \sigma^2 I)$ and obtain the corresponding multi-scale features $\mathbf{F}_n^i, i \in \{1, \cdots 4\}$ from the frozen RN50. We then compute the two-dimensional embeddings of these features using the t-SNE method [53] and present the result in Fig. 3. We find that the low-dimensional representation of $\mathbf{F}_n^i$ from different noisy image $\mathcal{I}_n^m$ showcases clear separation under different scale i and noise level σ, indicating a strong correlation between the image content and their multi-scale features from CLIP RN50.

3.2. Building a Generalizable Denoiser

Leveraging the favorable attributes of the frozen RN50 encoder from CLIP, we established a simple, effective, and generalizable denoiser, of which the architecture is depicted in Fig. 4. Our model mainly consists of the frozen RN50 image encoder and a 4-level learnable image decoder. Given the noisy input $\mathcal{I}_n \in \mathbb{R}^{H \times W \times 3}$, the multi-scale features $\mathbf{F}_n^i, i \in \{1, \cdots 4\}$ from the frozen RN50 are first extracted. The decoder takes $\mathbf{F}_n^4$ as input and progressively recovers the high-resolution features. During the upsampling, $\mathbf{F}_n^i, i \in \{1, 2, 3\}$ are concatenated with the decoder features to incorporate the multi-scale feature information into the restoration (level-4 to -2). Subsequently, the decoder at level-2 outputs the feature $\mathbf{F}_l \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times C}$. At level-1, $\mathbf{F}_l$ is then upsampled and concatenated with the noisy input $\mathcal{I}_n$, producing $\mathbf{F}_d \in \mathbb{R}^{H \times W \times C}$ after a final Conv-block. Ultimately, the denoised image $\mathcal{I}_d$ is obtained through a 3×3 convolution operation applied on $\mathbf{F}_d$. The learnable decoder, devised to align with the ResNet encoder, is convolution-based and comprises multiple Conv-blocks, each composed of Conv-ReLU-Conv-ReLU sequences. More details about the decoder are given in the Supplementary Material. It's worth noting that the model incorporates the noisy image $\mathcal{I}_n$ as an additional image feature into the decoder, which is important and reasonable as the noisy image $\mathcal{I}_n$ itself contains rich image details and can be considered as one distinctive dense feature. This operation will be analyzed further in Section 4.5.

During the training, we synthesize $\mathcal{I}_n$ from $\mathcal{I}_c$ based on a fixed noise type and level, and optimize the loss function,

$$\mathcal{L} = \mathrm{E}_{p(\mathcal{I}_c)} \|\mathcal{I}_d - \mathcal{I}_c\|_1 \tag{1}$$

Note that, we do not employ global residual learning as we aim to restore the high-quality image from its robust features rather than restore the residual noise. In inference, we directly evaluate our model on OOD noise. Ablations in Section 4.5 indicate that our simple baseline, i.e., CLIPDenoising, has already achieved good generalizability.

3.3. Progressive Feature Augmentation

By utilizing the frozen CLIP RN50 as the image encoder, the image denoising task turns into mapping invariant features to high-quality images. However, as images in the training dataset naturally exhibit certain degrees of similarities (e.g., similar textures in different images), the corresponding multi-scale features from CLIP RN50 tend to follow these similarities. This will reduce the feature diversity and can potentially lead to feature overfitting problems. To avoid this issue and enhance the decoder's robustness, we introduce the strategy of progressive feature augmentation, inspired by [29]. During the training phase, we apply random perturbations to the multi-scale features $\mathbf{F}^i$ as follows

$$\hat{\mathbf{F}}^i = \boldsymbol{\alpha}_i \odot \mathbf{F}^i, \boldsymbol{\alpha}_i \sim \mathcal{N}(\mathbf{1}, (\gamma \times i)^2 \boldsymbol{I}), i \in \{1, \cdots 4\} \tag{2}$$

where $\odot$ denotes the element-wise multiplication and $\boldsymbol{\alpha}_i$ has the same size as $\mathbf{F}^i$.

In Eq. (2), $\boldsymbol{\alpha}_i$ is sampled from the $i.i.d.$ Gaussian distribution with mean one and std $\gamma \times i$. For larger i, we inject more randomness to $\mathbf{F}^i$ as the deeper features tend to capture more semantic information and should also be more robust; Regarding smaller i, we inject less noise to preserve texture- and detail-related information contained in the shallow features. We note that such progressive feature augmentation is simple but effective.

4. Experiments

In this section, we first introduce experimental settings of denoising diverse OOD noises. Quantitative and qualitative results of our method and comparisons with other methods are then presented. The ablation is conducted in the last.

4.1. Experimental Settings

Synthetic noise. We choose $i.i.d.$ Gaussian noise with $\sigma = 15$ as the in-distribution noise and consider 5 kinds of synthetic OOD noise: (1) $i.i.d$ Gaussian noise with $\sigma \in \{25, 50\}$, (2) spatial Gaussian noise with $\sigma \in \{45, 50, 55\}$, (3) Poisson noise with levels $\alpha \in \{2.5, 3, 3.5\}$, (4) Speckle noise with levels $\sigma^2 \in \{0.02, 0.03, 0.04\}$ and (5) Salt&Pepper noise with levels $d \in \{0.012, 0.016, 0.02\}$. We follow MaskDenoising [6] to generate these OOD noises and adopt Kodak24 [14], McMaster [64], CBSD68 [40], and Urban100 [17] as test sets. Note that Gaussian and spatial Gaussian noises are generated in the intensity range of $[0, 255]$ while the rest uses the intensity range of $[0, 1]$, so as to be consistent with MaskDenoising.

Regarding the implementation details of our method, we build on CBSD432 dataset [40] and synthesize noisy images using $i.i.d.$ Gaussian noise with $\sigma = 15$ in the online fashion. Supervised training based on Eq. (1) and Fig. 4 is conducted. In the training phase, we utilize the AdamW [36] optimizer combined with the cosine-annealing learning rate. We conduct 300k training iterations with a batch size of 16, and the learning rate decreases from an initial value of $3e^{-4}$ to a final value of $1e^{-6}$. The training patch size is 128×128 and random geometric augmentations are applied to training patches. We set $\gamma = 0.025$ to augment the dense features from frozen RN50 of CLIP. We perform all experiments using PyTorch [47] and an Nvidia 2080ti GPU. Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) metrics are used to evaluate the denoising quality.

Real-world sRGB noise. For real-world sRGB noise, we consider SIDD validation dataset [1], PolyU [57] and CC [43] as test sets. These datasets comprise natural noisy sRGB images from smartphones and commercial cameras. During training, we simulate sRGB noise based on DIV2K dataset [2] and the image signal processing pipeline, following CBDNet [15]. In particular, we use a fixed level of Poisson-Gaussian noise, i.e., $\sigma_s = 0.04, \sigma_c = 0.03$ in the raw domain to generate noisy images in order to accentu-

ate the distribution gap of data between training and testing. The optimization here is identical to that of synthetic noise. **Low-dose CT image noise**. We utilize CLIPDenoising trained on $i.i.d$ Gaussian noise to remove real-world low-dose (LD) CT image noise, which is known to be complex and hard to model [12]. We use AAPM-Mayo Clinic Low Dose CT Grand Challenge dataset [3], which provides 1mm thickness abdomen slices with quarter-dose images (noisy) and corresponding normal-dose (ND) images (GT). We select 5410 NDCT images from nine patients combined with $i.i.d$ Gaussian noise with $\sigma = 5$ for training and use 526 LDCT images from patient L506 for test. To accommodate the one-channel CT image, we insert a learnable 1×1 convolution ahead of the CLIP RN50 encoder, which converts the single-channel image into a three-channel image. The optimization details remain the same as the above except that we set the total training iterations to 40k in this case.

4.2. Synthetic Noise Removal

Compared methods. We compare our method against three representative works on generalizable denoising, i.e., MaskDenoising [6], DIL [32], and HAT [58]. For MaskDenoising, we use the officially trained model. Regarding DIL and HAT, we follow their source codes and experimental settings to train denoisers based on $i.i.d.$ Gaussian noise with $\sigma \in \{5, 10, 15, 20\}$ and $\sigma \in [0, 25]$, respectively. Note that *multiple noise levels* are required by these two methods during the training phase to achieve generalization. We additionally evaluate DnCNN [62] and Restormer [60] trained on $i.i.d.$ Gaussian noise with $\sigma = 15$.

Results. We present comprehensive quantitative comparisons of various methods across diverse noise types, levels, and datasets in Table 1 and Table 9 (in the Supplementary Material). As indicated by Tables 1 and 9, our CLIPDenoising exhibits commendable in-distribution performance while demonstrating remarkable robustness against all considered OOD noises. In comparison, the compared methods merely excel in some specific noise types.

As shown in the first part of Table 9, Restormer achieves commendable performance in in-distribution noise. However, it struggles in tackling unseen noise levels and types (see Tables 1 and 9), indicating the overfitting to the noise in the training set. Compared with Restormer, DnCNN has a weaker modeling capacity but shows better robustness to OOD noise, which however is significantly behind the methods specialized in generalizable denoising. In comparisons among MaskDenoising, DIL, HAT, and CLIPDenoising, DIL and our method stand out as the only effective methods in eliminating Gaussian noise with higher noise level, i.e., $\sigma = 50$, and our method surpasses DIL in average as indicated in Table 1. Regarding unseen noise types, MaskDenoising outperforms HAT and DIL in spatial Gaussian noise; Conversely, HAT and DIL exhibit notable advan-

Table 1. Quantitative comparison (PSNR/SSIM) of different methods on CBSD68, McMaster, Kodak24 and Urban100 datasets with regard to diverse synthetic OOD noises. The best results are highlighted in **bold** and the second best is <u>underlined</u>. Note that *multiple noise levels* are required by HAT and DIL during the training to achieve generalization, while our method only needs *one noise level* for training

| Noise Types	Datasets	DnCNN [62]	Restormer [60]	MaskDenoising [6]	HAT [58]	DIL [32]	Ours
Gauss $\sigma = 50$	CBSD68	19.84/0.363	19.92/0.365	20.68/0.432	20.95/0.441	<u>26.43/0.717</u>	**26.69/0.731**
	McMaster	20.18/0.312	20.47/0.312	20.63/0.379	20.79/0.364	<u>26.61/0.669</u>	**27.43/0.727**
	Kodak24	19.78/0.301	20.12/0.321	20.72/0.368	21.04/0.390	**27.46/0.736**	<u>27.39/0.723</u>
	Urban100	19.62/0.420	19.36/0.437	20.51/0.485	20.80/0.492	25.89/0.768	**26.27/0.769**
Spatial Gauss $\sigma = 55$	CBSD68	25.91/0.699	23.51/0.595	<u>26.72/0.762</u>	26.39/0.713	24.61/0.630	**27.60/0.797**
	McMaster	26.18/0.649	24.01/0.539	<u>26.89/0.709</u>	26.62/0.665	24.82/0.574	**28.31/0.775**
	Kodak24	25.98/0.653	22.99/0.533	<u>27.28/0.745</u>	26.40/0.671	24.56/0.572	**28.29/0.786**
	Urban100	25.55/0.727	24.13/0.660	26.10/<u>0.788</u>	<u>26.48</u>/0.742	24.80/0.673	**27.68/0.822**
Poisson $\alpha = 3.5$	CBSD68	24.37/0.627	22.20/0.559	24.24/0.638	26.61/0.733	<u>27.64</u>/**0.819**	**27.67**/<u>0.818</u>
	McMaster	25.50/0.651	21.93/0.579	25.17/0.590	27.54/0.723	**28.91/0.825**	<u>28.81/0.820</u>
	Kodak24	24.49/0.560	22.55/0.517	24.30/0.572	27.10/0.695	<u>28.60</u>/**0.821**	**28.66**/<u>0.813</u>
	Urban100	23.57/0.649	21.08/0.584	23.90/0.669	25.95/0.746	<u>27.12</u>/**0.854**	**27.15**/<u>0.838</u>
Salt&Pepper $d = 0.02$	CBSD68	26.53/0.746	23.59/0.679	<u>29.74/0.843</u>	27.55/0.782	29.45/0.822	**29.81/0.844**
	McMaster	25.72/0.691	23.05/0.640	<u>29.28/0.773</u>	26.62/0.727	<u>29.28/0.773</u>	**29.79/0.807**
	Kodak24	27.10/0.723	23.81/0.639	<u>30.56</u>/**0.842**	28.19/0.766	29.99/0.810	**30.61**/<u>0.837</u>
	Urban100	25.61/0.777	23.51/0.734	28.43/<u>0.861</u>	26.88/0.792	<u>29.21</u>/0.841	**29.40/0.869**
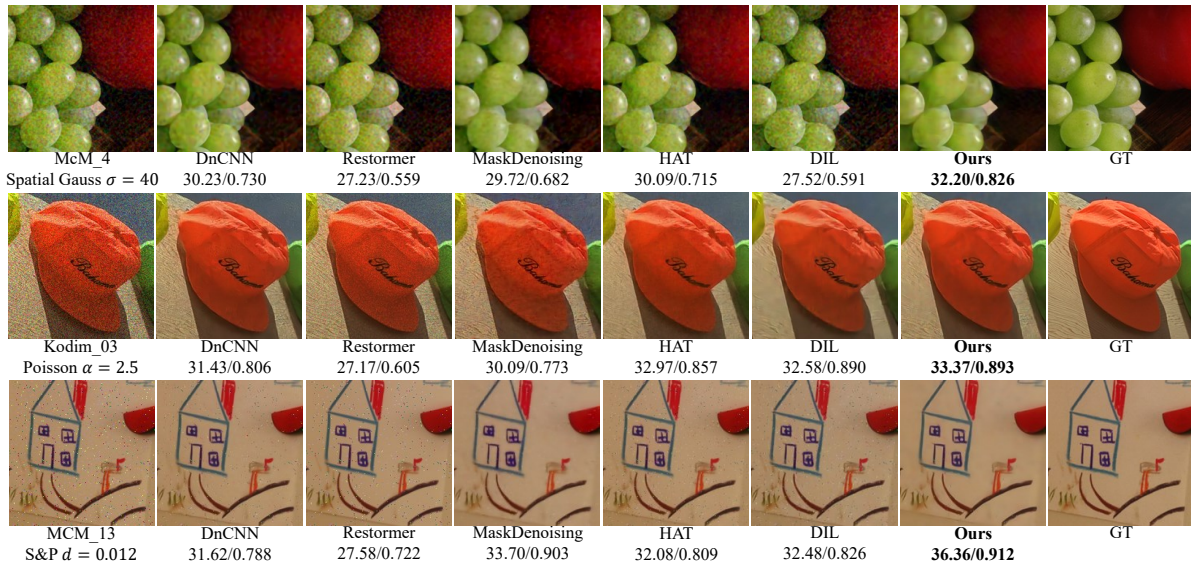
Figure 5. Qualitative denoising results on synthetic OOD noise. During the training, all the methods do not encounter the test noise types. PSNR/SSIM values are listed underneath the respective images. Zoom-in for a better comparison

tages over MaskDenoising in handling Poisson and Speckle noise. In contrast, our method demonstrates consistent and competitive performance across all types of OOD noise, suggesting the great benefits from the superior distortion-invariant property of the frozen CLIP ResNet encoder.

Note that our method, though exhibiting slightly lower performance than HAT and DIL in dealing with Speckle noise (see the middle part of Table 9), is accomplished with just one noise level, unlike these methods which necessitate multiple noise levels during training. Fig. 5 and Figs. 14, 15 (in the Supplementary Material) present the qualitative comparison among various methods. These visual results indicate that our method can effectively denoise OOD noise while preserving image contents and details.

4.3. Real-world sRGB Noise Removal

Compared methods. We again consider MaskDenoising, DIL, and HAT for comparison. We train MaskDenoising and DIL using their respective source codes with the same synthetic sRGB dataset used in our method. DIL mandates that the training set is categorized into four groups based on Bayer patterns. As for HAT, it's implemented based on the clean DIV2K dataset with $i.i.d.$ Gaussian noise within the range of $\sigma \in [0, 50]$ to ensure its standard functionality. In addition, we also consider two unsupervised denoising methods, i.e., CVF-SID [44] and LUD-VAE [68], which necessitate real-world noisy images for training.

Results. We present quantitative and qualitative results of different methods on real-world noise removal in Ta-
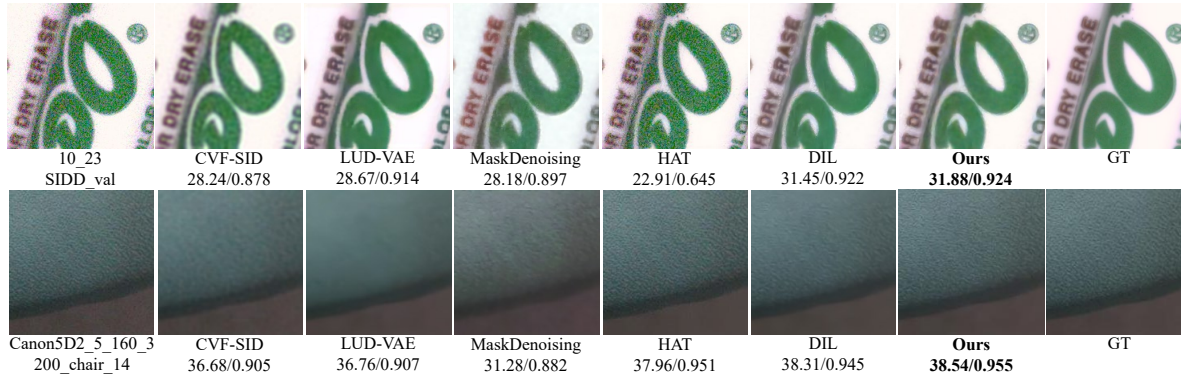
Figure 6. Qualitative results on real-world sRGB noise. PSNR/SSIM values are listed underneath the respective images.

Table 2. Quantitative comparison (PSNR/SSIM) of different methods on real-world sRGB datasets, i.e., SIDD Val, PolyU, and CC.

Methods	SIDD Val	PolyU	CC
MaskDenoising [6]	33.14/0.913	24.78/0.812	25.63/0.881
HAT [58]	28.58/0.570	37.25/0.948	35.27/0.901
DIL [32]	34.76/0.848	**37.65**/0.959	36.10/**0.948**
Ours	**34.79/0.866**	37.54/**0.960**	**36.30**/0.941
CVF-SID [44]	34.81/0.944	35.86/0.937	33.29/0.913
LUD-VAE [68]	34.91/0.892	36.99/0.955	35.48/0.941

ble 2, Fig. 6 and Fig. 16 (in the Supplementary Material), respectively. It is clear from Table 2 that our method nearly achieves the best performance among the methods of generalizable denoising. Notably, DIL stands on par with our CLIPDenoising, although it requires constructing *four distinct confounders*, i.e., four distortions during training, while our method does not have this requirement. In addition, compared with CVF-SID and LUD-VAE, our method shows competitive results on SIDD dataset and outperforms them on PolyU and CC datasets. As our method does not rely on real-world datasets, it emerges as a more practical and universal alternative compared to those noisy datasets-based unsupervised methods. The visual comparisons in Figs. 6 and 16 further underline the effectiveness of CLIP-Denoising in addressing real-world sRGB noise.

4.4. Low-dose CT Image Noise Removal

Compared methods. We compare our method against the representative works in LDCT image denoising, namely Noise2Sim [45] and ScoreSDE-CT [52]. Noise2Sim relies on the adjacent and similar LDCT images to train the denoiser, while ScoreSDE-CT learns score priors from NDCT images based on ScoreSDE and subsequently conducts posterior sampling for LDCT images. For both methods, we use their official codes and the same dataset as ours.

Results. We show the quantitative outcomes in Table 3 and the denoised images in Fig. 7. Our method outperforms ScoreSDE and achieves comparable performance to Noise2Sim, indicating the effectiveness of the robust CLIP
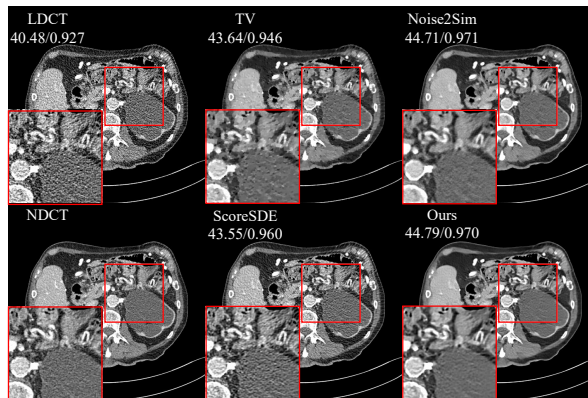

Figure 7. Denoising results on LDCT image. PSNR/SSIM are listed on the respective images. Display window: [-160, 240]HU.

Table 3. Quantitative comparison of different methods on low-dose CT abdomen slices. Note that Noise2Sim requires *adjacent LDCT images* for training, while we only demand *NDCT images*

TV	ScoreSDE [52]	Noise2Sim [45]	Ours
44.81/0.972	45.36/0.972	**45.98/0.978**	45.88/0.976

RN50 encoder in CT images. This holds substantial appeal as it enables direct knowledge transfer from models trained on natural images to medical domains, eliminating the need for, e.g., numerous adjacent low-dose CT images in Noise2Sim, especially beneficial in clinical scenarios.

4.5. Ablations

Here, we conduct ablation studies on synthetic noise to better evaluate the performance of our method.

Ablation on the use of noisy image and $\mathbf{F}^5$. Our model utilizes the noisy input as an extra dense feature to facilitate the transfer of image details and structures for restoration. Removing this operation significantly impacts the model's performance, as shown in the third row of Table 4. Without integrating the noisy image $\mathcal{I}_n$ into the decoder, the model's performance drastically declines, in both in-distribution and OOD cases. This highlights that although the dense features

Table 4. Effects of using the input image $\mathcal{I}_n$, features $\mathbf{F}^5$, different training types and progressive feature augmentation on synthetic OOD noise removal under Kodak24 dataset

	Gauss $\sigma = 15$	Gauss $\sigma = 50$	Sepckle $\sigma^2 = 0.04$
Baseline	34.69/0.922	26.87/0.692	30.60/0.871
w/o input $\mathcal{I}_n$	30.37/0.888	21.76/0.413	26.93/0.761
+ feature $\mathbf{F}^5$	34.03/0.916	25.61/0.645	30.39/0.865
+ *random*	34.89/0.925	20.87/0.342	29.61/0.821
+ *finetune*	34.91/0.926	19.03/0.276	27.37/0.694
+ PFA	**34.69/0.922**	**27.39/0.723**	**30.67/0.876**

Table 5. Effects of using different CLIP ResNet versions on synthetic OOD noise under Urban100 dataset (without PFA)

	Gauss $\sigma = 50$	Poisson $\alpha = 3.5$	Speckle $\sigma^2 = 0.04$
RN50	**25.86/0.743**	**27.11/0.837**	**28.61/0.871**
RN50x4	24.04/0.651	26.31/0.809	28.40/0.869
RN50x16	23.44/0.731	22.26/0.762	18.95/0.707

from the frozen CLIP RN50 exhibit favorable properties, they lack crucial details of the input image, rendering them ineffective for standalone restoration. In addition, utilizing feature $\mathbf{F}^5$, which does not hold distortion-invariant property as identified in Section 3.1, also leads to a decline in both in-distribution and OOD performance.

Ablation on training types. We evaluate two optional training types to optimize our model: (1) *random*, where the image encoder is randomly initialized and the entire network is trained from scratch, and (2) *finetune*, which initializes the image encoder using the pre-trained RN50 from CLIP and then finetunes the complete model. As observed in Table 4, both training types improve the in-distribution denoising performance since the denoiser begins to overfit the training noise. Consequently, the resultant models show poor robustness to OOD noise (also see Fig. 8) as the original property of the frozen RN50 has been broken.

Ablation on progressive feature augmentation (PFA). The inclusion of progressive feature augmentation is highlighted in the last row of Table 4. This strategy does not affect in-distribution performance but enhances the model's generalization to OOD noise, demonstrating its efficacy.

Ablation on ResNet versions. In Section 3.1, we observe that the first feature $\mathbf{F}^1$ from larger ResNet models of CLIP does not hold the distortion-invariant property. Here we ablate its effect on the generalization ability. We substitute the frozen RN50 model in Fig. 4 with frozen RN50x4 and RN50x16, respectively, and then train the corresponding denoisers. The outcomes, as depicted in Table 5, clearly indicate a substantial reduction in the generalization ability of denoisers under frozen RN50x4 and RN50x16. This underscores the significance of robust shallow features for achieving generalizable denoising.

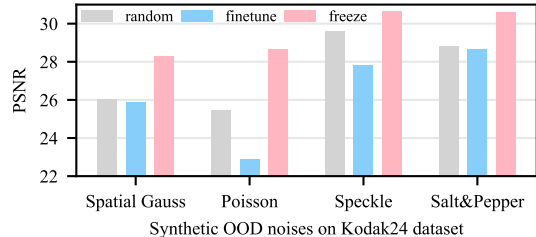More experiments on the computational cost (Table 6) and image deraining (Table 7) are provided in the supp.

Figure 8. Robustness of *random*, *fintune* and *freeze* (ours) to diverse OOD noises. Here, the highest OOD noise levels are used.

5. Discussions and Limitations

Why are dense features of CLIP RN50 encoder robust to noise? During the training of the CLIP model, the dense features of the CLIP image decoder might overlook high-frequency information within images, including noise and fine textures. Instead, they tend to prioritize image contents to align more effectively with text semantics. This hypothesis is reasonable as the text cannot provide dense and high-frequency details for images.

Is the robust property unique to CLIP RN50? We hypothesize that other well-established self-supervised representation learning may induce a similar property as CLIP. To validate this, we perform feature analysis and OOD experiments using the frozen RN50 pre-trained by MoCo-v3 [8], a prominent self-supervised pre-training method. The findings presented in Fig. 13 and Table 10 (in the Supplementary Material) suggest that a denoiser based on MoCo-v3 RN50 also exhibits a certain level of OOD robustness.

Transformer architectures for the denoiser. As discussed in Section 3.1 and verified in Table 8 of the supp, the CLIP ViT image encoder excessively downsamples image features, rendering it unsuitable for image restoration. We also attempt to implement a Transformer-based decoder by replacing the Conv-block with the Restormer-block. However, the resultant denoiser tends to overfit the training noise, leading to a decrease in its generalization. Additional strategies are needed to mitigate this overfitting.

6. Conclusion

This paper introduces a simple yet robust denoiser that can generalize to various OOD noises. Our method builds upon the discovery that the first four multi-scale dense features from CLIP frozen RN50 are distortion-invariant and content-related. By integrating these features and the noisy input image into the learnable image decoder, we construct a denoiser with generalization capabilities. Comprehensive experiments and comparisons on diverse OOD noises, including synthetic noise, real-world sRGB noise, and low-dose CT noise, demonstrate the superiority of our method.

References

[1] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1692–1700, 2018. 2, 5

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017. 5

[3] American Association of Physicists in Medicine. Low dose CT grand challenge. https://www.aapm.org/grandchallenge/lowdosect/, 2016. 5

[4] Joshua Batson and Loic Royer. Noise2self: Blind denoising by self-supervision. In *International Conference on Machine Learning*, pages 524–533. PMLR, 2019. 2

[5] Fei Chen, Lei Zhang, and Huimin Yu. External patch prior guided internal clustering for image denoising. In *Proceedings of the IEEE international conference on computer vision*, pages 603–611, 2015. 2

[6] Haoyu Chen, Jinjin Gu, Yihao Liu, Salma Abdel Magid, Chao Dong, Qiong Wang, Hanspeter Pfister, and Lei Zhu. Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1703, 2023. 1, 2, 5, 6, 7, 3

[7] Hao Chen, Chenyuan Qu, Yu Zhang, Chen Chen, and Jianbo Jiao. Multi-view self-supervised disentanglement for general image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12281–12291, 2023. 1, 2

[8] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 8

[9] Jun Cheng, Tao Liu, and Shan Tan. Score priors guided deep variational inference for unsupervised real-world single image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12937–12948, 2023. 3

[10] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on image processing*, 16(8):2080–2095, 2007. 4

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 3

[12] Wenchao Du, Hu Chen, and Hongyu Yang. Learning invariant representation for unsupervised image restoration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 14483–14492, 2020. 1, 2, 5

[13] Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, and Bo Dai. Generative diffusion prior for unified image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9935–9946, 2023. 3

[14] Rich Franzen. Kodak lossless true color image suite. http://r0k.us/graphics/kodak/, 1999. 5

[15] Shi Guo, Zifei Yan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1712–1722, 2019. 5

[16] 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. 3

[17] Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2015. 5

[18] Tao Huang, Songjiang Li, Xu Jia, Huchuan Lu, and Jianzhuang Liu. Neighbor2neighbor: Self-supervised denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14781–14790, 2021. 1, 2

[19] Junha Hyung, Sungwon Hwang, Daejin Kim, Hyunji Lee, and Jaegul Choo. Local 3d editing via 3d distillation of clip knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12674–12684, 2023. 1, 2

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

[21] Bahjat Kawar, Gregory Vaksman, and Michael Elad. Snips: Solving noisy inverse problems stochastically. *Advances in Neural Information Processing Systems*, 34:21757–21769, 2021. 3

[22] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. 2022. 3

[23] Kwanyoung Kim and Jong Chul Ye. Noise2score: tweedie's approach to self-supervised image denoising without clean images. *Advances in Neural Information Processing Systems*, 34:864–874, 2021. 2

[24] 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. *arXiv preprint arXiv:2304.02643*, 2023. 3

[25] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019. 2, 3

[26] Alexander Krull, Tim-Oliver Buchholz, and Florian Jug. Noise2void-learning denoising from single noisy images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2129–2137, 2019. 1, 2

[27] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photorealistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on*

computer vision and pattern recognition, pages 4681–4690, 2017. 2

[28] Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. Apbsn: Self-supervised denoising for real-world images via asymmetric pd and blind-spot network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17725–17734, 2022. 1, 2

[29] Pan Li, Da Li, Wei Li, Shaogang Gong, Yanwei Fu, and Timothy M Hospedales. A simple feature augmentation for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8886–8895, 2021. 4

[30] 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. 3

[31] Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, and Hao Su. Distilling large vision-language model with out-of-distribution generalizability. In *ICCV*, pages 2492–2503, 2023. 1, 2

[32] Xin Li, Bingchen Li, Xin Jin, Cuiling Lan, and Zhibo Chen. Learning distortion invariant representation for image restoration from a causality perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1714–1724, 2023. 1, 2, 5, 6, 7, 3

[33] Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, and Luc Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023. 1, 2, 3

[34] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 1, 2

[35] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844. IEEE Computer Society, 2021. 2, 3

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5

[37] Zhihe Lu, Zeyu Xiao, Jiawang Bai, Zhiwei Xiong, and Xinchao Wang. Can sam boost video super-resolution? *arXiv preprint arXiv:2305.06524*, 2023. 3

[38] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, pages 23033–23044. PMLR, 2023. 2

[39] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Controlling vision-language models for universal image restoration. *arXiv preprint arXiv:2310.01018*, 2023. 3

[40] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 416–423. IEEE, 2001. 5

[41] Sreyas Mohan, Zahra Kadkhodaie, Eero P Simoncelli, and Carlos Fernandez-Granda. Robust and interpretable blind image denoising via bias-free convolutional neural networks. In *International Conference on Learning Representations*, 2019. 1, 2

[42] Sreyas Mohan, Joshua L Vincent, Ramon Manzorro, Peter Crozier, Carlos Fernandez-Granda, and Eero Simoncelli. Adaptive denoising via gaintuning. *Advances in neural information processing systems*, 34:23727–23740, 2021. 2

[43] Seonghyeon Nam, Youngbae Hwang, Yasuyuki Matsushita, and Seon Joo Kim. A holistic approach to cross-channel image noise modeling and its application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1683–1691, 2016. 5

[44] Reyhaneh Neshatavar, Mohsen Yavartanoo, Sanghyun Son, and Kyoung Mu Lee. Cvf-sid: Cyclic multi-variate function for self-supervised image denoising by disentangling noise from image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17583–17591, 2022. 1, 2, 6, 7

[45] Chuang Niu, Mengzhou Li, Fenglei Fan, Weiwen Wu, Xiaodong Guo, Qing Lyu, and Ge Wang. Noise suppression with similarity-based self-supervised deep learning. *IEEE Transactions on Medical Imaging*, 42(6):1590–1602, 2023. 7

[46] Tongyao Pang, Huan Zheng, Yuhui Quan, and Hui Ji. Recorrupted-to-recorrupted: unsupervised deep learning for image denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2043–2052, 2021. 2

[47] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[48] 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. 1, 2

[49] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 1, 2

[50] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions*

on Pattern Analysis and Machine Intelligence, 45(4):4713–4726, 2022. 2

[51] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021. 1

[52] Yang Song, Liyue Shen, Lei Xing, and Stefano Ermon. Solving inverse problems in medical imaging with score-based generative models. In *International Conference on Learning Representations*, 2022. 7

[53] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 4

[54] Zichun Wang, Ying Fu, Ji Liu, and Yulun Zhang. Lg-bpn: Local and global blind-patch network for self-supervised real-world denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18156–18165, 2023. 1, 2

[55] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21970–21980, 2023. 2

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

[57] Jun Xu, Hui Li, Zhetong Liang, David Zhang, and Lei Zhang. Real-world noisy image denoising: A new benchmark. *arXiv preprint arXiv:1804.02603*, 2018. 5

[58] Hanshu Yan, Jingfeng Zhang, Jiashi Feng, Masashi Sugiyama, and Vincent Y. F. Tan. Towards adversarially robust deep image denoising. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 1516–1522, 2022. 1, 2, 5, 6, 7, 3

[59] Lijun Yu, Yong Cheng, Zhiruo Wang, Vivek Kumar, Wolfgang Macherey, Yanping Huang, David A Ross, Irfan Essa, Yonatan Bisk, Ming-Hsuan Yang, et al. Spae: Semantic pyramid autoencoder for multimodal generation with frozen llms. *arXiv preprint arXiv:2306.17842*, 2023. 3

[60] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *arXiv preprint arXiv:2111.09881*, 2021. 1, 2, 3, 5, 6

[61] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14821–14831, 2021. 3

[62] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 1, 2, 5, 6, 3

[63] 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, 2, 3

[64] Lei Zhang, Xiaolin Wu, Antoni Buades, and Xin Li. Color demosaicking by local directional interpolation and nonlocal adaptive thresholding. *Journal of Electronic imaging*, 20(2): 023016–023016, 2011. 5

[65] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6868–6874, 2022. 1

[66] Y Zhang, K Li, B Zhong, and Y Fu. Residual non-local attention networks for image restoration. In *International Conference on Learning Representations*, 2019. 2

[67] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on computational imaging*, 3(1):47–57, 2016. 2

[68] Dihan Zheng, Xiaowen Zhang, Kaisheng Ma, and Chenglong Bao. Learn from unpaired data for image restoration: A variational bayes approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 6, 7

[69] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 1, 2

[70] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[71] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185, 2023. 1, 2

[72] Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1229, 2023. 3