Implicit Prompt Learning for Image Denoising

Yao Lu[1], **Bo Jiang**[2*], **Guangming Lu**[1*] and **Bob Zhang**[3]
[1]Department of Computer Science, Harbin Institute of Technology at Shenzhen, China
[2]College of Mechanical and Electronic Engineering, Northwest A&F University, China
[3]Department of Computer and Information Science, University of Macau, China
luyao2021@hit.edu.cn, jiangbo_PhD@gmail.com, luguangm@hit.edu.cn, bobzhang@um.edu.mo

Abstract

Recently, various deep denoising methods have been proposed to solve the insufficient feature problem in image denoising. These methods can be mainly classified into two categories: (1) Injecting learnable tensors into denoising backbone to supplement feature, which is effective to some extent but may cause serious over-fitting. (2) Using diverse natural images from large image datasets to synthesize noisy images and pre-train denoising models, which can bring model generalization but require large model size and expensive training costs. To address these issues, this paper proposes Implicit Prompt Learning for Image Denoising (IPLID) method to flexibly generate adaptive prompts without meticulously designing them. Specifically, we first introduce an efficient Linear Prompt (LP) block with ultra-few parameters to produce dynamic prompts for both different stages and samples in denoising procedure. We further propose an efficient Compact Feature Fusion (CFF) block to process previous multi-level prompted denoising feature to reconstruct the denoising images. Finally, to further efficiently and effectively produce satisfactory prompt and denoising performance, a Gradient Accumulation (GA) learning scheme is proposed. Experiments on multiple benchmarks showed that the proposed IPLID achieves competitive results with only 1 percent of pre-trained backbone parameters, outperforming classical denoising methods in both efficiency and quality of restored images.

1 Introduction

Image denoising is a fundamental low-level vision task, which aims to reconstruct a noise-free image from a noisy image. Meanwhile, it is also a challenging inverse problem task [Jiang *et al.*, 2022a]. To elegantly address the inverse problem, currently, the end-to-end image denoising methods using **C**onvolutional **N**eural **N**etworks (CNNs) and Transformer structures [Liang *et al.*, 2021; Zamir *et al.*, 2021] solve this problem by learning the mapping between noisy and clean images. However, due to the limited training data, the reconstructed images may have insufficient feature, resulting in the excessive smoothness of denoised images.

For the past few years, several methods have been proposed for solving the problem of insufficient feature in the process of image denoising. These methods are mainly classified into two categories. (a) Learnable tensors are inserted into the Deep Neural Networks (DCNNs) for image denoising to supplement the feature within the image denoising process, which is an effective method to some extent. However, the information supplemented by this method is uncontrollable. Hence, it may be prone to over-fitting. For instance, APD-Net [Jiang *et al.*, 2022b] supplements the denoised images with prior information, while may cause the over-fitting. To effectively suppress over-fitting, APD-Net equips a corresponding regularization module. This results in the complex structure of the overall denoising network. Therefore, such methods of adding learnable tensors are inevitable to establish a difficult trade-off between the performance and the complexity of the DNNs for denosing. (b) Since the ImageNet dataset contains more than 1 million natural images from 1,000 different categories, which are highly diverse. Therefore, these images are used to manually synthesize noisy images using different noise levels to pre-train the denoising model to provide the initialization weights with abundant texture and color feature [Chen *et al.*, 2021]. Then, the denoising model is fine-tuned on the widely used image denoising training datasets (*e.g.*, in total only 1200 images on DIV2K and BSD400 datasets) to achieve image denoising performance correction.

Although these methods can bring generalization capabilities to image denoising models, they severely suffer from two main problems. **(a)** Compared to the methods that are directly trained on the noisy image datasets, using the ImageNet dataset to manually synthesize noisy images to pre-train the model introduces additional training cycles, resulting in greatly expensive training costs. **(b)** Due to the large number of images within the ImageNet datasets, the pre-trained model trained on such large dataset should have enough capacity to retrieve the abundant feature. This will introduce a huge number of model parameters and calculations in the fine-tune process.

In Natural Language Processing (NLP) tasks, prompt learning makes better use of the knowledge from the pre-

*Corresponding author

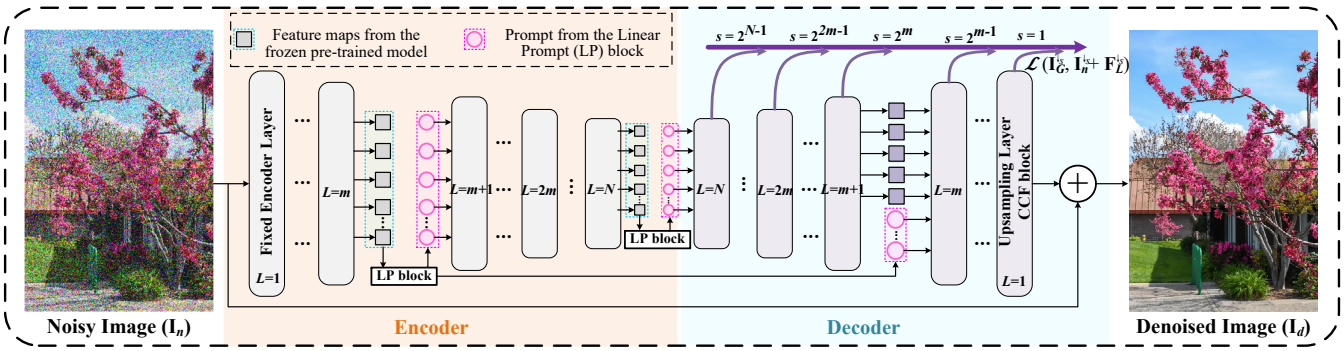

Figure 1: Overall structure of the proposed IPLID. Pink circles represent prompted denoising feature, and the gray squares represent feature maps generated by the pre-trained model.

trained model by adding additional texts to the input [Lester *et al.*, 2021]. Since prompt learning uses a pre-trained model with frozen weights, such a strategy does not retrain a pre-trained text model for downstream tasks, and can greatly shorten the training time and effectively improve the efficiency of the model. Inspired by the prompt mechanism, this paper applies the powerful prompt learning to image denoising task based on the pre-trained vision models. However, there are still some challenges in designing an effective and efficient prompt learning for image denoising.

Current prompt for pure vision task is still unexplored, and the noise within the real-word noisy images is very complex. Furthermore, different input noisy samples contain various noise. Therefore, ***(1) it is challenge to efficiently produce appropriate and individulized prompt for denoising pre-trained model without explicitly designing prompt scheme.*** In the stage of reconstructing denosing images, since the denoising feature generated using the denoising prompt has multiple levels, ***(2) it is also critical to efficiently fuse such multi-level prompted denoising information with effect to produce high-quality denoised images.*** In the learning process, due to the frozen pre-trained model in the entire denoising framework, the injected learnable blocks may not obtain satisfactory gradient. Thus, ***(3) it is also vital to effectively optimize the learnable blocks employing more efficacious learning scheme.***

To overcome above challenges, we propose the **I**mplicit **P**rompt **L**earning for **I**mage **D**enoising (**IPLID**), as shown in Fig. 1. Formally, different from explicit prompt learning, the implicit prompt learning generates prompt information more flexibly and conveniently without meticulously designing them. It is not necessary to feed hints and sample data together in the pre-trained model. The prompt information only needs to be generated according to the previous feature layer and adaptively acts on the subsequent layers. Specifically, we first propose the **L**inear **P**rompt (**LP**) block for image denoising, which is efficiently equipped in the pre-trained vision model with ultra-few training parameters. It is powerful enough to produce satisfactory prompt by transforming the feature retrieved from the pre-trained vision model. Additionally, we further propose an efficient and flexible **C**ompact **F**eature **F**usion (**CFF**) block as a decoder to comprehensively process the multi-level prompted denoising feature to reconstruct high-quality denoised images. Eventually, to effectively optimize the learnable modules, we propose the **G**radient **A**ccumulation (**GA**) learning scheme by introducing multi-stage losses to update the trainable parameters. Extensive experimental results show that the proposed IPLID can produce competitive performance using only 1% of the backbone parameters, greatly outperforming traditional deep denoising methods in both the quality and vision of restored images with higher efficiency. The contributions of this paper can be summarized as follows:

- We propose the **I**mplicit **P**rompt **L**earning for **I**mage **D**enoising (**IPLID**) to implicitly and flexibly produce adaptive prompt information for pre-trained vision model without meticulously design.
- We propose the **L**inear **P**rompt (**LP**) block, which is light-weight but powerful enough to produce satisfactory prompt using the retrieved pre-trained feature for different noisy samples and different stages within denoising procedure.
- We propose the **C**ompact **F**eature **F**usion (**CFF**) block, which is also efficient as a decoder to sufficiently retrieve multi-level prompted denoising feature to recover high-quality denoised images.
- We propose the **G**radient **A**ccumulation (**GA**) learning scheme to effectively update the trainable parameters within frozen pre-trained model. Extensive experiments validated the effectiveness and superiority of our IPLID.

2 Related Works

2.1 Image Denoising

With the development of CNNs, many efficient image denoising models based on CNNs have emerged, which significantly improve the performance of the image denoising task. Examples include DnCNN [Zhang *et al.*, 2017], which introduces a residual learning method to reconstruct noise maps that are then subtracted from noisy images. MWCNN [Liu *et al.*, 2018] introduces a new form of down-sampling and up-sampling layers in the discrete wavelet domain. Blind noise settings can be handled with FFDNet [Zhang *et al.*, 2018]

by employing a customized noise level map. Although these CNNs-based models improve image denoising to some extent, the vanilla convolution layer has a limitation in capturing long-range pixel dependencies.

As an alternative to CNNs, transformers [Kolesnikov *et al.*, 2021] can capture dependencies over long-range patches using a self-attention mechanism. Examples of transformer-based models for image denoising include Uformer [Wang *et al.*, 2021], SwinIR [Liang *et al.*, 2021], Restormer [Zamir *et al.*, 2021], and SCUNet [Zhang *et al.*, 2022]. Particularly, IPT [Chen *et al.*, 2021] and EDT [Li *et al.*, 2021] are the prior denoising neural networks built with transformer blocks, but require pre-training on large-scale datasets to provide the initialization weights with rich texture and color feature for the process of image denoising. However, this not only introduces a large number of model parameters and calculations in the fine-tune process, but also may limit the further improvement of image denoising performance due to the designed complex pre-trained structure.

2.2 Prompt Learning

The prompt learning is widely used in the field of Natural Language Processing (NLP) and has a large potential application to computer vision [Liu *et al.*, 2023]. In NLP, a prompt is a pre-defined statement or sequence of words that serves as a starting point for generating text [Fan *et al.*, 2017]. By reformulating NLP tasks as the text completion problems, it is possible to solve a variety of NLP tasks without fine-tuning the model on specific datasets [Su *et al.*, 2021]. This approach, known as prompting, has been shown to be effective in solving a range of NLP tasks and benchmarks. More recently, different approaches to prompting have been developed, including Prompt Engineering, which involves designing prompts that are optimized for specific tasks, and Prompt Ensembling, which uses a combination of multiple prompts to improve performance [Wang *et al.*, 2023]. Another Prompt Prefix Tuning involves adjusting the prefix of a prompt to optimize performance of specific task [Zhou *et al.*, 2022].

Despite prompt learning has been successful in NLP, it has not been widely explored in the field of computer vision. However, the idea of using prompt has a anticipating application in various tasks such as image denoising. This paper tries to explore the prompt potentiality in the image denoising and propose a prompt scheme to produce the appropriate prompt and employ them in image denoising.

3 Method

3.1 Overall Pipeline

As shown in Fig. 1, the overall framework of the proposed IPLID is constructed based on classical but powerful encoder-decoder pipeline, where the encoder is composed of a pre-trained model with frozen weights and equipping with **L**inear **P**rompt (LP) blocks, while the decoder is composed of **C**ompact **F**eature **F**usion (CFF) blocks and up-sampling layers. The input noisy image is first fed into the pre-trained model with frozen weights for extracting initial feature. Specifically, given a noisy input image $\mathbf{I}_n \in \mathbb{R}^{H\times W\times 3}$, the low-level feature map $\mathbf{F}_s \in \mathbb{R}^{H\times W\times C}$ is extracted by the first layer of the pre-trained model:

$$\mathbf{F}_s = \mathcal{F}_i\left(\mathbf{I}_n\right),\ i = 1, \tag{1}$$

where $\mathcal{F}_i$ denotes the i^{th} layer of the pre-trained model with frozen weights. C, H and W denote the numbers of channels, height and width of the noisy image $\mathbf{I}_n$, respectively.

Then, the captured feature maps are then fed into the LP block to generate prompt information, which is then added to the input feature maps to form the prompted denoising feature maps, as shown in Eqn. 2:

$$\mathbf{F}_p = \mathscr{H}\left(\mathbf{F}_s\right) + \mathbf{F}_s, \tag{2}$$

where, $\mathscr{H}$ represents the operation of the LP block, and $\mathbf{F}_p$ is the generated prompted feature maps. The above process constitutes one unit of the encoder. In the proposed IPLID, the encoder is composed of four sets of such units.

Similarly, the image prompted denoising feature with different levels generated by the encoder is fed into the decoder. The decoder consists of four sets of upsampling layers and CFF blocks. The decoder gradually recovers high-resolution image feature from the low-resolution prompted denoising feature maps $\mathbf{F}_{pe} \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 16C}$ (*i.e., the output prompted denoised feature at the end of decoder*) as they pass through the upsampling layers. During the upsampling process, a deconvolution operation with a stride of 2 and a kernel size of 2×2 is employed to expand the reconstructed denoised image, concomitantly decreasing the number of channels and augmenting the spatial resolution of the feature maps. Finally, the skip connection is employed to connect the generated prompted denoising feature between the encoder and decoder with same scales, formulated in Eqn. 3:

$$\mathbf{F}_d = \mathscr{C}\left(cat\left(\mathscr{U}^i\left(\mathbf{F}_{pe}\right),\ \mathbf{F}_p^i\right)\right), \tag{3}$$

where $\mathscr{U}$, cat, and $\mathscr{C}$ denote the up-sampling operation, concatenation operation along the channel dimension of tensors, and CFF block operation, respectively. $\mathbf{F}_p^i$ denotes the prompted denoising feature maps generated in the i^{th} encoder, and $\mathscr{U}^i\left(\mathbf{F}_{pe}\right)$ indicates the feature maps generated in the decoder with same size to $\mathbf{F}_p^i$. Finally, the feature maps $\mathbf{F}_{last} \in \mathbb{R}^{H\times W\times 3}$ generated from the last layer within decoder are element-wise added to the input noisy image for final denoised image $\mathbf{I}_d$, as shown in Eqn. 4:

$$\mathbf{I}_d = \mathbf{F}_{last} + \mathbf{I}_n. \tag{4}$$

3.2 Linear Prompt (LP) Block

To implicitly generate adaptive prompts for different denoised samples and different stages in the denoising procedure, an efficient transformation is introduced. For arbitrary feature maps $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ generated by a frozen pre-trained vision model *ResNet-152* [He *et al.*, 2016], we can generally use the transformation $\mathcal{P}$ to produce the prompt $\mathbf{F}_p$, shown as the following equation:

$$\mathbf{F}_p = \mathcal{P}\left(\mathbf{X}\right). \tag{5}$$

To preserve the high efficiency of the denoising process, the transformation $\mathcal{P}$ employs *separate linear transformation*. Thus, Eqn. 5 can be further formulated as follows:

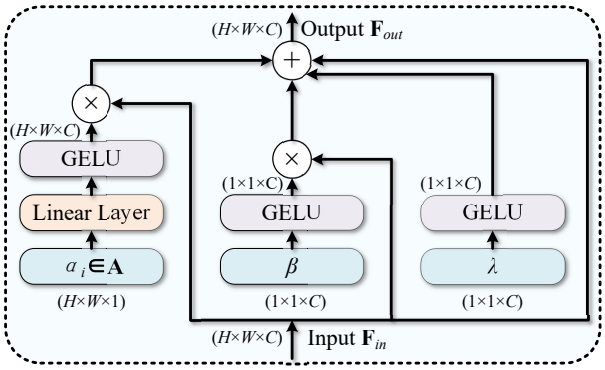

Figure 2: Structure of the proposed **L**inear **P**rompt (LP) block.

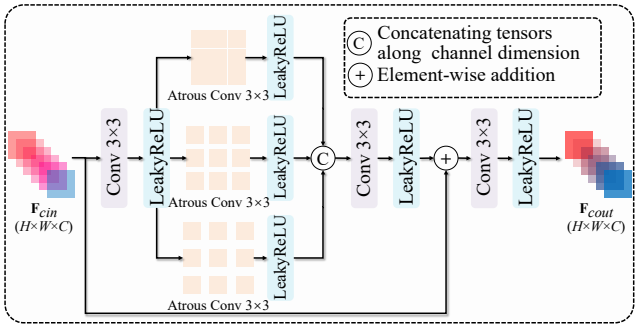

Figure 3: Structure of the proposed **C**ompact **F**eature **F**usion (CFF) block.

$$\mathbf{F}_p = \mathbf{W} \cdot (\mathbf{X}) + \beta \cdot (\mathbf{X}) + \lambda, \tag{6}$$

where $\mathbf{W} \in \mathbb{R}^{H\times W\times C}$, $\beta \in \mathbb{R}^{C\times 1\times 1}$, and $\lambda \in \mathbb{R}^{C\times 1\times 1}$, and $\cdot$ is the element-wise multiplication to decrease the computation complexity. Hence, $\mathbf{W}$ and β separately process the input on all dimensions and channel dimension, respectively. To further promote the hierarchical richness of the produced prompt information and decrease the number of parameters within the proposed LP block, $\mathbf{W}$ is reformulated to follows:

$$\mathbf{W} \approx \sum_{i=0}^{C-1} T_l^{\theta}(\alpha_i), \tag{7}$$

where T_l^{θ} is the general linear transformation with parameters $\theta \in \mathbb{R}^{1\times 1\times C}$, $\alpha_i \in \mathbb{R}^{H\times W\times 1}$ ($\alpha_i \in \mathbf{A}$, $\mathbf{A} \in \mathbb{R}^{H\times W\times C}$), and $\beta \in \mathbb{R}^{1\times 1\times C}$. Thus, the parameters decreases from $C \times H \times W + 2C$ to $C \times H + 3C$. Above transformation can make a better trade-off between the prompt richness and efficiency.

According to the above strategies, it is critical to design suitable transformation coefficients α_i, bias term β, and scalar offset λ. Thus, these necessary coefficients can be learned and optimized in training. Based on these principles, the **L**inear **P**rompt (LP) block with ultra-few parameters to generate adaptive denoising prompt is constructed in Fig. 2.

In detail, the proposed LP block consists of three learnable parameters, three GELU activation layers [Hendrycks and Gimpel, 2016], and one linear mapping function layer. Specifically, given the extracted feature maps $\mathbf{F}_{in} \in \mathbb{R}^{H\times W\times C}$ by the pre-trained model with frozen weights, according to the Eqn. 7, the proposed LP block generates the denoised prompt information as the following formulation:

$$\mathrm{F}_{out} = \varphi\left(\textstyle\sum_{i=0}^{C-1} T_l^{\theta}(\alpha_i)\right) \cdot \mathbf{F}_{in} + \varphi(\beta) \cdot \mathbf{F}_{in} + \varphi(\lambda) + \mathbf{F}_{in}, \tag{8}$$

where $\mathbf{F}_{out} \in \mathbb{R}^{H\times W\times C}$ is the prompted denoising feature map, φ denotes the GELU activation function.

3.3 Compact Feature Fusion (CFF) Block

To fully utilize guidance prompted denoising feature maps for reconstructing denoised images, we propose an elegant and efficient **C**ompact **F**eature **F**usion (CFF) block to build the decoder. The decoder not only effectively fuses multi-level guidance prompted denoising feature maps, but also strives to be as efficient as possible, as shown in Fig. 3. The CFF block is composed of convolutional layers with a kernel size of 3×3, LeakyReLU activation function, and dilated convolutional layers (with dilation rates of $d = 1$, 2, and 3). As multi-level prompted denoising feature has various scales, the higher levels correspond to coarser scales and vice versa. Thus, the effective multi-level learning is necessary for sufficient feature fusion. Therefore, the CFF block employs a multi-level learning strategy to enhance this fusion process.

Specifically, the input feature map is refined through convolutional and activation layers, as shown in Eqn. 9:

$$\mathbf{F}_0 = \psi\left(f_{conv}^{3\times 3}\left(\mathbf{F}_{cin}\right)\right), \tag{9}$$

where $\mathbf{F}_0 \in \mathbb{R}^{H\times W\times C}$ is the refined feature map, ψ denotes the LeakyReLU activation function, and f_{conv} is the convolutional layers with a kernel size of 3×3. To increase the receptive field, the CFF block employs dilated convolutions with a 3×3 kernel, which can expand the effective receptive field of the convolutional kernel without increasing the number of parameters, thereby capturing a wider range of contextual information. As a result, the refined feature map is obtained through dilated convolutions to obtain a broader range of contextual feature, as shown in Eqn. 10:

$$\mathrm{F}_1 = cat\left(\psi\left(f_{d=1}^{3\times 3}\left(\mathbf{F}_0\right)\right), \psi\left(f_{d=2}^{3\times 3}\left(\mathbf{F}_0\right)\right), \psi\left(f_{d=3}^{3\times 3}\left(\mathbf{F}_0\right)\right)\right), \tag{10}$$

where $\mathbf{F}_1 \in \mathbb{R}^{H\times W\times 3C}$ is the mixed feature map, $f_{d=1}^{3\times 3}$, $f_{d=2}^{3\times 3}$, and $f_{d=3}^{3\times 3}$ correspond to atrous convolutions with dilation rates of $d = 1$, $d = 2$, and $d = 3$, respectively. Then, the output feature map $\mathbf{F}_{out} \in \mathbb{R}^{H\times W\times C}$ of the CFF block can be expressed as follows:

$$\mathbf{F}_{out} = \psi\left(f_{\text{conv}}^{3\times 3}\left(\psi\left(f_{\text{conv}}^{3\times 3}\left(\mathbf{F}_1\right)\right) + \mathbf{F}_{cin}\right)\right), \tag{11}$$

Eqn. 11 employs the $\mathbf{F}_1$ mixed contextual information to readjust the fused feature maps.

3.4 Gradient Accumulation Learning Scheme

Owing to the frozen weights of the pre-trained vision model in the learning process, the learnable LP block injected into the pre-trained model may not achieve enough satisfactory and dynamic optimization. To provide more effective prompt information from pre-trained feature with frozen weights layer by layer, we propose a gradient accumulation learning

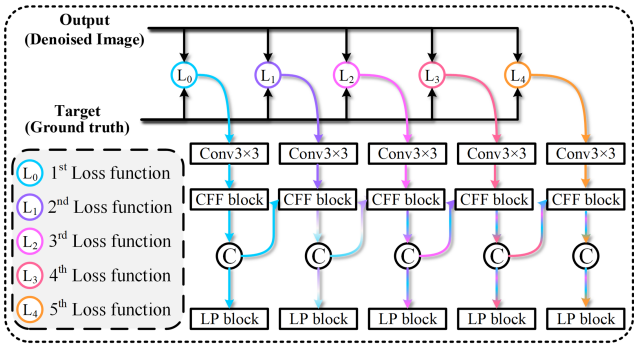

Figure 4: Schematic diagram of updating prompt parameters using gradient accumulation.

scheme to effectively update the parameters of the LP block, which is shown in Fig. 4.

In every level of reconstructed denoised image with different scales, a loss supervision will be exerted. Specifically, a 3×3 convolutional layer is connected after each CFF block to generate the denoised image of the corresponding size. To obtain the ground-truth image of corresponding size, we use bilinear interpolation to calculate the loss at different levels (scales), which is calculated as shown in Eqn. 12:

$$\mathcal{L} = \frac{1}{S+1}\frac{1}{N}\sum_{s=0}^{S}\sum\sqrt{\|\mathbf{I}_d^s - \mathbf{I}_{gt}^s\|^2 + \varrho^2}. \qquad (12)$$

where N denotes the number of training samples, s is the downsampling factor ($s = 0, 1, \cdots, S$), which denotes the loss calculation performed at a scale of image size $\frac{1}{2^s}$. $\mathbf{I}_d$ denotes the denoised image produced from the proposed IPLID, $\mathbf{I}_{gt}$ represents the ground-truth corresponding to the input noise image, and ϱ^2 is a constant that is empirically set to 1×10^{-6}.

By updating the parameters of the LP blocks layer by layer through backward gradient, the IPLID model can gradually adapt to noisy images. This allows the model to first learn more general lower-level representations before moving on to higher-level representations. This can improve the model's ability to transfer knowledge from pre-trained feature to the task of reconstructing denoised images.

4 Experiments

We use the Adam [Loshchilov and Hutter, 2017] optimizer to train our IPLID with setting β_1 and β_2 to 0.9 and 0.999, respectively. The learning rate is set to 1×10^{-4} in training.

4.1 Architecture Scales

To illustrate the effectiveness of the proposed IPLID architecture scale in image denoising, our experiments used three different scale parameters for IPLID. Therefore, we adjust different IPLID architectural scales by changing the number of feature channels C, and other settings remain unchanged. The three specific parameters of different scales are set as follows: IPLID-T (Tiny, $C = 16$), IPLID-S (Small, $C = 32$), and IPLID-B (Basic, $C = 64$).

Dataset	SIDD		Nam		PolyU	
Methods	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
DnCNN-B	38.56	0.910	36.08	0.903	35.74	0.878
FFDNet	38.60	0.909	37.85	0.938	37.19	0.939
TWSC	35.89	0.838	38.37	0.952	37.63	0.954
CBDNet	38.68	0.909	38.51	0.957	37.85	0.956
RIDNet	38.71	0.913	38.72	0.960	38.07	0.957
VDN	39.29	0.911	39.16	0.965	38.43	0.960
GCDN	38.93	0.910	38.96	0.962	38.21	0.958
PAN-Net	39.33	0.912	40.18	0.978	39.91	0.971
AINDNet	39.45	0.915	39.21	0.966	38.78	0.963
APD-Nets	39.75	0.959	40.36	0.989	N/A	N/A
MIRNet	39.71	0.959	39.88	0.973	39.25	0.971
HPDNet	39.72	0.958	40.26	0.979	39.89	0.970
Uformer	39.77	0.959	N/A	N/A	N/A	N/A
Restormer	40.02	0.960	N/A	N/A	N/A	N/A
IPLID-T	**39.73**	**0.959**	**39.80**	**0.971**	**39.72**	**0.966**
IPLID-S	**39.88**	**0.959**	**39.92**	**0.975**	**39.81**	**0.968**
IPLID-B	**40.05**	**0.960**	**40.39**	**0.989**	**39.92**	**0.971**

Table 1: Average PSNRs and SSIMs of the denoised real noisy images from Nam, PolyU and SIDD datasets. The values of PSNRs and SSIMs are positively correlated with visual quality.

4.2 Evaluation on Real-world Noisy Images

This section focuses on evaluating the effectiveness of the proposed IPLID in dealing with real-world noisy images with complex and unknown sources, which is crucial for practical applications. Table 1 presents the denoising results of the proposed IPLID on real noisy images from the SIDD [Abdelhamed *et al.*, 2018], PolyU [Xu *et al.*, 2018], and Nam [Nam *et al.*, 2016] datasets. The performance of the proposed IPLID is compared with that of fourteen state-of-the-art denoising methods. The obtained results indicate that the proposed IPLID generally outperforms the other fourteen state-of-the-art methods in terms of PSNR/SSIM on all three real noisy image datasets. Notably, on all real noisy image datasets, IPLID-B surpasses MIRNet with an PSNR of 1.52 dB, which confirms the effectiveness of the proposed IPLID structure in denoising real noisy images. These results further suggest that IPLID can more effectively remove complex real-world noise and produce superior denoising performance compared to all other methods.

To visually demonstrate the superior performance of the proposed IPLID method, we present a comparison of different denoising methods on various datasets using real noisy images in Figure 5. The results clearly show that our IPLID outperforms all other methods in terms of noise removal and detail preservation. Notably, Uformer and Restormer fail to preserve the fine details of the letterform, whereas our proposed IPLID-B successfully reconstructs and preserves the letterform. This exemplifies the efficacy of generating the prompted denoising feature and seamlessly fusing feature across multiple levels using multi-scale learning, which facilitates complex noise removal while preserving crucial details, thereby enhancing the image denoising performance both quantitatively and qualitatively.

Efficiency Comparison. This section is dedicated to contrasting our IPLID with the most recent cutting-edge image denoising methods in terms of their denoising efficiency. To ensure equitable comparisons of efficiency, we utilize FLOPs,

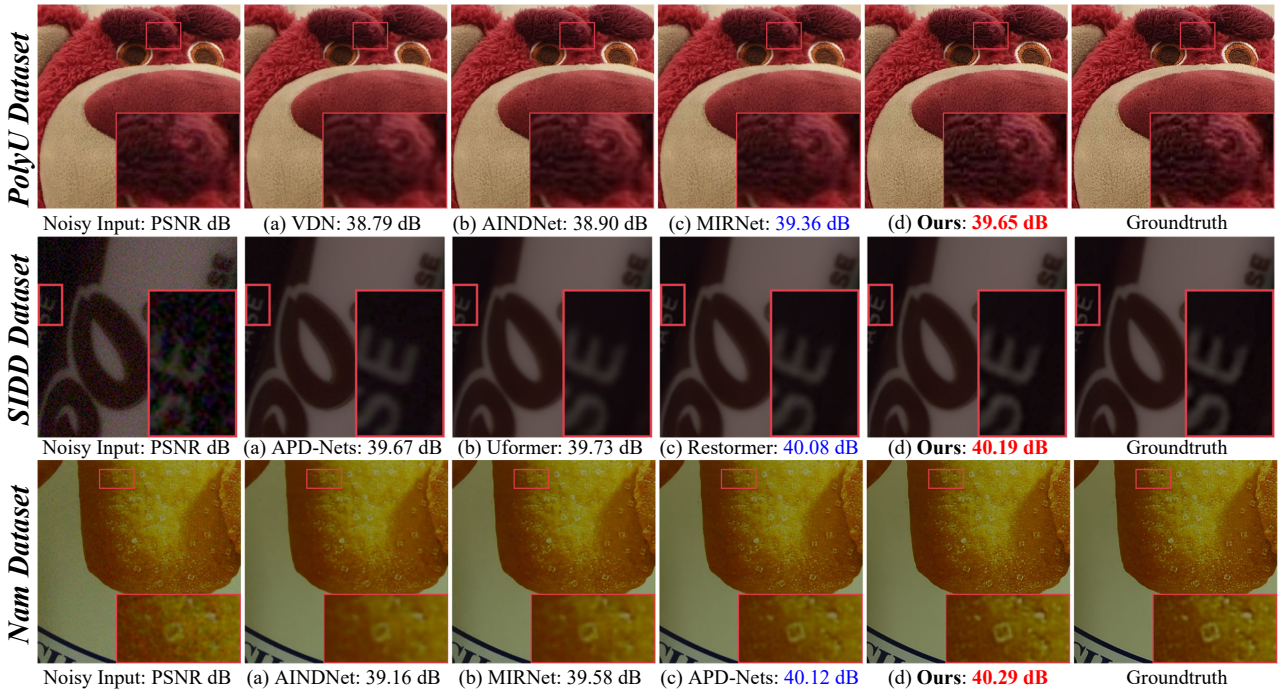

Figure 5: Visual comparisons between the proposed method and its competitors in the evaluation of real noisy image denoising.

Method	DRUNet	SwinIR	Uformer	SCUNet	Restormer	IPLID-T(Ours)	IPLID-S(Ours)	IPLID-B(Ours)
Runtime	0.092 s	0.132 s	0.254 s	0.098 s	0.153 s	**0.014 s**	0.016 s	0.018 s
FLOPs	143.63 G	503.95 G	84.76 G	79.84 G	140.9 G	**23.75G**	24.88G	28.98G
Params.	32.64 M	7.77 M	39.56 M	17.95 M	26.13 M	**0.70 M**	2.08 M	6.84 M

Table 2: Runtime, FLOPs, and trainable Parameters. comparisons on image sizes of 256×256 on an Nvidia RTX Titan GPU with state-of-the-art methods. The results from all the compared methods in this table are obtained by inferring an image with the size of 256×256 on the same GPU device (*i.e.*, an Nvidia RTX Titan GPU).

α_i		✗	✗	✗	✓	✗	✓	✓	✓
β		✗	✗	✓	✗	✓	✗	✓	✓
λ		✗	✓	✗	✗	✓	✓	✗	✓
↑ PSNR(in dB)	SIDD	38.95	39.53	39.69	39.74	39.86	39.92	39.97	40.06
	CBSD68	33.73	34.02	34.26	34.28	34.35	34.39	34.47	34.51

Table 3: Parameter effect of the LP Block on image denoising.

inference time, and trainable parameters as metrics. In particular, we perform the comparisons on identical computer equipment (*i.e.*, an Nvidia RTX Titan GPU) for efficiency. The results of image inference testing on a 256×256 image size, using the same GPU device across all compared methods, are reflected in the data presented in Table 2. For all the compared methods, the proposed IPLID stands out with the lowest FLOPs, which is in stark contrast to DRUNet. Despite its self-attention mechanism, SCUNet yields high FLOPs and long inference times. SwinIR and Uformer also exhibit intricate network structures that result in high FLOPs and long inference time. In contrast, our IPLID strikes an optimal balance between the performance and inference time with ultra-few learnable parameters (only about 1% parameters of pre-trained ResNet-152, *i.e.*, 0.7M Vs. 60.19M), making it stand-out in both efficiency and denoising performance.

4.3 Ablation Study

This section is to conduct a thorough examination and analysis of the image denoising effectiveness of the proposed IPLID, with particular emphasis on the ensuing three fundamental aspects: (1) the impact of the LP block, as implemented in the proposed IPLID framework, on the augmentation of image denoising performance. (2) The effect of the CFF block, as incorporated in the proposed IPLID framework, on the improvement of image denoising performance. (3) The validation of the IPLID learning scheme, *i.e.*, the impact of the layer-by-layer parameter updation with backward gradient accumulation. Our empirical experiment and analyses are conducted on real noisy datasets.

Ablation Study of the LP Block. From Table 3, it shows the performance effect of the learnable parameters including coefficient α_i, bias term β, and scalar offset λ in the LP block on image denoising. The results demonstrate that adding learnable parameters to the proposed LP block, including the transform coefficient α_i, bias term β, and scalar

dilation ratio dr = 1		✔	✗	✗	✗
dilation ratio dr = 2		✗	✔	✗	✗
dilation ratio dr = 3		✗	✗	✔	✗
CFF block		✗	✗	✗	✔
PSNR (in dB)	SIDD	39.62	39.70	39.75	40.06
	CBSD68	34.12	34.18	34.21	34.51

Table 4: Impact of CFF on the performance of image denoising.

L_0		✗	✗	✗	✗	✔
L_1		✗	✗	✗	✔	✔
L_2		✗	✗	✔	✔	✔
L_3		✗	✔	✔	✔	✔
L_4		✔	✔	✔	✔	✔
↑ PSNR(in dB)	SIDD	39.84	39.87	39.93	39.99	40.06
	CBSD68	34.39	34.42	34.46	34.49	34.51

Table 5: Impact of layer-wise parameter updation with backward gradients on the performance of image denoising.

offset λ, generally leads to improving performance in image denoising. The best performance is achieved when all learnable parameters are utilized. This indicates that the learnable parameters in the LP block can effectively generate prompted denoising feature and improve image denoising performance.

Ablation Study of the CFF Block. Since three branches are widely used to retrieve and fuse multi-level feature in vision tasks, we only conduct an ablation study on the dilation rates of the dilated convolutions within CFF block. To investigate the impact of dilation rates on image denoising performance, we set the dilation rates d of the dilated convolutions in all three branches to be the same. From Table 4, when different dilation rates in the CFF block are replaced with the same dilation rate (*i.e.*, the dilation rates d are the same in all branches), the image denoising performance significantly decreases. On the contrary, using branches with different dilation rates within our CFF block results in a noticeable improvement in image denoising performance. This suggests that capturing effective receptive fields of different sizes without increasing the number of parameters can capture abundant contextual information to enhance image denoising performance.

Ablation Study of GA learning scheme. As shown in Table 5, the image denoising performance of the proposed IPLID method improves with an increase in the loss functions applied to the decoder. The best denoising performance is achieved when the supervision signal is applied to all CFF blocks (L_0 to L_4), with PSNR values of 40.06 dB for the SIDD dataset and 34.51 dB for the CBSD68 dataset. This indicates that the proposed IPLID method is already learning more general lower-level representations and is now focusing on refining higher-level representations. To visually observe the differences between the predicted denoised images and ground-truth images resulting from different layer-wise parameter updation and backward gradients, we marked the difference values with green pixels for better illustration, as shown in Figure 6. As the number of loss functions on the decoder increases (from L_0 to L_4), the number of green pixels (indicating differences) will gradually decrease. This means the difference between the predicted denoised image and the ground-truth image is reduced, demonstrating the effectiveness of the GA learning scheme.

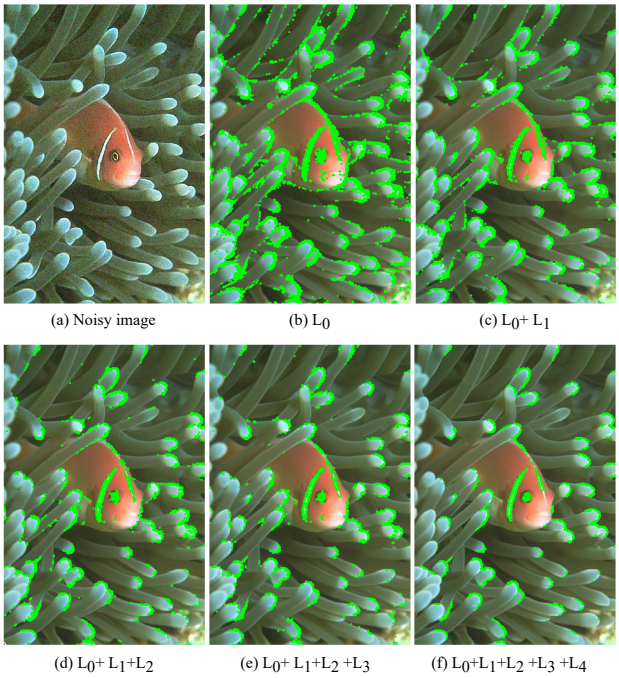

Figure 6: Influence of layer-wise parameter updation and backward gradients on the visualization of image denoising. Green pixels are used to highlight the differences between the denoised and ground-truth images, visually demonstrating the effects of parameter updation and backward gradients.

4.4 Further Study of the Implicit Prompt

To explore what our LP block prompts for image denoising, we visually compare the distributions of the ground-truth noisy maps, feature maps generated by pure pre-trained backbone and the denoised prompt information produced by our LP block, shown in Fig. 7. It is obvious that, compared to the shallow and deep feature maps (Fig. 7(b) and (C)) generated by the pre-trained model, the shallow and deep denoised prompt feature maps (Fig. 7(e) and (f)) both have closer distributions to the ground-truth noisy maps. This implies that our LP block may provide the noise distribution information in the denoising process for better denoising performance.

5 Conclusion

This paper proposed the **I**mplicit **P**rompt **L**earning for **I**mage **D**enoising (IPLID) method to implicitly generate adaptive prompt information conveniently without manual design, which can improve the visual effect and denoising performance. The proposed **L**inear **P**rompt (LP) block with ultra-few parameters is significantly efficient to produce denoising prompt feature for the pre-trained vision model. To fully uti-

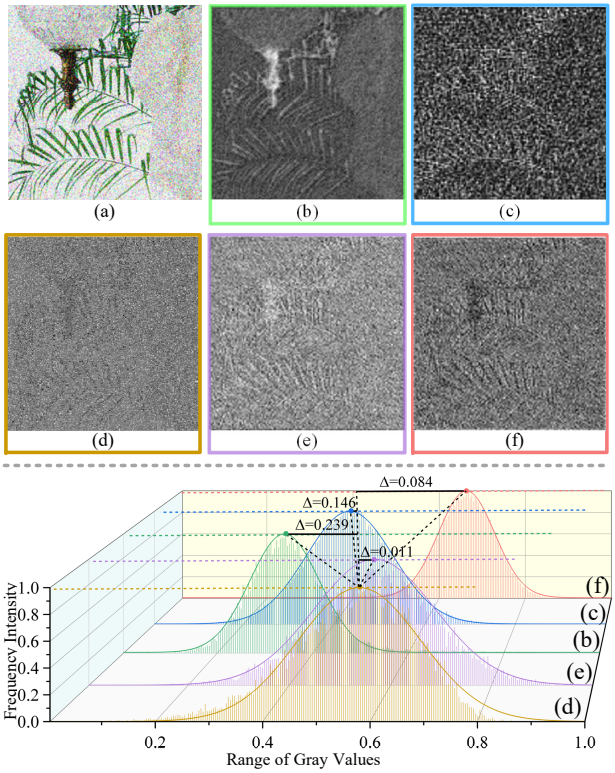

Figure 7: Visualization of distribution differences between noise maps, feature maps from pre-trained vision models, and the prompt feature maps from the proposed LP block. (a) Noisy image; (b) Shallow feature maps of backbone (1^{st} stage); (c) Deep feature maps of backbone (3^{rd} stage); (d) Ground-truth noisy maps; (e) Prompt information (1^{st} stage); (f) Prompt information (3^{rd} stage).

lize the multi-level guidance prompt feature maps for reconstructing denoised images, we have also proposed a **C**ompact **F**eature **F**usion (CFF) block. The CFF block employs a multi-level learning strategy to enhance the fusion process and improves the model's ability to transfer knowledge from pre-trained feature to the task of recovering denoised images. Finally, a **G**radient **A**ccumulation (GA) learning scheme was introduced to comprehensively impose the supervision in the optimization process. Experiment results showed that our IPLID can achieve satisfactory performance using only 1% of the backbone pre-trained parameters in both the quality of the denoised images and the practical efficiency. Further exploration also suggests that it is promising to introduce or learn the noise intrinsic attributes within the denoising prompt, providing an inspirational way for future denoising prompt.

Acknowledgements

This work was supported in part by the NSFC fund (NO. 62176077, 62206073), in part by the Shenzhen Key Technical Project (NO. JSGG20220831092805009, JSGG20220831105603006, JSGG20201103153802006, KJZD20230923115117033), in part by the Guangdong International Science and Technology Cooperation Project (NO. 2023A0505050108), in part by the Shenzhen Fundamental Research Fund (NO. JCYJ20210324132210025), in part by the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (NO. 2022B1212010005), and in part by the Guangdong Shenzhen joint Youth Fund under Grant 2021A151511074, in part by the Natural Science Foundation of Guangdong Province under Grant 2023A1515010893, in part by the Shenzhen Doctoral Initiation Technology Plan under Grant RCBS20221008093222010.

References

[Abdelhamed *et al.*, 2018] A. Abdelhamed, Stephen Lin, and M. S. Brown. A high-quality denoising dataset for smartphone cameras. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2018.

[Chen *et al.*, 2021] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12294–12305, 2021.

[Fan *et al.*, 2017] Xiangmin Fan, Wencan Luo, Muhsin Menekse, Diane Litman, and Jingtao Wang. Scaling reflection prompts in large classrooms via mobile interfaces and natural language processing. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 363–374, 2017.

[He *et al.*, 2016] 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.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[Jiang *et al.*, 2022a] Bo Jiang, Yao Lu, Jiahuan Wang, Guangming Lu, and David Zhang. Deep image denoising with adaptive priors. *IEEE Transactions on Circuits and Systems for Video Technology*, 32:5124–5136, 2022.

[Jiang *et al.*, 2022b] Bo Jiang, Yao Lu, Jiahuan Wang, Guangming Lu, and David Zhang. Deep image denoising with adaptive priors. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[Kolesnikov *et al.*, 2021] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

[Lester *et al.*, 2021] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[Li *et al.*, 2021] Wenbo Li, Xin Lu, Jiangbo Lu, X. Zhang, and Jiaya Jia. On efficient transformer and image pre-training for low-level vision. *ArXiv*, abs/2112.10175, 2021.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021.

[Liu *et al.*, 2018] Pengju Liu, Hongzhi Zhang, K. Zhang, Liang Lin, and Wangmeng Zuo. Multi-level wavelet-cnn for image restoration. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 886–88609, 2018.

[Liu *et al.*, 2023] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017.

[Nam *et al.*, 2016] 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. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1683–1691, 2016.

[Su *et al.*, 2021] Yusheng Su, Xiaozhi Wang, Yujia Qin, Chi-Min Chan, Yankai Lin, Huadong Wang, Kaiyue Wen, Zhiyuan Liu, Peng Li, Juanzi Li, et al. On transferability of prompt tuning for natural language processing. *arXiv preprint arXiv:2111.06719*, 2021.

[Wang *et al.*, 2021] Zhendong Wang, Xiaodong Cun, Jianmin Bao, and Jianzhuang Liu. Uformer: A general u-shaped transformer for image restoration. *arXiv preprint arXiv:2106.03106*, 2021.

[Wang *et al.*, 2023] Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Qiushi Yang, Yiheng Liu, Songyao Zhang, et al. Review of large vision models and visual prompt engineering. *arXiv preprint arXiv:2307.00855*, 2023.

[Xu *et al.*, 2018] Jun Xu, Hui Li, Zhetong Liang, David C. Zhang, and Lei Zhang. Real-world noisy image denoising: A new benchmark. *ArXiv*, abs/1804.02603, 2018.

[Zamir *et al.*, 2021] 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.

[Zhang *et al.*, 2017] K. Zhang, W. 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*, 26:3142–3155, 2017.

[Zhang *et al.*, 2018] K. Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27:4608–4622, 2018.

[Zhang *et al.*, 2022] K. Zhang, Yawei Li, Jingyun Liang, Jiezhang Cao, Yulun Zhang, Hao Tang, Radu Timofte, and Luc Van Gool. Practical blind denoising via swin-conv-unet and data synthesis. *ArXiv*, abs/2203.13278, 2022.

[Zhou *et al.*, 2022] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.