ProRes: Exploring Degradation-aware Visual Prompt for Universal Image Restoration

Jiaqi Ma[1]*, **Tianheng Cheng**[2]* , **Guoli Wang**[3] , **Qian Zhang**[3] , **Xinggang Wang**[2] , **Lefei Zhang**[1]

[1]School of Computer Science, Wuhan University
[2]School of EIC, Huazhong University of Science & Technology
[3]Horizon Robotics
{jiaqima,zhanglefei}@whu.edu.cn
{thch,xgwang}@hust.edu.cn
{guoli.wang,qian01.zhang}@horizon.ai

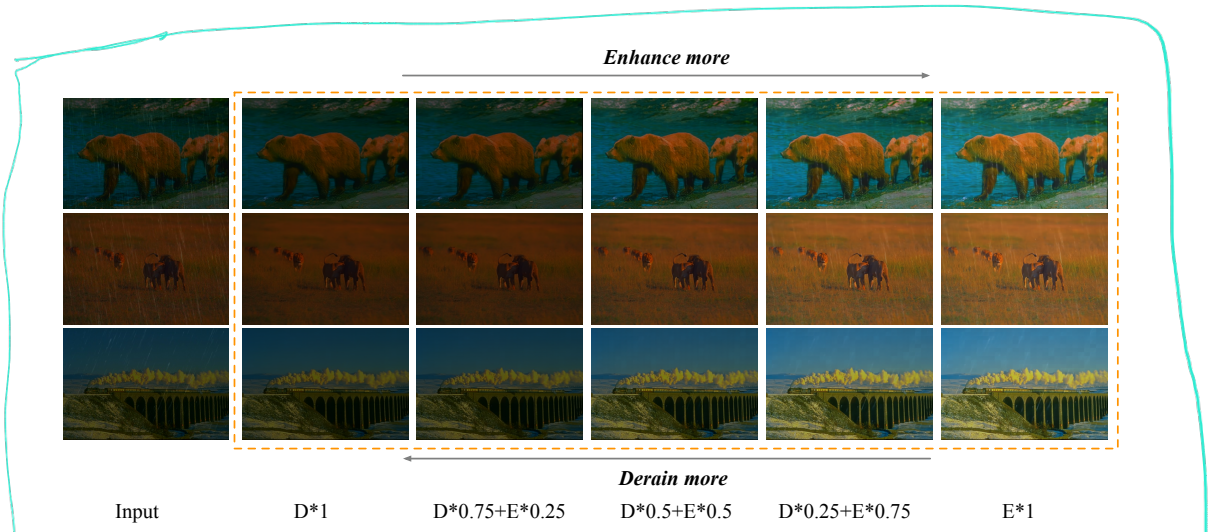


Figure 1: We show the visualization results processed by **ProRes** from images of mixed types of degradation, *i.e.*, low-light and rainy. ProRes adopts two visual prompts for low-light enhancement (E) and deraining (D) and combines the two visual prompts by linear weighted sum, *i.e.*, αD + $(1-\alpha)$E, to control the restoration process.

Abstract

Image restoration aims to reconstruct degraded images, *e.g.*, denoising or deblurring. Existing works focus on designing task-specific methods and there are inadequate attempts at universal methods. However, simply unifying multiple tasks into one universal architecture suffers from uncontrollable and undesired predictions. To address those issues, we explore prompt learning in universal architectures for image restoration tasks. In this paper, we present **Degradation-aware Visual Prompts**, which encode various types of image degradation, *e.g.*, noise and blur, into unified visual prompts. These degradation-aware prompts provide control over image processing and allow weighted combinations for customized image restoration as shown in Fig. 1. We then leverage degradation-aware visual **Pro**mpts to establish a controllable and universal model for image

*Equal contribution. This work is done when Jiaqi Ma was an intern at Horizon Robotics. Corresponding to Lefei Zhang <zhanglefei@whu.edu.cn>.

Preprint. Under review.

Restoration, called **ProRes**, which is applicable to an extensive range of image restoration tasks. ProRes leverages the vanilla Vision Transformer (ViT) without any task-specific designs. Furthermore, the pre-trained ProRes can easily adapt to new tasks through efficient prompt tuning with only a few images. Without bells and whistles, ProRes achieves competitive performance compared to task-specific methods and experiments can demonstrate its ability for controllable restoration and adaptation for new tasks. The code and models will be released in https://github.com/leonmakise/ProRes.

1 Introduction

Image restoration, as the fundamental challenge in the field of computer vision, aims to reconstruct high-quality images from degraded images which suffer from noise, blur, compression artifacts, and other distortions. Those kinds of low-level vision tasks are critical in a wide range of applications, including general vision perception, medical imaging, and satellite imaging.

In recent years, deep learning methods [1, 2, 3, 4] have revolutionized the field of image restoration, achieving high accuracy rates and improving the quality of restored images. Prevalent works concentrate on carefully designing task-specific methods and have shown promising results in various low-level image restoration tasks, *e.g.*, denoising, deraining, and deblurring. However, the task-specific methods have limited transfer ability on new datasets and still require specific designs for adapting to other tasks, as shown in Fig. 2 (a). Moreover, it's much more challenging for task-specific methods to process images of different corruptions. Developing general approaches for simultaneously processing different corruptions is necessary for nowadays applications.

Recently, multi-task learning [5, 6, 7] has been explored for processing images with different types of degradation by sharing the backbone and designing task-specific heads, as shown in Fig. 2 (b). Despite its success in image restoration, multi-task methods with shared parameters still suffer from the task-interference problem [8] that different tasks might have different optimization directions.

Inspired by the pioneering works on universal approaches for image recognition [9, 8] and image segmentation [10, 11], we aim to investigate the universal architectures for image restoration. In this paper, we formulate different image restoration tasks into a universal architecture, as illustrated in Fig. 2 (c), in which images from mixed tasks are fed into the universal model for unified training. Nonetheless, the inference using the universal model is uncontrollable or unpredictable, as it cannot guarantee that the output will meet our expectations without any task-specific indicator.

To mitigate the above issues, we present Degradation-aware Visual Prompts as parametric identifiers for different types of degradation, *e.g.*, a visual prompt for "low-light enhancement". Specifically, we define a series of visual prompts with the same size as the images. And then we add the selected visual prompt to the degraded image as a prompted image and feed it into the universal architecture for the desired restored image, as shown in Fig. 2 (d). Further, we incorporate the degradation-aware prompts into a simple universal architecture, *i.e.*, a vanilla Vision Transformer [12] with a pixel decoder, for universal image restoration, and present a novel and versatile framework named **ProRes**. Compared to previous task-specific or multi-task approaches, ProRes with task-agnostic designs can be trained with various tasks and can easily process various types of degradation. In this paper, we aim to demonstrate two superior capabilities of the proposed ProRes: (1) ProRes has strong control ability, allowing us to combine different prompts with different weights according to our needs and obtain the desired output; (2) ProRes possesses exceptional transferability, where we can quickly and cost-effectively adapt ProRes to new tasks or datasets using simple prompt tuning.

To verify the effectiveness of the proposed ProRes, we construct a joint image restoration dataset of several tasks including denoising, low-light enhancement, deraining, and deblurring, and train ProRes on it. Without bells and whistles, ProRes, as a universal approach, achieves competitive performance on various benchmarks, *e.g.*, SSID [13] and LOL [14], compared to well-established and carefully-designed methods. Surprisingly, qualitative and quantitative results can demonstrate the control ability of the proposed degradation-aware visual prompts, which can be combined with weights to generate desired restored images. Moreover, experimental results on haze removal show that the proposed ProRes with prompt tuning can be efficiently and effectively adapted to new tasks. And we believe large-scale datasets can further boost the capability of the proposed ProRes for

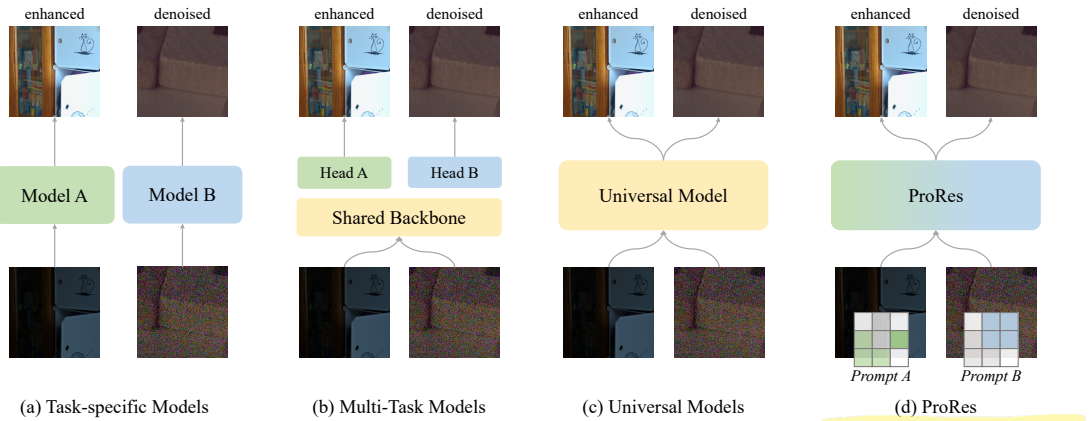


Figure 2: **Conceptual comparison with previous approaches.** (a) Task-specific models design specialized architectures and strategies for different tasks, *e.g.*, Model-A for low-light enhancement and Model-B for image denoising. (b) Multi-task models adopt a shared backbone for image feature extraction and leverage multiple task-specific heads for different tasks. (c) Universal models adopt mixed inputs without any task-specific indicator. (d) The proposed ProRes adopts input images with degradation-aware visual prompts for specific targets.

both controllable image restoration and new-task adaption. Our contribution can be summarized as follows:

- We present degradation-aware visual prompts for universal image restoration which provide control over image processing given any degraded image.
- We propose ProRes to address universal image restoration with degradation-aware prompts, which is the first prompt-based versatile framework for image restoration.
- We additionally propose the effective and efficient prompt tuning with ProRes to adapt for new tasks or new datasets without fine-tuning ProRes.
- The proposed ProRes obtains competitive results compared to task-specific methods on various benchmarks. We hope the simple yet effective ProRes can serve as a solid baseline for universal image restoration and facilitate future research.

2 Related Works

2.1 Multi-Task Learning for Image Restoration

Multi-task learning (MTL) has emerged as a promising paradigm by leveraging shared information across multiple related tasks. For image restoration tasks, MTL is employed to solve multiple related tasks simultaneously, such as denoising and super-resolution. The shared structure or feature representation across these tasks can often lead to enhanced performance compared to models trained on individual tasks. One notable work is the AIRNet [5], which incorporated MTL by feeding several corruption. Besides, [6] constrains relationships of tasks by the multi-contrastive regularization. Path-Restore [7] offers a multi-path CNN to select an appropriate route for each image region.

However, multi-task learning for image restoration is achieving a balance between learning shared features and preserving task-specific features. The common framework contains several encoders for task selection or decoders for variable outputs. Our work addresses this challenge by integrating degradation-aware visual prompts, enabling us to control the task-specific aspects while still benefiting from the shared features.

2.2 Universal Foundation Models

Foundation models aim to serve as a shared basis for multiple tasks. The motivation behind such universal models is to leverage the commonalities among various tasks and modalities to improve

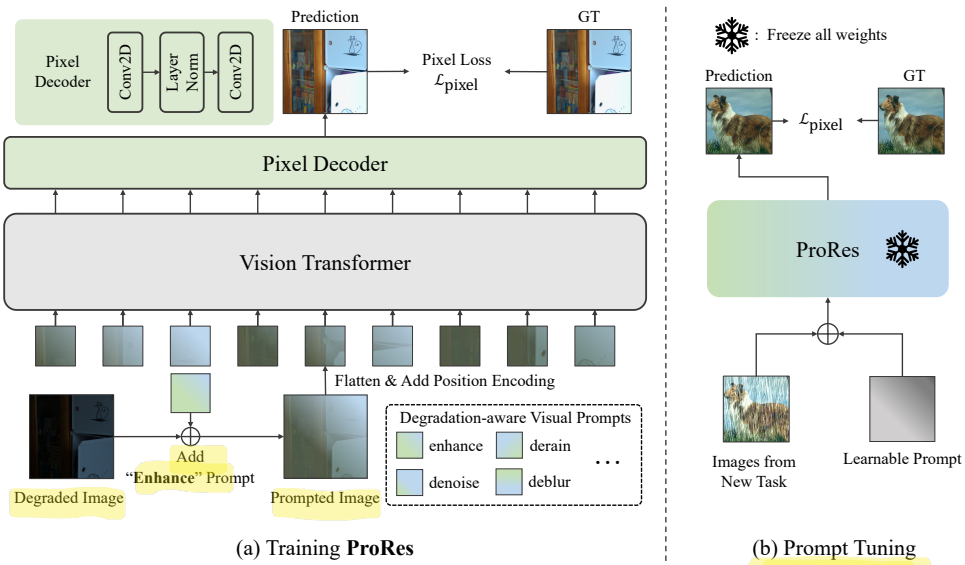


Figure 3: **Overall Pipeline of ProRes. (a) Training ProRes**: we add the target visual prompt to the input image and flatten the prompted image into patches. We leverage a vision transformer, *i.e.*, ViT-Large, as the image encoder and adopt a simple pixel decoder to generate the restored image. Then we adopt pixel loss to optimize ProRes. **(b) Prompt Tuning**: we freeze the weights of ProRes and randomly initialize the learnable prompts for new tasks or new datasets.

efficiency and performance. Notable instances include the Vision Transformer (ViT) [12] and BERT [15], which have been used across various vision or language tasks. Perceiver [16] and Perceiver-IO [17] provide universal solutions to natural language and visual understanding processing. Then Uni-Perceiver [18] is extended to generic perception tasks and Uni-Perceiver-MoE [19] discusses the main factor to performance degradation. Besides, for a series of similar tasks, there are some universal models based on various of network architectures [20, 21, 22, 23].

Those works also reveal the potential of the universal foundation model. Universal foundation models for low-level vision tasks, which can handle tasks such as image denoising, deraining, and low-light enhancement with a single model architecture, are still lacking in exploration. [6, 24, 25] involve in inpainting and denoising. However, there still exists no universal model, especially for low-level vision tasks, which can provide controllable predictions.

2.3 Visual Prompt Learning

Inspired by the success of prompt learning in natural language processing, recent works attempt to adapt prompt learning to several visual tasks. MAE-VQGAN [24] gives the grid-like input during the inference phase, and automatically inpaints blank areas by a pre-trained generative model. Then a supervised version of visual prompt learning named Painter [25] removes generative models like VQGAN, and calculates the regression loss by tokens from the vision transformer. These approaches usually involve providing the model with additional input or cues to guide the model's reasoning and attention. Some studies have demonstrated the effectiveness of visual prompts in tasks like image classification [26], object detection [27], and visual question answering [28].

3 Methodology

In this section, we begin by introducing degradation-aware visual prompts designed for various image restoration tasks. Then, we present our approach called ProRes, which integrates degradation-aware prompts to achieve universal image restoration. Furthermore, we demonstrate the versatility of ProRes by adapting it to new tasks through prompt tuning.

3.1 Degradation-aware Visual Prompts

Prompt Design. Previous methods [25, 24] utilize grid-like combinations of inputs and treat tasks as inpainting problems. However, the grid-like layout increases computational costs since the input becomes four times larger than the original image. Additionally, when using different predefined sets of prompts, the outputs may significantly vary for a pre-trained model.

Nevertheless, we concur [25] that the format of visual prompts should resemble images, even if they do not possess visual meaning. In line with this, we define each visual prompt with a shape of $H \times W \times 3$, which matches the shape of the input images. The image-like format can be seamlessly integrated with any natural image and provides ample information as a prompt.

Considering the limitations of the grid-like layout, the visual prompts are added directly to input images, which allows it to overcome the limitations of grid-like combinations and achieve more efficient and consistent results. For each task, a single visual prompt is utilized to capture and represent the task-specific information. And the image-like input can be easily incorporated into degraded images or network layers to facilitate degradation-aware restoration.

Prompt Initialization. The initialization of visual prompts can be approached in various ways. In our case, since we consider them as visible images with the additive property, we choose a simple initialization method using one `nn.Parameter()` layer. This allows us to treat the visual prompts as trainable parameters that can be optimized during the restoration process.

To ensure training stability and expedite convergence, we employ a lightweight pre-trained image restoration model, *i.e.*, MPRNet [3]. We utilize this model to optimize the visual prompts by training them alongside the corresponding task datasets. In this process, we freeze all the weights of the pre-trained model, except for the parametric prompts, which are allowed to be learnt during training. This approach enables efficient optimization of the prompts while leveraging the knowledge encoded in the pre-trained model.

Prompt as Control. As previously discussed, we leverage the benefits of image-like prompts, allowing us to easily incorporate them into the original degraded images. By simply adding the prompts to the degraded images, we can effectively control the restoration process. Specifically, the control ability of degradation-aware prompts is mainly reflected in three aspects: (1) the degradation-aware prompt guides the restoration for its corresponding degradation; (2) the irrelevant prompts have no impact on the input without corresponding degradation; (3) different prompts can be combined for complicated degradation, *e.g.*, the input contains several degradation. The capabilities in the above three aspects enable the proposed degradation-aware visual prompt to obtain the desired outputs.

In Fig. 4, we illustrate some samples of combined images with various types of degradation. With the utilization of ProRes, we can observe that the desired images are successfully generated based on the provided visual prompts. ProRes demonstrates its capability to handle degraded images with prompts that may seem irrelevant or combined. Furthermore, we also try combinations of two visual prompts, the results with different weights of combination in Fig. 1 confirm the control ability of ProRes. For additional results, please refer to Sec. 4.5.

3.2 ProRes

Architecture. The overall architecture of ProRes is illustrated in Fig. 3. Given the degraded image, we adopt a single task-specific visual prompt and add it to the input image and form the prompted image. We leverage a vanilla Vision Transformer (ViT) [12] pre-trained by MAE [29] as the encoder and a 2-conv pixel decoder to reconstruct the output RGB image. The intentionally simple design of the architecture is not intended to achieve top performance on various benchmarks, but rather to serve as a universal baseline for exploring the potential of degradation-aware visual prompts.

Training Loss. As the training process can be supervised by regression losses, we prefer simple losses such as ℓ_1, ℓ_2, and Smooth-ℓ_1. We use Smooth-ℓ_1 loss as $\mathcal{L}_{\text{pixel}}$ without sophisticated designs and it is effective enough for ProRes. For the performance of different training losses, we evaluate in ablation experiments in Sec. 4.8.

3.3 Adaption via Prompt Tuning

Previous works tend to fine-tune the whole models to adapt to new tasks or new datasets. However, directly fine-tuning a new task is prone to lose the model's ability on the original task, and fine-tuning the whole model brings significant training costs. In contrast, we freeze the weights of ProRes which is pre-trained on various tasks, and randomly initialize a new visual prompt for the new dataset or new task, as shown in Fig. 3 (b). During prompt tuning, we only update the learnable parameters of visual prompts by gradient descent, which is efficient without updating the entire model or long-schedule training. For prompt tuning, we can simply replicate the training settings used from scratch, ensuring that the prompts are optimized effectively without the need for extensive modifications or additional training procedures.

4 Experiments

4.1 Datasets and Pre-processing

Image restoration tasks are evaluated on several popular benchmarks, including SIDD [30] for image denoising, LoL [31] for low-light image enhancement, the merged deraining dataset [2] for deraining, and the merged deblurring dataset [32]. A brief description of each dataset is provided below, and their details are summarized in Tab. 1.

The Smartphone Image Denoising Dataset (SIDD) [13] is a large-scale dataset designed for image denoising. It contains both noisy and clean images captured by various smartphone cameras under different conditions, covering a diverse range of scenes and noise levels.

The Low-Light enhancement dataset (LOL) [14] is designed for the task of low-light image enhancement. It contains 500 image pairs, with each pair consisting of a low-light image and its corresponding well-exposed ground truth. The images in the LOL dataset cover various indoor and outdoor scenes, with different levels of lighting and noise.

The merged deraining dataset is obtained from the work of Li *et al*. [36], which combines three popular deraining datasets: Rain100H [37], Rain100L [37], and Rain800 [38]. It is a comprehensive benchmark for evaluating single-image deraining algorithms. The merged dataset covers a diverse range of rain densities and streak orientations, providing a robust evaluation platform for assessing the performance of deraining algorithms.

Dataset	Train set	Test set
Denoising		
SIDD [13]	320	1280
Low-light enhancement		
LOL [14]	485	15
Merged Deraining [2]		
Rain800	700	100
Rain1800	1800	-
Rain14000	11200	2800
Rain1200	-	1200
Rain12	12	-
Rain100H	-	100
Rain100L	-	100
Merged Deblurring [32]		
GoPro [33]	2103	1111
HIDE [34]	-	2025
RealBlur-R [35]	-	980
RealBlur-J [35]	-	980

Table 1: Summary of the datasets used for ProRes.

The merged deblurring dataset is sourced from the work of Zhang *et al*. [39], which combines the GoPro dataset [33] and three additional deblurring datasets: HIDE [34], RealBlur-R, and RealBlur-J [35]. This dataset contains various types and levels of blur caused by camera shake and object motion, making it a comprehensive and challenging benchmark for assessing the performance of deblurring algorithms.

4.2 Training Details

To learn prompts, we select a small image restoration model for quick fine-tuning. We use MPRNet [3] and utilize the pre-trained models offered by the authors. We add one layer (`nn.Parameter()`) to the inputs, and freeze all other layers to ensure the prompts can be learnt during the fine-tuning stage. We reduce the learning rate to 1e-4 and remain other settings the same as training ProRes.

To train the universal model, we choose a variant of ViT-Large according to [25]. We employ the AdamW optimizer [40] with a cosine learning rate scheduler, and train for 100 epochs. The training hyper-parameters are: the batch size as 160, base learning rate as 1e-3, weight decay as 0.05, $\beta_1 = 0.9$, $\beta_2 = 0.999$, drop path [41] ratio as 0.1, a warm-up for 2 epochs. We follow a light data augmentation

Method	Denoising SIDD		Deraining 5 datasets		Enhancement LoL		Deblurring 4 datasets	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Task-specific models								
Uformer [4]	39.89	0.960	-	-	-	-	32.31	0.941
MPRNet [3]	39.71	0.958	32.73	0.921	-	-	33.67	0.948
MIRNet-v2 [2]	39.84	0.959	-	-	**24.74**	0.851	-	-
Restormer [42]	**40.02**	0.960	**33.96**	**0.935**	-	-	32.32	0.935
MAXIM [43]	39.96	0.960	33.24	0.933	23.43	0.863	**34.50**	**0.954**
Universal models								
Painter [25]	38.88	0.954	29.49	0.868	22.40	0.872	-	-
ProRes	39.28	**0.967**	30.67	0.891	22.73	**0.877**	28.03	0.897

Table 2: Comparison with the universal models, and the recent best task-specific models on four representative low-level image restoration tasks. The backbone of ProRes and Painter is ViT-Large.

strategy: random resize cropping with a scale range of [0.3, 1] and an aspect ratio range of [3/4, 4/3], with a random flipping. To make use of the pre-trained ViT-Large model, we resize the input image to 448×448. Considering that there are four different tasks, the sampling weight for each task is 0.3 (image deraining), 0.3 (low-light enhancement), 0.1 (image denoising), and 0.3 (image deblurring). Essential ablation studies are conducted to explore the effectiveness of some training strategies.

4.3 Performance on Image Restorations Tasks

With the corresponding task prompts, we compare our approach, namely ProRes, with recent best universal models and task-specific models on four representative image restoration vision tasks, shown in Tab. 2. Without task-specific design and only utilizing Smooth-ℓ_1 loss for supervision, ProRes outperforms the universal models such as Painter [25] and also gets closer results compared with the state-of-the-art task-specific model on several tasks. There is still much room for boosting our approach compared to other well-designed task-specific models. For example, our default training iteration number is 81k (50 epochs), while those task-specific models use a much larger number for training, *e.g.*, 300k in MIRNet-v2 for image denoising, 150k for low-light enhancement, and 400k in MPRNet for all tasks. Achieving state-of-the-art performance on every task is not the goal of this paper, the unified model is expected to yield superior performance with more comprehensive training.

4.4 Universal Models *v.s.* Task-specific Models

Tab. 3 compares the performance between universal models and task-specific models. For a fair comparison, we adopt the same architecture of ProRes and train a series of task-specific models along with a prompt-free universal model. All models are based on ViT-Large with MAE pre-trained weights and trained with the same setting (50 epochs). In comparison to task-specific models, ProRes can achieve competitive or better performance, especially on low-light enhancement (LoL).

In comparison to the vanilla universal baseline, *i.e.*, ViT-Large trained with the joint dataset, ProRes still obtains similar results on denoising and deraining. However, ProRes achieves significantly better performance on the other two tasks, *i.e.*, enhancement and deblurring, which shows the superiority of ProRes for learning with different tasks. Note that we can leverage visual prompts in ProRes for specific restoration tasks and generate the desired outputs, which is infeasible for the vanilla universal model.

4.5 Control Ability

Compared to the vanilla universal model, one core advantage of ProRes is to control desirable outputs with given prompts. To reveal the control ability of ProRes, we adopt the following settings to evaluate ProRes. *Setting1*: for images with single degradation, we apply single task-aware prompts; *Setting2*: for images containing single degradation, we offer single task-irrelevant prompts; *Setting3*:

Method	Denoising SIDD		Deraining 5 datasets		Enhancement LoL		Deblurring 4 datasets	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Task-specific models								
ViT-Large	**39.74**	0.969	-	-	-	-	-	-
	-	-	29.95	0.879	-	-	-	-
	-	-	-	-	18.91	0.741	-	-
	-	-	-	-	-	-	27.51	0.882
Universal models								
ViT-Large	39.28	**0.967**	**30.75**	**0.893**	21.69	0.850	20.57	0.680
ProRes	39.28	**0.967**	30.67	0.891	**22.73**	**0.877**	**28.03**	**0.897**

Table 3: Comparison between ProRes and the vanilla task-specific models based on ViT-Large.

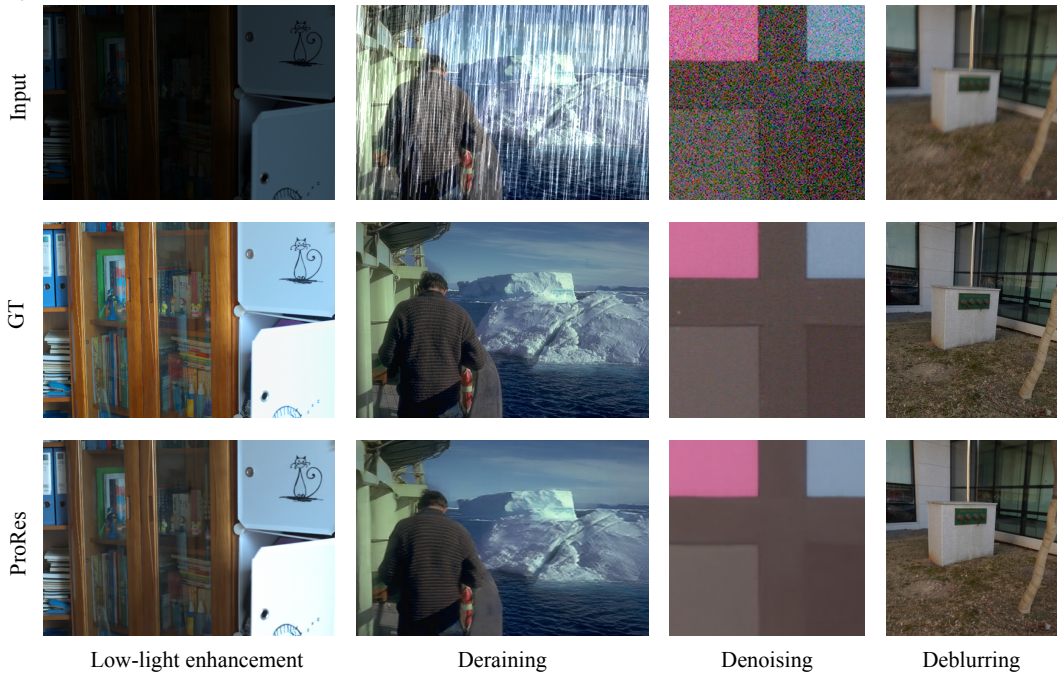


Figure 4: Visualization results processed from images of different corruptions. Compared with the original inputs, the outputs are consistent with the given visual prompts.

for images containing several complex degradation, we linearly combine prompts for different tasks for combined task-relevant prompts.

Independent Control. For *Setting1*, we directly feed the original degraded images into our unified model with relevant prompts. As is shown in Fig. 4, for those degraded images, we can easily control the outputs of our ProRes model by providing specific single prompts. It reflects the ProRes model's inherent ability for independent control.

Sensitive to Irrelevant Task-specific Prompts. In *Setting2*, we directly feed the original degraded images into our unified model with random irrelevant prompts. As is shown in Fig. 5, the rainy images can still keep the same with the denoising prompt. It reflects that the ProRes model is sensitive to prompts from irrelevant tasks for each input image.

Tackle Complicated Corruptions. Regarding *Setting3*, we explore the fusion of prompts and the ability to tackle complicated corruptions. We generate a subset from the Rain100L dataset, which

Low-light input | Derain prompt | Deblur prompt | Rainy input | Denoise prompt | Deblur prompt

Figure 5: Visualization results processed by different prompts. Compared with the original inputs, the outputs remain unchanged with irrelevant visual prompts.

Method	Enhancement FiveK		Dehazing RESIDE-6K	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
ProRes w/o Prompt Tuning	18.94	0.815	-	-
ProRes w/ Prompt Tuning	22.78	0.839	21.47	0.840

Table 4: Experimental results of ProRes with prompt tuning on the FiveK and RESIDE-6K datasets.

is originally a deraining dataset. To ensure the subset contains several degradation, we follow the procedures of generating a synthetic low-light environment and making it a low-light rainy dataset. Considering that the subset does not exist paired ground truth for single degradation, hence we solely provide the qualitative evaluation. We combine those prompts by weights, and the aim is to jointly deal with those complicated corruptions.

In Fig. 1, we illustrate the results of different weights. We can find that the outputs can be controlled by adjusting the weight. Increasing the weight of low-light enhancement prompts will result in brighter output while increasing the weight of the deraining prompts will enhance the deraining effect. By combining multiple prompts, ProRes can enhance the restoration performance and handle the challenging and diverse forms of corruption present in the images.

4.6 Adaptation via Prompt Tuning

Here, we evaluate the generalization ability and transferring ability of ProRes on new datasets or new tasks. Specifically, we adopt the FiveK dataset [44] for low-light enhancement and the RESIDE-6K dataset [45] for image dehazing, which is a new task for ProRes. To adapt ProRes on new tasks or new datasets, we freeze all parameters of ProRes but the learnable visual prompts. Specifically, we randomly initialize these two visual prompts and individually train on the two datasets for 50 epochs with all parameters of ProRes frozen, as shown in Fig. 3 (b).

In addition, we compare the performance of directly using ProRes and ProRes with prompt tuning. As shown in Tab. 4, directly applying ProRes can achieve good performance on FiveK dataset, showing the generalization ability of ProRes on unseen samples. Further, using prompt tuning brings remarkable improvements, which can demonstrate the effectiveness of prompt tuning and the transferring ability for new datasets (FiveK for low-light enhancement) or tasks (RESIDE-6K for dehazing).

4.7 Training ProRes with Degradation-aware Prompts

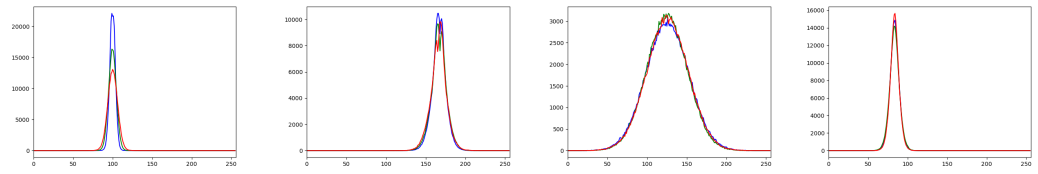
In Tab. 5, we study the different strategies of training ProRes with degradation-aware visual prompts, *i.e.*, initialization of visual prompts and prompt update. We initialize ProRes with MAE [29] pre-trained weights and train ProRes from scratch. We adopt two initialization methods for visual prompts: (1) random initialization (using `torch.nn.init.normal_(self.prompt, std=0.1)`) and (2) pre-trained weights from light-weight image restoration models (default). During training, the visual prompts can be detached without parameter update or learnable for the parameter update. As shown in Tab. 5, using random initialization is inferior to pre-trained prompts in several tasks, *e.g.*, deraining and enhancement. However, training ProRes with learnable prompts brings negative impacts, and even leads to collapse in enhancement, which can be attributed to the limited training samples in enhancement. Tab. 5 can demonstrate that using detached pre-trained prompts is more stable and performs better.

Figure 6: Visualization results of ProRes on the FiveK dataset. We adopt two settings, *i.e.*, direct inference and prompt tuning, to evaluate ProRes on the FiveK dataset (a new dataset for low-light enhancement).

Figure 7: Visualization results of ProRes on the RESIDE-6K dataset via prompt tuning for image dehazing (a new task).

Prompt		Denoising SIDD		Deraining 5 datasets		Enhancement LoL		Deblurring 4 datasets	
Initialization	Learnable	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Random	Learnable	39.24	0.966	29.98	0.881	10.60	0.417	26.19	0.844
Random	Detached	39.14	0.966	29.98	0.877	22.02	0.819	**28.10**	**0.898**
Pre-trained	Learnable	39.26	**0.967**	30.20	0.884	22.47	0.876	27.83	0.891
Pre-trained	Detached	**39.28**	**0.967**	**30.67**	**0.891**	**22.73**	**0.877**	28.03	0.897

Table 5: Comparison of using different training strategies for ProRes with degradation-aware visual prompts.



(a) Denoising (b) Deraining (c) Enhancement (d) Deblurring

Figure 8: Histogram distribution of degradation-aware visual prompts. The red, blue, and green lines denote the 3 channels, *i.e.*, RGB channels in input. It's clear to see that different visual prompts have different distributions.

4.8 Training Loss

Although ProRes is utilized for tackling low-level vision tasks, we opt to use straightforward regression losses rather than complex combinations of losses (Charbonnier loss [46] or Perceptual loss [47]). Therefore, we only compare ℓ_1, ℓ_2, and Smooth-ℓ_1 losses to provide clear guidance. As shown in Table 6, we find that Smooth-ℓ_1 slightly outperforms ℓ_1, and ℓ_2 experiences a performance decline in most image restoration tasks. Hence, we use Smooth-ℓ_1 as the training loss.

Loss	Denoising SIDD		Deraining 5 datasets		Enhancement LoL		Deblurring 4 datasets	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
ℓ_2	38.93	0.956	**30.69**	**0.893**	22.48	0.873	27.91	0.896
ℓ_1	39.27	**0.967**	30.63	0.890	22.60	0.875	**28.03**	**0.898**
Smooth-ℓ_1	**39.28**	**0.967**	30.67	0.891	**22.73**	**0.877**	**28.03**	0.897

Table 6: Ablation on different losses for ProRes.

4.9 Visualization of Degradation-aware Visual Prompts

Here, we reveal discriminative visual prompts by visualizing the histogram distribution of each visual prompt. Fig. 8 illustrates the distinct distributions of visual prompts for different degradation types. These notable differences in distribution highlight the control ability of visual prompts and the potential for weighted combinations of prompts in hybrid restoration tasks. By leveraging these distinct distributions, we can effectively manipulate the restoration process and achieve desired outcomes based on the specific degradation types and requirements.

4.10 Integrate Prompts in Different Layers

To better reveal the compatibility of prompts in different layers of the model, we design several settings which are trained from scratch. Besides directly adding visual prompts to input, we try to integrate them into layers in ProRes network. We integrate visual prompts in all layers, equal spacing, the first 5 layers, and the last 5 layers.

Layers	Denoising SIDD		Deraining 5 datasets		Enhancement LoL		Deblurring 4 datasets	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Add to input	**39.28**	**0.967**	30.67	0.891	**22.73**	0.877	**28.03**	**0.897**
All	39.16	0.966	30.31	0.884	22.67	0.875	26.83	0.867
0, 5, 11, 17, 23	39.25	**0.967**	30.57	0.888	22.47	0.873	27.95	0.896
0, 1, 2, 3, 4	39.26	**0.967**	30.49	0.888	23.15	0.882	27.99	0.896
19, 20, 21, 22, 23	39.27	**0.967**	**30.71**	**0.892**	22.70	**0.884**	20.57	0.680

Table 7: Quantitative results of different prompt integration strategies.

In Tab. 7, we can find that ProRes can be effective in almost all settings, except the last setting, *i.e.*, the last 5 layers. With only the last five layers integration, the performance on deblurring drops dramatically. This phenomenon is also revealed in other settings compared with only adding visual prompts to input. Hence, we uphold the idea that simplicity is the best and choose the simplest strategy, *i.e.*, add to input.

5 Conclusion

In this paper, we propose a universal framework named ProRes for versatile image restoration, which leverages the presented degradation-aware visual prompts as task identifiers. ProRes adopts the simple transformer architecture without task-specific designs and obtains competitive performance on various benchmarks for image restoration. Extensive experiments demonstrate the control ability with combined prompts and transfer ability on new tasks through prompt tuning. We believe ProRes is a significant step towards controllable and versatile image restoration and can motivate future research.

References

[1] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *ECCV*, pages 17–33, 2022. 2

[2] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for fast image restoration and enhancement. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(2):1934–1948, 2023. 2, 6, 7

[3] Armin Mehri, Parichehr B. Ardakani, and Ángel D. Sappa. Mprnet: Multi-path residual network for lightweight image super resolution. In *WACV*, pages 2703–2712, 2021. 2, 5, 6, 7

[4] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17662–17672, 2022. 2, 7

[5] Boyun Li, Xiao Liu, Peng Hu, Zhongqin Wu, Jiancheng Lv, and Xi Peng. All-in-one image restoration for unknown corruption. In *CVPR*, pages 17431–17441, 2022. 2, 3

[6] Wei-Ting Chen, Zhi-Kai Huang, Cheng-Che Tsai, Hao-Hsiang Yang, Jian-Jiun Ding, and Sy-Yen Kuo. Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In *CVPR*, pages 17632–17641, 2022. 2, 3, 4

[7] Ke Yu, Xintao Wang, Chao Dong, Xiaoou Tang, and Chen Change Loy. Path-restore: Learning network path selection for image restoration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):7078–7092, 2022. 2, 3

[8] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. In *NeurIPS*, 2022. 2

[9] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *CVPR*, 2022. 2

[10] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 2

[11] Jitesh Jain, Jiacheng Li, Man Chun Chiu, Ali Hassani, Nikita Orlov, and H. Shi. Oneformer: One transformer to rule universal image segmentation. In *CoRR*, volume abs/2211.06220, 2022. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 4, 5

[13] Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, pages 1692–1700, 2018. 2, 6

[14] Chen Wei, Wenhan Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, pages 1–12, 2018. 2, 6

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 4

[16] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *ICML*, pages 4651–4664, 2021. 4

[17] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *ICLR*, 2022. 4

[18] Xizhou Zhu, Jinguo Zhu, Hao Li, Xiaoshi Wu, Hongsheng Li, Xiaohua Wang, and Jifeng Dai. Uni-perceiver: Pre-training unified architecture for generic perception for zero-shot and few-shot tasks. In *CVPR*, pages 16783–16794, 2022. 4

[19] Jinguo Zhu, Xizhou Zhu, Wenhai Wang, Xiaohua Wang, Hongsheng Li, Xiaogang Wang, and Jifeng Dai. Uni-perceiver-moe: Learning sparse generalist models with conditional moes. In *NeurIPS*, 2022. 4

[20] Soon Yau Cheong, Armin Mustafa, and Andrew Gilbert. UPGPT: universal diffusion model for person image generation, editing and pose transfer. *CoRR*, abs/2304.08870, 2023. 4

[21] Dan Song, Tianbao Li, Wenhui Li, Wei-Zhi Nie, Wu Liu, and An-An Liu. Universal cross-domain 3d model retrieval. *IEEE Trans. Multim.*, 23:2721–2731, 2021. 4

[22] Yutong Lin, Ze Liu, Zheng Zhang, Han Hu, Nanning Zheng, Stephen Lin, and Yue Cao. Could giant pretrained image models extract universal representations? *CoRR*, abs/2211.02043, 2022. 4

[23] Zhiheng Ma, Xiaopeng Hong, Xing Wei, Yunfeng Qiu, and Yihong Gong. Towards A universal model for cross-dataset crowd counting. In *ICCV*, pages 3185–3194, 2021. 4

[24] Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei A. Efros. Visual prompting via image inpainting. In *NeurIPS*, 2022. 4, 5

[25] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. *CoRR*, abs/2212.02499, 2022. 4, 5, 6, 7

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763, 2021. 4

[27] Ziyang Cao, Lijun Yang, and Dahua Zhang. Open compound domain adaptation. In *CVPR*, pages 12807–12816, 2020. 4

[28] Yunpeng Li, Lida Wang, Liqiang Li, Hang Guo, Lei Zhang, and Jun Zhu. Align before fuse: Vision and language representation learning with momentum distillation. In *CVPR*, pages 12727–12737, 2021. 4

[29] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 5, 9

[30] Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, pages 1692–1700, 2018. 6

[31] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. In *BMVC*, page 155, 2018. 6

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

[33] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, pages 3883–3891, 2017. 6

[34] Chao Kim, Seungryong Choi, Jinhyeok Oh, Hyoseop Kim, and In So Kweon. Hide: A hierarchical image dataset for deblurring. In *CVPRW*, 2019. 6

[35] Chanwoo Rim, Haerin Lee, Seungjun Baek, Bumsub Kim, Minsu Kim, and In So Kweon. Real-world blind image deblurring using an adaptive activation function. In *ECCV*, pages 3–19, 2020. 6

[36] Yawei Li, Shuhang Gu, Christoph Meyer, and Radu Timofte Wang. Learning enriched features for real image restoration and enhancement. In *ECCV*, pages 3–19, 2020. 6

[37] Wenzhao Yang, Robby T Tan Zhang, Jiashi Wang, and Jian Xiong. Deep joint rain detection and removal from a single image. In *CVPR*, pages 1357–1366, 2017. 6

[38] He Zhang, Vishwanath A Sindagi, and Vishal M Patel. Density-aware single image de-raining using a multi-stream dense network. In *CVPR*, pages 695–703, 2018. 6

[39] Kai Zhang, Yawei Liang, Yi Yang, and Radu Timofte Wang. Multi-stage progressive image restoration. In *CVPR*, pages 2616–2625, 2021. 6

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[41] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger. Deep networks with stochastic depth. In *ECCV*, pages 646–661, 2016. 6

[42] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, pages 5718–5729, 2022. 7

[43] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan C. Bovik, and Yinxiao Li. MAXIM: multi-axis MLP for image processing. In *CVPR*, pages 5759–5770, 2022. 7

[44] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *CVPR*, pages 97–104, 2011. 9

[45] Codruta O. Ancuti, Cosmin Ancuti, and Radu Timofte. NH-HAZE: an image dehazing benchmark with non-homogeneous hazy and haze-free images. In *CVPR*, pages 1798–1805, 2020. 9

[46] Pierre Charbonnier, Laure Blanc-Féraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*, pages 168–172, 1994. 11

[47] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, volume 9906, pages 694–711, 2016. 11