UGPNet: Universal Generative Prior for Image Restoration

Hwayoon Lee[1,*] Kyoungkook Kang[2] Hyeongmin Lee[2] Seung-Hwan Baek[2] Sunghyun Cho[2]

[1]GENGENAI
hwayoon.lee@gengen.ai

[2]POSTECH
{kkang831, hmin970922, shwbaek, s.cho}@postech.ac.kr

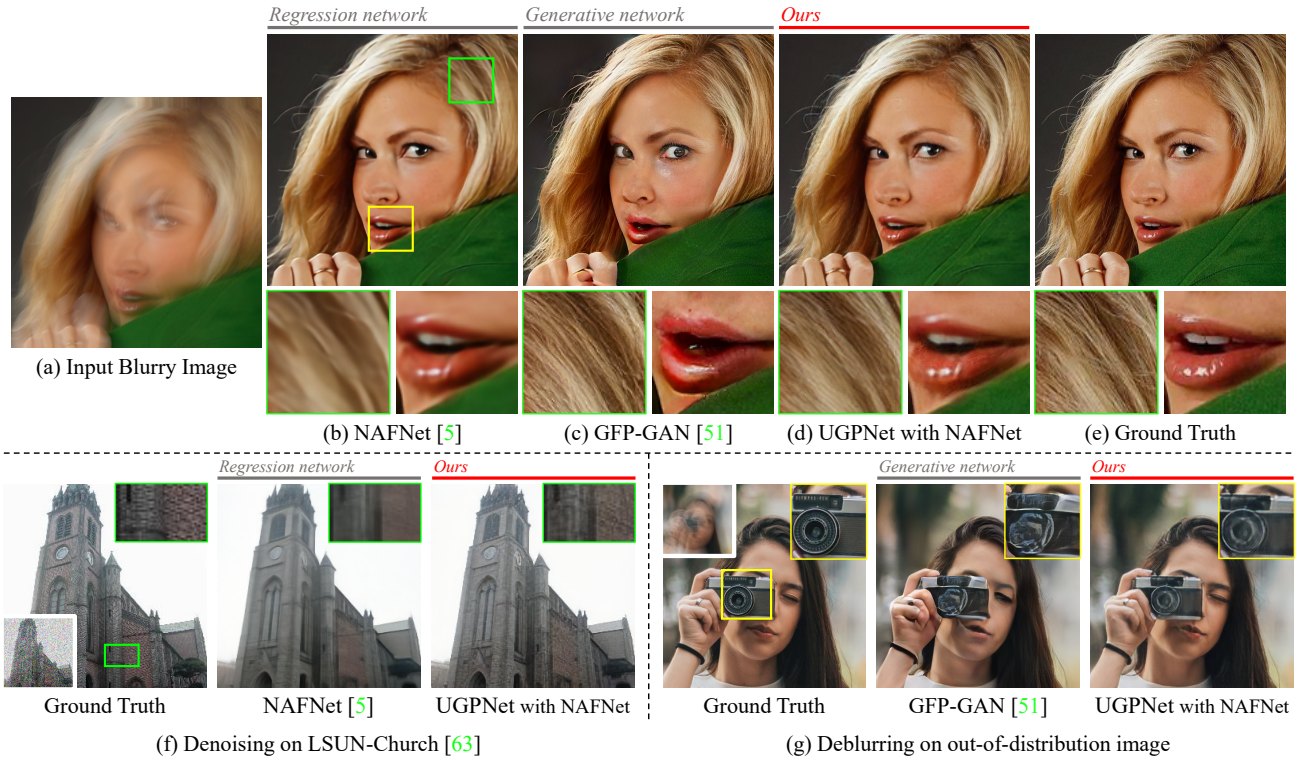


Figure 1. We present UGPNet, a universal image restoration framework that combines the benefits of an existing regression-based restoration network and a generative prior-based network. (a) Given degraded images, e.g. a blurry one, (b) a regression network [5] fails to recover perceptually-realistic details while it recovers the coarse structure of the original image. (c) In contrast, a generative network [51] synthesizes perceptually-realistic high-frequency details while sacrificing structural consistency with the input image. (d) UGPNet allows us to maintain the original structure of the input image and synthesize perceptually-realistic high-frequency details. As a universal framework, (f) UGPNet is applicable to natural images [63]. In addition, (g) it is robust against catastrophic failures that generative prior-based methods encounter when restoring images outside the training distributions.

Abstract

Recent image restoration methods can be broadly categorized into two classes: (1) regression methods that recover the rough structure of the original image without synthesizing high-frequency details and (2) generative methods that synthesize perceptually-realistic high-frequency details even though the resulting image deviates from the original structure of the input. While both directions have been extensively studied in isolation, merging their benefits with a single framework has been rarely studied. In this paper, we propose UGPNet, a universal image restoration framework that can effectively achieve the benefits of both approaches by simply adopting a pair of an existing regression model and a generative model. UGPNet first restores the image structure of a degraded input using a regression

*This work was done at POSTECH.

model and synthesizes a perceptually-realistic image with a generative model on top of the regressed output. UGPNet then combines the regressed output and the synthesized output, resulting in a final result that faithfully reconstructs the structure of the original image in addition to perceptually-realistic textures. Our extensive experiments on deblurring, denoising, and super-resolution demonstrate that UGPNet can successfully exploit both regression and generative methods for high-fidelity image restoration.

1. Introduction

Image restoration [3, 9, 11, 15] has been studied for decades in computer vision and graphics. To recover natural structure from unwanted image degradation such as blur, noise, and low resolution, various non-learning and learning-based approaches have been proposed.

Since the advent of deep neural networks, *regression methods* [5,6,10,50,57,65,67] have emerged as an effective approach for image restoration. Regression methods commonly use convolution neural networks and directly map an input degraded image to a clean image via regression losses such as mean-squared error (MSE). While faithful reconstruction of the original image structure can be achieved, regression methods tend to enforce the resulting image to follow the average image of all potential natural images corresponding to the input. This results in a blurry image without high-frequency details, which harms the perceptual realism of restored images. See Fig. 1 (b) and (f).

Generative approach is another major principle for image restoration that tackles the high-frequency detail problem. Learning natural image statistics using generative models [14, 24, 25] allows synthesizing perceptually-realistic high-frequency image details from degraded input images [4,17,18,51,62]. Exploiting the latent space of generative models via inversion [1, 43, 49] further improves the naturalness of synthesized details. Generative methods, however, heavily depend on the synthesis ability of their generators, which is often limited in recovering the exact structure of the input images. As a result, they suffer from the deviation of synthesized image structures from the original image, resulting in the loss of the identity of the portrait image (Fig. 1 (c)) or severe artifacts (Fig. 1 (g)).

Another important limitation of the existing generative approaches is their limited extendability, which prevents them from enjoying the rapid advance of regression-based methods. The existing generative approaches rely on their own fixed network architectures tightly coupled with generative priors. Thus, it is hard to extend them for other restoration tasks such as deblurring as will be shown in Sec. 4 or to combine with other restoration networks.

In this paper, we present UGPNet, a universal generative prior framework for image restoration that can enjoy the restoration power of state-of-the-art regression-based methods and perceptually-realistic high-frequency details from generative priors. UGPNet is designed as a flexible framework that can plug-and-play an arbitrary regression-based image restoration module. Thanks to its flexibility, we can replace the restoration module with a more suitable one for different tasks or a more effective architecture in the future.

UGPNet is designed with simplicity and effectiveness in mind to make the framework easily adaptable to various restoration tasks and produce high-quality results. Specifically, UGPNet consists of three modules: restoration, synthesis, and fusion. The restoration module is a neural network for regression-based image restoration, whose architecture can be flexibly chosen by users. The synthesis module plays the role of a generative prior and synthesizes high-frequency details suitable for the output of the restoration module. Finally, the fusion module takes the outputs of both restoration and synthesis modules and produces a final result of high fidelity and high perceptual quality (Fig. 1(d)).

We evaluate the effectiveness of UGPNet on multiple restoration tasks, including deblurring, super-resolution, and denoising. We demonstrate that UGPNet successfully brings the generative power of a generative prior to state-of-the-art regression approaches, enabling faithful restoration of image structures as well as the synthesis of high-frequency details of high perceptual quality.

2. Related Work

Regression-based Restoration Networks Deep neural networks have found their applications in image restoration. Various network architectures have been proposed for each task, such as denoising [8, 30, 34, 35, 54, 64, 68, 69], deblurring [10, 21, 29, 37, 41, 47, 48, 61], and super-resolution [7, 13, 36, 39, 42, 71, 72], and also for multiple tasks [5, 6, 50, 57, 65–67]. This flood of research is still ongoing, rapidly breaking performance records every year. On the other hand, regression-based methods commonly suffer from blurry textures. Their regression losses that minimize the distortion between the restored and ground-truth images via a distance metric such as mean-absolute error (MAE) and mean-squared error (MSE) lead to restored results close to an average of all possible realistic images. Unfortunately, such an average image inherently has blurry textures [33].

Our proposed framework, UGPNet, is particularly designed to resolve the blurry texture problem of the regression-based methods by synthesizing realistic textures on top of their results using a generative prior. At the same time, UGPNet is designed to be flexible to allow the plug-and-play of regression-based methods of different tasks. Thanks to the flexibility of UGPNet, we can enjoy the state-of-the-art restoration quality of recent and even future regression-based methods and realistic textures.

Regression Networks with Adversarial Losses To achieve perceptually pleasing restoration results with realis-

tic high-frequency textures, recent methods [12, 31–33, 52, 53, 60] adopt adversarial losses with discriminators. However, adopting only an adversarial loss without exploiting pretrained generator networks of GANs tends to produce unrealistic textures that do not fit the context compared to the synthesis approaches that leverage the network architectures and pretrained prior knowledge of existing generative models, which will be discussed in the following.

Synthesis Networks with Generative Prior To benefit from the remarkable synthesis capability of existing generative models [2, 14, 24, 25], the generative priors learned in generative models have recently been exploited, and enabled high-quality image restoration including blind face restoration [17, 51, 62, 73], super-resolution [4, 18], and colorization [27, 59]. To exploit the generative prior learned in a pretrained GAN model, early works adopt a GAN inversion approach that optimizes the latent code by iteratively minimizing the discrepancy between the input and generated images in consideration of image degradation [16, 40]. Instead of directly optimizing the latent code, recent works utilize encoder networks that estimate the latent code, which is subsequently fed to a GAN generator for synthesizing a clean image [4, 17, 18, 51, 62]. These methods embed a generator into their networks and inherit their ability to synthesize realistic details.

Although these generative-prior-based methods have shown to be able to produce perceptually-realistic textures for image restoration, they suffer from limited representation power. We propose an effective solution that introduces a generative prior on top of high-fidelity regression methods for faithful restoration.

More recently, a few works have explored the use of diffusion models [20] for image restoration. However, these models cannot benefit from regression-based methods like other generative-prior-based methods. Furthermore, the slow inference speed [26, 46, 55, 58] and the strict assumption on the degradation model [26, 55] pose significant challenges to their practical use.

3. UGPNet

Fig. 2 illustrates an overview of UGPNet, which consists of restoration, synthesis, and fusion modules. UGPNet takes a degraded image x as input, and feeds it to the restoration module. The restoration module first recovers the original image structure exploiting an existing regression-based restoration network. The synthesis module then synthesizes high-frequency details by inverting and regenerating the restored image through a generative network. The fusion module combines the latent features from the restoration and synthesis modules to generate a final restored image $\hat{x}$ that maintains both high-frequency details as well as the structure of the original input. In the following, we describe each module in detail.

3.1. Restoration Module

Given a degraded input image x, the restoration module aims to recover authentic image structures via an off-the-shelf pretrained regression-based restoration network. In order to enable flexible selection among diverse regression networks, we propose simple strategies for handling two types of CNN-based regression approaches: direct and residual approaches. Direct approaches such as RRDB-Net [53] directly regress the pixel values of a clean image from an input degraded image. In contrast, residual approaches [5, 6, 28], which are gaining popularity thanks to their effectiveness in handling various restoration tasks, learn to predict a residual image that is later added to the input image to restore a clean image.

In the case of direct approaches, we adopt a direct regression network into our restoration module without any modification, and connect its output x_{reg} to the synthesis module, and its last-layer feature map f_{reg} to the fusion module (Fig. 2).

In the case of residual approaches, we introduce a slight modification for additional quality improvement and for the connection with the fusion module that takes features. Specifically, a residual regression network predicts residual information and adds them to the input image in the image domain at its last stage, i.e., $x_{reg} = R(x) + x$ where $R(x)$ is a residual image predicted by the regression network R. Instead, we modify this stage to perform in the feature domain to extract a feature map on the final output that will be fed to the fusion module. Specifically, our restoration module obtains a restored output x_{reg} as $x_{reg} = R_{mg}(R'(x) + R_{se}(x))$ where R_{mg} is a merging network, and R_{se} is a structure encoder network that embeds an input image x into the feature domain. $R'(x)$ is the last-layer feature map of x estimated by R.

This modification enables us to use the last-layer feature map of R_{mg} as the feature map f_{reg} for the fusion module. Moreover, the residual information is estimated and added back to the input image in a higher-dimensional feature space where different image characteristics can be better represented. As a result, the restoration quality can be further improved, as will be demonstrated in Sec. 4.

3.2. Synthesis Module

The synthesis module takes the regression result x_{reg} from the restoration module and synthesizes a clean image that has similar image structures to x_{reg} and realistic textures. To this end, our synthesis module adopts the GAN inversion approach that embeds an image into the latent space of a pretrained GAN model. Specifically, the synthesis module is composed of an encoder E and a generator G. The encoder estimates the latent code of x_{reg} in the GAN latent space. Then, the generator synthesizes an image with realistic textures from the latent code. Finally, the last-layer

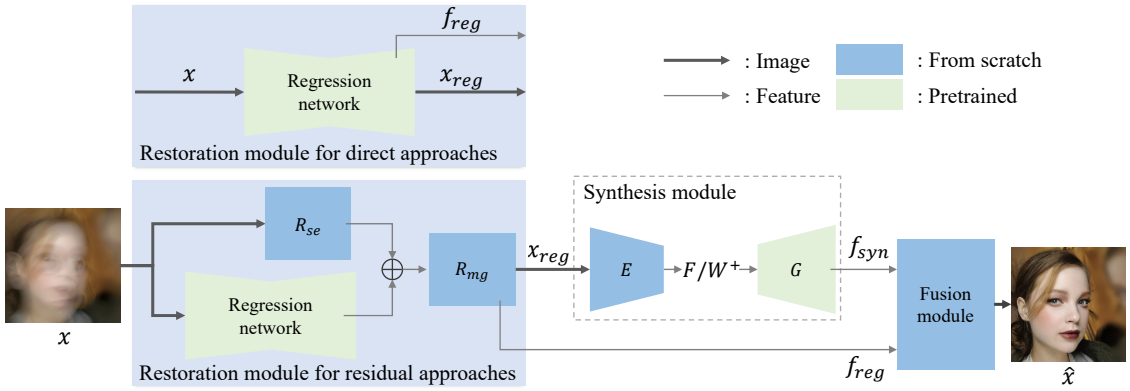
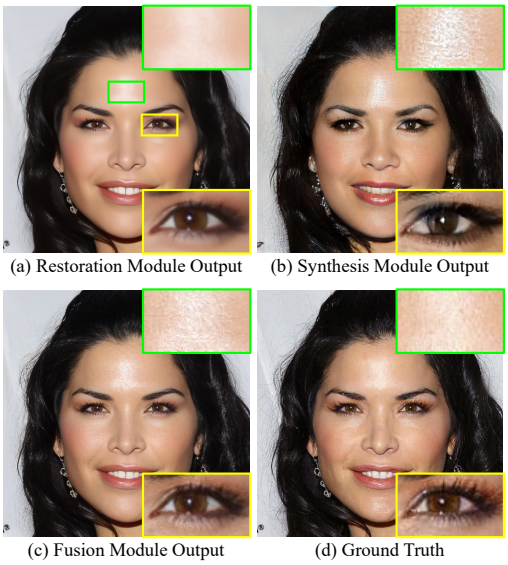


Figure 2. UGPNet consists of three sub-modules: restoration, synthesis, and fusion modules. Given a degraded input image x, the restoration module first recovers the original image structure exploiting a regression network. On top of the regressed output, the synthesis module synthesizes high-frequency details exploiting a generative network. Lastly, the fusion module combines the latent features from both modules to generate a final restored image $\hat{x}$.



(a) Restoration Module Output (b) Synthesis Module Output

(c) Fusion Module Output (d) Ground Truth

Figure 3. An example of image outputs of UGPNet's modules and corresponding ground-truth image. On top of the image structure of restoration module output (a) x_{reg}, the fusion module brings high-frequency details of synthesis module output (b) x_{syn} to generate the final output (c) $\hat{x}$.

feature of the generator f_{syn} is extracted and fed to the fusion module.

For high-quality synthesis, we adopt a pretrained generator of StyleGAN2 [25] for our generator G. Also, to achieve a high-fidelity result that faithfully reconstructs the input restored image x_{reg}, our encoder E embeds x_{reg} into the latent space $\mathcal{F}/\mathcal{W}^+$ proposed by BDInvert [22], which supports GAN inversion of a wide range of out-of-distribution images. While BDInvert directly estimates the latent code in the $\mathcal{F}$ space in a feed-forward manner, it uses an iterative optimization to estimate the latent code in the $\mathcal{W}^+$ space, resulting in significant computational overhead. To resolve this, our encoder employs an additional CNN inspired by the map2style network of pSp [43] to directly estimate the latent code in the $\mathcal{W}^+$ space. The latent code in the $\mathcal{F}/\mathcal{W}^+$ space is then fed to the StyleGAN2 generator G to produce a final synthesis result x_{syn} and its feature map f_{syn}. For f_{syn}, we use the feature map before the last toRGB layer of the StyleGAN2 generator.

The original pSp method uses multiple map2style networks for high-fidelity reconstruction, which incurs significant overhead in terms of model size [43]. On the other hand, as we employ the $\mathcal{F}/\mathcal{W}^+$ space guaranteeing coarse-level alignment and adjust the remaining spatial misalignment in the subsequent fusion module, we could share a single map2style network, which only requires approximately 10% of the parameters without quality degradation. Refer to the supplementary material for the detailed architecture of the encoder.

3.3. Fusion Module

The fusion module combines the authentic image structure of x_{reg} and realistic texture of x_{syn} to generate the final output $\hat{x}$, as shown in Fig. 3. To this end, our fusion module combines the feature outputs f_{reg} and f_{syn} instead of image outputs x_{reg} and x_{syn}. This feature domain fusion helps circumvent potential misalignment between the restoration and synthesis outputs. We design the fusion module with residual blocks and convolution layers as:

$$\hat{x} = Conv(R_{fusion}\,(Conv(f_{syn}) + f_{reg})), \qquad (1)$$

where R_{fusion} is a CNN consisting of eight residual blocks and $Conv$ denotes a 3×3 convolution layer.

3.4. Training

We train each module of UGPNet separately; we first train the restoration module, then the synthesis module, and finally the fusion module. In this section, we describe the training strategy for each module.

Restoration Module In the case of direct regression methods, we use them without any modification as dis-

cussed in Sec. 3.1. Thus, we can assume that they are already carefully trained to restore high-fidelity results, which leaves us nothing to train further. On the other hand, in the case of residual regression methods, we adopt a slight modification including an additional structure encoder and a merging network. Thus, we train the restoration module only in this case. We train the restoration module using a loss $\mathcal{L}_{res}$ to fuse the structural information and residual information in the feature domain. Specifically, we use the same loss function originally used to train the regression network to encourage the output of the merging network x_{reg} to be close to its corresponding ground-truth image. As UGPNet adopts a pretrained regression network, its weights are initialized with its own pretrained weights and further finetuned during the training. The structure encoder and the merging network are trained from scratch.

Synthesis Module To train the synthesis module, we employ a discriminator D and jointly train the encoder E, generator G, and discriminator D. We initialize G and D using the weights of a pretrained StyleGAN2 model, while training E from scratch. The encoder and generator are trained using a loss $\mathcal{L}_{syn}$ defined as:

$$\mathcal{L}_{syn} = \mathcal{L}_1 + \lambda_{per}\mathcal{L}_{per} + \lambda_{adv}\mathcal{L}_{adv}, \tag{2}$$

where $\mathcal{L}_1$ and $\mathcal{L}_{per}$ are an L^1 loss and an LPIPS loss [70] between x_{syn} and its corresponding ground-truth image, respectively. $\mathcal{L}_{adv}$ is an adversarial loss. λ_{per} and λ_{adv} are weights for $\mathcal{L}_{per}$ and $\mathcal{L}_{adv}$, respectively. We adopt the non-saturating loss for $\mathcal{L}_{adv}$ [25]. To train the discriminator, we use a logistic loss following StyleGAN2 [25].

Fusion Module The goals of the fusion module are twofold. The first goal is to produce restored images that faithfully reconstruct the ground-truth clean images. The second goal is to produce restored images with realistic textures by using the output of the synthesis module. Based on these goals, the fusion module is trained using a loss $\mathcal{L}_{fusion}$ defined as:

$$\mathcal{L}_{fusion} = \mathcal{L}_1 + \lambda_{per}\mathcal{L}_{per} + \lambda_{cf}\mathcal{L}_{cf}, \tag{3}$$

where $\mathcal{L}_1$ and $\mathcal{L}_{per}$ are an L^1 loss and an LPIPS loss [70] between the output of the fusion module $\hat{x}$ and its corresponding ground-truth image, respectively. The two loss terms are used for the faithful reconstruction of the ground-truth clean image. $\mathcal{L}_{cf}$ is a patch-wise contextual loss [38] that measures the average distance between the closest feature of x_{syn} for each feature of $\hat{x}$ in a patch-wise manner. $\mathcal{L}_{cf}$ transfers the textures in x_{syn} to the fusion result $\hat{x}$. As x_{syn} is a synthesized image, it may have structures and details unaligned with those of $\hat{x}$. The patch-wise contextual loss handles such an alignment issue by searching the closest feature in a local patch. The mathematical definitions of the losses are provided in the supplementary material.

4. Experiments

Implementation and Evaluation Details For the evaluation of UGPNet, we use NAFNet [5] for denoising and deblurring, and RRDBNet [53] for super-resolution as the regression network unless otherwise noted. We train UGPNet on 70,000 face images of size 512×512 in the FFHQ dataset [24] and test on 3,000 images of the CelebA-HQ dataset [23]. We synthesize a degraded version of a clean image as follows. For denoising, we synthesize noisy images by adding Gaussian ($\mu = 0$, $\sigma = 0.3$) and Poisson noise ($k = 30$). For deblurring, we apply random motion blur sampled from 1,000 motion blur kernels of size 71×71 synthesized following Rim *et al.* [44]. For super-resolution, we apply $8\times$ downsampling with bicubic interpolation. For evaluation, we use PSNR, SSIM [56], LPIPS [70], and FID [19]. PSNR, SSIM, and LPIPS measure how similar the restored image and the ground-truth image are in terms of pixel values, structural similarity, and perceptual similarity, respectively, while FID evaluates the perceptual quality of the restored image. More details can be found in the supplementary material.

4.1. Flexible Selection of Regression Methods

UGPNet supports flexible selection of diverse regression networks. To validate its flexibility, we employ four different regression networks in the restoration module: UNet [45], HINet [6], NAFNet [5], and RRDBNet [53]. Fig. 4 shows the deblurring and super-resolution results of each network and UGPNet equipped with each regression network. UGPNet successfully brings restoration power from the task-specialized regression networks and combines it with a generative prior. Such flexible choice of regression networks allows us to easily equip existing and future regression models with our universal generative prior.

4.2. Comparison with Restoration Methods

We compare UGPNet with recent restoration methods whose source codes are publicly released. Among regression-based methods, we compare UGPNet with Uformer [57] and NAFNet [5] for denoising and deblurring, and RRDBNet [53] for super-resolution. Among generative-prior-based methods, we compare UGPNet with VQFR [17], GFP-GAN [51] and GPEN [62] for denoising and deblurring. We compare two more networks, GLEAN [4] and GCFSR [18] for super-resolution, both of which are specifically designed for super-resolution. We train all the methods from scratch on our dataset using the authors' code except GPEN. For GPEN, we finetuned an officially released model as we found it performs better.

Fig. 5 shows qualitative comparisons on denoising and deblurring. The regression-based methods (Uformer [57] and NAFNet [5]) fail to synthesize sharp details compared to the generative-prior-based methods. The generative methods (GFP-GAN [51], GPEN [62], and VQFR [17])

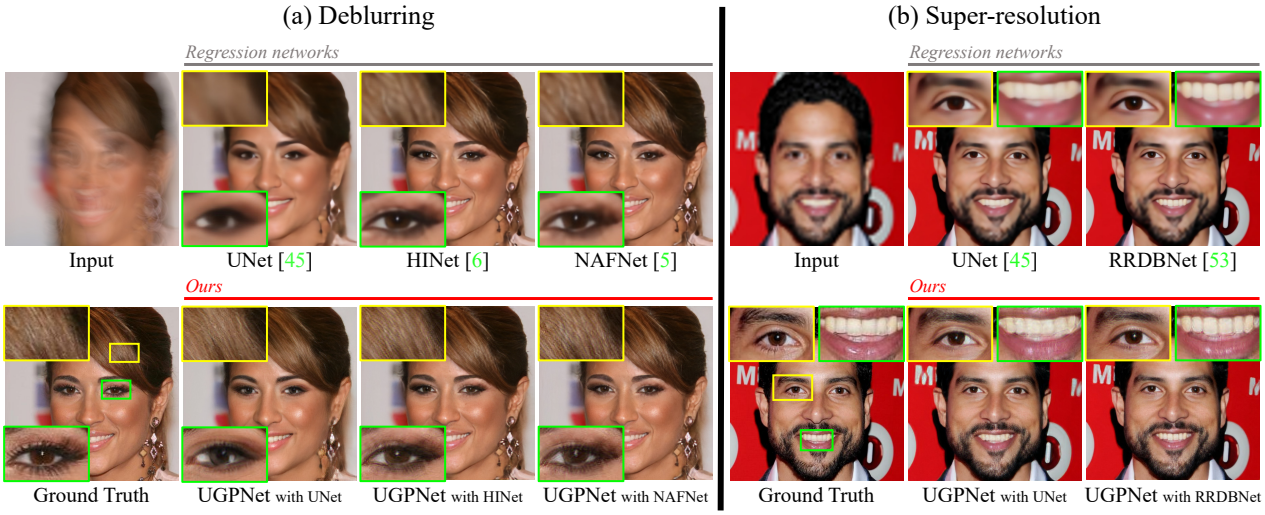


Figure 4. UGPNet allows flexible selection of diverse regression networks in the restoration module. We show restoration results using regression models (UNet [45], HINet [6], NAFNet [5] and RRDBNet [53]) on the top row for (a) deblurring and (b) super-resolution. We can equip any regression models into UGPNet that synthesizes perceptually-realistic high-frequency details, as shown in the bottom row.

	(a) Denoising				*(b) Deblurring*					*(c) Super-resolution*			
Method	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**FID ↓**	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**FID ↓**	Method	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**FID ↓**
NAFNet [5]	30.23	**0.82**	**0.30**	**39.32**	29.53	0.81	0.31	42.18	RRDBNet [53]	**30.03**	**0.80**	0.32	52.42
Uformer [57]	**30.24**	**0.82**	0.31	40.50	**30.38**	**0.83**	**0.29**	**37.02**	ESRGAN [53]	26.88	0.69	**0.31**	**8.92**
GFP-GAN [51]	**27.96**	**0.76**	**0.31**	11.66	**23.09**	**0.63**	**0.34**	**10.73**	GFP-GAN [51]	27.38	0.72	0.28	6.88
GPEN [62]	27.50	0.73	0.35	11.84	20.60	0.56	0.43	15.48	GPEN [62]	26.35	0.69	0.32	9.89
VQFR [17]	27.45	0.74	**0.31**	**11.64**	21.80	0.59	0.35	10.81	VQFR [17]	27.16	0.71	0.27	6.41
UGPNet	**29.20**	**0.78**	**0.31**	**10.24**	**28.64**	**0.76**	**0.31**	**8.02**	GLEAN [4]	27.74	0.72	0.28	6.47
									GCFSR [18]	**28.16**	**0.74**	**0.26**	**5.72**
									UGPNet	**28.70**	**0.74**	**0.30**	**6.70**

Table 1. Quantitative comparison of the restoration quality of different methods. The methods marked in yellow are regression methods, and the methods marked in orange are generative methods. The best scores in each category are marked in bold. In the case of (a) denoising and (b) deblurring, UGPNet records significantly higher PSNR and SSIM scores than the generative methods but slightly lower scores than the regression methods. In LPIPS, UGPNet achieves the second best scores. In FID, UGPNet achieves the best scores. In the case of (c) super-resolution, UGPNet reports higher PSNR and SSIM scores with comparable LPIPS and FID scores compared to all the generative models except for GCFSR [18], which is specifically designed for super-resolution. Compared to GCFSR, UGPNet shows comparable results with a higher PSNR score.

are unable to restore faithful image structures. UGPNet succeeds in the faithful restoration and high-frequency synthesis of image structures and details. We further report quantitative evaluation in Tab. 1 (a) and (b). In PSNR and SSIM, UGPNet records higher scores than the generative methods, indicating structural consistency. UGPNet has a slightly lower PSNR scores than the best regression methods, due to the synthesized high-frequency details. In LPIPS, UGPNet achieves the second best scores. In FID, UGPNet achieves the best scores, demonstrating that it generates perceptually-plausible high-frequency details.

Fig. 6 and Tab. 1 (c) show the qualitative and quantitative comparisons on super-resolution. As the super-resolution requires more aggressive high-frequency generation, a regression method (RRDBNet [53]) produces perceptually low-quality blurry images with high FID scores. Adopting an adversarial loss (ESRGAN [53]) produces sharp images and lowers the FID score, but fails to synthesize realistic textures as the generative prior-based methods. Both generative methods and ours succeed in synthesizing realistic high-frequency details showing comparable results thanks to powerful generative priors. Quantitatively, UGPNet achieves higher PSNR and SSIM scores with comparable LPIPS and FID scores compared to all the generative methods except GCFSR [18], which is specifically designed for super-resolution. Compared to GCFSR, UGPNet achieves similar performance, recording a slightly better PSNR score and slightly worse LPIPS and FID scores.

Model Complexity We compare the number of parameters and the inference speed on deblurring in Tab. 2. Although UGPNet combines a regression network (NAFNet [5]) and a generative network, it has a comparable number of parameters with the recent models. This is because the other methods require a lot of parameters in their encoders or decoders, whereas our lightweight model design achieves better performance with fewer parameters.

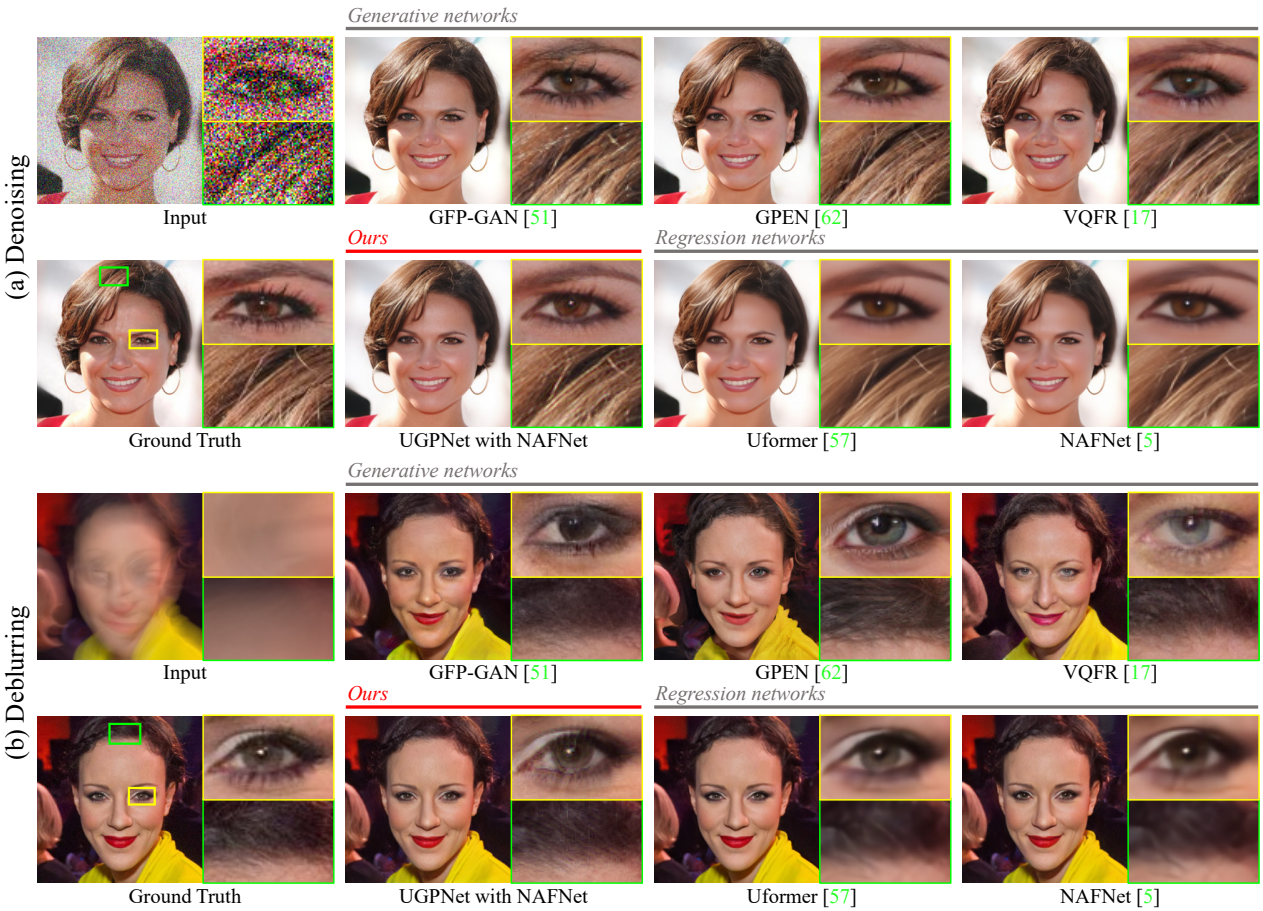


Figure 5. Qualitative comparison of (a) denoising and (b) deblurring methods: regression methods (Uformer [57] and NAFNet [5]), generative methods (GFP-GAN [51], GPEN [62], VQFR [17]), and UGPNet with NAFNet [5]. UGPNet recovers authentic image structure and colors compared to generative methods while synthesizing sharp high-frequency details compared to regression methods.

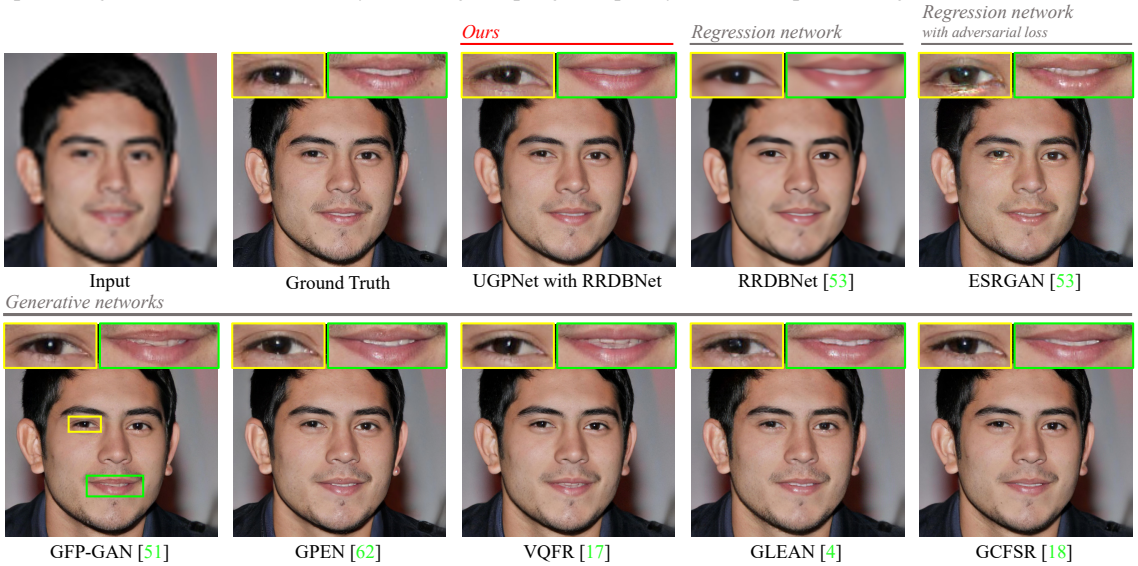


Figure 6. Qualitative comparison of super-resolution methods: regression methods (RRDBNet [53] and ESRGAN [53]), generative methods (GFP-GAN [51], GPEN [62], VQFR [17], GLEAN [4], GCFSR [18]), and UGPNet with RRDBNet [53]. UGPNet succeeds in generating sharp, realistic high-frequency details compared to regression methods, and shows comparable performance to generative methods.

	NAFNet [5]	GFP-GAN [51]	VQFR [17]	Uformer [57]	UGPNet
Param (M)	17.1	76.2	76.6	50.9	69.6
Time (ms)	45.2	22.7	181.2	170.6	90.5

Table 2. Comparison of the parameter numbers and the inference times with an NVIDIA GeForce RTX 3090 GPU. UGPNet has a comparable number of parameters with the others and it is faster than transformer-based [57] and dictionary-based models [17].

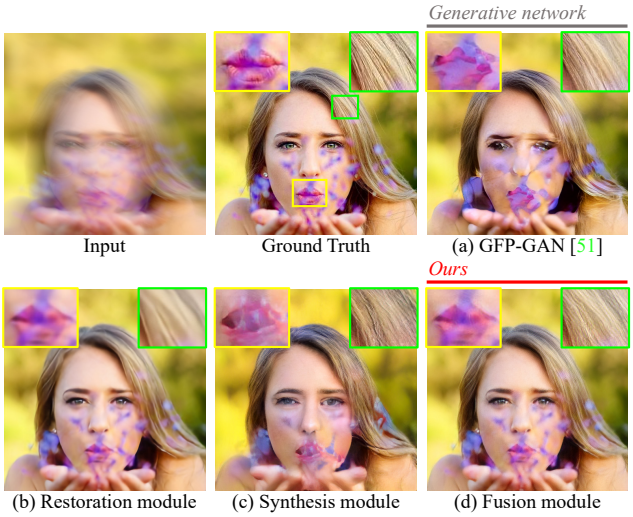


Figure 7. UGPNet supports a wider range of images compared to the generative methods. (a) A generative method fails to synthesize images outside the training distribution and produces artifacts on the face. While (c) our synthesis module also introduces artifacts, (d) our fusion module successfully recovers a clean image without artifacts as it selectively uses the synthesized high-frequency details on top of (b) the result of the restoration module.

In terms of inference speed, UGPNet is slower than the lightweight models, but it is faster than computationally heavy models such as the transformer-based model [57] and the dictionary-based model [17].

Restoration of Out-of-Distribution Images UGPNet is robust against catastrophic failures that generative prior-based methods suffer from when restoring images outside the training distributions, as shown in Fig. 1 (g) and Fig. 7. These robustness of UGPNet stems from the fusion module that adaptively combines features from the restoration and synthesis modules to achieve faithful and natural-looking restoration. For a deeper analysis, Fig. 7 compares outputs of different modules on a degraded input image outside the training distribution. In the figure, the restoration module robustly restores the image structures thanks to the well-generalized regression network. The synthesis module synthesizes high-frequency details from the output of the restoration module but it also introduces artifacts in the face as the image is outside the training distribution of the generative prior. Despite such synthesis artifacts, the fusion module selectively uses the synthesized high-frequency details on top of the result of the restoration module and produces the final result that is faithfully reconstructed without artifacts and has natural-looking high-frequency details.

Configurations	w/ NAFNet [5]			w/ HINet [6]		
	PSNR ↑	**SSIM** ↑	**FID** ↓	**PSNR** ↑	**SSIM** ↑	**FID** ↓
(a) w/o R_{se} and R_{mg}	27.43	0.75	8.33	27.30	0.72	9.70
(b) w/ fusion module combining x_{reg} and x_{syn}	28.49	**0.76**	8.54	28.16	0.75	9.68
(c) UGPNet	**28.64**	**0.76**	**8.02**	**28.49**	**0.76**	**8.34**

Table 3. We validate network components of UGPNet on deblurring. (a) Without the structure encoder and the merging network, the restoration module has difficulty in extracting features with structural information, leading to performance drop. (b) UGPNet combining images rather than (c) features also leads to slight performance drop.

4.3. Ablation Study

Network Components Tab. 3 quantitatively validates the network components of UGPNet on deblurring. In the restoration module, introducing the structure encoder and the merging network for the residual regression network leads to performance improvement (Tab. 3 (a) and (c)). This improvement demonstrates that conveying additional information from the input helps the restoration feature f_{reg} contain authentic image structures. In the fusion module, combining the restoration feature f_{reg} and the synthesis feature f_{syn} rather than the image alternatives, x_{reg} and x_{syn}, leads to slight performance improvement (Tab. 3 (b) and (c)).

5. Conclusion

This paper proposed UGPNet, a universal generative prior framework for image restoration. UGPNet supports diverse regression networks developed for each task and brings the generative power of recent generative prior on top of it. Through extensive experiments on deblurring, denoising, and super-resolution, we demonstrated that UGPNet succeeds in high-quality image restoration, enabling faithful restoration with realistic high-frequency details.

Limitation Although UGPNet is robust against failure of the generative prior, it may fail if the regression method fails, as shown in UGPNet with UNet in Fig. 4. Also, UGPNet is less favorable to PSNR and SSIM scores compared to the regression method, and produces less sharp images than its backbone generative model (StyleGAN2 [25]), as it aims at higher-fidelity results. It would be an interesting future direction to address these challenges, e.g., we may measure the uncertainty of the regression result and use it to adaptively synthesize necessary details for higher fidelity.

Acknowledgement This work was supported by the NRF grants (No.2023R1A2C200494611, No.2018R1A5A1060031) and IITP grant (No.2019-0-01906, Artificial Intelligence Graduate School Program(POSTECH)) funded by the Korea government (MSIT) and Samsung Electronics Co., Ltd. Seung-Hwan Baek is partly supported by Korea NRF (RS-2023-00211658, 2022R1A6A1A03052954), Korea MOTIE (NTIS1415187366-20025752) and Samsung Research Funding Center (SRFCIT1801-52).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019. 2

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3

[3] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 2, pages 60–65. Ieee, 2005. 2

[4] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14245–14254, 2021. 2, 3, 5, 6, 7

[5] Liangyu Chen, Xiaojie Chu, Xiangyu Zhang, and Jian Sun. Simple baselines for image restoration. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 17–33. Springer, 2022. 1, 2, 3, 5, 6, 7, 8

[6] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 182–192, 2021. 2, 3, 5, 6, 8

[7] Xiangyu Chen, Xintao Wang, Jiantao Zhou, and Chao Dong. Activating more pixels in image super-resolution transformer. *arXiv preprint arXiv:2205.04437*, 2022. 2

[8] Shen Cheng, Yuzhi Wang, Haibin Huang, Donghao Liu, Haoqiang Fan, and Shuaicheng Liu. Nbnet: Noise basis learning for image denoising with subspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4896–4906, 2021. 2

[9] Sunghyun Cho and Seungyong Lee. Fast motion deblurring. In *ACM SIGGRAPH Asia 2009 papers*, pages 1–8. 2009. 2

[10] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4641–4650, 2021. 2

[11] 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. 2

[12] Xin Deng, Ren Yang, Mai Xu, and Pier Luigi Dragotti. Wavelet domain style transfer for an effective perception-distortion tradeoff in single image super-resolution. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3076–3085, 2019. 3

[13] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014. 2

[14] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 2, 3

[15] Rob Fergus, Barun Singh, Aaron Hertzmann, Sam T Roweis, and William T Freeman. Removing camera shake from a single photograph. In *Acm Siggraph 2006 Papers*, pages 787–794. 2006. 2

[16] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3012–3021, 2020. 3

[17] Yuchao Gu, Xintao Wang, Liangbin Xie, Chao Dong, Gen Li, Ying Shan, and Ming-Ming Cheng. Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, pages 126–143. Springer, 2022. 2, 3, 5, 6, 7, 8

[18] Jingwen He, Wu Shi, Kai Chen, Lean Fu, and Chao Dong. Gcfsr: a generative and controllable face super resolution method without facial and gan priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1889–1898, 2022. 2, 3, 5, 6, 7

[19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017. 5

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[21] Xiaobin Hu, Wenqi Ren, Kaicheng Yu, Kaihao Zhang, Xiaochun Cao, Wei Liu, and Bjoern Menze. Pyramid architecture search for real-time image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4298–4307, 2021. 2

[22] Kyoungkook Kang, Seongtae Kim, and Sunghyun Cho. Gan inversion for out-of-range images with geometric transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13941–13949, 2021. 4

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2, 3, 5

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2, 3, 4, 5, 8

[26] Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *arXiv preprint arXiv:2201.11793*, 2022. 3

[27] Geonung Kim, Kyoungkook Kang, Seongtae Kim, Hwayoon Lee, Sehoon Kim, Jonghyun Kim, Seung-Hwan Baek, and Sunghyun Cho. Bigcolor: Colorization using a generative color prior for natural images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 350–366. Springer, 2022. 3

[28] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 3

[29] Kiyeon Kim, Seungyong Lee, and Sunghyun Cho. Mssnet: Multi-scale-stage network for single image deblurring. *arXiv preprint arXiv:2202.09652*, 2022. 2

[30] Yoonsik Kim, Jae Woong Soh, Gu Yong Park, and Nam Ik Cho. Transfer learning from synthetic to real-noise denoising with adaptive instance normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3482–3492, 2020. 2

[31] Orest Kupyn, Volodymyr Budzan, Mykola Mykhailych, Dmytro Mishkin, and Jiří Matas. Deblurgan: Blind motion deblurring using conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8183–8192, 2018. 3

[32] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8878–8887, 2019. 3

[33] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic 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, 3

[34] Wooseok Lee, Sanghyun Son, and Kyoung Mu Lee. Ap-bsn: 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. 2

[35] Yang Liu, Zhenyue Qin, Saeed Anwar, Pan Ji, Dongwoo Kim, Sabrina Caldwell, and Tom Gedeon. Invertible denoising network: A light solution for real noise removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13365–13374, 2021. 2

[36] Yuqing Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. Progressive multi-scale residual network for single image super-resolution. *arXiv preprint arXiv:2007.09552*, 2020. 2

[37] Xintian Mao, Yiming Liu, Wei Shen, Qingli Li, and Yan Wang. Deep residual fourier transformation for single image deblurring. *arXiv preprint arXiv:2111.11745*, 2021. 2

[38] Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 768–783, 2018. 5

[39] Yiqun Mei, Yuchen Fan, Yuqian Zhou, Lichao Huang, Thomas S Huang, and Honghui Shi. Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5690–5699, 2020. 2

[40] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 3

[41] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 2

[42] Ben Niu, Weilei Wen, Wenqi Ren, Xiangde Zhang, Lianping Yang, Shuzhen Wang, Kaihao Zhang, Xiaochun Cao, and Haifeng Shen. Single image super-resolution via a holistic attention network. In *European conference on computer vision*, pages 191–207. Springer, 2020. 2

[43] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021. 2, 4

[44] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, pages 184–201. Springer, 2020. 5

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5, 6

[46] 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*, 2022. 3

[47] Maitreya Suin, Kuldeep Purohit, and AN Rajagopalan. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3606–3615, 2020. 2

[48] Xin Tao, Hongyun Gao, Xiaoyong Shen, Jue Wang, and Jiaya Jia. Scale-recurrent network for deep image deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8174–8182, 2018. 2

[49] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2

[50] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5769–5780, 2022. 2

[51] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021. 1, 2, 3, 5, 6, 7, 8

[52] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1905–1914, 2021. 3

[53] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 3, 5, 6, 7

[54] Yuzhi Wang, Haibin Huang, Qin Xu, Jiaming Liu, Yiqun Liu, and Jue Wang. Practical deep raw image denoising on mobile devices. In *European Conference on Computer Vision*, pages 1–16. Springer, 2020. 2

[55] Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022. 3

[56] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[57] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17683–17693, 2022. 2, 5, 6, 7, 8

[58] Jay Whang, Mauricio Delbracio, Hossein Talebi, Chitwan Saharia, Alexandros G Dimakis, and Peyman Milanfar. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16293–16303, 2022. 3

[59] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14377–14386, 2021. 3

[60] Jun Xiao, Tianshan Liu, Rui Zhao, and Kin-Man Lam. Balanced distortion and perception in single-image super-resolution based on optimal transport in wavelet domain. *Neurocomputing*, 464:408–420, 2021. 3

[61] Dan Yang and Mehmet Yamac. Motion aware double attention network for dynamic scene deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1113–1123, 2022. 2

[62] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021. 2, 3, 5, 6, 7

[63] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1

[64] Zongsheng Yue, Hongwei Yong, Qian Zhao, Deyu Meng, and Lei Zhang. Variational denoising network: Toward blind noise modeling and removal. *Advances in neural information processing systems*, 32, 2019. 2

[65] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 2

[66] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *European Conference on Computer Vision*, pages 492–511. Springer, 2020. 2

[67] 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. 2

[68] 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. 2

[69] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for cnn-based image denoising. *IEEE Transactions on Image Processing*, 27(9):4608–4622, 2018. 2

[70] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[71] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 2

[72] Yulun Zhang, Donglai Wei, Can Qin, Huan Wang, Hanspeter Pfister, and Yun Fu. Context reasoning attention network for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4278–4287, 2021. 2

[73] Feida Zhu, Junwei Zhu, Wenqing Chu, Xinyi Zhang, Xiaozhong Ji, Chengjie Wang, and Ying Tai. Blind face restoration via integrating face shape and generative priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7662–7671, 2022. 3