BIGPrior: Towards Decoupling Learned Prior Hallucination and Data Fidelity in Image Restoration

Majed El Helou, *Member, IEEE,* Sabine Süsstrunk, *Fellow, IEEE.*

***Abstract*—Classic image-restoration algorithms use a variety of priors, either implicitly or explicitly. Their priors are hand-designed and their corresponding weights are heuristically assigned. Hence, deep learning methods often produce superior image restoration quality. Deep networks are, however, capable of inducing strong and hardly predictable hallucinations. Networks implicitly learn to be jointly faithful to the observed data while learning an image prior; and the separation of original data and hallucinated data downstream is then not possible. This limits their wide-spread adoption in image restoration. Furthermore, it is often the hallucinated part that is victim to degradation-model overfitting.**

We present an approach with decoupled network-prior based hallucination and data fidelity terms. We refer to our framework as the Bayesian Integration of a Generative Prior (BIGPrior). Our method is rooted in a Bayesian framework and tightly connected to classic restoration methods. In fact, it can be viewed as a generalization of a large family of classic restoration algorithms. We use network inversion to extract image prior information from a generative network. We show that, on image colorization, inpainting and denoising, our framework consistently improves the inversion results. Our method, though partly reliant on the quality of the generative network inversion, is competitive with state-of-the-art supervised and task-specific restoration methods. It also provides an additional metric that sets forth the degree of prior reliance per pixel relative to data fidelity.

***Index Terms*—Deep image restoration, data fidelity, network hallucination, learned prior.**

I. Introduction

IMAGE restoration recovers original images from degraded observations. It is based on two fundamental aspects, specifically, the relation to the observed data and the additional assumptions or image statistics that can be considered for the restoration. The relation to the observed data is referred to as *data fidelity*. The remaining information, brought in by the restoration method based on prior assumptions, is referred to as *prior hallucination*. It is termed hallucination because the added information is derived from a general model or assumption and might not faithfully match the sample image.

The data fidelity and prior terms emerge theoretically in the Maximum A Posteriori (MAP) formulation, but can also be implicitly induced by the restoration algorithms. For instance, non-local means [1] and Block-Matching and 3D filtering (BM3D) [2] utilize the prior assumption that there exists different similar patches within an image. Diffusion [3] methods build on local smoothness assumptions. Data fidelity is typically enforced through the squared norm [4] that is equivalent to a MAP-based Gaussian noise model.

Classic image-restoration algorithms often rely on optimizations over explicit priors. An advantage of explicitly defined priors is the ability to easily control the relative relation between the weight of the data fidelity term and the weight (β) of the prior term. The general approach consists of an optimization

$$\arg\min_x \psi_d(f'(x), y) + \beta \cdot \psi_p(f''(x)), \tag{1}$$

where y is the observation, f' and f'' are various manipulation functions to match the degradation model (for f') and for instance to extract specific components such as frequency bands (for f''), ψ_d enforces the data fidelity, and ψ_p enforces the prior information. The optimal point is the estimate of the original image x. By making the prior term explicit, it is possible to have control over its contribution hence often better intuition and understanding of the reliability of the final restoration result. However, we note two shortcomings of these methods and we expand upon them in the following: (1) β is not adapted based on the confidence in the fitness of the prior, and (2) the priors are hand-designed heuristics.

(1) The parameter β should be inversely related to the quality of the observed degraded signal, but it should also be directly related to how well the assumed prior corresponds to the input image distribution or statistics. Although some methods, discussed in the section on related work, adjust their priors to the input data, they do not control β based on the confidence in the fitness of the prior to the current sample. (2) Recent methods with implicit data-learned priors, notably relying on deep Convolutional Neural Networks (CNNs), outperform the classic methods with hand-designed priors on various image restoration tasks. This is due to the rich prior learned by discriminative networks or generative networks that, with adversarial training, can even learn image distributions to synthesize new realistic photos [5]–[7]. It is worth noting, however, that domain-specific prior information can still be explicitly enforced to improve the performance of the networks [8], [9].

One shortcoming of the deep learning methods is the loss of interpretability and control between data fidelity and prior-based hallucination. Given an image restored by a network, it

Both authors are with the School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland.

Contact author's email address: majed.elhelou@epfl.ch

Our code and models are publicly available at https://github.com/majedelhelou/BIGPrior

is not possible to know how faithful it is to the observed signal versus how much prior-based hallucination was integrated in the image. We call prior-based hallucination the information coming from the learned prior rather than being derived from the observed data. These prior-based hallucinations are not always reliable and can be prone to overfitting [10]. Hence, it is important to have a grasp of the prior hallucination taking place in the restoration process.

To obtain decoupled prior-based hallucination and data fidelity terms, we propose a novel framework that we call the Bayesian Integration of a Generative Prior (BIGPrior framework). We replace the implicit data prior learned in feed-forward restoration networks with an explicit generative-network prior. This prior is then integrated following a MAP setting, where the data fidelity and prior terms are combined with a fusion weight that is adaptive to both. The BIGPrior framework is a generalization of a large family of classic restoration methods where the prior and its contribution weight are both learned, and the weight can adapt to both the signal quality *and* the fitness of the prior to the observed data.

Our framework structurally provides a reliable metric for per-pixel data fidelity in the final output to answer the question, "How much hallucination is there - at worst - in the output?". We present and analyze this metric by using blind denoising experiments. We also apply our method to various image restoration tasks and show consistent improvements, notably over the direct use of the generative prior, while additionally providing our faithfulness, a.k.a. data fidelity, metric.

II. Related Work

A. Classic Image Restoration

A variety of classic restoration methods, such as Non-Local Means (NLM) [1], BM3D [2], their variants [12], [13] or combinations with sparse coding [14], [15], and diffusion-based methods [3], [16], make use of various prior assumptions on self-similarity or frequency-content distribution. Other algorithms formulate the prior explicitly. For instance, dictionary-based methods [17] that assume images can be well represented by a fixed set of elements, which we discuss in the next section. Other examples are shrinkage methods [18], [19]. They can be directly connected with the family of MAP estimators, by deriving from the foundational work of Bayes and Laplace [20]. Considering an example with a hyper-Laplacian prior on image gradients, originally used in the context of deblurring in [4], optimizing the MAP negative log-likelihood

$$\arg\min_x -\log(P_{Y|X}(y|x)P_X(x)), \tag{2}$$

where $P_{Y|X}(\cdot,\cdot)$ is the conditional probability distribution of the random variable Y given X and $P_X(\cdot)$ is the probability distribution of X, yields the estimator

$$\hat{x} = \arg\min_x ||x-y||_2^2 + \beta \cdot \sum_{j=1}^{J} |x \otimes f_j|^\gamma, \tag{3}$$

where y is the signal we observe, $\hat{x}$ is the estimate of the target x, $\{f_j\}$ are J first-order derivative filters and β is a weight parameter. Setting γ to one, with the corresponding filters, gives the special case of total-variation methods [21]. Generally these approaches are an optimization of the form

$$\hat{x} = \arg\min_x \psi_d(x,y) + \beta \cdot \psi_p(\mathcal{T}(x)), \tag{4}$$

where ψ_d is the data-fidelity loss term, and ψ_p incorporates the prior information on a transformation $\mathcal{T}$ of x that could be the identity. $\mathcal{T}$ can also be based on derivatives [4], or wavelet [22] and other sparsifying transformations. For instance, Weighted Nuclear Norm Minimization (WNNM) [23] assumes that subsets of similar image patches are low-rank and uses a weighted nuclear norm for the low-rank minimization problem on similar patch groups. As with many classic image denoising methods, WNNM adapts β based on the noise level and controls the data fidelity weight as such. However, as we noted in our introduction, these methods face two shortcomings. First, β is not adapted based on the confidence in the prior given the degraded observation, but only on the quality of the latter. Hence, it is adapted based on the signal quality, such as the noise level, but also only following certain heuristics. Second, the prior itself is fixed based on hand-designed heuristics. We preserve an interpretable control over the contribution of the prior and decouple it from data fidelity, and we exploit learned network priors and increasing the flexibility in the fusion weight. Therefore, this weight is *learned* in our framework and can adapt both to the quality of the observed data, as well as the fitness of the prior, given the test observation.

B. Deep Neural Networks

Convolutional neural networks are able to learn rich image priors. These learned priors have improved image restoration results for various applications [24]–[27]. These methods use sample-based learning and can extract prior information from large image datasets. Further improvements are achieved with discriminative learning, notably discriminators that take into account the image degradation model [28]. This has enabled these deep learning methods to improve the state of the art on many restoration tasks [29]. However, the learned priors are implicit, meaning neither the prior nor its contribution can be disentangled from the data fidelity component in the final restored output. In a concurrent paper, Dong *et al.* propose to learn the data fidelity term and the prior term with separate neural network components [30]. The learning, and the final restoration output, are however carried out within a unified optimization. This makes the disentanglement of the two components and their relative contribution weights not possible downstream in the spatial domain. As recently shown for super-resolution and denoising tasks [10], neural networks can learn a frequency-conditional hallucination that is prone to overfitting to the training degradation models. Another recent example is in 3D reconstruction [31], where networks learn to recognize observations and to use memorized data samples, rather than to perform the reconstruction. In other words, the prior contribution can dominate over the data fidelity. Despite its importance, controlling this trade-off is, however, not attainable within the neural-network-based solutions. Our

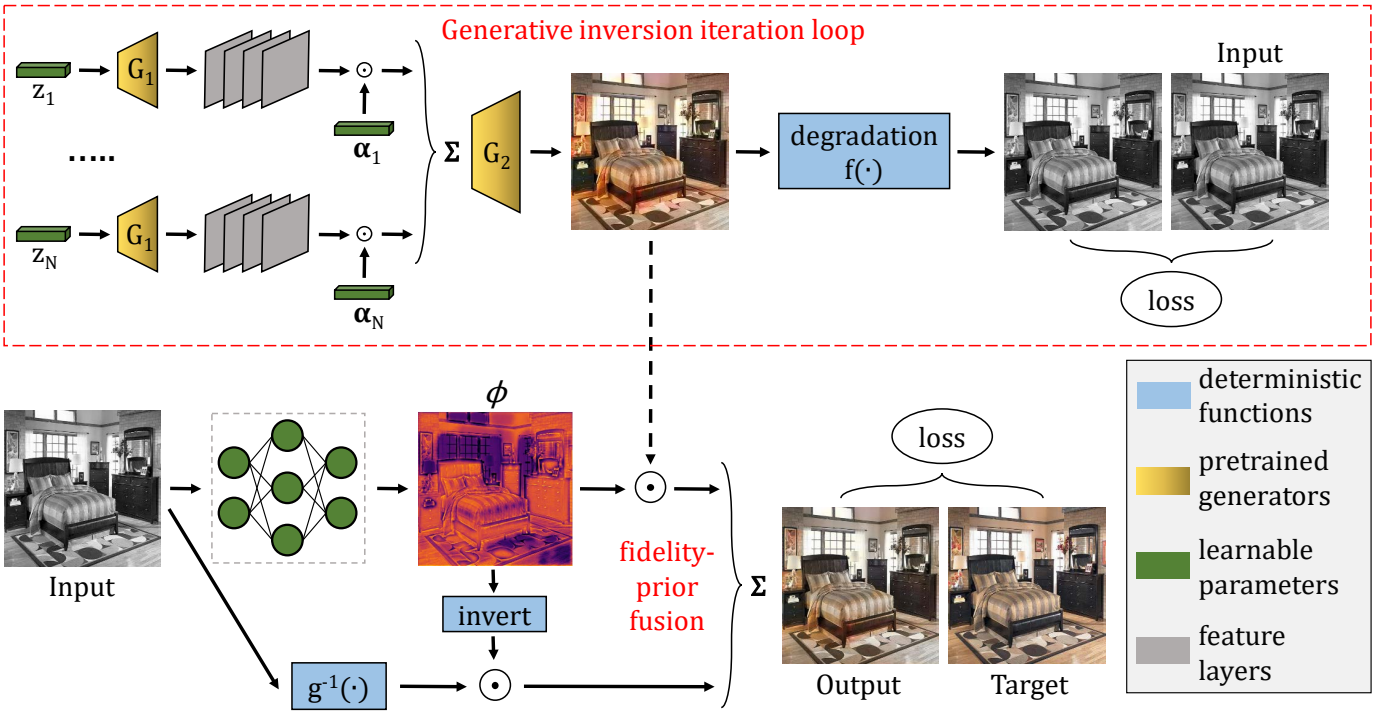


Fig. 1. Weights that are optimized are shown in green, and the sub-networks of the pre-trained generative network are shown in yellow. The generative network inversion process is optimized over a fixed set of iterations, which regularizes the output [11]. The final output is obtained through the fusion of the prior-based hallucination and the signal information, based on our ϕ map estimation.

proposed framework enables us to exploit the strength of learned network priors and keep both control and insight over the data fidelity and prior trade-offs.

Extracting prior information from neural networks is possible through an inversion process [32], [33]. By searching the network-learned space of image distributions, it is possible to project on it in a fashion similar to dictionary-based methods. Generative networks are sufficiently powerful to be trained to learn different distributions, such as image or noise distributions [34]. A network inversion is carried out in [35], where it is used for different image processing tasks. We discuss this in more detail in the following section. A generative network inversion is also conducted in [36]. However, the method performs a fine-tuning of the pre-trained generator that goes against our objective to project on a fixed learned space. We also emphasize that our goal is not to improve such priors, rather to use them in our framework as image-projection spaces.

C. Signal Adaptation of Priors

As described in our discussion on classic methods, some of them [23] adapt the weight assigned to their prior term according to the quality of the observed signal. However, the fitness of the chosen prior can itself be image dependent. In other words, the prior can be accurate on certain images, but not as fit to be applied to others. Yet this is rarely accounted for in the literature. In the content-aware image prior presented in [37], although the weight of the prior itself is not adaptive, the hyper-Laplacian prior used is tweaked to adjust to the texture in the observed signal. Similarly, the method in [38] carefully selects its filters upon processing of the observed signal, hence altering its implicit priors. Also in the same spirit, some recent deep learning methods have tried to adapt to the observed inputs, through self-supervised weight modification [39], or novel learning [40]. This approach has even appeared in recent classification work to adjust to distribution shifts [41]. Such methods address the issue of the fitness of the prior to the given input by modifying the former on the fly. However, once a prior is selected, its fitness relative to the observed signal's quality is dismissed. The weight of the prior term is therefore not adaptive, and the prior's contribution cannot be decoupled from data fidelity.

III. Method

In designing our method, we address the shortcomings discussed in the introduction. We present a framework where the prior and the data-fidelity terms are explicit. This enables us to exploit the modeling strength of deep neural networks for the prior and enables us to learn a weight between the prior and the data fidelity that is doubly adaptive to the quality of the observation and to the fitness of the prior to the input's distribution. Rather than combining the contribution of the prior and the data-fidelity terms through an optimization, we explicitly enforce their fusion in the final output. This explicit decoupling of the two terms enables us, as well as downstream applications, to gain in restoration interpretability. In this

section, we present the mathematical details of our proposed method and its relation to classic families of restoration algorithms. We also present a network-based prior that relies on generative-network projection and introduce our approach for learning the adaptive weight without guided supervision.

A. Mathematical Formulation

Given an observed signal y that is a degraded version of the image x, our restoration estimate $\hat{x}$ is formulated as

$$\hat{x} = \underbrace{(1-\phi(y;\theta_1)) \odot g^{-1}(y)}_{data\ \ fidelity} + \underbrace{\phi(y;\theta_1) \odot G(z^*;\theta_2)}_{prior}, \quad (5)$$

where $g^{-1}(\cdot)$ is a bijective function that we discuss in what follows, $\phi(\cdot;\theta_1)$ is an estimator for the fusion factor, parameterized by θ_1, and that assigns adaptive weights to the prior-based hallucination and the data fidelity. It is a generalization of β that we learn internally from sample-based training. $G(z^*;\theta_2)$ is the prior-based hallucination, parameterized by θ_2, described in detail in the following, and $\odot$ is the pixel-wise multiplication operator. To ensure a very strict lossless data-fidelity term, we restrict $g^{-1}(\cdot)$ to the set of bijective functions. We can choose it such that $g(\cdot)$ is close to the degradation model $f(\cdot)$ of the restoration task, as described in our experimental setup. We note that this formulation is closely related to the classic restoration methods based on explicit prior optimizations discussed in our related work. The difference is that our prior is based on a trainable neural network G, and that our fusion factor is also learned to be adaptive, per sample, both to the quality of the observed data and to the fitness of the prior.

We present the **relation to MAP estimation** in connection with the work in [8]. The authors derive a MAP estimate for Additive White Gaussian Noise (AWGN) removal. This MAP estimate is the mode of the posterior probability distribution. The solution is derived where the additive noise ($y_i = x_i + n_i$) follows the normal distribution $\mathcal{N}(0, \sigma_n)$, and an explicit image prior is enforced. More precisely, the solution is derived with the assumption of a Gaussian prior [42] *on the pixel distribution*. With this model, the prior distribution for a pixel value x_i follows $\mathcal{N}(\bar{x}_i, \sigma_{x_i})$ with mean $\bar{x}_i$ and standard deviation σ_{x_i}, and this yields a MAP estimate

$$\hat{x}_i = \arg\max_{x_i} P_{X_i|Y_i}(x_i|y_i) = \frac{y_i}{1+1/S_i} + \frac{\bar{x}_i}{1+S_i}, \quad (6)$$

with S_i being the signal-to-noise ratio defined as

$$S_i \triangleq \frac{\sigma_{x_i}^2}{\sigma_n^2}. \quad (7)$$

Note how S_i is, in fact, dependent on signal quality (through the noise standard deviation σ_n), as well as the confidence in the prior (through the pixel standard deviation σ_{x_i}). Indeed, intuitively the larger σ_{x_i} is, the less reliable the prior term $\bar{x}_i$ is; and the smaller it is, the more reliable the prior term is. In this special case of our general formulation,

$$\phi(y_i) = \frac{1}{1+S_i}, \quad (8)$$

	ϕ explicitly known	ϕ relation to *data fidelity*
Colorization	✗	Luminance and edge related
Inpainting	✓	Binary mask based
Denoising	✗	Noise-level adaptive

TABLE I
THE ϕ MAP VALUES ARE ONLY EXPLICITLY KNOWN FOR INPAINTING, BUT ARE ALWAYS RELATED TO THE DATA-FIDELITY AND PRIOR-CONFIDENCE TERMS DISCUSSED IN OUR MATHEMATICAL FORMULATION. INDEED, IN COLORIZATION THERE EXIST STRONG RELATIONS BETWEEN LUMINANCE AND THE FIDELITY OF THE OBSERVED DATA, IN INPAINTING THIS DIRECTLY MATCHES THE APPLIED MASK, AND IN DENOISING THE NOISE LEVEL DETERMINES THE FIDELITY OF THE OBSERVATION. THE ϕ MAP ALSO, ACROSS ALL TASKS, DEPENDS ON THE CONFIDENCE IN THE PRIOR.

$g(\cdot)$ is the identity mapping, and the prior is the expected value over the distribution of the input $\mathbb{E}_{X_i}[x_i]$. Our formulation in Equation (5) generalizes this solution to non-Gaussian, as well as image-wise prior distributions, while taking into account signal quality and prior confidence.

We also describe the **relation to dictionary-based methods**. Dictionary-based methods [17], [43] generally follow the formulation

$$\hat{x} = \arg\min_{x, d(x,Dv)<\epsilon} \psi_d(x,y) + \beta \cdot \psi_p(v), \quad (9)$$

where D is the dictionary, specifically, a vector set that spans the dictionary space, v holds the coordinates of a point in that space, $d(\cdot,\cdot)$ is a distance function, and ϵ is a small value in $\mathbb{R}_+$. It is typical to use a ψ_p that encourages sparsity, thus to assume that the dictionary captures the main directions of variation in an image. This sparsity of v parallels restrictions on the generative latent space. Effectively, enforcing

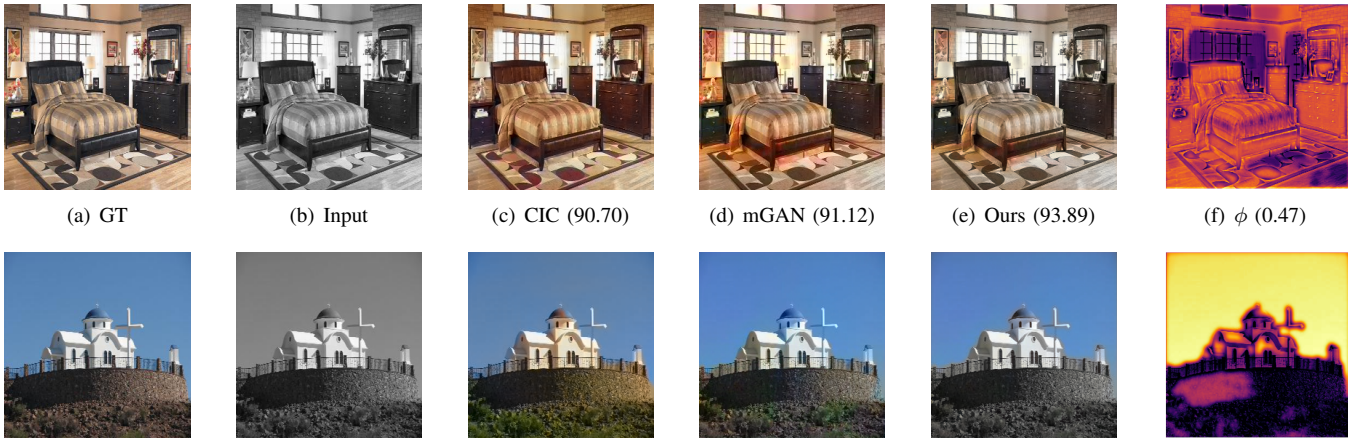
$$d(x, Dv) < \epsilon \quad (10)$$

is a subtle relaxation of the constraint $x \in \text{span}(D)$, which enforces the prior assumption that the image must belong to the dictionary space. This would correspond in our formulation of Equation (5) to $x \in \text{span}(G)$, where in our case the dictionary space is instead the learned space of a generative network. In our formulation, the restriction is enforced only on our decoupled prior element, rather than having to enforce it on x itself and then relaxing it through a tweaking of ϵ.

In summary, our formulation can be viewed as a general framework of MAP estimation and as a generalization of dictionary-based methods. We choose a strict data-fidelity term that preserves a bijective relation to the observed signal, and a fusion factor that takes into account both signal quality and prior confidence. The following two sections discuss in more detail the prior term and the ϕ fusion factor learning.

B. Generative-Space Projection Prior

Theoretically, an inference method can be used to replace the prior term. For instance, a feed-forward network's output can replace $G(z^*)$ in Equation (5). However, such a network trained with supervision takes into account both the data-fidelity and prior terms, albeit without any insight as to how much prior-based hallucination occurs or any control over the different contributions. Therefore, in order to best



(a) GT (b) Input (c) CIC (90.70) (d) mGAN (91.12) (e) Ours (93.89) (f) ϕ (0.47)

(g) GT (h) Input (i) CIC (94.17) (j) mGAN (93.87) (k) Ours (94.75) (l) ϕ (0.52)

Fig. 2. From left to right are the ground-truth image (GT), the grayscale input, the results of colorful image colorization (CIC) [26], mGAN [35], and ours, with the AuC (%), and our channel-averaged ϕ map (with global average between parentheses). The darker colors indicate values of ϕ closer to 0, whereas bright yellow indicates those closer to 1.

Method	Bedroom set AuC [26] ↑	Church set AuC [26] ↑
Colorful colorization [26]	88.55	89.13
Deep image prior [11]	84.33	83.31
Feature map opt. [40]	85.41	86.10
mGAN prior [35]	88.52	89.69
Ours	**89.27**	**90.64**

TABLE II
QUANTITATIVE AUC (%) RESULTS FOR IMAGE COLORIZATION ON THE BEDROOM AND CHURCH TEST SETS. THE HIGHER THE VALUE IS, THE LOWER THE CUMULATIVE COLORIZATION ERROR CURVE IS. WE HIGHLIGHT, WITH BACKGROUND SHADED IN GRAY, THE WIDELY USED TASK-SPECIFIC SUPERVISED METHOD. THE BEST RESULTS ARE SHOWN IN BOLD, AND THE SECOND BEST ARE UNDERLINED.

decouple data fidelity from prior hallucination, we opt for a pre-trained generative network inversion to act as the learned prior. Effectively, this is a better strategy for decreasing the upper bound on a worst-case hallucination contribution. The inversion produces a sampling from the generative space, or a projection on that space as in dictionary-based projections discussed earlier. The latent code z^* for the generative-space projection is obtained as

$$z^* = \arg\min_z \mathcal{L}_G(f(G(z)), y), \tag{11}$$

where $\mathcal{L}_G$ can be a weighted average of ℓ_1, ℓ_2, and perceptual losses, and $f(\cdot)$ is the degradation model of a restoration task. When using a single latent code, very limited information can be encoded, which yields a coarse prior, notably for high-resolution images. To avoid this loss of expressiveness, we use the recent multi-code Generative Adversarial Network (GAN) inversion method that splits the generative network G into two stages, at layer l [35]. The first stage $G_1^{(l)}$ generates multiple feature space representations, each corresponding to one of N latent codes $\{z_n^*\}_{n=1}^N$, where every α is a vector of length equal to the number of feature-space channels. The second stage $G_2^{(l)}$ generates the output image based on a fused feature representation by using adaptive channel weights $\{\alpha_n^*\}_{n=1}^N$. The latent codes and adaptive weights are obtained, as in Equation (11), by an inversion optimization

$$\{z_n^*\}_{n=1}^N, \{\alpha_n^*\}_{n=1}^N = \arg\min_{\{z_n\}_{n=1}^N, \{\alpha_n\}_{n=1}^N} \mathcal{L}(f(x^{inv}), x), \tag{12}$$

where the inverted image x^{inv} is given by

$$G(z; \alpha, \theta_2) \triangleq x^{inv} = G_2^{(l)}\left(\sum_{n=1}^N G_1^{(l)}(z_n) \cdot \alpha_n\right). \tag{13}$$

Our image prior term in Equation (11) is then given by $G(z^*; \alpha^*, \theta_2)$, where θ_2 are the frozen weights of the generative sub-networks G_1 and G_2. We also note that randomly traversing the latent space of a generative network can potentially produce hallucinated images that lie outside the natural image manifold [44]. This is averted by the guided inversion loss that maps the generative output, through the degradation model, to the observed image. The case of the generative projection being outside the natural-image manifold, which can occur when the degradation is extreme, still does not pose an issue in our framework. Indeed, this projection is already treated in our approach as a prior hallucination that might not be faithful to the original image.

C. Guide-Free ϕ Learning

A guided learning of the parameters θ_1 to predict ϕ is possible for a task such as inpainting but impossible for other tasks. This is simply because inpainting is the extreme case where signal quality is binary, specifically zero at the masked areas. For other tasks, a target ϕ cannot be readily obtained. We thus train a network with weights θ_1 to predict ϕ in an end-to-end manner, with ϕ effectively being an intermediate feature space having no explicit learning loss. Our mini-batch training loss $\mathcal{L}(x, y; \theta_1)$ for learning θ_1 is given by (we use ϕ to also denote the network outputting it, for better readability)

$$\begin{aligned}\mathcal{L}(x, y; \theta_1, \theta_2) = &|| \left(1 - \phi(y, \theta_1)\right) \odot g^{-1}(y) \\ &+ \phi(y, \theta_1) \odot G(z^*; \alpha^*, \theta_2) - x||_2^2 + \rho \cdot ||\phi(y, \theta_1)||_1,\end{aligned} \tag{14}$$

Method	PSNR ↑	SSIM ↑	LPIPS ↓
DeepFill v2 [27], [45]	**26.56**	**0.9555**	**0.0191**
Feature map opt. [40]	14.75	0.4563	-
Deep image prior [11]	17.92	0.4327	-
mGAN prior [35]	20.55	0.5823	0.2070
Ours	<u>25.32</u>	<u>0.9240</u>	<u>0.0376</u>

TABLE III
QUANTITATIVE PSNR (dB), SSIM, AND LPIPS RESULTS FOR CENTRAL IMAGE INPAINTING. WE MASK OUT A 64×64 PATCH FROM THE CENTER OF EACH INPUT IMAGE. THE TASK-SPECIFIC STATE-OF-THE-ART METHOD IS HIGHLIGHTED WITH BACKGROUND SHADED IN GRAY. THE BEST RESULTS ARE SHOWN IN BOLD, AND THE SECOND BEST ARE UNDERLINED.

Test	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Bed.	mGAN prior [35]	20.34	0.5902	0.2134
	Ours	**23.22**	**0.8598**	**0.0775**
Chu.	mGAN prior [35]	19.33	0.5359	0.2235
	Ours	**21.94**	**0.8509**	**0.0855**
Conf.	mGAN prior [35]	19.38	0.5641	0.2062
	Ours	**22.20**	**0.8318**	**0.0785**

TABLE IV
QUANTITATIVE PSNR (dB), SSIM, AND LPIPS RESULTS FOR RANDOMIZED-MASKING INPAINTING ON THE BEDROOM, CHURCH (OUTDOOR), AND CONFERENCE TEST SETS. THE RANDOMIZED MASKING INCREASES THE DIFFICULTY OF PREDICTING OUR ϕ MAPS. TO ANALYZE THE EFFECT OF MASK RANDOMIZATION ON THE PERFORMANCE OF OUR ϕ PREDICTION COMPARED TO THE CENTRAL INPAINTING TASK, WE COMPARE THE PRIOR-BASED RESULTS TO OURS.

where ρ is a scalar weight that we discuss next, and the parameters θ_2 of the generative network are the frozen weights of a *pre-trained* generative network. This end-to-end training enables the network predicting ϕ to learn to assess, based on the observation y, the quality of that observed image, as well as the fitness of the prior to this observation.

Fidelity-Bias Balance. For certain image test cases, the quality of the data-fidelity term can be very similar to that of the learned prior, at least over some subsets of pixels. This would induce no change in the loss term for varying values of our fusion factor ϕ, as all would result in similar final outputs. However, for these cases, it is not necessary to *hallucinate* information as the data fidelity is also just as accurate. By hallucinated information, we mean that which is not in direct relation with the observed data, but rather comes from previously learned priors. Therefore, we address these edge cases by adding an auxiliary loss on the ℓ_1 norm of ϕ in Equation (14), which can additionally regularize the feature learning process [46]. This term enforces that the training favors smaller values of ϕ such that the overall contribution of the data fidelity term is maximized when this is not detrimental to the quality of the final output. This fidelity-bias term is weighted by the scalar ρ in Equation (14).

IV. EXPERIMENTS

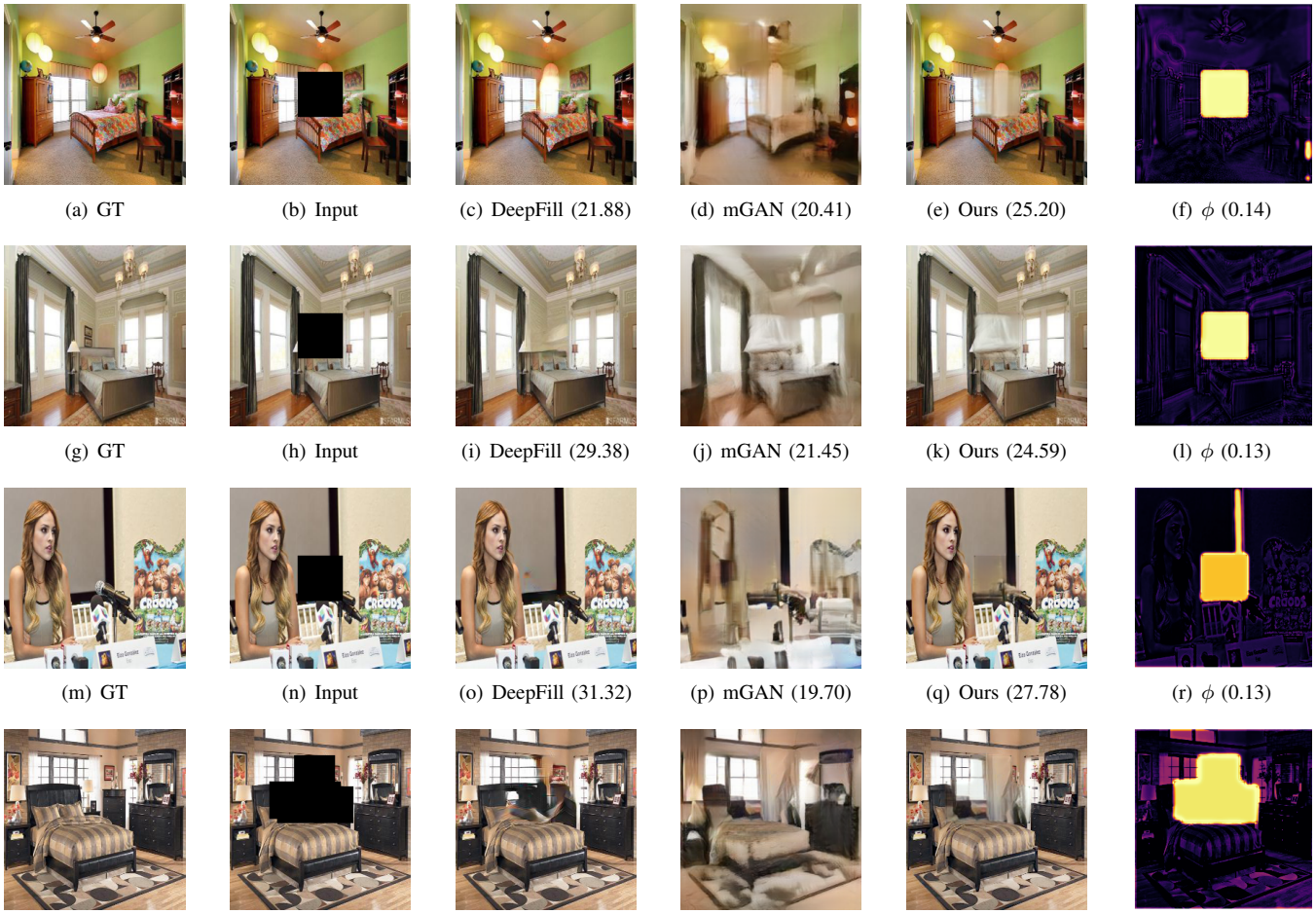
We conduct experiments on image colorization, inpainting, and blind AWGN removal. When comparing to other methods in our experiments, we use the setups recommended by the authors in their original papers and public codes. For Colorization does not induce an explicit solution for ϕ, aside for certain exceptions that we discuss in the next section, such as edges and extreme luminance areas. Inpainting induces an explicitly known solution for ϕ. Whereas, AWGN does not have an explicit solution for ϕ, as the image prior is not explicitly formulated. However, the AWGN experimental setup enables us to analyze the guide-free learning of ϕ, which would intuitively fluctuate mainly with the noise level (direct relation), but also marginally with the uncertainty in the prior (opposite relation), as described in our discussion section. This is summarized in Table I and analyzed in the following sections.

A. Experimental Setup

As described in Section III, we use the multi-code GAN inversion approach for our generative-space projection prior. The pre-trained GAN models, which correspond to each dataset used, are all different versions of the Progressive Growing of GANs (PGGAN) [5] network. They are pre-trained on the Bedroom, Church (Outdoor), and Conference room datasets taken from the LSUN database [47]. The details for each experiment follow the settings given by the authors of [35] and are given in the following sections. We note that any generative network, such as DCGAN [48], LR-GAN [49], CVAE-GAN [50], StyleGAN [6], StyleGAN2 [7], or even any future method allowing projections or sampling from a learned image distribution, can be used for the projection prior of our method. To enable direct comparisons with with mGAN [35], we use the PGGAN in our experiments. We use AuC [26], [51], PSNR, SSIM [52], and the perceptual metric LPIPS [53] in our quantitative evaluations.

For our fusion factor learning, we train the same backbone network with the same settings for all of our experiments. The architecture is inspired by [25] and is a residual learning made up of a sequence of convolutional, batch normalization, and ReLU blocks. We omit further architecture details that can be found in our code. We use a batch size of 8 except for the real denoising experiment where the batch size is 4, a starting learning rate of 0.01, and a fidelity-bias balancing weight $\rho = 1e{-}5$. A very small balancing weight is sufficient ($1e{-}5$), because our objective is to favor fidelity over hallucination only when they are deemed to have the same accuracy. We train for 25 epochs with random shuffling and update the learning rate following a cosine annealing with warm restarts scheduler [54]. The restart period is adaptive to the batch size such that it is always 4 epochs. We also note for reproducibility that training with images that are normalized to $[0, 1]$ and then zero-centered is empirically observed to improve the final results. The same normalization is then performed before inference and inverted once the output is obtained. We train our model with the loss of Equation (14) on a subset of the LSUN validation set that corresponds to each of the large training sets used for pre-training the PGGAN models, and we test on the remaining subset.

mGAN [35]: https://github.com/genforce/mganprior
PGGAN [5]: https://github.com/tkarras/progressive_growing_of_gans



(a) GT (b) Input (c) DeepFill (21.88) (d) mGAN (20.41) (e) Ours (25.20) (f) ϕ (0.14)

(g) GT (h) Input (i) DeepFill (29.38) (j) mGAN (21.45) (k) Ours (24.59) (l) ϕ (0.13)

(m) GT (n) Input (o) DeepFill (31.32) (p) mGAN (19.70) (q) Ours (27.78) (r) ϕ (0.13)

(s) GT (t) Input (u) DeepFill (17.95) (v) mGAN (16.04) (w) Ours (17.98) (x) ϕ (0.22)

Fig. 3. From left to right are the ground-truth image (GT), the masked input, the results of DeepFill v2 [27], [45], mGAN [35], and ours, with the PSNR in dB, and our channel-averaged ϕ map (with global average between parentheses). The first three rows show example images from the standard central-inpainting benchmark, and the last row is an example from our randomized-inpainting experiment.

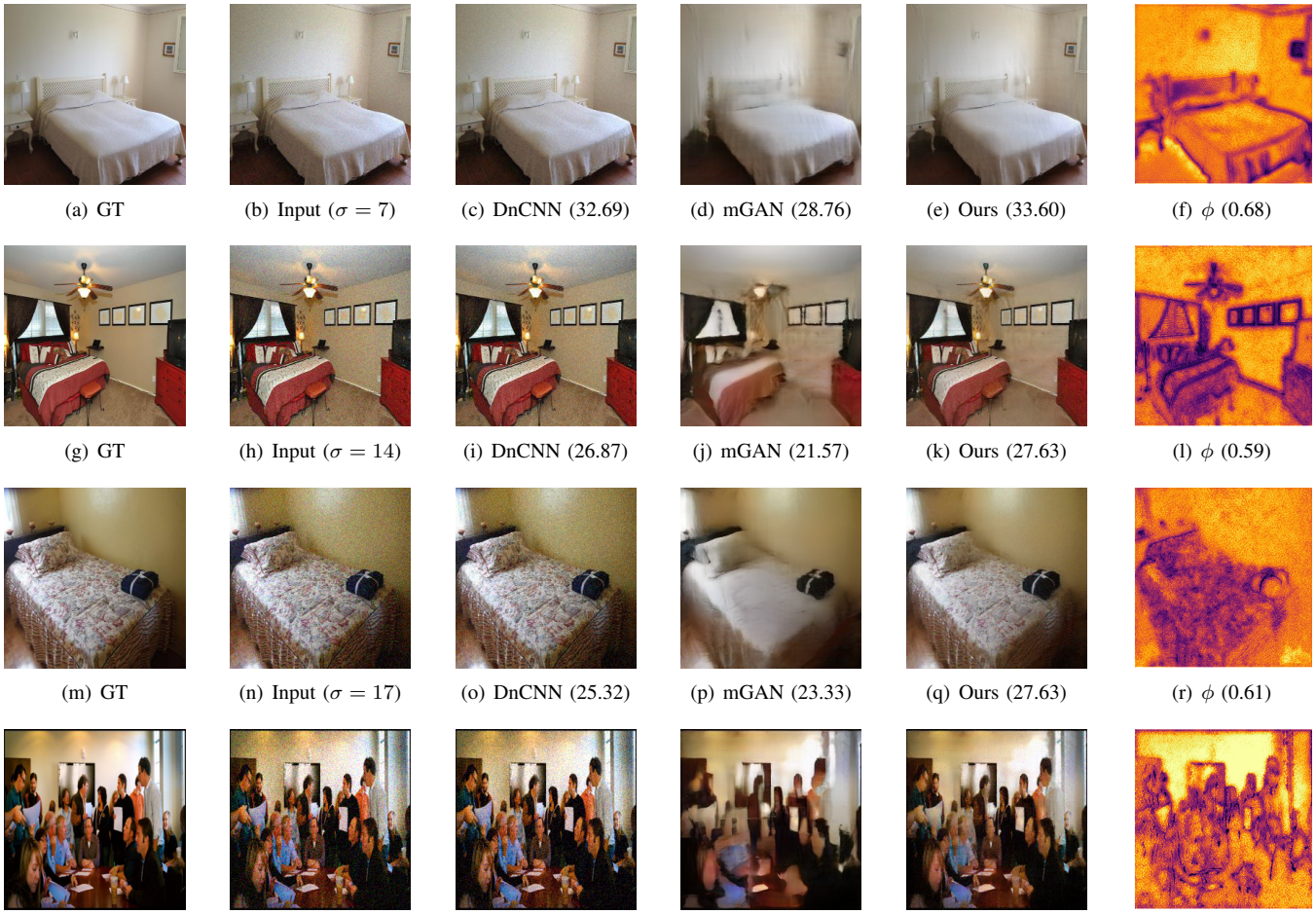
Test	Method	PSNR ↑	SSIM ↑	LPIPS ↓
Bed.	DnCNN† [25]	24.96	0.5804	0.1859
	mGAN prior [35]	22.72	0.6257	0.1978
	Ours	**26.80**	**0.7279**	**0.0998**
Church	DnCNN† [25]	22.40	0.5166	0.2046
	mGAN prior [35]	21.12	0.5643	0.2065
	Ours	**23.38**	**0.5959**	**0.1435**
Conf.	DnCNN† [25]	22.81	0.5310	0.2167
	mGAN prior [35]	21.49	0.5962	0.1968
	Ours	**24.70**	**0.6578**	**0.1192**

TABLE V
PSNR (dB), SSIM, AND LPIPS RESULTS FOR AWGN REMOVAL ON THE BEDROOM, CHURCH, AND CONFERENCE SETS. THE NOISE FOLLOWS A GAUSSIAN DISTRIBUTION WITH STANDARD DEVIATION SAMPLED UNIFORMLY AT RANDOM FROM [5,50] PER IMAGE. †WE RETRAIN AND TEST DNCNN WITH THE SAME DATA AND SETUP AS OURS.

B. Colorization

For the colorization of a grayscale input image, unlike inpainting for example, it is much less predictable what an ideal ϕ map would be. We conduct colorization experiments, where the grayscale input is the luminance channel, and we evaluate the error on the ab color space. The AuC metric [26], [51] computes the area under the cumulative percentage ℓ_2 error distribution curve in the ab space. The percentage is that of pixels lying within an error threshold that is swept over $[0, 150]$ in steps of one. For generative network inversion, we use the sixth layer of the PGGAN for the feature composition, with 20 latent codes, and ℓ_2 and VGG-16 perceptual loss [55], optimized with gradient descent for 1500 iterations, following [35]. Our $g^{-1}(y)$ function duplicates the grayscale channel over each of the color channels. The remaining details follow the experimental setup of Section IV-A.

We present our quantitative image colorization results in Table II, along with those of the deep image prior [11], the feature map optimization [40], the colorful image colorization [26], which is a feed-forward method supervised specifically for colorization, and the mGAN prior [35]. We note the considerable improvement of our method, despite the restriction of enforcing a strict data fidelity. Relative to the mGAN results, we improve by +0.75AuC on the Bedroom set reaching 89.27AuC, and by +0.95AuC on the Church set reaching 90.64AuC. These results even exceed the performance of the task-specific colorful image colorization method [26] on the two test sets, by +0.72AuC and +1.51AuC,



(a) GT (b) Input ($\sigma = 7$) (c) DnCNN (32.69) (d) mGAN (28.76) (e) Ours (33.60) (f) ϕ (0.68)

(g) GT (h) Input ($\sigma = 14$) (i) DnCNN (26.87) (j) mGAN (21.57) (k) Ours (27.63) (l) ϕ (0.59)

(m) GT (n) Input ($\sigma = 17$) (o) DnCNN (25.32) (p) mGAN (23.33) (q) Ours (27.63) (r) ϕ (0.61)

(s) GT (t) Input ($\sigma = 24$) (u) DnCNN (23.00) (v) mGAN (20.06) (w) Ours (23.70) (x) ϕ (0.65)

Fig. 4. From left to right are the ground-truth image (GT), the noisy input with the AWGN standard deviation, the results of DnCNN [25], mGAN [35], and ours, with the PSNR in dB, and our channel-averaged ϕ map (with global average between parentheses).

respectively.

Visual results are shown in Figure 2 for the different image colorization methods. We can observe that ϕ is lower on image edges, which indeed generally constitute information that is not lost by the grayscale degradation. We observe as well that ϕ tends to be low when the luminance is around extreme values, as in such cases the grayscale images are faithful to the original color images. In both of these cases, it is the confidence in the data fidelity that is adapted to the observation. We also note, for instance in the sample of the second row in Figure 2, that ϕ can be very insightful. It indicates that the color of the sky was heavily hallucinated, whereas the bottom half and the church dome use almost no prior hallucination. This is advantageous for downstream tasks as the dome was, in fact, incorrectly hallucinated by the generative-network projection prior. This is similar for the grass part of that same image in the second row. The generative prior incorrectly adds a green color for the grass as can often be expected. This hallucination is however deemed incorrect by our ϕ map, i.e., the added fake color is not correct. Therefore, our final output relies more heavily on its data fidelity term and achieves a more accurate colorization. In parallel, this weighting is illustrated by our ϕ map and can be beneficial for users or downstream tasks such as computer vision ones.

Visual results are shown in Figure 3 for the different methods. For our method, there is little flexibility in terms of the fusion factor ϕ for the inpainting tasks, which are tasks with binary degradation, i.e., the signal is either perfectly available or not at all. The ϕ map effectively predicts the inpainting mask, a mask that is taken as input in the DeepFill method, and the quality of our results is tied mostly to those of the generative-network inversion, as can be visually observed.

C. Inpainting

We present results on the standard central-crop inpainting task in Table III. A 64×64 patch is masked from the test image, and the task is to recover the hidden crop. For generative-network inversion, we use the fourth layer of the PGGAN for the feature composition, with 30 latent codes, and ℓ_2 and VGG-16 perceptual loss [55], optimized with gradient descent for 3000 iterations, following [35]. We use an identity function $g^{-1}(y) = y$ for the data fidelity, and the remaining setup follows that of Section IV-A. The PSNR, SSIM and LPIPS results show a significant improvement of our approach, due to the use of the data fidelity, over the mGAN

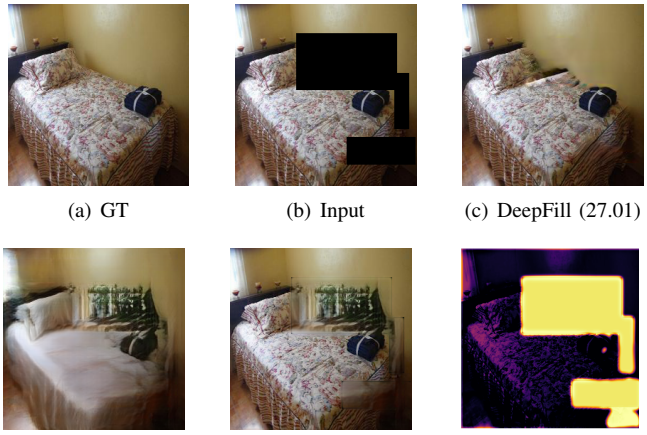
(a) GT (b) Input ($\sigma = 10$) (c) DnCNN (29.70)

(d) mGAN (19.86) (e) Ours (26.32) (f) ϕ (0.52)

(g) GT (h) Input ($\sigma = 6$) (i) DnCNN (34.09)

(j) mGAN (25.89) (k) Ours (31.85) (l) ϕ (0.66)

Fig. 5. Failure cases of AWGN removal. The quality of the generative-network inversion, which remains a very challenging task, is detrimental to our final results. Although our results significantly improve on the prior by exploiting the input observation by using our fusion weight, they still fall short of the task-specific DnCNN denoiser's results.



(a) GT (b) Input (c) DeepFill (27.01)

(d) mGAN (18.91) (e) Ours (20.26) (f) ϕ (0.33)

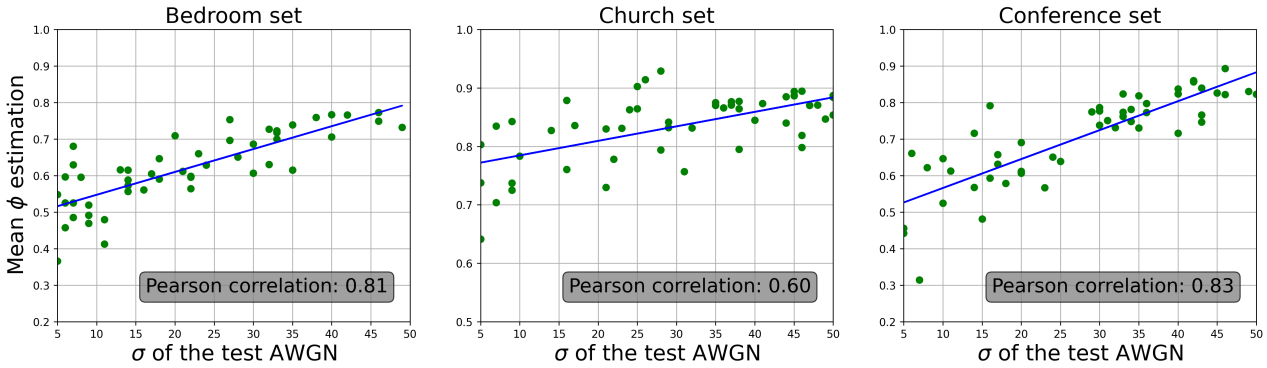
Fig. 6. Failure case in a randomized inpainting experiment. We note a misprediction in the ϕ mask in (f), in the bottom right corner. Our network mistakenly assumed the very dark region was a masked region. Note that in inpainting, data fidelity cannot be used in the masked area, and our results become directly dependent on the quality of the prior that, in this example, is not high.

prior results ($+4.77dB$ in terms of PSNR). The inpainting results are averaged across the Bedroom, Church (Outdoor), and Conference datasets. We compare them with the deep-image prior method [11] and with the recent feature map optimization approach [40] that is a method using GAN priors with test-image specific adaptation. For reference, we compare the results with a task-specific supervised inpainting method, namely, the most recent version [27] of DeepFill [45], trained on the Places2 dataset. DeepFill takes the mask as input and uses gated convolutions to account for invalid pixel locations, and contextual attention [45] to exploit similar patches across the image. The output is refined by using an adversarial GAN loss on every neuron in the feature space [27]. For inpainting, our approach cannot use anything out of the signal over the masked area hence is dependent on the prior hallucination.

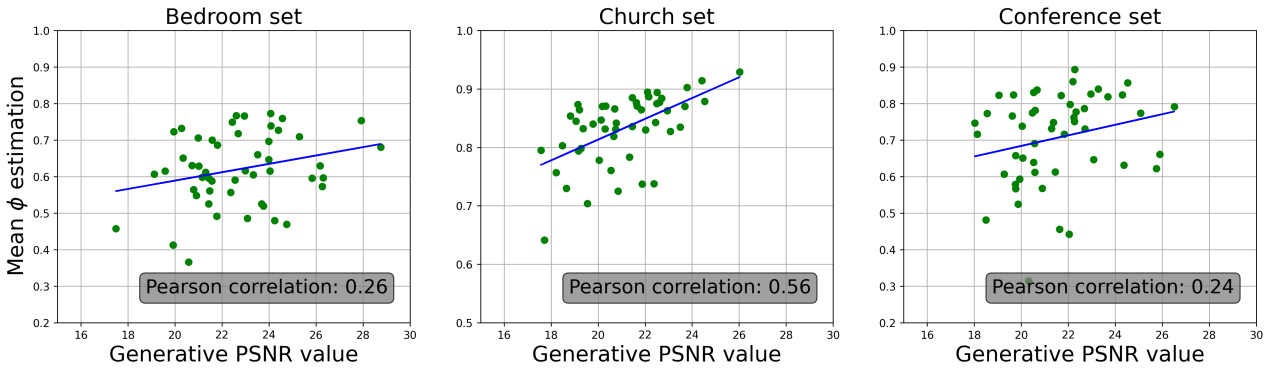
The aforementioned benchmarking setup, however, makes the task simpler for our method in terms of predicting ϕ. Therefore, we design a randomized-masking inpainting setup and present experimental results on it in Table IV. Our randomized-masking algorithm selects uniformly at random a number of patches to be masked, in $[2, 4]$. Then, per patch, a random pixel location for the corner of that patch is selected. The algorithm samples from a normal distribution $\mathcal{N}(64, 32)$, truncated to $[9, +\infty)$, a width and a height for each patch, with re-sampling in case the patch extends beyond the image coordinates. We compare the mGAN prior results with ours in Table IV. We omit the other methods because the purpose of this randomized-masking experiment is specifically to analyze the effect of randomizing the mask on our ϕ prediction, and to analyze how the incurred errors in ϕ affect the performance relative to the prior. We can first note that the mGAN performance decreases, by almost 0.02 SSIM on average. With the randomization of the mask, our performance decreases even more, by almost 0.1 SSIM. However, we still significantly improve over the mGAN results, by +2.88, +2.62, and +2.82 PSNR on the Bedroom, Church, and Conference test sets, respectively, and +0.270, +0.315, and +0.268 in terms of SSIM on those same datasets. This comparison highlights the increased difficulty of our internal ϕ prediction when the mask is randomized relative to the central inpainting task where the mask location is immutable.

D. Denoising

Blind AWGN Removal. We conduct experiments on blind denoising, specifically on AWGN removal. For blind denoising, we follow the standard setup [8], [10], [25] of sampling a noise level, uniformly at random over the range $[5, 50]$. This level is the standard deviation of the AWGN. For generative network inversion, we use the fourth layer of the PGGAN for the feature composition, with 30 latent codes, and ℓ_2 and VGG-16 perceptual loss [55], optimized with gradient descent for 3000 iterations, following [35]. We set $f(\cdot)$ (Equation (11)) to the identity. Our $g^{-1}(y)$ function is also the identity function as the noise is zero-mean. Generally, $g^{-1}(\cdot)$ can be the subtraction of the noise mean value. For the remaining setup details, we follow the experimental setup of Section IV-A.



(a) Correlation between ϕ and the AWGN σ



(b) Correlation between ϕ and the generative PSNR

Fig. 7. (a) Shows across three datasets the relation between the AWGN standard deviation in test images, which is directly related to signal quality, and our corresponding mean estimations for ϕ. (b) Shows the same analysis but with respect to the PSNR of the generative network inversion results, which is directly related to the fitness of the prior. The results show a strong Pearson correlation factor between ϕ and signal quality (a), with the remaining factor of variation explained by the fitness of the prior (b) (e.g., Church set).

We present the AWGN removal results in Table V, along with those of DnCNN [25], which we retrain on the same data as ours. Our approach achieves the best performance, consistently across the different datasets and evaluation metrics. For instance, our smallest improvement in terms of PSNR is +2.26dB relative to the mGAN outputs on the Church dataset, and of +0.98dB relative to DnCNN on that same test set. Similarly we also improve the perceptual metrics SSIM and LPIPS on the different test sets.

Visual results are shown in Figure 4 for the different methods. We observe that DnCNN preserves details well, but at the cost of poorer denoising on low-frequency regions (e.g., walls). The mGAN results are worse than DnCNN, but with our framework the final results become more visually pleasing and accurate. The ϕ map illustrates, per pixel, the contribution of hallucination relative to data fidelity and is again lower around edges, as with colorization. We analyze ϕ in more detail, in the context of AWGN removal, in the next section.

Real Denoising on SIDD. We extend our denoising experiments to the Smartphone Image Denoising Dataset (SIDD) benchmark [56]. We present in Table VI the results of standard classic and learning-based denoising methods, namely, Trainable Nonlinear Reaction Diffusion (TNRD) [57], BM3D [2], WNNM [23], KSVD [58], Expected Patch Log Likelihood (EPLL) [59], DnCNN [25], and CBDNet [60]. We also report the results of HI-GAN [61] and the Variational Denoising Network (VDN) [62]. These methods are designed specifically to address the problem of handling real image data, adapting to the distribution of the noise in the SIDD data. For mGAN [35] and Ours, we present the results with the different PGGAN generators pre-trained on the Church, Conference, and Bedroom datasets and following the same experimental procedure as for blind AWGN removal. We begin by noting that these two approaches use the small version of the SIDD dataset (SIDD-Small) because of the high computational power required to carry out the generative network inversion and the high resolution of SIDD images. Our denoising performance is competitive with standard denoising methods, but it does not reach that of the state-of-the-art task-specific methods HI-GAN and VDN. However, our method additionally provides the ϕ map that determines the degree of contribution of prior-based hallucination relative to data fidelity. We can observe that our final PSNR results improve when our method can rely more on its prior (larger ϕ in Table VI). This indicates that with better priors, either improved methods or richer pre-training datasets, our method can provide increasingly better denoising results.

V. Discussion

	Mean PSNR ↑	Mean ϕ
TNRD [57]	24.73	–
BM3D [2]	25.65	–
WNNM [23]	25.78	–
KSVD [58]	26.88	–
EPLL [59]	27.11	–
DnCNN [25]	32.59	–
CBDNet [60]	33.28	–
†(Church PGGAN) mGAN [35]	32.66	–
(Church PGGAN) Ours	34.01	0.6703
(Conference PGGAN) mGAN [35]	33.26	–
(Conference PGGAN) Ours	34.30	0.6924
(Bedroom PGGAN) mGAN [35]	33.14	–
(Bedroom PGGAN) Ours	34.46	0.7318
HI-GAN [61]	38.88	–
VDN [62]	39.26	–

TABLE VI
QUANTITATIVE IMAGE DENOISING RESULTS ON THE REAL IMAGE SIDD BENCHMARK [56]. THE BEST RESULTS ARE OBTAINED BY HI-GAN [61] AND VDN [62] THAT ARE SPECIFICALLY DESIGNED FOR REAL IMAGE DENOISING AND CAPABLE OF LEARNING THE NOISE DISTRIBUTION FOR EACH DATASET (VDN). †WE NOTE THAT MGAN AND OURS USE CENTRAL CROPS OF THE SMALL VERSION OF THE SIDD DATASET (SIDD-SMALL) DUE TO THE COMPUTATIONAL COMPLEXITY OF THE UNDERLYING GENERATIVE NETWORK INVERSION, AS WE DISCUSS IN SECTION V.

A. Computation Times

Prior extraction. We choose to exploit in our experimental design a generative network inversion for our prior extraction. The mGAN [35] inversion on the PGGAN generator takes on average ≈0.15sec per inverse iteration on one Titan X GPU. That is ≈3.75min for the experiments with 1,500 iterations, and ≈7.5min for the 3,000 iteration experiments.

BIGPrior fusion. Our fusion pipeline with BIGPrior involves efficient operations for applying the bijective $g^{-1}(\cdot)$ function to reverse the degradation, and for applying the ϕ map weight to both the data fidelity term and the prior-based term. BIGPrior also requires the prediction of the ϕ map, which is carried out using a deep network. The architecture we use is similar to DnCNN [25] and requires ≈0.1sec on a Titan X GPU for inference.

We can therefore note that although our ϕ prediction requires similar computation times as standard feed-forward CNN solutions, the computational bottleneck is due to the network inversion for extracting the prior. Improving the computational efficiency of our presented solution requires more efficient network inversions that are outside the scope of this manuscript, or the use of other prior extraction techniques, for instance, using dictionary-based methods.

B. Analysis of ϕ

The AWGN experiments provide the ideal setup for an analysis of ϕ that we carry out in this section. We know that ϕ should be inversely related to the signal quality, the poorer the signal is, the higher the ϕ values are. And ϕ is then also directly related to the confidence in the prior, or the fitness thereof. With AWGN, the quality of the signal is also inversely related to the noise level, in this case, to the standard deviation of the Gaussian noise. We analyze the correlation between the mean ϕ value for a test image, and the standard deviation of the noise in this test image. Results are shown in Figure 7(a), with the Pearson correlation factor, for three datasets. We can clearly observe the positive correlation between the two variables, with a factor of 0.83 and 0.81 for the bedroom and conference sets, respectively. The correlation is lower, at 0.6, for the church dataset. The remaining factor of variability in ϕ is the fitness of the prior, which we analyze in Figure 7(b). The correlation between the average ϕ value and the generative PSNR is the highest for the church set, reaching 0.56, and supporting our claims with regard to ϕ. Indeed, we observe that ϕ is well-correlated with the signal's quality, and when that correlation is somewhat lower it is matched with a higher correlation between ϕ and the fitness of the prior, exactly as expected from the MAP framework's perspective. To summarize, we make two supporting observations from our aforementioned analysis. First, the ϕ estimation, which is learned with no guide in our framework, strongly correlates with the signal quality. Second, a lower correlation with signal quality, as in the church set, is directly justified by a higher correlation between ϕ and the fitness of the prior to the test data. These two observations align exactly with the intuitions derived from the MAP estimation framework, as presented in Section III-A.

The framework we present can be a novel basis for image restoration as it can counter the obstacle of degradation-model overfitting, common in image-restoration tasks. This is because hallucination is the key part prone to overfitting to the chosen model. Our framework can guard against this type of overfitting by relying on decoupled data fidelity and prior hallucination, and by using a pre-trained and frozen generative network, independent of the degradation model, for the hallucination part. Our fusion factor could also be used to increase the robustness and reliability of down-stream computer vision tasks, by making the latter aware of the extent of per-pixel hallucination in the restoration result. For instance, when a computer-vision algorithm deals with degraded images, rather than training the downstream network only on the restoration output, further information regarding the degree of hallucination can be used to increase robustness, notably against adversarial attacks. Our fusion map ϕ conveys such hallucination information, which can also be used for better interpretability of the results by human users.

C. Limitations

As mentioned earlier in this section, one of the limitations of our current implementation of the BIGPrior framework is the computational overhead required by the prior extraction. However, our framework can be used with any future extraction technique, one option being the learning of the prior projection step to replace it by an efficient feed-forward solution.

Another limitation, also related to the current implementation of the prior projection part, is due to the generative network itself. The training of such networks in an adversarial setup is expensive and requires significant amounts of data to achieve good image distribution learning and in turn good image quality. However, future generative solutions can be integrated into our framework in a straight-forward manner, simply by replacing the pre-trained generator.

VI. Conclusion

We have presented a framework for image restoration that enables the use of deep networks for extracting an image prior while decoupling prior-based hallucination and data fidelity. We have shown how our framework is a generalization of a large family of classic restoration methods, notably of Bayesian MAP estimation setups, and of dictionary-based restoration methods.

We have conducted experiments on image colorization, inpainting, and Gaussian and real denoising. Our results, which structurally come with a pixel-wise map indicating data fidelity versus prior hallucination contributions, outperform prior-based methods and are even competitive with state-of-the-art task-specific supervised methods. We have also presented an analysis of our fusion factor ϕ map estimation that supports the claims we make.

References

[1] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 60–65. 1, 2

[2] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080–2095, 2007. 1, 2, 10, 11

[3] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990. 1, 2

[4] D. Krishnan and R. Fergus, "Fast image deconvolution using hyper-Laplacian priors," in *Neural Information Processing Systems (NeurIPS)*, 2009, pp. 1033–1041. 1, 2

[5] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations (ICLR)*, 2018. 1, 6

[6] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410. 1, 6

[7] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8110–8119. 1, 6

[8] M. El Helou and S. Süsstrunk, "Blind universal Bayesian image denoising with Gaussian noise level learning," *IEEE Transactions on Image Processing*, vol. 29, pp. 4885–4897, 2020. 1, 4, 9

[9] V. A. Sindagi, P. Oza, R. Yasarla, and V. M. Patel, "Prior-based domain adaptive object detection for hazy and rainy conditions," in *European Conference on Computer Vision (ECCV)*, 2020. 1

[10] M. El Helou, R. Zhou, and S. Süsstrunk, "Stochastic frequency masking to improve super-resolution and denoising networks," in *European Conference on Computer Vision (ECCV)*, 2020. 2, 9

[11] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 9446–9454. 3, 5, 6, 7, 9

[12] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "BM3D image denoising with shape-adaptive principal component analysis," in *Signal Processing with Adaptive Sparse Structured Representations*, 2009. 2

[13] M. Lebrun, A. Buades, and J.-M. Morel, "A nonlocal Bayesian image denoising algorithm," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1665–1688, 2013. 2

[14] W. Dong, G. Shi, X. Li, Y. Ma, and F. Huang, "Compressive sensing via nonlocal low-rank regularization," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3618–3632, 2014. 2

[15] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Non-local sparse models for image restoration," in *International Conference on Computer Vision (ICCV)*, 2009, pp. 2272–2279. 2

[16] Y. Chen, W. Yu, and T. Pock, "On learning optimized reaction diffusion processes for effective image restoration," in *Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5261–5269. 2

[17] R. Rubinstein, T. Peleg, and M. Elad, "Analysis K-SVD: A dictionary-learning algorithm for the analysis sparse model," *IEEE Transactions on Signal Processing*, vol. 61, no. 3, pp. 661–677, 2012. 2, 4

[18] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995. 2

[19] J. Xu and S. Osher, "Iterative regularization and nonlinear inverse scale space applied to wavelet-based denoising," *IEEE Transactions on Image Processing*, vol. 16, no. 2, pp. 534–544, 2007. 2

[20] P.-S. Laplace, *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*. Springer Science & Business Media, 1998, vol. 13. 2

[21] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: nonlinear phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992. 2

[22] S. Sardy, P. Tseng, and A. Bruce, "Robust wavelet denoising," *IEEE Transactions on Signal Processing*, vol. 49, no. 6, pp. 1146–1152, 2001. 2

[23] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 3, 10, 11

[24] V. Jain and S. Seung, "Natural image denoising with convolutional networks," in *Neural Information Processing Systems (NeurIPS)*, 2009, pp. 769–776. 2

[25] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3142–3155, 2017. 2, 6, 7, 8, 9, 10, 11

[26] R. Zhang, P. Isola, and A. Efros, "Colorful image colorization," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 649–666. 2, 5, 6, 7

[27] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 4471–4480. 2, 6, 7, 9

[28] J. Pan, J. Dong, Y. Liu, J. Zhang, J. Ren, J. Tang, Y.-W. Tai, and M.-H. Yang, "Physics-based generative adversarial models for image restoration and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2449–2462, 2020. 2

[29] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," in *International Conference on Learning Representations (ICLR)*, 2019. 2

[30] J. Dong, S. Roth, and B. Schiele, "Learning spatially-variant MAP models for non-blind image deblurring," in *Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4886–4895. 2

[31] M. Tatarchenko, S. R. Richter, R. Ranftl, Z. Li, V. Koltun, and T. Brox, "What do single-view 3D reconstruction networks learn?" in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3405–3414. 2

[32] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," in *International Conference on Learning Representations (ICLR)*, 2017. 3

[33] M. Albright and S. McCloskey, "Source generator attribution via inversion." in *Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 3

[34] J. Chen, J. Chen, H. Chao, and M. Yang, "Image blind denoising with generative adversarial network based noise modeling," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3155–3164. 3

[35] J. Gu, Y. Shen, and B. Zhou, "Image processing using multi-code GAN prior," in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3012–3021. 3, 5, 6, 7, 8, 9, 10, 11

[36] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, and P. Luo, "Exploiting deep generative prior for versatile image restoration and manipulation," in *European Conference on Computer Vision (ECCV)*, 2020. 3

[37] T. S. Cho, N. Joshi, C. L. Zitnick, S. B. Kang, R. Szeliski, and W. T. Freeman, "A content-aware image prior," in *Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 169–176. 3

[38] S. Choi, J. Isidoro, P. Getreuer, and P. Milanfar, "Fast, trainable, multiscale denoising," in *International Conference on Image Processing (ICIP)*, 2018, pp. 963–967. 3

[39] S. Lee, D. Cho, J. Kim, and T. H. Kim, "Self-supervised fast adaptation for denoising via meta-learning," *arXiv preprint arXiv:2001.02899*, 2020. 3

[40] D. Bau, H. Strobelt, W. Peebles, B. Zhou, J.-Y. Zhu, A. Torralba *et al.*, "Semantic photo manipulation with a generative image prior," *arXiv preprint arXiv:2005.07727*, 2020. 3, 5, 6, 7, 9

[41] Y. Sun, X. Wang, Z. Liu, J. Miller, A. A. Efros, and M. Hardt, "Test-time training with self-supervision for generalization under distribution shifts," in *International Conference on Machine Learning (ICML)*, 2020. 3

[42] S. Romdhani and T. Vetter, "Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior," in

Computer Vision and Pattern Recognition (CVPR), 2005, pp. 986–993. 4

[43] R. Giryes and M. Elad, "Sparsity-based Poisson denoising with dictionary learning," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5057–5069, 2014. 4

[44] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "PULSE: Self-supervised photo upsampling via latent space exploration of generative models," in *Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2437–2445. 5

[45] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5505–5514. 6, 7, 9

[46] M. El Helou, F. Dümbgen, and S. Süsstrunk, "AL2: Progressive activation loss for learning general representations in classification neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4007–4011. 6

[47] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015. 6

[48] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015. 6

[49] J. Yang, A. Kannan, D. Batra, and D. Parikh, "LR-GAN: Layered recursive generative adversarial networks for image generation," in *International Conference on Learning Representations (ICLR)*, 2017. 6

[50] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "CVAE-GAN: fine-grained image generation through asymmetric training," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 2745–2754. 6

[51] A. Deshpande, J. Rock, and D. Forsyth, "Learning large-scale automatic image colorization," in *International Conference on Computer Vision (ICCV)*, 2015, pp. 567–575. 6, 7

[52] Z. 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*, vol. 13, no. 4, pp. 600–612, 2004. 6

[53] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 586–595. 6

[54] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations (ICLR)*, 2017. 6

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015. 7, 8, 9

[56] A. Abdelhamed, S. Lin, and M. S. Brown, "A high-quality denoising dataset for smartphone cameras," in *Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1692–1700. 10, 11

[57] Y. Chen and T. Pock, "Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1256–1272, 2016. 10, 11

[58] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006. 10, 11

[59] D. Zoran and Y. Weiss, "From learning models of natural image patches to whole image restoration," in *International Conference on Computer Vision (ICCV)*, 2011. 10, 11

[60] S. Guo, Z. Yan, K. Zhang, W. Zuo, and L. Zhang, "Toward convolutional blind denoising of real photographs," in *Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1712–1722. 10, 11

[61] D. M. Vo, D. M. Nguyen, T. P. Le, and S.-W. Lee, "HI-GAN: A hierarchical generative adversarial network for blind denoising of real photographs," *Information Sciences*, vol. 570, pp. 225–240, 2021. 10, 11

[62] Z. Yue, H. Yong, Q. Zhao, D. Meng, and L. Zhang, "Variational denoising network: Toward blind noise modeling and removal," *Neural Information Processing Systems (NeurIPS)*, vol. 32, pp. 1690–1701, 2019. 10, 11