Natural Language Offers Disambiguation in Extreme Image Recovery

Abhay Kumar, Kalyani Unnikrishnan, Kriti Goyal, Varun Sundar *
University of Wisconsin Madison
{abhay.kumar,kunnikrishna,kgoyal6,vsundar4}@wisc.edu

Abstract

Existing deep learning approaches cannot effectively recover high-fidelity reconstructions when faced with scenarios having extreme image degradation. An alternative paradigm is to traverse a parameterized subset of the image manifold and pick a single reconstruction consistent with the observed measurement. Such a framework is inherently ambiguous and requires additional information to produce a unique output. In this work, we propose to use natural language as an added modality for disambiguation. Our method requires only a single image along with its associated caption and can provide reconstructions across four example tasks—64× super-resolution, large kernel blurring (Gaussian and motion blur), and image inpainting. We demonstrate near-exact reconstruction across these tasks when a sufficiently detailed caption is provided. Code and model artifacts will be made publicly available at https://github.com/varun19299/clip-prior.

1. Extreme Image Recovery

Image recovery is a core aspect of most computational imaging problems with research spanning several decades [8, 10, 22, 23, 28, 51]. Across problems, given a noisy, low-resolution or corrupted measurement, we wish to recover an output closely resembling the measurement when degraded itself. Typically, the degradation or forward operator is known a priori and can be computed in closed form. The challenge in image recovery arises from its ill-conditioned nature that warrants regularization. In the past, such regularization techniques were hand crafted based on heuristics such as sparsity in gradient domain [42], bounds on total variation [13, 16] or frequency domain constraints [57].

Deep neural networks which replace these heuristics with data-driven priors or regularizers have witnessed rapid adoption in the image recovery community. Among these, a class of techniques using convolutional neural networks (CNNs) operate in a feed-forward fashion to produce clean outputs from corrupted inputs. Such networks are trained in either a fully supervised or semi-supervised manner requiring a large dataset of groundtruth and measurement images. While these deep learning techniques have enjoyed greater success than their traditional counterparts, high fidelity recovery remains elusive—typically an artifact of the L2 norm or mean squared error (MSE) objective employed. Auxiliary objectives such as perceptual loss [30] and adversarial loss [25, 41] can potentially improve the sharpness of outputs, but at the cost of producing hallucinating artifacts.

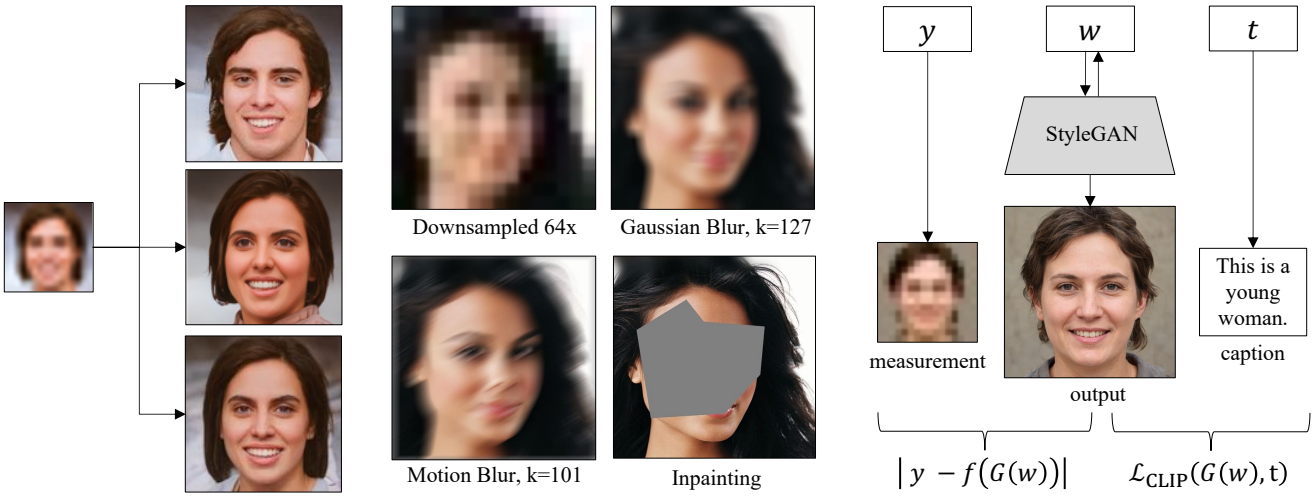
Extreme degradation scenarios. The aforementioned shortcomings of existing deep learning approaches in image recovery is exacerbated in certain "extreme" scenarios—when the forward operators lead to highly degraded measurements thereby resulting in poorly conditioned problems. In extreme scenarios, multiple clean images correspond to the provided measurement—even a perfectly trained feed-forward network would output the average of such images, thereby precluding photorealistic recovery. Similar to Menon et al. [47], we argue that the inductive bias of trained deep architectures alone is not sufficient to regularize these problems.

Traversing the image manifold. An effective alternative is to instead directly traverse the image manifold and choose a photorealistic image which is close to the measurement in forward space. While simple to conceptualize, this paradigm requires solving the much harder problem of parametrizing the manifold of natural images. Fortunately, deep generative models, whose expressivity and output fidelity have increased dramatically in the last few years are good enabling candidates—atleast when restricted to a object-specific sets of natural images, such as facial images, images of buildings and monuments, or images of cars, etc. Specifically, in this work, we use StyleGAN-v2 [34], a state-of-the-art generative adversarial model (GAN) that provides fine-grained control on the visual attributes of its generated images. For brevity, we hereafter refer to the model as StyleGAN.

Persisting ambiguity in reconstruction. Explicitly constraining solutions to the natural image manifold guarantees photorealism, but not necessarily uniqueness. Importantly, the degradation tasks we consider induce multiple solutions on the image manifold. *Can we provide disambiguation?*

Natural language as a disambiguation tool. Our key ob-

*Project report as a part of CS762, Fall 2021.



(a) Inherent ambiguity (b) Extreme scenarios (c) Proposed framework

Figure 1. **Natural language provides disambiguation in extremely ill-posed image recovery.** *(left)* Ambiguity, even when constrained to the set of natural images, is inherent in severe degradation tasks. *(center)* We consider several extreme tasks in this work such as $64\times$ downsampling, large kernel blurring (Gaussian blur with kernel size 127 and motion blur with kernel size 101 with respect to a megapixel groundtruth), and large mask inpainting. *(right)* Our proposed framework optimizes StyleGAN's latent vector over its $\mathcal{W}$-space to satisfy forward operator consistency along with an auxiliary loss term based on the CLIP embedding of the provided caption.

servation is that textual inputs often co-occur today with images in a myriad different ways—prominently in captions and hashtags on social media, headings and bulletins on news media, and website metadata. We posit that providing text captions as additional information can mitigate the disambiguation problem in extreme image recovery. To imbibe this multimodal information, we turn to CLIP [56]—a large-scale contrastive model trained on over 400 million image-caption pairs.

Contributions. We propose an optimization framework that facilitates image recovery across four different tasks (Sec. 3.2) by explicitly parameterizing solutions using StyleGAN (Sec. 4), given a measurement and an accompanying caption. Our method requires only a single measurement and knowledge of the corresponding degradation operator. We demonstrate the role played by text in manipulating the final output conditioned on the provided description (Sec. 5.2). Further, we show that near-exact recovery is possible even in these extreme degradation scenarios—provided a sufficiently descriptive caption is present (Sec. 5.3). Finally, while a rigorous quantitative evaluation remains challenging owing to the mismatch in compatible datasets—domains that have accessible StyleGAN and rich captions—we provide preliminary comparisons based on alignment-free metrics such as LPIPS [76] and face ID similarity [19] for two degradation tasks on a small (~30 images) collection of facial images.

Limiations and future outlook. The proposed method inherits many limitations endemic to StyleGAN-v2 and CLIP encoders. We chose image sets in this work based on the availability of StyleGAN checkpoints. Additionally, CLIP embeddings seem to contain more information when using longer sentences—even if semantically redundant. Further, our understanding of visual and language concepts captured by CLIP at this point is still limited; while we have demonstrated text-conditional outputs and their utility in image recovery, precise control over the severity and nature of these attributes remains elusive.

Fortunately, our framework is general and can be applied to other generative models [20, 21] and multi-modal encoders [29, 50, 68]. With the flurry of recent work in both these domains, there is potential for our work to be used in less restrictive settings.

2. Related Work

Image restoration encompasses tasks like image denoising, inpainting, deblurring and super resolution [4, 7, 45, 49]. In recent years, convolutional neural networks (CNNs) have rapidly become the method of choice across a variety of image restoration problems [46, 60, 64, 74]. Typically, a low-resolution or corrupted input is fed through multiple layers of the CNN and the final output is treated as the reconstruction—which can be used to train the network either using an available set of groundtruth images or in its absence, forward operator consistency. Of these techniques, residual dense connections [77] exploiting hierarchical features have garnered interest with subsequent works in specific restoration tasks [17, 36, 55, 73, 77].

Loss functions in image recovery. Early works in the deep learning era considered L2 norm or MSE loss between the

groundtruth and generated output as their objective. While correlating strongly with a high peak signal-to-noise ratio (PSNR), visually pleasing outputs are not guaranteed [38, 41]. Ledig et al. [41] introduce a combination of perceptual and adversarial losses to produce sharper images, but at the cost of PSNR and hallucinating artifacts [12]. In contrast, our approach achieves high-fidelity outputs by explicitly constraining solutions to the image manifold.

Latent space optimization in GANs. The workhorse of our proposed recovery framework is latent optimization—which has the general goal of finding latent inputs that best represent a given image when fed through the generative model. A related problem is GAN inversion [31, 59, 71], where emphasis is also given to the editability of such representations. Popular approaches include direct optimization [1, 3, 18], training an explicit encoder [26, 54, 58] or a hybrid combination of the two [9, 79]. Tov et al. [66] analyze the distortion-editability and distortion-perception tradeoff within the StyleGAN space by adopting an encoder-decoder approach to facilitate user-driven control.

Closest to our work is Menon et al. [48] which solves single-image super resolution by searching the latent space for a high-resolution version of a given low-resolution image using direct optimization. In comparison, we consider a larger set of degradation tasks and incorporate multi-modal information for near-exact recovery.

3. Background

In this section, we describe image recovery as an optimization problem that subsumes a class of restoration problems where the degradation or forward operator is known beforehand. We further enumerate the analytic forms of four forward operators considered in this work. Subsequently, we use this formulation to intuitively elucidate the drawbacks of existing feed-forward deep learning models trained in either a fully-supervised or semi-supervised manner. This serves as a motivation for the paradigm shift of explicitly constraining solutions to lie on (a portion of) the natural image manifold.

3.1. Optimization framework

Given a noisy, low-resolution or corrupted measurement y, the objective of image recovery is to find image x that closely matches y when degraded itself. The degradation or forward operator f may or may not be known, termed as non-blind or blind image recovery accordingly. For the purpose of this work, we assume that f is (a) known, (b) efficient to compute, and (c) differentiable. Such assumptions are satisfied in many common degradation tasks, some of which we detail in Sec. 3.2. We can now mathematically describe the optimization problem representing image recovery as

$$x^* = \text{argmin}_{x \in \mathcal{M}} |y - f(x)| \tag{1}$$

Since f is a many-to-one mapping, inversion is a non-trivial task—and requires regularization, such as constraining to the image manifold $\mathcal{M}$, which is strictly smaller than the high-dimensional space containing these images.

3.2. Examples of Forward Operators

The degradation operators we consider—viz., downsampling, gaussian blur, motion blur and inpainting—can be easily computed and are differentiable. We choose their parameters (such as downsampling factor or kernel size) to yield highly degraded measurements.

Downsampling. We use the popular bicubic downsampling which can be interpreted as a convolution followed by naive subsampling

$$f_{\text{down, t}}(x) = (x * k_{\text{bicubic}}) \downarrow_t \tag{2}$$

where x is the 2D image, t is the downsampling factor (we use $t = 64$), and k_{bicubic} is the multidimensional separable bicubic kernel [35], comprising of piece-wise cubic functions. When dealing with RGB images, this process is performed channel-wise.

Gaussian blur. This forward operator is convolutional as well, and for a zero-centered Gaussian kernel with width σ ($k_{\text{gaussian, }\sigma}$, with $\sigma = 127$ considered in this work), the expression for a single-channel given as

$$f_{\text{gaussian, }\sigma}(x) = (x * k_{\text{gaussian, }\sigma}) \tag{3}$$

Motion blur. Here, the forward operator closely follows Eq. (3), with the sole exception of a non-separable kernel k_{motion} being the convolutional operand. Since we employ a large kernel width of 101, the convolution is performed as a Hadamard product in Fourier space for efficiency.

Inpainting. Given a binary mask M and Hadamard operator $\odot$, the inpainting forward operator is given by

$$f_{\text{inpainting}}(x) = x \odot M \tag{4}$$

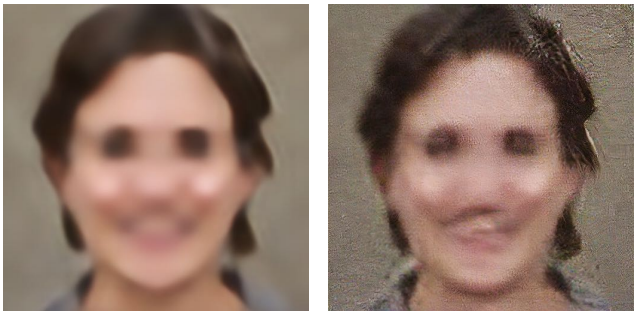
3.3. Drawbacks in Existing Approaches

In the fully-supervised setting, deep models utilize a paired dataset $\mathcal{D} := \{y, x_{\text{gt}}\}$ comprising of measurements y and clean groundtruth x_{gt} to learn a mapping g_θ (weights θ) from measurements to clean images. Effectively, this captures a prior embodied by the dataset and parameterized by the network's weights. With the conventional $L2$ norm objective, the learnt mapping minimizes

$$\mathbb{E}_{(y, x_{\text{gt}}) \sim \mathcal{D}} [\|x - g_\theta(y)\|_2^2] \tag{5}$$

In the limit of infinite data, the optimal mapping can be shown [47] to be

$$g_\theta^*(y) = \mathbb{E}_{x \in \mathcal{M}} [x \mid f(x) = y] \tag{6}$$



(a) MSE (b) MSE +VGG + GAN

Figure 2. **Existing deep learning approaches are not suited extreme scenarios,** such as $64x$ super-resolution shown here using SRResNet and SRGAN [41]. Part of this limiation arises from the mismatch in commonly considered upsampling factors (2-4×). Even a network trained perfectly on a dataset of clean and corrupted image pairs precludes high-quality outputs (Sec. 3.3).

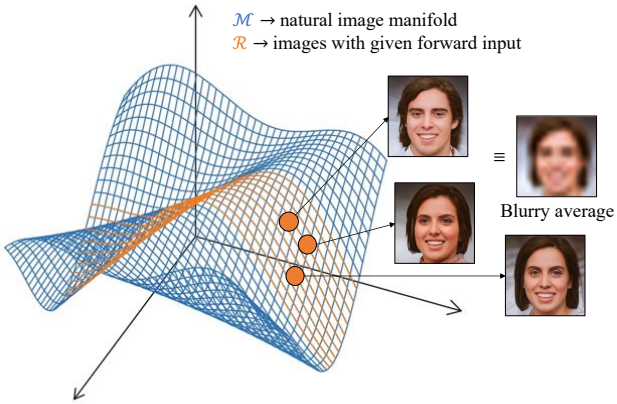
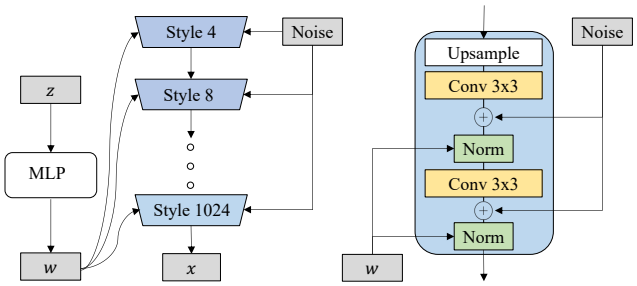


Figure 3. **Manifold perspective for image restoration.** We wish to find a single point on the image manifold most consistent with the given measurement, instead of averaging over multiple points.

Thus, even a fully trained deep network will output the average of multiple images, leading to blurry results (Fig. 1(a)). While we shown this only for the MSE objective, there is no guarantee that auxiliary losses can mitigate this issue. Indeed, in Fig. 2, we show that even with the addition of perceptual and adversarial losses, SRGAN [41] results in blurry reconstructions at $64\times$ super-resolution.

4. Generative Models can represent Image Manifolds

Instead of relying on the priors captured by traditional methods or trained deep models, *can we directly traverse the image manifold?* Doing this would allow us to pick a single, high-fidelity image that closely satisfies the forward operator constraints (Fig. 3); but requires some method to explicitly parameterize the image manifold. Up until recently [47], such a paradigm was perceived as significantly more challenging in realization. However, with the recent explosion of high-resolution and high-quality deep generative models—such as the recent line of work in variational autoencoders (VAEs) [67], generative adversarial networks (GANs) [15, 32, 34], and diffusion models [20, 21]—it is possible to parameterize at least a region of the natural image manifold, such as images of a particular type. Particularly, we consider StyleGAN in this work, a state-of-the-art GAN that offers fine-grained control over the visual (or style) attributes of its generated images.



(a) StyleGAN-v2 overview (b) Detailed architecture

Figure 4. **StyleGAN-v2 architecture.** *(a)* The generative model consists of a latent vector z fed through a multi-layer perceptron (MLP) to produce vector w—which is then replicated and fed across all Style blocks corresponding to increasing tensor resolution. **(b)** Each Style block consists of upsampling, 3×3 convolution and normalization (or demodulation) [61] blocks, with the intermittent incorporation of trainable noise and w tensors.

4.1. StyleGAN Architecture

For completeness, we summarize the StyleGAN [34] architecture used in this work (illustrated in Fig. 4). The generator consists of two main components: (a) an 8 layer multi-layer perceptron (MLP) that maps a random Gaussian vector $z \in \mathbb{R}^{512} \sim \mathcal{N}(0, \sigma\mathbb{I}_{512})$ to a *style code* $w \in \mathbb{R}^{512}$, (b) several *style blocks* that take as input replicated copies of vector w and optionally noise-tensors to generate output images at varying resolutions—from 4×4 to 1024×1024.

Trainable layers. It has been shown that latent space optimization benefits from considering multiple independent style vectors instead of a single vector w replicated multiple times [2, 70]. Likewise, we also find it beneficial to include a subset of the input noise tensors as trainable variables.

4.2. Disambiguating Image Recovery

Owing to the severe degradation of measurements, even when the image manifold is completely captured by a generative model, exact recovery may not be possible. Indeed when PULSE [48] is run across multiple random seeds, non-identical outputs ensue. We therefore seek to imbibe an additional modality to better condition these problems. Natural language—which co-occurs today with images in a myriad of ways across social media, news media, and the internet in general—is a strong candidate.

To incorporate text inputs, we use the Contrastive Lan-

guage Image Pre-training model [56] or CLIP, which can map image tensors and text tokens to the same embedding space—using a similarity metric on these embeddings can then serve as a proxy for the "closeness" of an image to the supplied caption. We note that prior work StyleCLIP [53] also incorporates a CLIP similarity term in its latent optimization objective—but in the context of image manipulation, where additional information of the source input, such as its inverted latent representation is assumed to be known.

4.3. Optimization Objective

Our optimization objective comprises of three terms and reads analytically as

$$\mathcal{L} = \lambda_{\text{forward}}\mathcal{L}_{\text{forward}} + \lambda_{\text{CLIP}}\mathcal{L}_{\text{CLIP}} + \lambda_{\text{geocross}}\mathcal{L}_{\text{geocross}} \quad (7)$$

Forward loss. Computed as the L1-norm between the input measurement y and high-fidelity output x when degraded itself

$$\mathcal{L}_{\text{forward}}(y, x) = |y - f(x)| \quad (8)$$

CLIP loss. Computed using the cosine similarity of the CLIP embeddings of output x, denoted $\mathbf{v}_x$, and supplied caption t, denoted $\mathbf{v}_t$

$$\mathcal{L}_{\text{CLIP}}(t, x) = 1 - \frac{\mathbf{v}_t \cdot \mathbf{v}_x}{\|\mathbf{v}_t\| \, \|\mathbf{v}_x\|} \quad (9)$$

Geocross loss. Style space $\mathcal{W}$, the set of style vectors w that lead to photorealistic StyleGAN outputs, is a strict subset of $\mathbb{R}^{512}$. Following Menon et al. [48], we incorporate the geodesic distance ($\mathcal{D}_{\text{geodesic}}$) on the hypersphere $\mathbf{S}^d$

$$\mathcal{L}_{\text{geocross}}(\{w_i\}_{i=1}^{k}) = \sum_{i,j} \mathcal{D}_{\text{geodesic}}\left(\frac{w_i}{\|w_i\|}, \frac{w_j}{\|w_j\|}\right) \quad (10)$$

where $\{w_i\}$ is the set of k style vectors treated as optimization variables. Coupled with projected gradient descent, this encourages style vectors to remain in regions of high prior probability—thereby w.h.p producing high-fidelity images.

Based on visual inspection of preliminary results, we set the value of λ_{forward} and λ_{CLIP} as 10 and 0.1 respectively. For $\lambda_{\text{geocross}}$, we adopt the hyperparameter value from PULSE [47]—and find it to transfer reasonably well to our setting. Owing to the difficulty finding of a non-reference metric that reliable captures perceptual performance, we do not perform a rigorous hyperparameter search using grid search or pruning frameworks (Optuna [6]).

5. Experimental Results

5.1. Implementation Details

Datasets Considered. Majority of our experiments are conducted on the CelebA-HQ dataset [33], a megapixel reconstruction of the original CelebA dataset [43] comprising of 30,000 images. The dataset also contains 5-10 text captions describing each image. For non-face datasets, we consider LSUN Church [72] and Stanford cars [37] which have images at a resolution of 256×256 and 512×512 respectively. For these datasets, we are unable to find annotated captions and instead supply them manually.

To ensure output images bearing semblance to the groundtruth exist in StyleGAN, we perform GAN inversion [59, 71] using the recently proposed e4e architecture [66] to choose images. However, this may equivalently be done by reporting failure if the forward loss at the end of the optimization procedure exceeds a certain threshold.

Training Description. We employ the Adam optimizer [44] using a learning rate of 0.3, decayed linearly to 0—modified to perform projected gradient descent onto a 512 dimensional hypersphere. Optimization requires ~1000 steps for a typical image and a runtime of around 5 minutes on a NVIDIA 2080Ti GPU. All neural networks are implemented using the Pytorch framework [52].

5.2. Text Conditional Outputs

To elucidate the expressivity of text, we first consider conditional outputs—where we try producing diverse outputs that are still consistent by the forward operator. Fig. 5 shows that the CLIP loss term finds outputs faithful to the caption while being consistent with the provided measurement. This experiment demonstrates image manipulation capabilities akin to StyleCLIP [53] while only using the forward operator constraint, and not a known latent vector.

5.3. Towards Exact Recovery

Motivated by the manipulation capacity of text inputs, we explore if a sufficiently detailed caption yields exact recovery. Fig. 6 contains encouraging results in this direction across all degradation tasks. Notably, this differentiates our work from PULSE [47], where exact recovery of groundtruth remains elusive.

Quantitative Evaluation. Our method's ability to nearly reconstruct the original image paves the way for qualitative evaluation—that considers not just the fidelity of outputs, but also their deviation from groundtruth.

Evaluation metrics. Conventional metrics such as PSNR and SSIM [78] are heavily biased towards per-pixel precise recovery and may not reflect perceptual quality. High-fidelity outputs can be penalized as heavily as blurry or artifact-ridden outputs—clearly, the latter is less preferable. For instance, most of the outputs in Fig. 6 have a PSNR score of just 25 dB, and are not reflective of their utility.

Instead, we opt to use the recently proposed LPIPS [76] that leverages the effectiveness of CNNs such as Alexnet [39] and VGG [63] for judging perceptual quality. Despite its shortcomings to adversarially perturbations [40], LPIPS

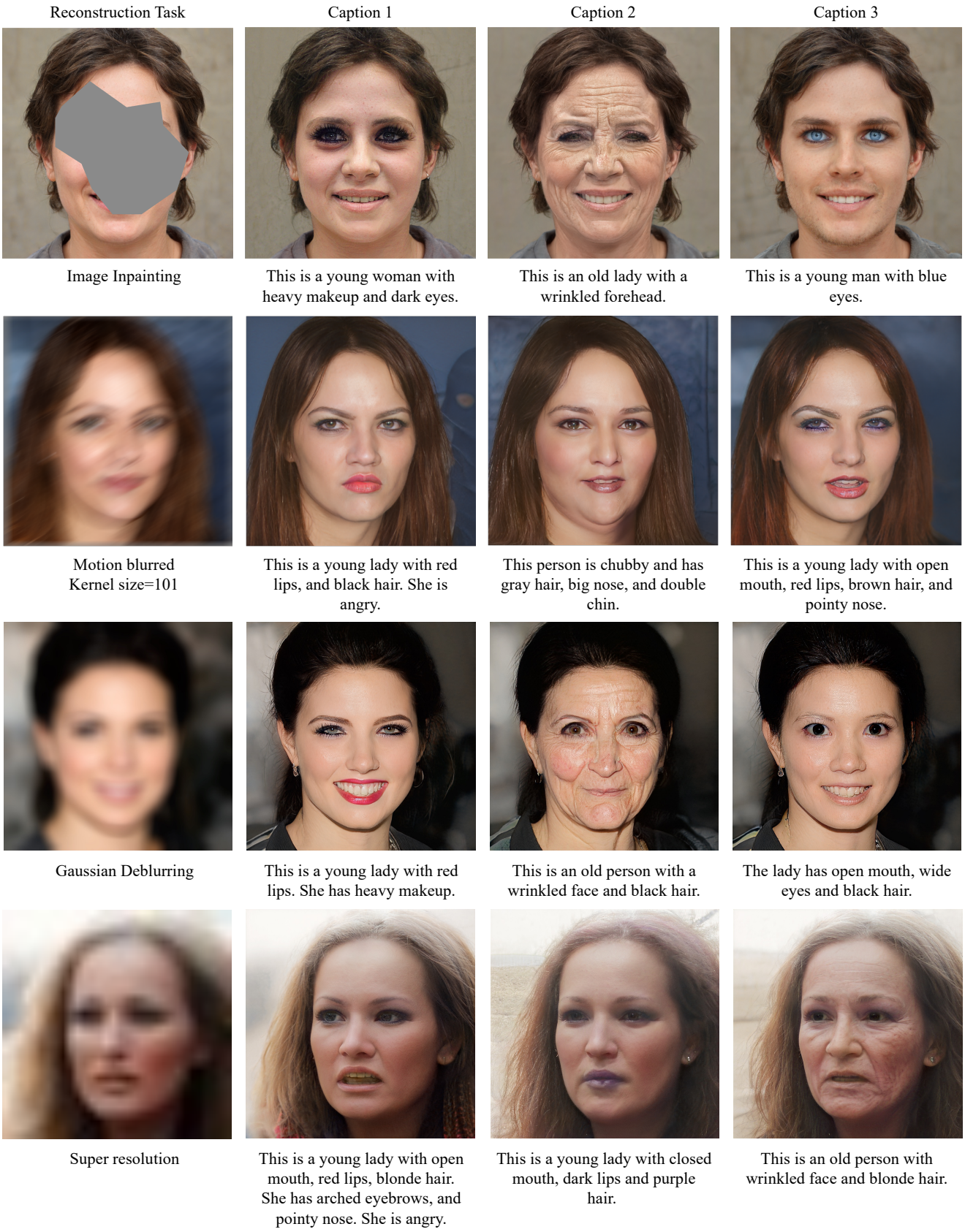
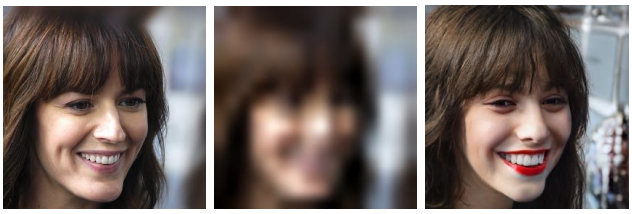


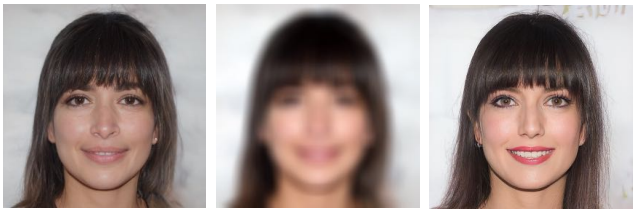
Figure 5. **Text conditional outputs,** shown across the four degradation tasks we consider. The outputs demonstrate image editing capabilities with access only to degraded measurements. Such manipulations benefit from detailed captions providing rich semantic input.

(a) Inpainting



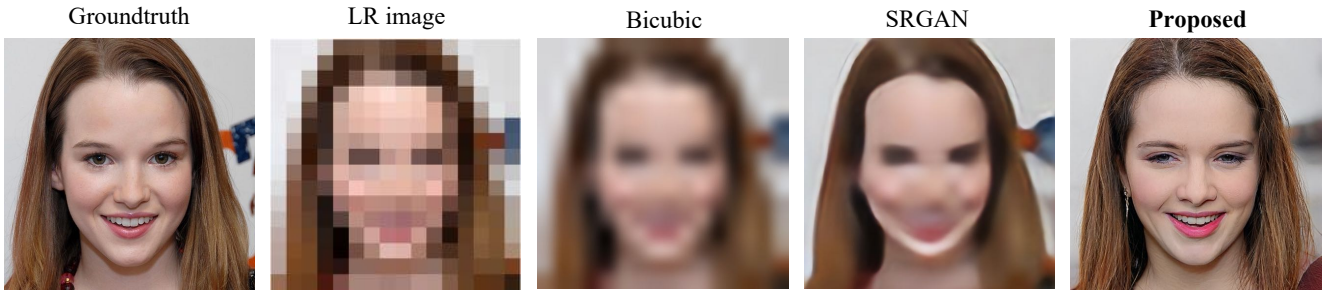
(b) Super Resolution

(c) Motion Blurring

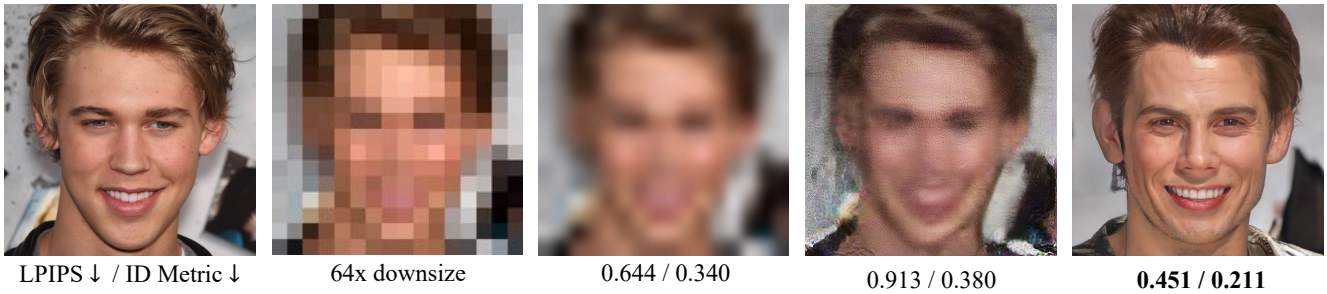


(d) Gaussian Blurring

Figure 6. **Towards exact recovery.** Using a descriptive caption, an output closely resembling groundtruth can be recovered. Shown across tasks of inpainting *(top left)*, super-resolution *(top right)*, motion deblurring *(bottom left)*, and Gaussian deblurring *(bottom right)*.



(a) This person is young and has mouth slightly open, brown hair, and oval face. She wears lipstick.



(b) This young man has mouth slightly open, high cheekbones, bags under eyes, and brown hair. He has no beard.

Figure 7. **Quantitative evaluation for super-resolution.** The proposed method recovers clean, photorealistic images that closely resemble the groundtruth justifying its lower LPIPS and face ID metrics when compared to methods such as bicubic interpolation and SRGAN.

better correlates with perception. Similarly, we also include the face identity (ID) metric based on Arcnet [19]. Tables 1 and 2 show the superiority of our method over existing baselines on super-resolution and inpainting tasks. For both metrics, outputs closer to groundtruth result in lower values.

Baselines considered. We perform metric-based evaluation on the super-resolution and inpainting tasks using 30 randomly sampled images (for each task) from the Celeba-HQ dataset. For the former, we consider bicubic interpolation as a classic method, and SRResNet and SRGAN [41] as a feed-forward deep learning method. For the inpainting task, we compare Telea [65], a traditional method implemented us-

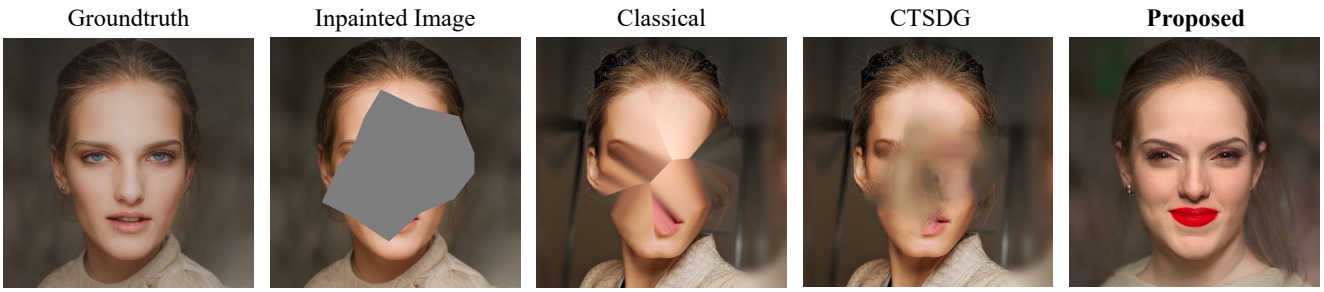


Figure 8. **Quantitative evaluation for inpainting,** with a visual comparison of the proposed method against Telea [65] (classical method) and CTSDG [27]. To highlight the distorted nature of the baselines, we compute metrics exclusively on the masked out (gray) region.

Metric	Bicubic	SRResNet	SRGAN	Proposed
LPIPS ↓	0.658	0.611	0.860	**0.469**
ID metric ↓	0.359	0.368	0.336	**0.227**

Table 1. **Metrics evaluated for** $64\times$ **super-resolution.** Baselines include bicubic upsampling, SRResNet and SRGAN [41].

Metric	Classical	CTSDG	Proposed
LPIPS ↓	0.149	0.282	**0.128**
ID metric ↓	0.459	0.431	**0.233**

Table 2. **Metrics evaluated for inpainting task,** compared using LPIPS and ID metric against Telea [65] and CTSDG [27].

ing OpenCV [14], and CTSDG [27], a text-agnostic learning based state-of-the-art technique. Unfortunately, we did not find existing text-guided inpainting works [24, 69, 75] that were trained on the Celeba-HQ dataset—which could have led to stronger baselines.

6. Conclusion and Discussion

In this work, we have proposed a novel image restoration approach that extends the paradigm proposed by Menon et al. [47] to a larger set of reconstruction tasks while incorporating multi-modal information in the form of natural language to effectively facilitate this. We argue that textual information that co-occurs with image acquisition is largely commonplace today, with prominent examples including social and news media, websites etc. Our approach produces photorealistic outputs while mitigating the ambiguity in outputs arising whilst using generative models to solve inverse problems [47].

Bias considerations. Our framework is potentially susceptible to the same bias problem as the underlying networks StyleGAN and CLIP—we caution usage in sensitive applications. Salminen et al. [62] recently published an exhaustive study on the bias in StyleGAN highlighting the skew

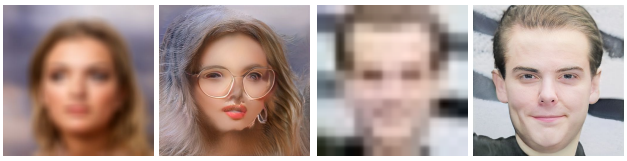
This church is a white edifice with a tall tower.

(a) LSUN church, Motion deblurring

This is a white BMW sedan. It is front facing, with number plate visible.

(b) Stanford car, Gaussian deblurring

Figure 9. **Examples on non-facial datasets,** shown for deblurring tasks on instances from LSUN church and Stanford car datasets.



The person has wavy hair, and big lips and is wearing heavy makeup, necklace, and lipstick. She is young.

(a) Not photorealistic

The man is wearing necktie. He has straight hair, bags under eyes, and mouth slightly open.

(b) CLIP does not capture text

Figure 10. **Failure modes,** predominantly occur when: (a) the optimized latent vectors do not lie on the image manifold, or (b) when CLIP fails to capture relevant text information.

of image generation towards white faces when compared to people of color. Agarwal et al. [5] find inherent racial and gender bias in zero-shot classification using CLIP, while Birhane et al. [11] draw attention to the curation process in pre-training CLIP, resulting in problematic content.

7. Acknowledgements

We thank Prof. Sharon Li and Yifei Ming for the guidance provided throughout the course, and the compute resources used to run the experiments comprising this paper.

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 (ICCV)*, October 2019. 3

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019. 4

[3] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[4] Abdelrahman Abdelhamed, Mahmoud Afifi, Radu Timofte, and Michael S. Brown. Ntire 2020 challenge on real image denoising: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[5] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: Towards characterization of broader capabilities and downstream implications, 2021. 8

[6] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2019. 5

[7] Codruta O. Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2020 challenge on nonhomogeneous dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[8] M Salman Asif, Ali Ayremlou, Aswin Sankaranarayanan, Ashok Veeraraghavan, and Richard G Baraniuk. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Transactions on Computational Imaging*, 3(3):384–397, 2017. 1

[9] Baylies. stylegan-encoder, 2019. Accessed: January 2021. 3

[10] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 1

[11] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 8

[12] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 3

[13] Peter Blomgren, Tony F Chan, Pep Mulet, and Chak-Kuen Wong. Total variation image restoration: numerical methods and extensions. In *Proceedings of International Conference on Image Processing*, volume 3, pages 384–387. IEEE, 1997. 1

[14] Gary Bradski and Adrian Kaehler. Opencv. *Dr. Dobb's journal of software tools*, 3, 2000. 8

[15] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 4

[16] Tony Chan, Selim Esedoglu, Frederick Park, A Yip, et al. Recent developments in total variation image restoration. *Mathematical Models of Computer Vision*, 17(2):17–31, 2005. 1

[17] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua. Gated context aggregation network for image dehazing and deraining. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, March 2018. 2

[18] Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974, 2018. 3

[19] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 2, 7

[20] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021. 2, 4

[21] Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-Based Generative Modeling with Critically-Damped Langevin Diffusion. *arXiv:2112.07068*, 2021. 2, 4

[22] Marco F Duarte, Mark A Davenport, Dharmpal Takhar, Jason N Laska, Ting Sun, Kevin F Kelly, and Richard G Baraniuk. Single-pixel imaging via compressive sampling. *IEEE signal processing magazine*, 25(2):83–91, 2008. 1

[23] William T Freeman, Thouis R Jones, and Egon C Pasztor. Example-based super-resolution. *IEEE Computer graphics and Applications*, 22(2):56–65, 2002. 1

[24] Stanislav Frolov, Tobias Hinz, Federico Raue, Jörn Hees, and Andreas Dengel. Adversarial text-to-image synthesis: A review. *arXiv preprint arXiv:2101.09983*, 2021. 8

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[26] Shanyan Guan, Ying Tai, Bingbing Ni, Feida Zhu, Feiyue Huang, and Xiaokang Yang. Collaborative learning for faster stylegan embedding. *arXiv preprint arXiv:2007.01758*, 2020. 3

[27] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14134–14143, October 2021. 8

[28] Robert A Hummel, B Kimia, and Steven W Zucker. Deblurring gaussian blur. *Computer Vision, Graphics, and Image Processing*, 38(1):66–80, 1987. 1

[29] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 2

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In

Proceedings of the European Conference on Computer Vision (ECCV), October 2016. 1

[31] 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 (ICCV)*, pages 13941–13949, October 2021. 3

[32] 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. 4

[33] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017. 5

[34] 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 (CVPR)*, June 2020. 1, 4

[35] Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981. 3

[36] Dong-Wook Kim, Jae Ryun Chung, and Seung-Won Jung. Grdn:grouped residual dense network for real image denoising and gan-based real-world noise modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019. 2

[37] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013. 5

[38] Dilip Krishnan, Terence Tay, and Rob Fergus. Blind deconvolution using a normalized sparsity measure. In *CVPR 2011*, pages 233–240. IEEE, 2011. 3

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 5

[40] Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations*, 2021. 5

[41] 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. 1, 3, 4, 7, 8

[42] Chengbo Li, Wotao Yin, Hong Jiang, and Yin Zhang. An efficient augmented lagrangian method with applications to total variation minimization. *Computational Optimization and Applications*, 56(3):507–530, 2013. 1

[43] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5

[44] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, April 2019. 5

[45] Andreas Lugmayr, Martin Danelljan, and Radu Timofte. Ntire 2020 challenge on real-world image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[46] Xiaojiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections. In *Advances in neural information processing systems*, December 2016. 2

[47] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. *arXiv:2003.03808 [cs, eess]*, July 2020. 1, 3, 4, 5, 8

[48] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 4, 5

[49] Seungjun Nah, Sanghyun Son, Radu Timofte, and Kyoung Mu Lee. Ntire 2020 challenge on image and video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 2

[50] Pandu Nayak. Mum: A new ai milestone for understanding information. https://blog.google/products/search/introducing-mum/, December 2021. (Accessed on 12/19/2021). 2

[51] Shree K Nayar and Moshe Ben-Ezra. Motion-based motion deblurring. *IEEE transactions on pattern analysis and machine intelligence*, 26(6):689–698, 2004. 1

[52] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, December 2019. 5

[53] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 5

[54] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 3

[55] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, February 2020. 2

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transfer-

able Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]*, Feb. 2021. 2, 5

[57] Dikpal Reddy, Ashok Veeraraghavan, and Rama Chellappa. P2c2: Programmable pixel compressive camera for high speed imaging. In *CVPR 2011*, pages 329–336. IEEE, 2011. 1

[58] 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. *arXiv preprint arXiv:2008.00951*, 2020. 3

[59] 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 (CVPR)*, pages 2287–2296, June 2021. 3, 5

[60] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015. 2

[61] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29:901–909, 2016. 4

[62] Joni Salminen, Soon-gyo Jung, Shammur Chowdhury, and Bernard J. Jansen. Analyzing demographic bias in artificially generated facial pictures. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–8, New York, NY, USA, 2020. Association for Computing Machinery. 8

[63] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[64] 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 (CVPR)*, June 2018. 2

[65] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004. 7, 8

[66] 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. 3, 5

[67] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*, 2020. 4

[68] Wikipedia. Wu Dao — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Wu%20Dao&oldid=1045892362, 2021. [Online; accessed 26-September-2021]. 2

[69] Xingcai Wu, Yucheng Xie, Jiaqi Zeng, Zhenguo Yang, Yi Yu, Qing Li, and Wenyin Liu. Adversarial learning with mask reconstruction for text-guided image inpainting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3464–3472, 2021. 8

[70] Jonas Wulff and Antonio Torralba. Improving inversion and generation diversity in stylegan using a gaussianized latent space. *arXiv preprint arXiv:2009.06529*, 2020. 4

[71] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *arXiv preprint arXiv:2101.05278*, 2021. 3, 5

[72] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, 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. 5

[73] He Zhang and Vishal M. Patel. Density-aware single image de-raining using a multi-stream dense network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[74] Jiawei Zhang, Jinshan Pan, Wei-Sheng Lai, Rynson W. H. Lau, and Ming-Hsuan Yang. Learning fully convolutional networks for iterative non-blind deconvolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[75] Lisai Zhang, Qingcai Chen, Baotian Hu, and Shuoran Jiang. Text-guided neural image inpainting. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1302–1310, 2020. 8

[76] 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 (CVPR)*, June 2018. 2, 5

[77] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2020. 2

[78] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2016. 5

[79] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049*, 2020. 3