TextIR: A Simple Framework for Text-based Editable Image Restoration

Yunpeng Bai[1], Cairong Wang[1], Shuzhao Xie[1], Chao Dong[2,3], Chun Yuan[1,4], Zhi Wang[1,4]

[1] Tsinghua University, [2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3] Shanghai AI Laboratory, China, [4]Peng Cheng Laboratory, Shenzhen, China

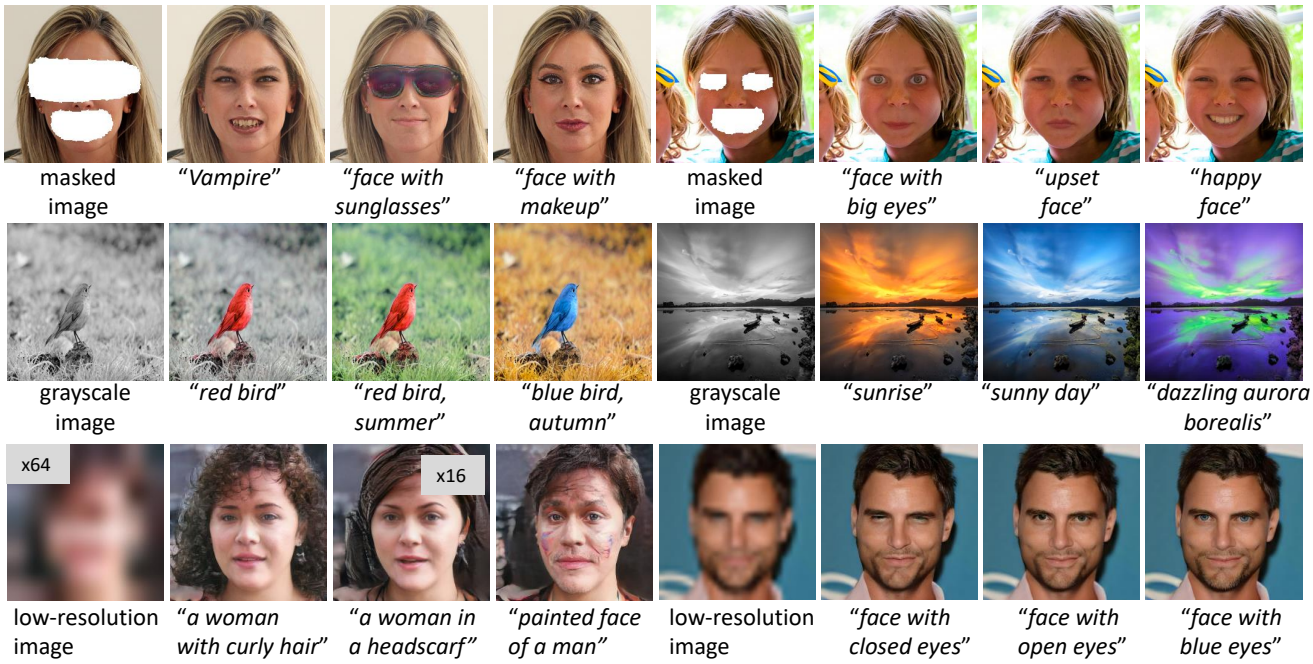

Figure 1. **Overview of image restoration results of the proposed framework.** The framework can be applied to various restoration tasks. For a degraded image, users can use different text inputs to get the restored image they want. Our framework can accept a wide range of text input and can precisely locate the area to be restored.

Abstract

Most existing image restoration methods use neural networks to learn strong image-level priors from huge data to estimate the lost information. However, these works still struggle in cases when images have severe information deficits. Introducing external priors or using reference images to provide information also have limitations in the application domain. In contrast, text input is more readily available and provides information with higher flexibility. In this work, we design an effective framework that allows the user to control the restoration process of degraded images with text descriptions. We use the text-image feature compatibility of the CLIP to alleviate the difficulty of fusing text and image features. Our framework can be used for various image restoration tasks, including image inpainting, image super-resolution, and image colorization. Extensive experiments demonstrate the effectiveness of our method. This new framework also provide a good starting point for the text-based image restoration task.

1. Introduction

Image restoration is a fundamental computer vision problem, which takes a degraded image (e.g., grayscale, damaged, low-resolution image) as input and reconstructs

the corresponding high-quality image. Due to its ill-posed nature, most existing works propose complex deep models to learn strong image priors from massive data to fill in the lost information. However, both convolutional neural networks (CNNs) and Transformer architectures struggle to deal with cases that contain severe information deficits. For example, when the input contains large holes or has complex content (e.g., semantic layout, texture, and depth) to be restored, these methods cannot generate visual-pleasant images.

To further improve the performance of extreme cases, some external priors are explored and introduced into the image restoration model to provide additional guidance. These priors usually include generative prior [47, 53] and structural prior [5, 8], but they can only be applied to a particular domain, such as face, and cannot be generalized to other image contents. Some other works have attempted to solve these hard cases by using another reference image with some desired content that is useful to restore the degraded image. However, they still require users to find a suitable reference image first, which will limit its application scenarios.

Compared to images, text descriptions can easily represent the concept of the image that matches our imagination. It can effortlessly depict an image's global style, local property, and abstract concepts, such as color, shape, emotion, etc. Therefore, we can use text descriptions to provide the information needed for the restoration process more flexibly, improve the controllability of image restoration methods, and achieve editable restoration effects to meet diverse requirements. Some approaches attempt to use text to inpaint [26,59] or colorize [50] images. However, these methods have strict requirements for the dataset, such as images with text that accurately describe their colors. Additionally, their models can only be applied to a specific task.

How to effectively fuse the features of the two data modalities of text and image is another challenge. The Contrastive Language-Image Pre-training (CLIP) model [35] is a recent advance that connects the image and text data by training two encoders on an Internet-scale dataset. CLIP learns a multi-modal embedding space shared by text and image features and contains a wide range of visual concepts. In this work, we utilize the properties of CLIP and propose the first text-based image restoration framework — TextIR. Specifically, TextIR uses the text-image feature compatibility of CLIP to convert text descriptions into the corresponding image features, which are then merged with the degraded image features for restoration. It is worth noting that the training procedure does not require additional text-image pairs. This framework allows users to control the restoration process with text descriptions and can be used for various image restoration tasks, including image inpainting, image super-resolution, image colorization, etc.

To summarize, our main contributions are as follows:

- We design a simple and effective framework that allows the user to use text input to get desired image restoration results.
- We take advantage of CLIP's text-image feature compatibility to enable a better fusion of image and text features.
- Our framework can be used for various image restoration tasks, including image inpainting, image super-resolution, and image colorization.

2. Related Works

2.1. Image Restoration

Image restoration aims to remove the effects of degradation from the degraded image input and reconstructs the original high-quality image . In recent years, CNN-based architectures [13, 56], along with spatial & channel attention modules [15, 29, 56, 61] and skip connection-based approaches [24, 56], have achieve significant breakthrough in this task. In addition, encoder-decoder based U-Net architectures [2, 9] have been predominantly studied for restoration due to their hierarchical multi-scale representation while remaining computationally efficient. So far, replacing the CNN with Transformers [45] that have the capacity to capture long-range dependencies in the data enables researchers [25, 48, 55] to achieve better performance. However, these methods still perform poorly on severely degraded images.

To tackle this drawback, subsequent works intend to introduce additional prior as guidance. As a pioneer in the use of generative priors, GFP-GAN [47] utilizes rich and diverse priors encoded in a pre-trained StyleGAN for blind face restoration. Thereafter, GPEN [53] trains a GAN for generating high-quality face images and embeds it into a U-shaped network as the prior decoder. Contemporary works integrate face structure priors [5, 8] into restoration. For instance, FSRNet [8] uses a prior estimation network to ensure that the spatial information at different scales is preserved in the process of face super-resolution. DeepSEE [5] explores the application of semantic maps in the face super-resolution method. Nevertheless, these prior-based methods are only applicable to images of a particular domain, which limits the application scenarios. Our framework uses text descriptions to provide the information needed for the restoration process more flexibly and can be used on a variety of data categories.

2.2. Text-driven Image Manipulation

With the successful development of cross-modal visual and linguistic representations [30, 42, 43, 54], especially the

omnipotent CLIP [35], many efforts [7, 18, 23, 34, 46, 49, 51] have recently started investigating text-driven image manipulation. However, there are no existing methods specifically for image restoration. Among these works, the most relevant ones are StyleCLIP [34], HairCLIP [49], and CLIPStyler [23]. StyleCLIP performs attribute manipulation with exploring learned latent space of StyleGANv2 [21]. However, it can only edit the original image and is limited in the specific domain. Therefore, CLIPStyler proposes a framework that enables a style transfer only with a text description of the desired style. Besides, HairCLIP introduces a method tailored for hair editing, which can manipulate hair attributes individually or jointly based on the texts or reference images. These methods can only edit images according to the attributes in the text description, and cannot use the information in the text to restore the degraded image.

2.3. Text-based Image Generation

In recent years, there has been a rapid rise in text-based image generation works. Early RNN-based works [31] were quickly superseded by generative adversarial approaches [39]. The latter was further improved by multi-stage architectures [57, 58] and an attention mechanism [52]. DALL-E [36, 37] introduced a GAN-free two-stage approach. First, a discrete VAE [38, 44] is trained to reduce the context for the transformer. Next, a Transformer [45] is trained autoregressively to model the joint distribution over the text and image tokens. TediGAN [51] proposes to generate an image corresponding to a given text by training an encoder to map the text into the StyleGAN latent space. Several recent works [1, 10] jointly utilize pre-trained generative models [4, 12, 14] and CLIP to steer the generated result towards the desired target description. More recently, diffusion models (DM) [17, 33, 41] achieves state-of-the-art results on test-to-image synthesis by decomposing the image formation process into a sequential application of denoising autoencoders. Previous DMs operate directly in pixel space, often consuming hundreds of GPU training days. Latent diffusion models [40] are then proposed to enable DM training on limited computational resources while retaining their quality and flexibility by applying them to latent space. Later, [3, 28, 32] fill in the blank that previous DMs are mainly used to create abstract artworks from text descriptions and cannot edit parts of an actual image while preserving the rest.

3. Proposed Method

Restoring a degraded image (e.g., masked image, grayscale image, low-resolution image) is an ill-posed problem because the missing information is uncertain. Our goal is to train an efficient model allowing the user to provide this information in text format and obtain a restored image corresponding to the text description. Text provides a highly intuitive user interface for describing the desired result. The most straightforward way to obtain such a model is to use text as conditional input during training and supervise the restored results with the corresponding image data. However, obtaining such text-image paired data is typically costly. Since CLIP's image and text feature spaces are semantically aligned, we can use the image features extracted by CLIP instead of text features as conditional inputs during training. Then, the conditional input image can be used as ground truth for supervision.

3.1. Overall Pipeline

TextIR takes as input the degraded image $I_d \in \mathbb{R}^{H_{in} \times W_{in} \times C_{in}}$ and a text condition c that controls the attributes of the result, and outputs the restored image $I_r \in \mathbb{R}^{H_{out} \times W_{out} \times C_{out}}$ that satisfies the condition: $I_r = G(I_d, c)$. In the training process, we utilize the text-image shared space of CLIP to simulate texts with images, where the input of an image is similar to the input of the text. In this way, we can translate I_{gt} into the most suitable "text" condition by CLIP's image encoder. The embeddings encoded by the CLIP encoder are taken as conditional input c to the TextIR. After training, the input condition can be converted into text embeddings. The overall pipeline can be formalized as:

$$\begin{aligned} &\text{training: } I_r = G(I_d, E_I(I_{gt})), \\ &\text{inference: } I_r = G(I_d, E_T(\text{text})), \end{aligned} \tag{1}$$

where E_I and E_T denote the image and text encoder of CLIP, respectively.

3.2. Network Architecture

TextIR consists of an encoder and a style-based generator. The encoder network is a simple convolution neural network (CNN) that takes the degraded image I_d as input and extracts its multi-scale features $\{f^0, ..., f^{l-1}, f^l\}$. The feature map f^l from the last layer is used as the "constant" input of style-based generator architecture. Encoded features from other layers are passed to the generator to be fused with the generated features of the same shape through skip connections. This practice ensures that the generated results match the input degraded images. The expression intensities of encoded and generated features are adjusted by a feature fusion module at each level under a conditional strength factor s. The StyleConv layers [21] used in the generator also receive style code $\boldsymbol{w}$ for the modulate operation. Since our model is a conditional generator, the style code $\boldsymbol{w}$ is obtained from the input condition c.

Style modulation. StyleGANv2 [21] proposes style-modulated convolution to eliminate the drop-like artifacts in StyleGAN [20]. The adaptive instance normalization

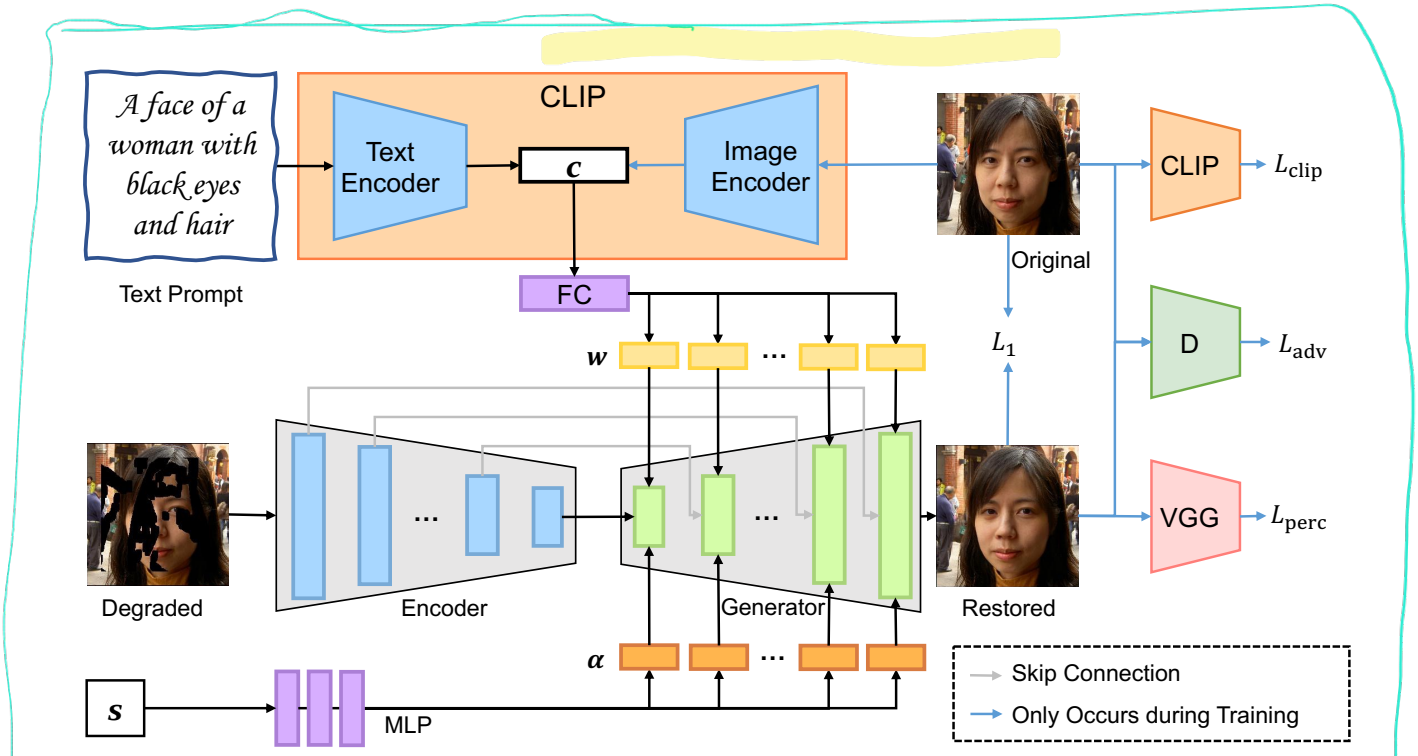

Figure 2. **The proposed TextIR framework.** Our framework consists of an encoder and a generator. The encoder is used to extract features of the degraded images for fusion with the generated features. The generator is used to generate the restoration results. During training, we use the ground truth as the condition. With the help of CLIP's text-image shared feature space, we can use text as a condition to obtain results that match the description during inference.

(AdaIN) on feature maps is replaced by a weaker demodulation technique based on statistical assumptions of the input signal. It uses a style code to modulate the convolution kernel on the input channel dimension. Before conducting the convolution, the kernel is normalized channel-wisely to approximately preserve the standard deviation between the input and output features. We denote such a convolution operator as StyleConv.

We apply StyleConv to inject the conditions into the network. The style code $\boldsymbol{w} = [w^0, ..., w^{l-1}_{1,2}, w^l_{1,2}]$ is obtained from the input text (inference) or image (training) embeddings c by a fully-connected (FC) layer:

$$w = \text{Reshape}(\text{FC}(c)),\ c = E_I(I_{gt}) \text{ or } c = E_T(\text{text}). \quad (2)$$

Unlike StyleGANv2, our generator does not start from a constant value, but takes the feature f^l from the encoder as input. Moreover, to fully utilize the information of the input image I_d, we fuse the multi-level features $\{f^0, ..., f^{l-1}, f^l\}$ output by encoder with their corresponding same shape generated features $\{g^0, ..., g^{l-1}, g^l\}$ by the feature fusion module to obtain x^i at level i. Then, the generator can generate the appropriate restoration part according to the degraded input. Formally, the style modulation is defined as:

$$g^i = \begin{cases} \text{StyleConv}(f^{l-i}, w^i) & i = 0, \\ \text{StyleConv}(\text{StyleConv}(\uparrow_2 (x^{i-1}), w^i_1), w^i_2) & i > 0, \end{cases} \quad (3)$$

where $\uparrow_2$ denotes $2\times$ upsampling.

Feature fusion. As mentioned before, we consider incorporating multi-level features from the encoder into the generator with skip connections. Since I_d differ in the degree of degradation and the contribution of their features to the generation process, strength factors s are employed to flexibly control the expression intensity of encoded and generated features [16]. Therefore, we use a channel-wise weighted sum for the feature fusion instead of simply concatenating or adding them together. Strength factors s are first converted into a series of channel-wise weighting vectors by an MLP:

$$\boldsymbol{\alpha} = \{\alpha^0_{1,2}, ..., \alpha^{l-1}_{1,2}, \alpha^l_{1,2}\} = \text{Reshape}(\text{MLP}(s)). \quad (4)$$

At the i-th level, α^i_1 and $\alpha^i_2 \in \mathbb{R}^{chan(i)}$ are used as weights to fuse f^{l-i} and g^i, where $chan(i)$ indicates the channel dimension. To avoid undesirable effects in the statistics of the output features of StyleConv, these weighting vectors are normalized to be positive and to have channel-wise unit

Figure 3. **The inpainting results.** Compared to blended-diffusion, our method can produce a more natural and realistic result for masked images.

L_2 norm. It is defined formally as:

$$\alpha^i_{enc/gen} = \frac{|\alpha^i_{1/2}|}{\sqrt{{\alpha^i_1}^2 + {\alpha^i_2}^2 + \epsilon}}, \quad \alpha^i_{1/2} \in \boldsymbol{\alpha}, \\ x^i = \alpha^i_{enc} \cdot \text{Conv}(f^{l-i}) + \alpha^i_{gen} \cdot g^i, \tag{5}$$

where ϵ equals to 10^{-8} and $\text{Conv}(\cdot)$ denotes the convolution operator for the initial adjustment of encoded features. In the super-resolution task, we set the strength factor s as the downscaling factor. In other experiments, we also use the encoder to obtain s by another MLP.

3.3. Objective Functions

We demonstrate the capability of our framework on three image restoration tasks: (**a**) image inpainting, (**b**) super-resolution, and (**c**) colorization. For training, we add the corresponding degradation on the ground-truth image I_{gt} to get the degraded image I_d. In the case of inpainting, a mask M is sampled from a free-form mask dataset [27] to get $I_d = [I_{gt} \odot M, M] \in \mathbb{R}^{4\times256\times256}$, where $[\cdot,\cdot]$ denotes concatenation in the channel dimension and $\odot$ denotes Hadamard product. In the case of super-resolution, we downsample I_{gt} and then upsample it to the original resolution to get $I_d \in \mathbb{R}^{3\times512\times512}$. The downsampling factor s is randomly sampled from $\{4, 8, 16, 32, 64\}$. In the case of colorization, we use the L channel of the I_{gt} in *Lab* color space (composed of *L*, *a*, and *b* channels) as $I_d \in \mathbb{R}^{1\times256\times256}$, where the output of the network I_o is the value of *ab* channels. The output is concatenated together with I_d and then converted to *RGB* color space to get $I_r = lab\_to\_rgb([I_d, I_o])$.

We train all tasks using non-saturating adversarial loss:

$$\begin{aligned}\mathcal{L}_{adv,D} &= \mathbb{E}[\log(1+\exp(-D(I_{gt})))\\ &\quad + \log(1+\exp(D(G(I_d,c))))],\\ \mathcal{L}_{adv,G} &= \mathbb{E}[\log(1+\exp(-D(G(I_d,c))))],\end{aligned} \tag{6}$$

and introduce the CLIP loss to guide the restored result to satisfy the condition. We define $\mathcal{L}_{clip}$ as 1 minus the cosine similarity of the result with I_{gt} in the CLIP embedding space:

$$\mathcal{L}_{clip} = 1 - \frac{E_I(I_r)\cdot E_I(I_{gt})}{|E_I(I_r)||E_I(I_{gt})|}. \tag{7}$$

In combination with the L_1 loss (smooth L_1 in colorization) and the perceptual loss [19], the total loss is:

$$\begin{aligned}\mathcal{L}_D &= \lambda_{adv}\mathcal{L}_{adv,D},\\ \mathcal{L}_G &= \lambda_{adv}\mathcal{L}_{adv,G} + \lambda_{clip}\mathcal{L}_{clip} + \lambda_{l_1}\mathcal{L}_1 + \lambda_{perc}\mathcal{L}_{perc}.\end{aligned} \tag{8}$$

We set $\lambda_{adv} = \lambda_{perc} = 0.01, \lambda_{l_1} = 1$ in all experiments, for inpainting, we set $\lambda_{clip} = 0.5$, for super-resolution and colorization, we set $\lambda_{clip} = 0.1$. For CLIP, we use the ViT-B/32 model to extract text prompt embeddings.

4. Experiments and Analysis

4.1. Implementation details

We implement our model using the PyTorch framework. The optimizer we use is Adam [22] and the learning rate for both encoder and generator is 2×10^{-3}. All models are trained on 2 NVIDIA Tesla V100 GPUs with mini-batch size of 16 for 300k iterations. Image inpainting and image super-resolution experiments are performed on the FFHQ

Figure 4. **The colorization results.** Compared to L-CoDe, our method is able to locate the target to be colored more precisely. Our results are colorful and match the text descriptions.

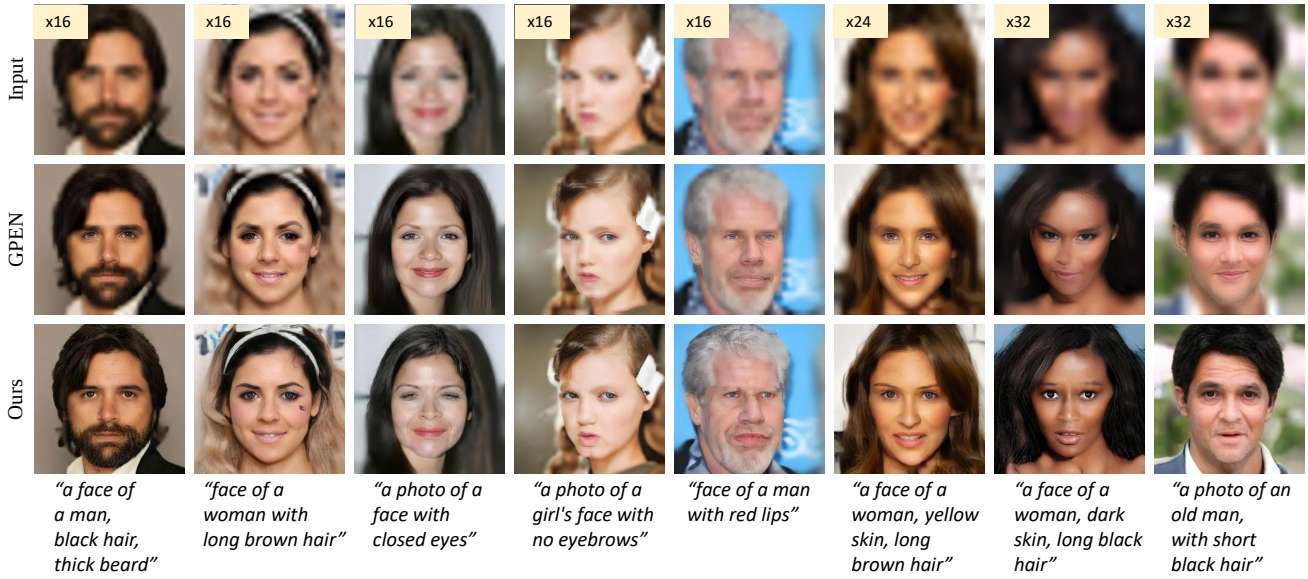

Figure 5. **The super-resolution results.** Compared to GEPN, our method is able to recover more details and these details match the text description.

[20] dataset with resolutions of 256×256 and 512×512, respectively. We use the ImageNet [11] with resolution of 256×256 for image colorization experiment. Image inpainting and super-resolution models are evaluated on the Multi-Modal-CelebA-HQ [51] dataset. Image colorization model is evaluated on COCO-Stuff [6]. For evaluation, we adopt the widely used pixel-wise metrics (PSNR and SSIM) and the perceptual metric (LPIPS [60]).

4.2. Image Inpainting

In the experiments on image inpainting, there is no direct text-based image inpainting work available for comparison. Blended-diffusion [3] is a similar work. They use a denoising diffusion probabilistic model (DDPM) to generate natural-looking results and use CLIP to steer the edit results toward a target text description. They also use a mask to remove the content of the original image and replace it with the edited effect, but they have to use the original image as

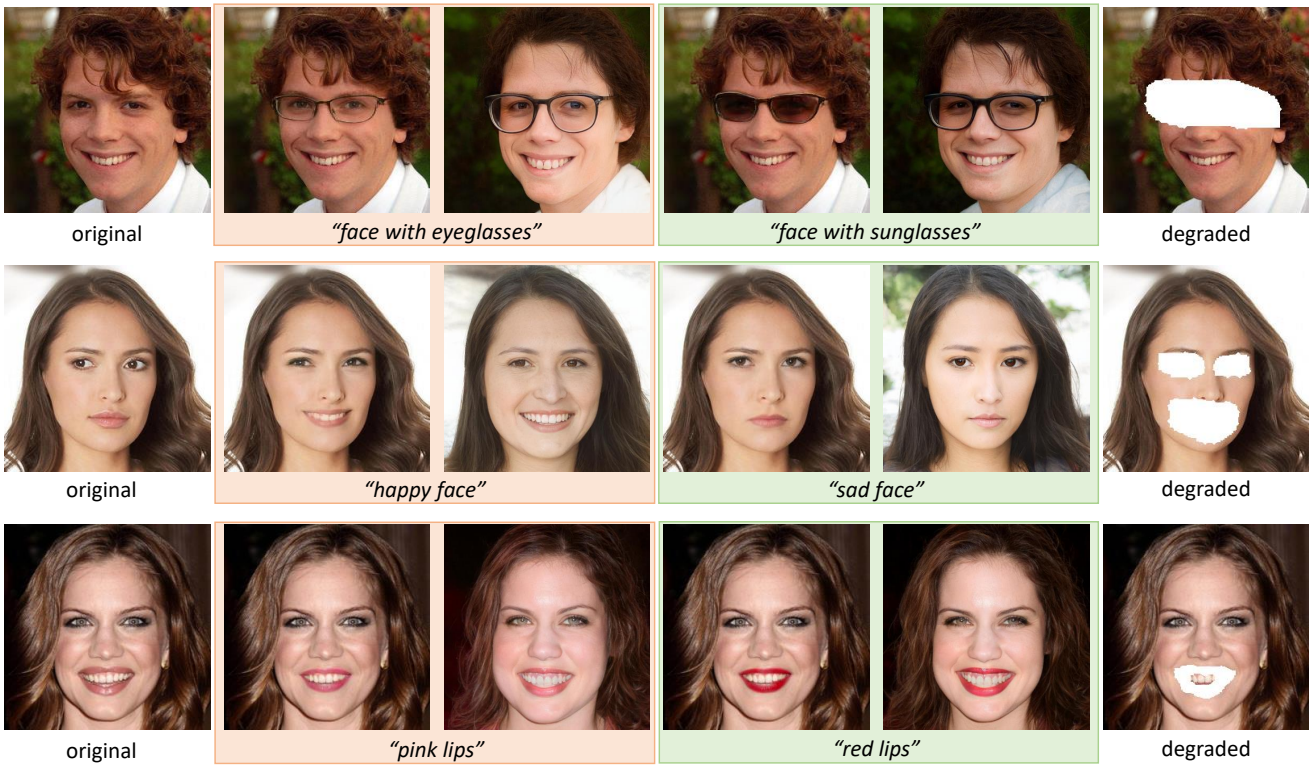

Figure 6. **Comparison with StyleCLIP on face editing**. We first mask out selected regions of the original face and then obtain the edited result by text-based image inpainting. Compared with GAN inversion-based methods like StyleCLIP, our method, **left side** of each rectangle, can specify local editing regions, thus keeping other regions unchanged and maintaining identity perfectly in face editing.

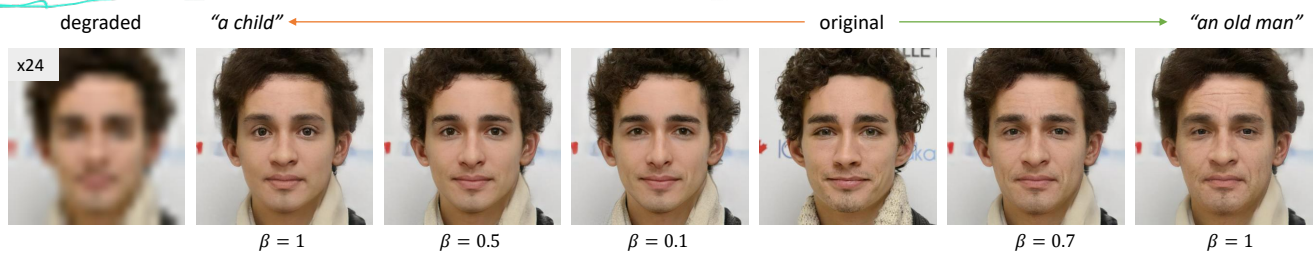

Figure 7. **Conditional strength interpolation.** We do interpolation between the CLIP embeddings of the edited text and the original image to get the control conditions of different intensities.

input. By replacing their image input with a masked image, their method can somehow have the capability of text-based inpainting. Then, we compare our method with theirs in this way.

We first qualitatively compare our method with the blended-diffusion. The faces with different regions masked out and a paired text input are fed into different models. The images restored according to the text are shown in Figure 3. The result of blended-diffusion looks less natural. For example, when inpainting the description of a bald head, the area inpainted by blended-diffusion is different from the original human skin color. When inpainting a man's face and the masked area is large, they do not make any meaningful content. In contrast, our results are more realistic and natural than their results, and all match the description of the text.

We quantitatively evaluate the two methods using the Multi-Modal-CelebA-HQ dataset, in which each image has several corresponding text labels. Using this label as input for the text condition, then the original image can be used as ground truth to evaluate the results. Test mask sampled from the irregular mask dataset [27]. We use the average of these text labels for each image as condition input. The comparison results are shown in Table 1. It can be seen that our method is much better in terms of these metrics compared to blended-diffusion.

Table 1. Quantitative evaluation of inpainting experiment.

Methods	Metrics		
	PSNR ↑	SSIM↑	LPIPS↓
Blended-diffusion [3]	23.16	0.901	0.226
Ours	**29.83**	**0.932**	**0.068**

Table 2. Quantitative evaluation of colorization experiment.

Methods	Metrics		
	PSNR ↑	SSIM↑	LPIPS↓
L-CoDe [50]	24.965	0.916	0.169
Ours	**26.70**	**0.923**	**0.124**

Table 3. Quantitative evaluation of super-resolution experiment.

Methods	Metrics		
	PSNR ↑	SSIM↑	LPIPS↓
GPEN [53]	26.82	0.704	0.273
Ours	**27.41**	**0.784**	**0.227**

4.3. Image Colorization

In the experiments of colorization, we compare our method with a previous language-based colorization method, L-CoDe [50]. We also make a qualitative comparison first, and the comparison results are shown in Figure 4. L-CoDe always fails to match the color to the target object. In the case of colorizing the car yellow, most of the car's body is still gray. When colorful colors were given to the tapes, their results were only grayish without distinct colors. In the last column, when the shoes were colorized with red, the red color was not obvious in their results. In contrast, our method colorizes the target image according to the text and accurately locates the target object. In the quantitative evaluation, we use the caption in COCO-Stuff as the text input and the corresponding image as the ground truth. The comparison results are shown in Table 2. Our method exceed L-CoDe in all three metrics.

4.4. Image Super-resolution

For the experiment on image super-resolution, since there is no similar text-based method for comparison, we compare our methods with a blind face restoration method GPEN [53]. The results of the restored images are shown in Figure 5. Compared to the GPEN, the results of our method are clearer. The details are also consistent with the text descriptions. We also perform a quantitative comparison using the Multi-Modal-CelebA-HQ dataset in the same manner as above. The qualitative results for $16\times$ SR task are shown in Table 3. The introduction of text information in the super-resolution process shows a significant improvement.

4.5. Image Editing via Degradation

TextIR is a text-based controllable restoration framework, so that we can perform text-based image editing by degrading first and restoring later. In contrast to traditional image editing paradigms, our degradation-based approach allows users to specify the area and degree of image degradation to preserve the specific information we want to keep. We compare our method with a previous text-based image editing method StyleCLIP [34]. Figure 6 shows the face editing comparison of our approach with StyleCLIP's global direction method. Since our method only change the local area to be edited, it can perfectly preserve the identity information. While StyleCLIP can make edits based on text, it also changes other attributes and the identity.

4.6. Ablation Studies

Interpolation. By interpolating between the CLIP emdeddings of the original image and the text we want to edit, we can get a condition that is controlled in intensity with β: $c = \beta \cdot E_T(\text{text}) + (1-\beta) \cdot E_I(I_{gt})$. Figure 7 shows an example, the results are obtained by image super-resolution. We can achieve fine-grained image manipulation by this way.

Usage of "s" & the help of text prior. "s" is used in all experiments to control the amount of the degraded image in the result according to the degradation level. As shown in Figure 8. When the degradation is not severe, texts do not affect the image's identity. When the image degradation gets more severe, the role of text prior comes to the fore.

5. Conclusion

In this work, we manage to use text information to assist in image restoration because text input is more readily available and provides information with higher flexibility. To achieve a text-based image restoration method, we utilize the recent CLIP model and design a simple and effective framework, which allows the user to use text input to get desired image restoration results. The framework utilize CLIP's text-image feature compatibility to enable a better fusion of image and text features. Our framework can be used for various image restoration tasks, including image inpainting, image super-resolution, and image colorization.

References

[1] Rameen Abdal, Peihao Zhu, John Femiani, Niloy Mitra, and Peter Wonka. Clip2stylegan: Unsupervised extraction of stylegan edit directions. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 3

[2] Abdullah Abuolaim and Michael S Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, pages 111–126. Springer, 2020. 2

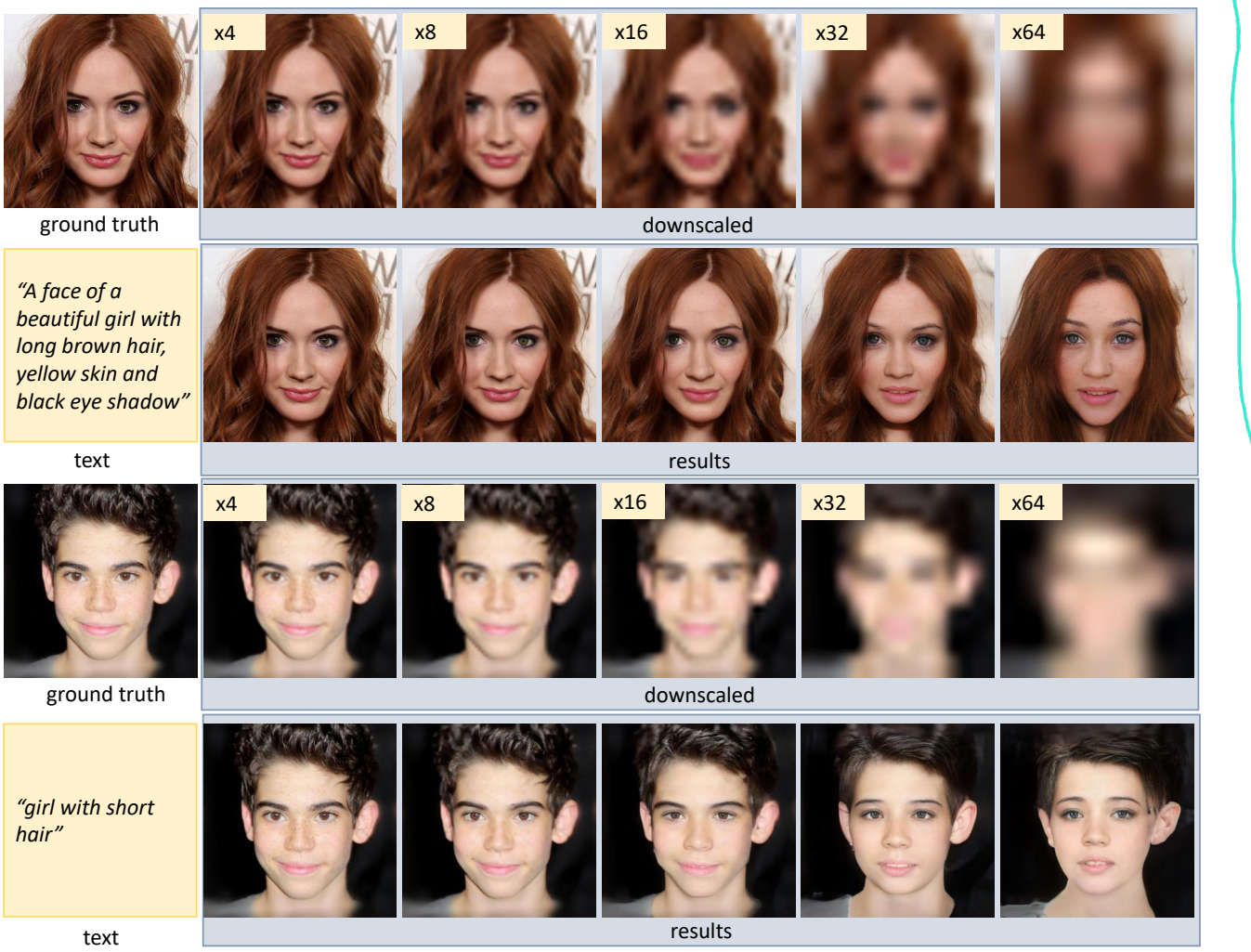

Figure 8. **The ablation study of "s".**

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022. 3, 6, 8

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

[5] Marcel C Buhler, Andrés Romero, and Radu Timofte. Deepsee: Deep disentangled semantic explorative extreme super-resolution. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 6

[7] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8721–8729, 2018. 3

[8] Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2492–2501, 2018. 2

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

[10] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022. 3

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li,

and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[13] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5759–5768, 2022. 2

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

[15] Shuhang Gu, Yawei Li, Luc Van Gool, and Radu Timofte. Self-guided network for fast image denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2511–2520, 2019. 2

[16] Jingwen He, Wu Shi, Kai Chen, Lean Fu, and Chao Dong. Gcfsr: a generative and controllable face super resolution method without facial and gan priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1889–1898, 2022. 4

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

[18] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13799–13808, 2021. 3

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3, 6

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

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[23] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 3

[24] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 254–269, 2018. 2

[25] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 2

[26] Qing Lin, Bo Yan, Jichun Li, and Weimin Tan. Mmfl: Multimodal fusion learning for text-guided image inpainting. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1094–1102, 2020. 2

[27] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 5, 7

[28] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! image synthesis with semantic diffusion guidance. *arXiv preprint arXiv:2112.05744*, 2021. 3

[29] Xing Liu, Masanori Suganuma, Zhun Sun, and Takayuki Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7007–7016, 2019. 2

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[31] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015. 3

[32] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[33] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 3

[34] 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*, pages 2085–2094, 2021. 3, 8

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[37] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever.

Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[38] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 3

[39] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 3

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[41] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 3

[42] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[44] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[46] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 3

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

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

[49] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. Hairclip: Design your hair by text and reference image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18072–18081, 2022. 3

[50] Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi. L-code: Language-based colorization using color-object decoupled conditions. 2022. 2, 8

[51] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021. 3, 6

[52] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018. 3

[53] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021. 2, 8

[54] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 2

[55] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 2

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

[57] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 3

[58] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018. 3

[59] 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. 2

[60] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition,CVPR*, pages 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. 6

[61] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*, 2019. 2