Real image Inversion by learning classifier-free guidance in text-driven diffusion model

Boshi Li
Zhengzhou University of Aeronautics
1602522393boxili@gmail.com

Abstract

One appealing feature of diffusion models is their exceptional ability to generate diverse and high-quality images. Consequently, significant efforts have been invested in editing real images using these pretrained diffusion models. These efforts typically involve finetuning the pretrained model or inverting the image within the latent space of the frozen pretrained model. However, these methods encounter two challenges: (I) They demand users to provide a complete text prompt accurately describing every visual object in the input image. (II) They result in unsatisfactory outcomes for selected regions and unexpected changes in non-selected regions. To tackle these issues, we propose two enhancements for editing real images with a frozen pretrained diffusion model: (I) We invert the real image, and learn a CFG w embedding. This facilitates learning more precise structure maps and an approximate trajectory for reconstructing the real image. Extensive experimental results on various images and prompt editing demonstrate, both qualitatively and quantitatively, that our method achieves more powerful editing capabilities compared to existing and current works.

1. Introduction

Large-scale models, such as those highlighted in the citations [28, 31, 29], have made significant strides owing to their exceptional realism and diversity. Current research delves into the exploration of the text-guided diffusion model for image editing. SDEdit [24], based on a diffusion model generative prior, introduces noise to the input, followed by denoising the resulting image to enhance generative image realism. Despite these efforts, the generated image falls short of accurately preserving input image details. Several studies [26, 3, 2] leverage the mask mechanism for performing mask-specific image editing, allowing users to achieve precise edits. However, the requirement for additional masks makes the editing process less intuitive, necessitating users to provide a perfect mask and limiting their flexibility. P2P [13] innovates prompt-to-prompt image editing by exploring the cross-attention layer, eliminating the need for extra mask information. Meanwhile, certain works concentrate on optimizing textual embedding for image editing, categorized into global editing [9, 21, 19] and local editing [4]. Despite these endeavors, complex image editing remains a challenge, attributed to the fact that the applied regularization is performed globally for the entire image.

The transfer of diffusion model knowledge to real image domains has been explored, focusing on finetuning either the entire [18, 34, 30] or specific parts [20] of the network to manipulate real images while preserving high semantic and visual fidelity. Nevertheless, finetuning with only a few examples, whether for the entire or a part of the generative model, faces challenges such as the cumbersome tuning of model weights and catastrophic forgetting [38]. Recent works [13, 11, 25] address these challenges by preventing the updating of the pre-trained model, focusing on optimizing conditional or unconditional inputs of the cross-attention layers in the classifier-free diffusion model [15] (e.g., Stable Diffusion model [29]). Textual Inversion [11] optimizes the textual embedding of the conditional branch given a few content-similar images, while Null-text optimization [25] modifies the unconditional textual embedding of the unconditional branch. However, these approaches face challenges, including unsatisfactory results for selected regions and unexpected changes in non-selected regions, as well as the need for a user to provide an accurate text prompt describing every visual object and their relationships in the input image.

To address the aforementioned challenges, our approach involves analyzing the role of the classifier-free guidance scale (CFG Scale) mechanism. This analysis reveals that the CFG dominates the output image structure. As a solution, we propose learning the CFG embedding, focusing on CFG. Our method is built upon Stable Diffusion [29], and we conduct experiments across various images and prompt editing scenarios.

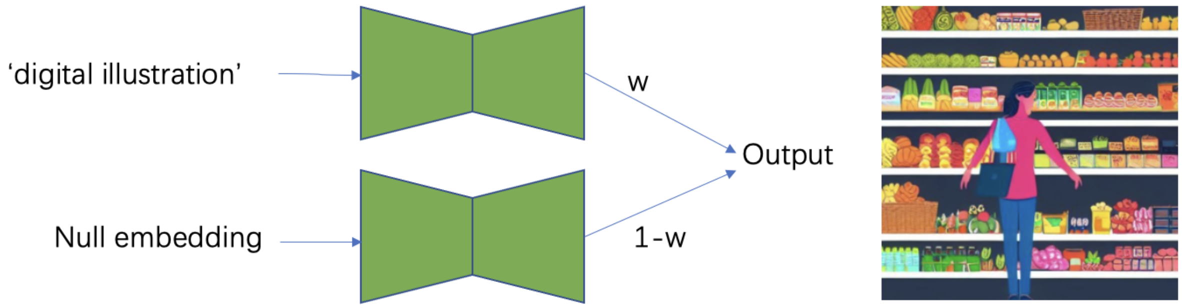

Figure 1: Overview of the proposed method. our method is to learn cfg w when editing a real image.

2. Related work

Knowledge Transfer in Diffusion Models Several recent studies have explored the realm of knowledge transfer within diffusion models [18, 20, 30, 34] using a limited number of images. Existing research, such as [30, 18, 34, 20], either fine-tunes pre-trained models or employs image inversion in the latent space of the pre-trained model. For instance, Dreambooth [30] suggests that training a diffusion model on a small dataset (3-5 images) benefits significantly from a pre-trained model, preserving text editing capabilities. Similarly, Imagic [18] and UniTune [34] rely on interpolation weights or classifier-free guidance during inference, except during fine-tuning. Another approach, presented by Kumari et al. [20], focuses on updating specific parameters of the pre-trained model, specifically the key and value mappings in the cross-attention layers. However, updating the diffusion model inevitably sacrifices the text editing capability of the pre-trained model. In our work, we concentrate on real image editing using a frozen diffusion model.

GAN-based Image Inversion with Knowledge Transfer Early works [37, 17, 35, 42, 36] train a custom GAN and perform image inversion with transfer learning. Image inversion, which aims to project real images into latent spaces for manipulation, is a well-explored concept with various approaches [5, 8, 12, 16, 23, 39, 40, 44]. These methods leverage pre-trained GANs for image manipulation, altering output images based on target semantic attributes. Some approaches [1, 43] reverse images into the input latent space of a pre-trained GAN, often StyleGAN, by optimizing latent representations to reconstruct the target image. These techniques involve fixing or updating the generator for reconstruction, yielding diverse outcomes in image restructuring.

Diffusion Model-based Inversion Inversion techniques for diffusion models can be performed by optimizing latent representations [10]. For instance, DDIM [32] sampling, as demonstrated by [10], can effectively reconstruct real images. Other works [2, 3, 26] assume user-provided masks to control applied changes, achieving both meaningful edits and background preservation. P2P [13] introduces a mask-free editing method, but it may lead to unexpected results when applied to real images. Recent investigations focus on text embedding in the conditional input [11] or null-text optimization in the unconditional input (Null-Text Inversion [25]). Stylediffusion [22] optimizes the input of the value linear network in the cross-attention layers. Despite the editing capabilities afforded by combining new prompts, challenges persist, including unsatisfactory results in selected regions and unexpected changes in non-selected regions. Moreover, these methods require meticulous text prompt editing, demanding accurate inclusion of all visual objects in the input image. Recent work by Parmar et al. [27] introduces pix2pix-zero, aiming to enhance the accurate editing capabilities of real images. However, this approach initially requires computing the textual embedding direction with a thousand sentences in advance, adding a preliminary computational step to the editing process.

3. Method

3.1. Diffusion Model

Text-driven diffusion models on a large scale, as exemplified by references [28, 29, 31], represent a category of conditional generative models designed to approximate the distribution of training data. Typically, these diffusion models optimize a denoiser network ϵ_θ based on UNet to predict Gaussian noise ϵ. This optimization follows a defined objective:

$$\min_{\theta} E_{\mathbf{z}_0,\epsilon\sim N(0,I),t\sim[1,T]} \left\|\epsilon-\epsilon_\theta(\mathbf{z}_t,t,\mathbf{c})\right\|_2^2$$

Here, z_t signifies a noise sample corresponding to timestamp $t \sim [1,T]$, and T denotes the number of timesteps. The text embedding $\mathbf{c}$ is derived by a Clip-text Encoder Γ with a given prompt $\mathbf{p}$: $\mathbf{c} = \Gamma(\mathbf{p})$. Gaussian noise ϵ is introduced to the image feature z_0 [1].

Our work builds upon the Stable Diffusion model [29]. Initially, both the encoder E and decoder D undergo training. Subsequently, the diffusion process unfolds in the latent space. The encoder maps the image $\mathbf{x}$ to the latent representation $\mathbf{z_0} = E(\mathbf{x})$, and the decoder D endeavors to invert the latent representation $\mathbf{z_0}$ back to the image $\mathbf{x} = D(\mathbf{z_0})$. The sampling process is given by:

$$\mathbf{z}_{t-1} = \sqrt{\tfrac{\alpha_{t-1}}{\alpha_t}}\mathbf{z}_t + \sqrt{\alpha_{t-1}}\left(\sqrt{\tfrac{1}{\alpha_{t-1}}-1}-\sqrt{\tfrac{1}{\alpha_t}-1}\right)\cdot\epsilon_\theta(\mathbf{z}_t,t,\mathbf{c}), \quad (1)$$

where α_t is a scalar function. During inference, a random noise image z_T is denoised sequentially for a fixed number of timesteps T (i.e., $T = 50$ in this paper) using the optimized model ϵ_θ.

DDIM inversion. In the realm of real-image editing employing a pretrained diffusion model, the task involves reconstructing a given real image by identifying its initial noise. Drawing inspiration from the relevant study [13], our approach, P2P [13], leverages the deterministic DDIM model for image inversion. The generation of latent noises follows a similar methodology, encapsulated by the process defined as:

$$\mathbf{z}_{t+1} = \sqrt{\tfrac{\alpha_{t+1}}{\alpha_t}}\mathbf{z}_t + \sqrt{\alpha_{t+1}}\left(\sqrt{\tfrac{1}{\alpha_{t+1}}-1}-\sqrt{\tfrac{1}{\alpha_t}-1}\right)\cdot\epsilon_\theta(\mathbf{z}_t,t,\mathbf{c}). \quad (2)$$

The DDIM inversion process generates latent noise that, when fed into the diffusion process, approximates the input image. While DDIM-based reconstruction may lack precision, it serves as a solid starting point for training, facilitating the efficient attainment of high-fidelity inversion [13]. Employing the intermediate results of DDIM inversion, a method introduced by [25] optimizes the embedding in the unconditional part of the Stable Diffusion Model. Specifically, during the inference stage, it aligns the denoised sample with the one produced by DDIM inversion at the corresponding timestep. In our work, we adopt a similar mechanism to train our model, drawing parallels to [7, 25].

In this paper, we introduce a novel use of DDIM sampling [10, 32] for processing a given real image. This approach generates latent noises that, when introduced into the diffusion process, yield an approximation of the input image.

[1] Our focus in this paper is on the Stable Diffusion Model, which operates in the image feature space.

Algorithm 1 Our algorithm

Require: the features of the training images and the prompt embeddings: $\{\mathbf{z}_0, \mathbf{c}_0\}$.
Middle results: With guidance scale $w = 1$ for the classifier-free diffusion model, we use DDIM inversion to produce $\{\hat{\mathbf{z}}_j\}(j = 1, ..., T)$.
Output: CFG w.

Set guidance scale $w = 7.5$;
Initializing $\widetilde{\mathbf{z}}_T \leftarrow \hat{\mathbf{z}}_T$;
for $t = T, T-1, \ldots, 1$ **do**
  for $k = 0, \ldots, K-1$ **do**
    $\mathbf{z}_{t-1} \leftarrow \widetilde{\mathbf{z}}_t$;
    $\omega \leftarrow \omega - \eta\nabla_\omega\mathcal{L}$;(Eq. **??**)
  end
  Synthesizing $\widetilde{\mathbf{z}}_{t-1}$;(Eq. 4)
end
Return CFG w

3.2. CFG W optimization

Method overview. For a given real image, our goal is to obtain more accurate editing capabilities with a frozen pretrained model. We invert a real image into a textual embedding $\mathbf{c}$ which is fed into the cross-attention layers. Given the pair image feature $\mathbf{z}_0$ and textual embedding $\mathbf{c}_0$, We learn the CFG embedding $\widetilde{\mathbf{w}}$. In addition, for the inverted image we further improve the editing technique which is used for the unconditional branch of classifier-free guidance, as well as the conditional one, like P2P [13]. Our method is illustrated in Fig. 1

Reconstruction Loss. Since the noise representations $(\{\hat{\mathbf{z}}_1, \cdots \hat{\mathbf{z}}_T\})$ provide an initial trajectory which is close to the real image, we train the mapping network M_{t-1} to output the noise, which is close to the noise representations $(\hat{\mathbf{z}}_t)$ with Eq. 1 [25]. The objective is

$$\mathcal{L}_{rec} = \min_{M_{t-1}} \left\|\hat{\mathbf{z}}_{t-1} - \mathbf{z}_{t-1}\right\|^2, \quad (3)$$

$$\mathbf{z}_{t-1} = \sqrt{\tfrac{\alpha_{t-1}}{\alpha_t}}\widetilde{\mathbf{z}}_t + \sqrt{\alpha_{t-1}}\left(\sqrt{\tfrac{1}{\alpha_{t-1}}-1}-\sqrt{\tfrac{1}{\alpha_t}-1}\right)\cdot\epsilon_\theta(\widetilde{\mathbf{z}}_t,t-1,\mathbf{c}_0), \quad (4)$$

$$\widetilde{\mathbf{z}}_t = \sqrt{\tfrac{\alpha_t}{\alpha_{t+1}}}\widetilde{\mathbf{z}}_{t+1} + \sqrt{\alpha_t}\left(\sqrt{\tfrac{1}{\alpha_t}-1}-\sqrt{\tfrac{1}{\alpha_{t+1}}-1}\right)\cdot\epsilon_\theta(\widetilde{\mathbf{z}}_{t+1},t,\mathbf{c}_0), \quad (5)$$

At inference time, the initial input is $\widetilde{\mathbf{z}}_T = \hat{\mathbf{z}}_T$.

4. Experimental setup

Training details and datasets. We implement the pretrained Stable Diffusion model in our approach. For detailed network information and additional results, refer to Supplementary Material A. Our dataset comprises 50 randomly collected image and caption pairs (with a resolution of 512×51) from Unsplash (https://unsplash.com/) and COCO [6]. The evaluation metric *Clipscore* [14] gauges the quality of a prompt-edited image pair.

Figure 2: Visualization of our method.

Metric	Structure-dist↓	NS-LPIPS↓	Clipscore↑
*DDIM	0.094	0.3408	**84.2**%
SDEdit	0.044	0.2046	80.1%
Null-text	0.028	0.1114	77.8%
StyleDiffusion	0.022	0.0845	79.3%
Ours	**0.021**	**0.0840**	81.3%

Table 1: Comparison with baselines on three metrics. NS-LPIPS: non-selected LPIPS. *DDIM: DDIM inversion with word swap.

To assess the preservation of structural information post-editing, we employ Structure Dist [33] for computing the structural consistency of the edited image.

In this study, our focus is on modifying the selected region corresponding to the target prompt while preserving the non-selected region. Consequently, evaluating changes in the non-selected region post-editing becomes crucial. To automatically obtain the non-selected region of the edited image, we employ a binary method to generate the raw mask using the attention map, followed by reversal to derive the non-selected region mask. Utilizing this mask, we calculate the non-selected region LPIPS [41] between a pair of real and edited images, referred to as *NS-LPIPS*. A lower score in NS-LPIPS indicates greater similarity between the non-selected region and the input image.

Baselines. We conduct comparisons with the following baseline models. *Null-text* [25] transforms real images along with corresponding captions into the text embedding of the unconditional part of the classifier-free diffusion model. *SDEdit* [24] introduces a stochastic differential equation for generating realistic images through an iterative denoising process. *Pix2pix-zero* [27] (concurrent work) edits real images to identify potential directions from source to target words. In addition, we compare our method with *DDIM + word swap* [27], which involves DDIM sampling using an edited prompt generated by swapping the source word with the target. For the comparisons, we utilize the official codes of the baseline models.

5. Experiments

Qualitative and quantitative results. Fig. 2 presents a our qualitative method. We first invert the real image, and edit the given image by different prompt based on P2P. Our method manages to generate high-quality images, such as dog or cat faces (second column). For example, we are able to the input image (left column) into different target images. We could generate the target image with a similar pose as the input images. Also, the background information is still preserved. Our method successfully edits the target-specific object resulting in a high-quality image, indicating that the proposed method has more accurate editing capabilities.

We assess the effectiveness of the proposed approach using the gathered dataset. As shown in Table 1, the proposed method attains the highest scores for both Structure distance and NS-LPIPS, highlighting its superior ability to preserve structural information. Regarding Clipscore, our method outperforms StyleDiffusion and shows comparable results to SDEdit. Specifically, *DDIM with word swap* achieves the highest Clipscore. Notably, we observe that *DDIM with word swap* not only alters the background but also modifies the structure within the selected region.

6. Conclusions and Limitations

We present a novel approach for editing real images. In this method, we transform the real image by feeding it into CFG w embedding. This strategy allows us to preserve the structure information and an approximate trajectory for reconstructing the real image. Extensive experimental results on various images and prompt editing demonstrate, both qualitatively and quantitatively, that our method achieves more powerful editing capabilities compared to existing and current works.

References

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

[2] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *arXiv preprint arXiv:2206.02779*, 2022. 1, 2

[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. 1, 2

[4] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. *arXiv preprint arXiv:2204.02491*, 2022. 1

[5] Amit H Bermano, Rinon Gal, Yuval Alaluf, Ron Mokady, Yotam Nitzan, Omer Tov, Oren Patashnik, and Daniel Cohen-Or. State-of-the-art in the architecture, methods and applications of stylegan. In *Computer Graphics Forum*, volume 41, pages 591–611. Wiley Online Library, 2022. 2

[6] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 3

[8] 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. 2

[9] 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. *arXiv preprint arXiv:2204.08583*, 2022. 1

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2, 3

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 2

[12] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. pages 5744–5753, 2019. 2

[13] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 2, 3

[14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 3

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1

[16] Ali Jahanian, Lucy Chai, and Phillip Isola. On the"steerability" of generative adversarial networks. 2020. 2

[17] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 2

[18] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Hui-Tang Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *ArXiv*, abs/2210.09276, 2022. 1, 2

[19] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022. 1

[20] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022. 1, 2

[21] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. *arXiv preprint arXiv:2112.00374*, 2021. 1

[22] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing. *arXiv preprint arXiv:2303.15649*, 2023. 2

[23] Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782*, 2017. 2

[24] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021. 1, 4

[25] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022. 1, 2, 3, 4

[26] 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. 1, 2

[27] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 2, 5

[28] 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. 1, 2

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 3

[30] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 1, 2

[31] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2

[32] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 2, 3

[33] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022. 4

[34] Dani Valevski, Matan Kalman, Yossi Matias, and Yaniv Leviathan. Unitune: Text-driven image editing by fine tuning an image generation model on a single image. *arXiv preprint arXiv:2210.09477*, 2022. 1, 2

[35] Yaxing Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shahbaz Khan, and Joost van de Weijer. Minegan: Effective knowledge transfer from gans to target domains with few images. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[36] Yaxing Wang, Abel Gonzalez-Garcia, Chenshen Wu, Luis Herranz, Fahad Shahbaz Khan, Shangling Jui, Jian Yang, and Joost van de Weijer. Minegan++: Mining generative models for efficient knowledge transfer to limited data domains. *International Journal of Computer Vision*, pages 1–25, 2023. 2

[37] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and B. Raducanu. Transferring gans: generating images from limited data. In *ECCV*, 2018. 2

[38] Chenshen Wu, Luis Herranz, Xialei Liu, Joost Van De Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. *Advances in Neural Information Processing Systems*, 31, 2018. 1

[39] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey, 2021. 2

[40] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with deep generative models, 2017. 2

[41] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4

[42] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *arXiv preprint arXiv:2006.10738*, 2020. 2

[43] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. *arXiv preprint arXiv:2004.00049*, 2020. 2

[44] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European conference on computer vision*, pages 597–613. Springer, 2016. 2