Real image editing with SDS

Boshi Li,haowanming
Zhengzhou University of Aeronautics, Zhengzhou University
1602522393boxili@gmail.com, iewmhao@zzu.edu.cn

Abstract

The emergence of large-scale text-to-image generative models represents a groundbreaking advancement in the evolution of generative AI. These models can synthesize a wide range of images encompassing highly complex visual concepts.As a result, substantial endeavors have been dedicated to editing real images through the utilization of these pre-trained diffusion models. However, these methods encounter two challenges. Users are required to provide accurate text descriptions in the input image. The existing methods still fail to accurately process the given image, resulting in low-quality images, and the editing outcomes are unsatisfactory. To address this issue, we propose a new approach. We employ the SDS loss function to update the target image. Simultaneously, we extract the attention map to constrain the image updates. Our method exhibits superior editing capabilities compared to existing and contemporary works, as evidenced by comprehensive experimental results on diverse images and prompt editing. The evidence, both qualitative and quantitative, supports the effectiveness of our approach.

1. Introduction

Significant advancements have been made by large-scale models, as exemplified in the citations [31, 34, 32], owing to their exceptional realism and diversity. Current research is exploring the text-guided diffusion model for image editing, as demonstrated by SDEdit [25]. This approach, based on a diffusion model generative prior, introduces noise to the input and then denoises the resulting image to enhance generative image realism. Despite these efforts, the generated image falls short of accurately preserving input image details. Several studies [28, 3, 2] leverage the mask mechanism for mask-specific image editing, allowing users to achieve precise edits. However, the need for additional masks makes the editing process less intuitive, requiring users to provide a perfect mask and limiting their flexibility. P2P [14] innovates prompt-to-prompt image editing by exploring the cross-attention layer, eliminating the need for extra mask information. Meanwhile, some works focus on optimizing textual embedding for image editing, categorized into global editing [9, 22, 20] 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, with a focus on finetuning either the entire [19, 39, 33] or specific parts [21] 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 [43]. Recent works [14, 11, 27] 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 [16] (e.g., Stable Diffusion model [32]). Textual Inversion [11] optimizes the textual embedding of the conditional branch given a few content-similar images, while Null-text optimization [27] 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 attention map. We introduce a novel methodology. Our approach involves utilizing the SDS loss function for updating the target image. Additionally, we incorporate the extraction of the attention map to impose constraints on the image updates. This integrated strategy aims to enhance the efficiency and effectiveness of the target image refinement process.Our method is built upon Stable Diffusion [32], and we conduct experiments across various images and prompt editing scenarios.

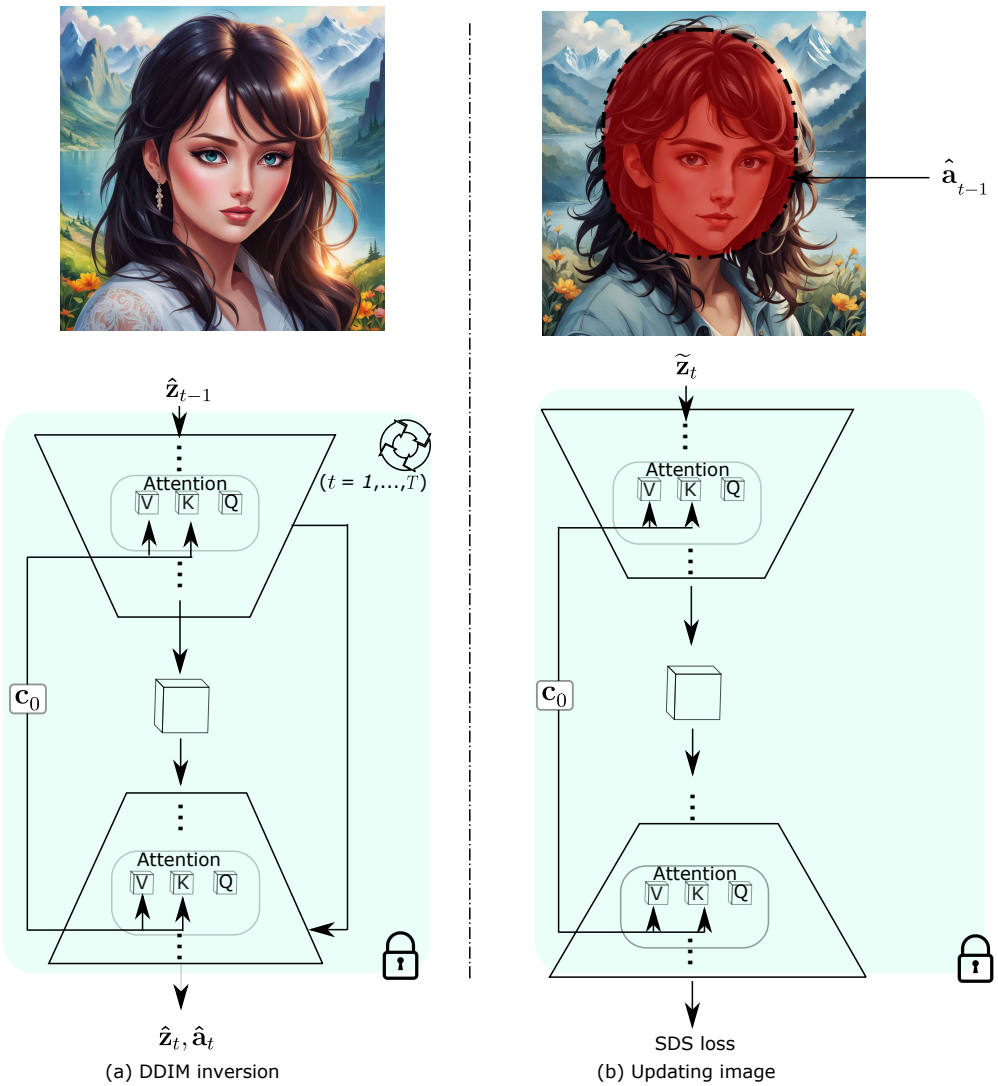

Figure 1: Overview of the proposed method. (a) DDIM inversion. We are able to generate attention map $\hat{a}_t$ as well as the latent code $\hat{z}_t$ (b) we use SDS loss to update a specific area decided by mask $\hat{a}_t$.

2. Related work

GAN-based Image Inversion with Knowledge Transfer In the initial studies [42, 18, 40, 47, 41, 42, 23], researchers employed a customized Generative Adversarial Network (GAN) and engaged in image inversion through transfer learning. The concept of image inversion, aiming to map real images into latent spaces for subsequent manipulation, has been extensively explored, with various approaches available [5, 8, 12, 17, 24, 44, 45, 13, 49]. These methods utilize pre-trained GANs to manipulate images, modifying the output based on specific semantic attributes.

Several strategies [1, 48] involve reversing images into the input latent space of a pre-trained GAN, often StyleGAN. This is achieved by optimizing latent representations to faithfully reconstruct the target image. These techniques may include fixing or updating the generator for the reconstruction process, resulting in diverse outcomes in terms of image restructuring. The utilization of pre-trained GANs for image inversion not only enhances efficiency but also provides a foundation for the exploration of various image manipulation tasks based on well-established semantic attributes.

Knowledge Transfer in Diffusion Models Recent studies have delved into the domain of knowledge transfer within diffusion models, as evidenced by works such as Imagic [19], MultiTune [21], Dreambooth [33], and UniTune [39], all of which have primarily operated with a limited dataset of images. The existing body of research, including Dreambooth, Imagic, UniTune, and MultiTune,

has predominantly focused on either fine-tuning pre-trained models or utilizing image inversion within the latent space of such models.

For instance, Dreambooth [33] advocates for training a diffusion model on a small dataset (3-5 images) with the assistance of a pre-trained model, emphasizing the preservation of text editing capabilities. Similarly, Imagic [19] and UniTune [39] rely on interpolation weights or classifier-free guidance during inference, with the exception being the fine-tuning stage. Another perspective, presented by Kumari et al. [21], revolves around updating specific parameters of the pre-trained model, particularly the key and value mappings in the cross-attention layers. However, a notable drawback of this approach is that updating the diffusion model inevitably comes at the cost of sacrificing the text editing capability of the pre-trained model.

In contrast, our work places a focal point on genuine image editing while utilizing a frozen diffusion model. This approach diverges from the mainstream strategies of updating pre-trained models, allowing us to explore the potential of knowledge transfer within diffusion models without compromising the pre-existing text editing capabilities.

Diffusion Model-based Inversion Several approaches exist for inverting diffusion models, such as optimizing latent representations [10]. One notable example is DDIM [36], which, as illustrated by [10], effectively reconstructs real images through sampling. Alternatively, some methods [2, 3, 28] leverage user-provided masks to control applied changes, achieving both meaningful edits and preserving the background. While P2P [14] introduces a mask-free editing method, it might yield unexpected results when applied to real images.

Recent research has explored incorporating text embeddings in the conditional input [11] or employing null-text optimization in the unconditional input, as demonstrated by Null-Text Inversion [27]. Stylediffusion [23] optimizes the input of the value linear network in the cross-attention layers. Despite advancements in combining new prompts for editing, challenges persist, including unsatisfactory results in selected regions and unexpected changes in non-selected regions. Furthermore, these methods necessitate meticulous text prompt editing, requiring the accurate inclusion of all visual objects in the input image.

In a recent contribution, Parmar et al. [29] introduced pix2pix-zero, aiming to enhance the accurate editing capabilities of real images. However, this approach initially involves 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

Large-scale text-driven diffusion models, illustrated in references [31, 32, 34], fall within the realm of conditional generative models crafted to model the distribution of training data. In general, these diffusion models fine-tune a denoiser network, denoted as ϵ_θ and structured on UNet, for the purpose of predicting Gaussian noise ϵ. The optimization process adheres to a specific objective:

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

In this context, z_t represents a noise sample corresponding to the timestamp $t \sim [1, T]$, where T is the total number of timesteps. The text embedding $\mathbf{c}$ is generated using a Clip-text Encoder Γ with a specified prompt $\mathbf{p}$: $\mathbf{c} = \Gamma(\mathbf{p})$. Gaussian noise ϵ is added to the image feature z_0.

Our research is based on the foundation laid by the Stable Diffusion model [32]. Initially, both the encoder E and the decoder D undergo a training phase. Following this, the diffusion process unfolds within the latent space. The encoder transforms the input image $\mathbf{x}$ into the latent representation $\mathbf{z_0} = E(\mathbf{x})$, and the decoder D strives to reverse this latent representation $\mathbf{z_0}$ back to the original image $\mathbf{x} = D(\mathbf{z_0})$. The sampling process is defined as follows:

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

In this context, α_t represents a scalar function. In the inference phase, an initial random noise image z_T undergoes sequential denoising for a specified number of timesteps T (in this paper, $T = 50$) utilizing the optimized model ϵ_θ.

DDIM inversion. Within the domain of real-image editing using a pre-trained diffusion model, the objective is to reconstruct a provided real image by discerning its initial noise. Taking cues from the pertinent research [14], our method, P2P [14], utilizes the deterministic DDIM model for image inversion. The generation of latent noises follows a similar methodology, encapsulated by the process outlined as:

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

The inversion process in DDIM produces latent noise that, when incorporated into the diffusion process, approximates the input image. While reconstructions based on DDIM may lack precision, they serve as a robust starting point for training, enabling the efficient achievement of high-fidelity inversion [14]. Building on the intermediate outcomes of DDIM inversion, a technique proposed by [27] optimizes the embedding in the unconditional segment of the Stable

Diffusion Model. Specifically, during the inference stage, it aligns the denoised sample with the one generated by DDIM inversion at the corresponding timestep. In our study, we employ a similar mechanism for model training, drawing parallels to [7, 27].

In this paper, we present an innovative application of DDIM sampling [10, 36] for processing a given real image. This method generates latent noises that, when introduced into the diffusion process, produce an approximation of the input image.

3.2. Our method

SDS. Score Distillation Sampling (SDS) stands as an innovative approach crafted by DreamFusion [30] to distill wisdom from a previously trained diffusion model into a differentiable 3D representation generator, such as NeRF [26] or DMTet [35]. Designated as g, the 3D representation generator is characterized by parameters $\theta \in \Theta$, where Θ delineates the realm of θ under the Euclidean metric. For a specified camera c, the derivation of the rendered image I involves the expression $I = g(\theta, c)$. Subsequently, we introduce stochastic perturbations to z decoded from the image I. The diffusion model is then leveraged to forecast the introduced noise ϵ, utilizing a pre-established denoising function ϵ_ϕ, given the noisy image z_t, text embedding y, and noise timestep t.

The SDS approach not only furnishes gradients for the adjustment of the generator, parameterized by θ, but also elucidates this updating procedure through the following formulation:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, \theta) = \mathbb{E}_{\epsilon,t}[w(t)(\epsilon - \epsilon_\phi(z_t, y, t)\frac{\partial z}{\partial I}\frac{\partial I}{\partial \theta}], \quad (3)$$

where $\epsilon \sim \mathcal{N}(0, I)$, $t \sim \mathcal{U}(0.02, 0.98)$. In this paper, instead of optimizing a generator, we directly update the given image.

Cross-attention. SD models achieve image generation based on textual prompts by employing the cross-attention layer. By inputting a prompt, the text embedding $\mathbf{c}$, and the image feature representation $\mathbf{f}$, we derive the key matrix $\mathbf{k}$ as $\Psi_K(\mathbf{c})$, the value matrix $\mathbf{v}$ as $\Psi_V(\mathbf{c})$, and the query matrix $\mathbf{q}$ as $\Psi_Q(\mathbf{f})$ through linear networks Ψ_K, Ψ_V, Ψ_Q. The attention maps are then computed as follows:

$$\mathbf{a} = \text{Softmax}(\frac{\mathbf{q}\mathbf{k}^T}{\sqrt{\mathbf{d}}}), \quad (4)$$

Here, $\mathbf{d}$ signifies the projection dimension of both keys and queries. Ultimately, we represent the cross-attention output as $\hat{\mathbf{f}} = \mathbf{a}\mathbf{v}$, which is subsequently employed as input in the ensuing convolution layers.

Method overview. Our goal is to refine the accuracy of editing functionalities applied to an existing authentic image while keeping a pretrained model constant. We feed an authentic image into the SD model alongside DDIM, leading to the extraction of a latent code $\hat{z}_t$ and an attention map $\hat{a}_t$. Using the attention map $\hat{a}_t$, we create a mask by implementing a threshold. Following this, we utilize the SDS loss for the step-by-step enhancement of the authentic image. It is crucial to emphasize that this enhancement process singularly impacts the particular region defined by the acquired mask.

4. Experimental setup

Training details and datasets. We integrate our methodology with the pretrained Stable Diffusion model. For a detailed exposition of network intricacies and supplementary results, please refer to Supplementary Material A. Our dataset comprises 50 pairs of images and captions, chosen randomly, each with a resolution of 512×51, drawn from Unsplash (https://unsplash.com/) and COCO [6]. To gauge the quality of a prompt-edited image pair, we employ the evaluation metric *Clipscore* [15]. The assessment of structural information preservation post-editing is conducted using Structure Dist [37] to compute the structural consistency of the edited image.

This investigation primarily centers on modifying the selected region corresponding to the target prompt while upholding the integrity of the non-selected region. Hence, evaluating changes in the non-selected region after editing becomes imperative. To automatically derive the non-selected region of the edited image, we utilize a binary method to create the initial mask using the attention map. Subsequently, inversion is applied to generate the mask for the non-selected region. Using this mask, we calculate the non-selected region LPIPS [46] between a pair of real and edited images, denoted as *NS-LPIPS*. A lower NS-LPIPS score signifies a higher similarity between the non-selected region and the input image.

Baselines. We contrast our methodology against several benchmark models. The *Null-text* paradigm [27] morphs genuine images and their associated captions into the textual embedding of the unconditional segment of the classifier-free diffusion model. Another reference, *SDEdit* [25], introduces a stochastic differential equation to iteratively denoise and generate lifelike images. Additionally, we take into account *Pix2pix-zero* [29] (an ongoing study), which modifies actual images to discern potential directions from source to target words.

Moreover, we incorporate a juxtaposition with *DDIM + word swap* [29], wherein DDIM sampling is executed using a prompt altered by interchanging the source word with the target word. To carry out these assessments, we leverage the official codebases of the foundational models.

Figure 2: Visualization of our method. Our results are generated by P2P

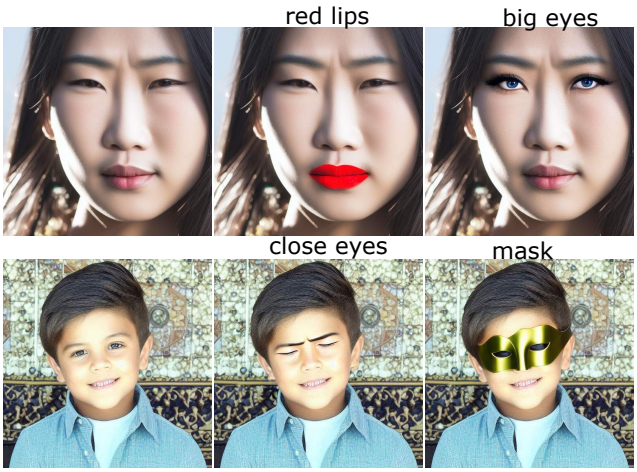

Figure 3: Visualization with PNP [38]

5. Experiments

Qualitative and quantitative results. In the portrayal presented in Fig. 2, we elucidate our qualitative methodology. Commencing with the inversion of the authentic image, we apply modifications based on diverse prompts employing the P2P (Point-to-Point) technique. Our approach excels in the creation of images of exceptional quality, encompassing portrayals of canine or feline countenances (second column). Specifically, we showcase the adeptness to metamorphose the original image (left column) into an array of distinct target images, preserving analogous poses and background details from the source images. This exhibition underscores the proficiency of our methodology in meticulously editing target-specific objects, yielding images of superior quality and attesting to its superior editing capabilities.

As illustrated in Fig. 3, we employ PNP (Point-to-Plane) for the manipulation of the inverted image. Evidently, our methodology seamlessly integrates with PNP, yielding edited images that uphold a commendable level of quality.

6. Conclusions

Presenting a groundbreaking technique for enhancing authentic images, our method entails manipulating real images by implementing SDS loss and attention maps.

Through the strategic application of these elements, we adeptly preserve essential structural details and an approximate trajectory pivotal for reconstructing the genuine image. Our comprehensive array of experimental outcomes, encompassing diverse images and prompt editing scenarios, conclusively illustrates the superior editing capabilities of our approach, surpassing those of current and prior methodologies, both qualitatively and quantitatively.

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, 3

[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, 3

[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. 4

[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. 4

[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. 3, 4

[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, 3

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

[13] Wanming Hao, Ming Zeng, Zheng Chu, and Shouyi Yang. Energy-efficient power allocation in millimeter wave massive mimo with non-orthogonal multiple access. *IEEE Wireless Communications Letters*, 6(6):782–785, 2017. 2

[14] 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, 3

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: a reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 4

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

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

[18] 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

[19] 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, 3

[20] 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

[21] 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, 3

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

[23] 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, 3

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

[25] 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

[26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421, 2020. 4

[27] 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, 3, 4

[28] 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, 3

[29] 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. 3, 4

[30] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 4

[31] 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, 3

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

[33] 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, 3

[34] 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, 3

[35] Tianchang Shen, Jun Gao, Kangxue Yin, Ming-Yu Liu, and Sanja Fidler. Deep marching tetrahedra: a hybrid representation for high-resolution 3d shape synthesis. *Advances in Neural Information Processing Systems*, 34:6087–6101, 2021. 4

[36] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 3, 4

[37] 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

[38] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 5

[39] 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, 3

[40] 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

[41] 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

[42] 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

[43] 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

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

[45] 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

[46] 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

[47] 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

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

[49] 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