DiffMorpher: Unleashing the Capability of Diffusion Models for Image Morphing - *Supplementary Materials*

Kaiwen Zhang[1,2*] Yifan Zhou[3] Xudong Xu[2] Bo Dai[2 ✉] Xingang Pan[3 ✉]

[1]Tsinghua University [2]Shanghai AI Laboratory [3]S-Lab, Nanyang Technological University

Project page: https://kevin-thu.github.io/DiffMorpher_page/

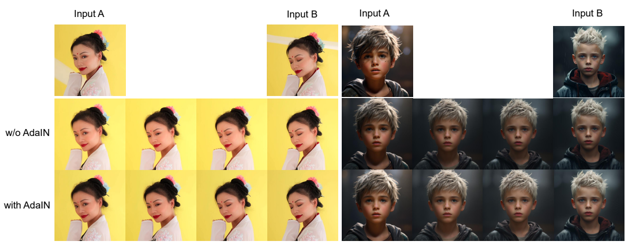

Figure 1. Effects of AdaIN adjustment.

1. Supplementary Method: AdaIN Adjustment

While our approach significantly surpasses previous methods both qualitatively and quantitatively, we occasionally observe color and brightness incoherence between generated images and input images, especially for animation of the same objects. To mitigate this minor problem, we additionally introduce Adaptive Instance Normalization (AdaIN) [6] adjustment for interpolated latent noise $\mathbf{z}_{0\alpha}(\alpha \in (0,1))$ before denoising.

Specifically, we calculate the mean μ_i and standard deviation $\sigma_i (i = 0, 1)$ for each channel of latent noises $\mathbf{z}_{00}, \mathbf{z}_{01}$, and interpolate between μ_i, σ_i as the adjustment target of intermediate noises:

$$\mu_\alpha = (1-\alpha)\mu_0 + \alpha\mu_1 \tag{1}$$

$$\sigma_\alpha = (1-\alpha)\sigma_0 + \alpha\sigma_1 \tag{2}$$

$$\tilde{\mathbf{z}}_{0\alpha} = \sigma_\alpha \left(\frac{\mathbf{z}_{0\alpha} - \mu(\mathbf{z}_{0\alpha})}{\sigma(\mathbf{z}_{0\alpha})} \right) + \mu_\alpha \tag{3}$$

and replace the intermediate latent noise $\mathbf{z}_{0\alpha}$ with the adjusted one $\tilde{\mathbf{z}}_{0\alpha}$ in the denoising process. As demonstrated in Fig. 1, the color and brightness are more coherent after AdaIN adjustment

*Work done during internship at Shanghai AI Laboratory.
✉ Corresponding Author.

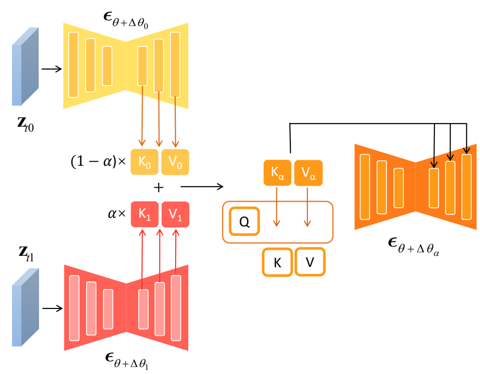

Figure 2. Illustration of Self-Attention Interpolation and Replacement.

2. Implementation Details

In all of our experiments, we use the publicly available state-of-the-art Stable Diffusion v2.1-base as our diffusion model. When training LoRA, to achieve a balance between efficiency and quality and avoid overfitting the single image, we only fine-tune the projection matrices Q, K, V in the attention modules of the diffusion UNet. Additionally, we set the rank of LoRA to 16, and train for 200 steps using AdamW optimizer [8] with a learning rate of 2×10^{-4}. In this setting, training a LoRA for a 512×512 image requires only $\sim$ 20s on a NVIDIA A100 GPU.

During the inversion and denoising process, we adopt the DDIM schedule of 50 steps distilled from entire diffusion steps $T = 1000$. It's noteworthy that we do not apply classifier-free guidance (CFG) [5] in both DDIM inversion and denoising. This is because CFG tends to accumulate numerical errors and cause supersaturation problems, which is also observed in [9, 14]. For attention control, we only perform the feature injection in the upsampling blocks in the self-attention module of the diffusion UNet, and set the hyperparameter λ to 0.6 by default.

3. MorphBench

Conventional image morphing techniques in computer graphics generally require tedious manual labeling of correspondences, and general image morphing is rarely explored in depth in the area of generative models. Therefore, there is a lack of specific evaluation benchmarks for this task. To comprehensively evaluate the effectiveness of our methods, we present *MorphBench*, the first benchmark dataset for assessing image morphing of general objects.

We collect 90 pairs of pictures of diverse content and styles, and divide them into two categories: i) *metamorphosis* between different objects (66 pairs) and ii) *animation* of the same objects (24 pairs). The latter is obtained using off-the-shelf image editing tools such as DragDiffusion [14], Imagic [7], and MasaCtrl [4]. We hope *MorphBench* can also promote future studies on this important problem.

4. More Details of Baselines

In Sec.5, we comprehensively compare our method with previous state-of-the-art methods, including graphical, GAN-based and diffusion-based techniques. We offer more details of the baselines that we use here:

- Warp & Blend [1, 18, 21]: Conventional graphical techniques usually involve bidirectional image warping based on correspondence point pairs with blending operations to achieve morphing effects. We select the representative triangulation-based method [2] as our baseline, which is also widely used in standard libraries such as OpenCV. It divides the images into triangles by performing Delaunay triangulation on user-defined corresponding points, and then morphs between the triangle pairs. Thus, the quantity and quality of the manually labeled pair of points greatly affect the generated results. Since all the other methods do not require correspondence annotations, for the sake of fairness, we adopt the automatic version of this approach https://github.com/jankovicsandras/autoimagemorph that selects 50 morph-points automatically using OpenCV.
- Deep Generative Prior (DGP) [10]: DGP is an image manipulation method based on BigGAN [3], which is suitable for general image morphing. We adopt the official code https://github.com/XingangPan/deep-generative-prior with its default hyperparameters and the pretrained BigGAN model trained on ImageNet [3] as our baseline.
- StyleGAN-XL [13]: Since the pretrained checkpoint of StyleGAN-T [12] is not publicly available, we use the alternative state-of-the-art GAN model StyleGAN-XL https://github.com/autonomousvision/stylegan-xl as our another baseline. Similarly to DGP, the model is trained on ImageNet. We obtain the latent codes of input images by GAN inversion [19] and tune the generator by PTI [11] for better reconstruction results, and interpolate both the latent codes and the generator parameters to get intermediate images. For both GAN-based methods, we use the ImageNet classifier DeiT [16] to automatically determine the class label.
- Denoising Diffusion Implicit Model (DDIM) [15]: We implement a naive diffusion-based interpolation method through DDIM inversion and latent interpolation as our baseline, as discussed in the DDIM paper. As with our approach, the underlying model used is also Stable Diffusion v2.1-base https://huggingface.co/stabilityai/stable-diffusion-2-1-base.
- Diff.Interp. [17]: *Interpolating between Images with Diffusion Models* is a recent state-of-the-art image interpolation method based on diffusion models. Besides latent interpolation, it further introduced pose guidance based on ControlNet [20] to encourage more reasonable intermediate results. However, the smoothness of the morphing video was not considered in this work, and the generated video is full of flickering artifacts. We employ the official code https://github.com/clintonjwang/ControlNet with default settings and pretrained Stable Diffusion v2.1-base model as our baseline. For all three diffusion-based methods, the prompts for each test case are shared.

5. User Study

To assess the quality of image morphing from a human perspective, we invite 40 volunteers to conduct a user study. Each participant are shown 20 groups of morphing videos created by our approach and five baseline methods, chosen at random. They are asked to evaluate the image morphing quality from the perspective of intermediate image fidelity and video smoothness, and to select the one with the best quality for each question. An example of the questionnaire is shown in Fig. 8. In total, we collect 800 responses and summarize the results in Fig. 3. As we can see, our approach is significantly more preferred by users than any of the prior methods.

6. Limitations

One of the limitations of our approach is that we have to train a LoRA for each input image before morphing, which costs additional time ($\sim$ 20 s on a single NVIDIA A100 GPU for a 512×512 image). Another limitation of text-guided diffusion models is that the user must input aligned text prompts in addition to images. Besides, our approach occasionally fails in difficult cases where the correspondence between two input images is not clear enough, and produces relatively unreasonable intermediate images, as shown in Fig. 4. Lastly, although most output images maintain a high level of quality similar to that of the input images, some cases of blurry output can be attributed to the

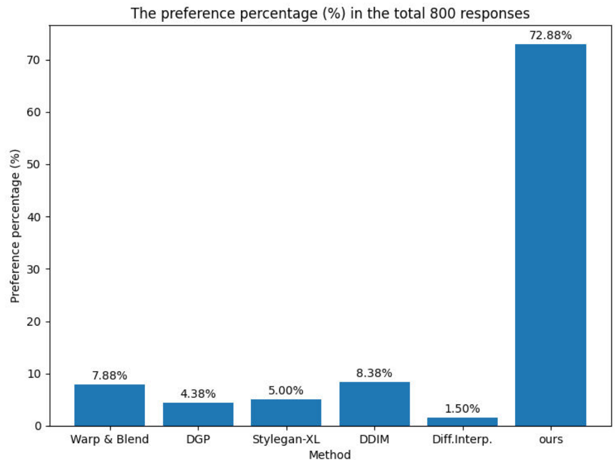

Figure 3. User study result. Our method surpasses all the previous methods by a large margin in terms of user preference.

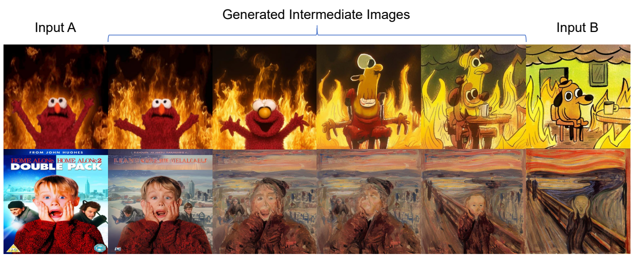

Figure 4. Some relatively unsuccessful cases where the correspondence between two images is not clear enough.

suboptimal selection of the hyperparameter λ. A larger λ can improve video smoothness but is more likely to create blurry textures, as shown in Fig. 8. To reduce the blurry textures, we can select a lower λ (*e.g.* $0.2 \sim 0.4$).

7. More Qualitative Results

Here we present more qualitative results to demonstrate the effectiveness of our *DiffMorpher*. Fig. 5 gives more examples to illustrate the superiority of our approach compared to previous methods in diverse scenarios, and Fig. 6 and Fig. 7 provide additional qualitative results generated by our method that further demonstrate its versatility in real-world applications.

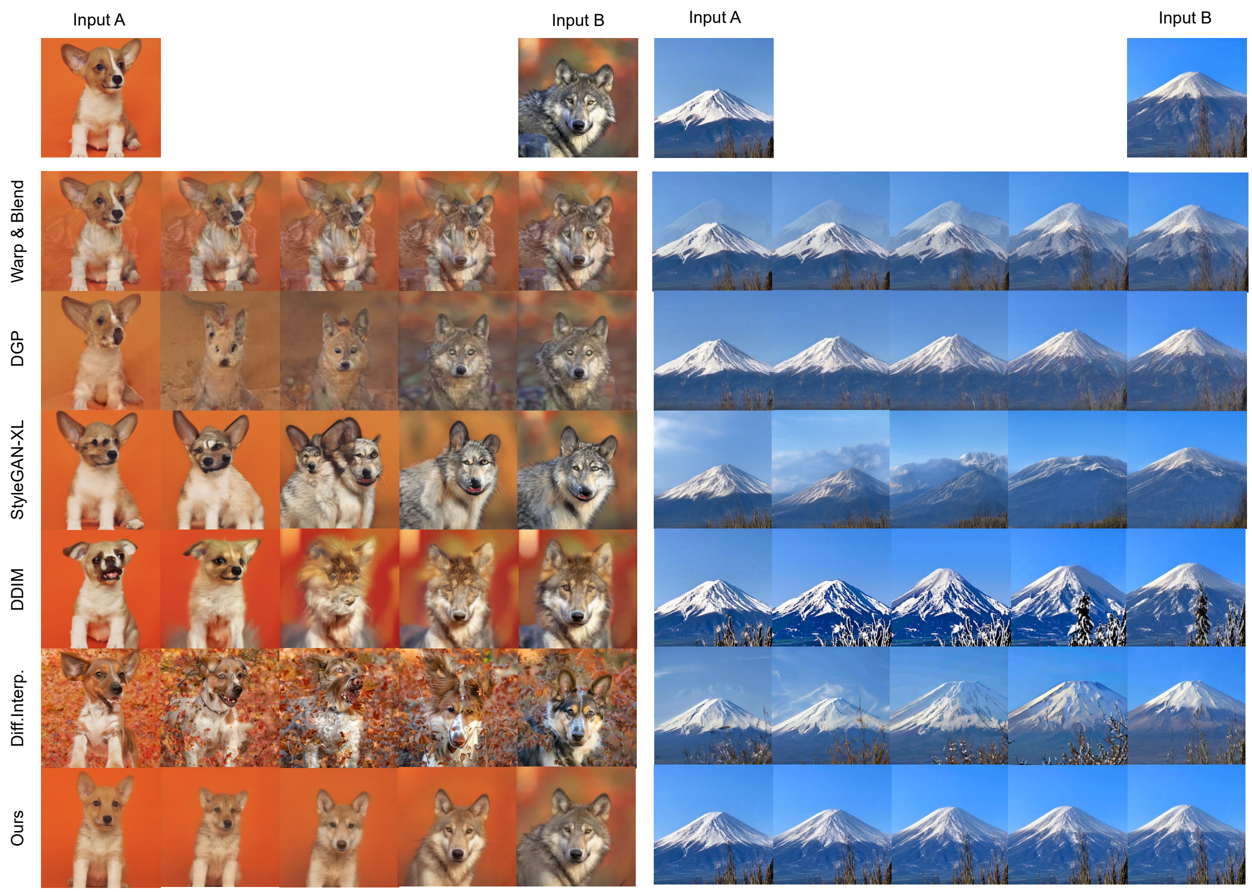

Figure 5. More qualitative comparison results.

Figure 6. More qualitative results of our approach.

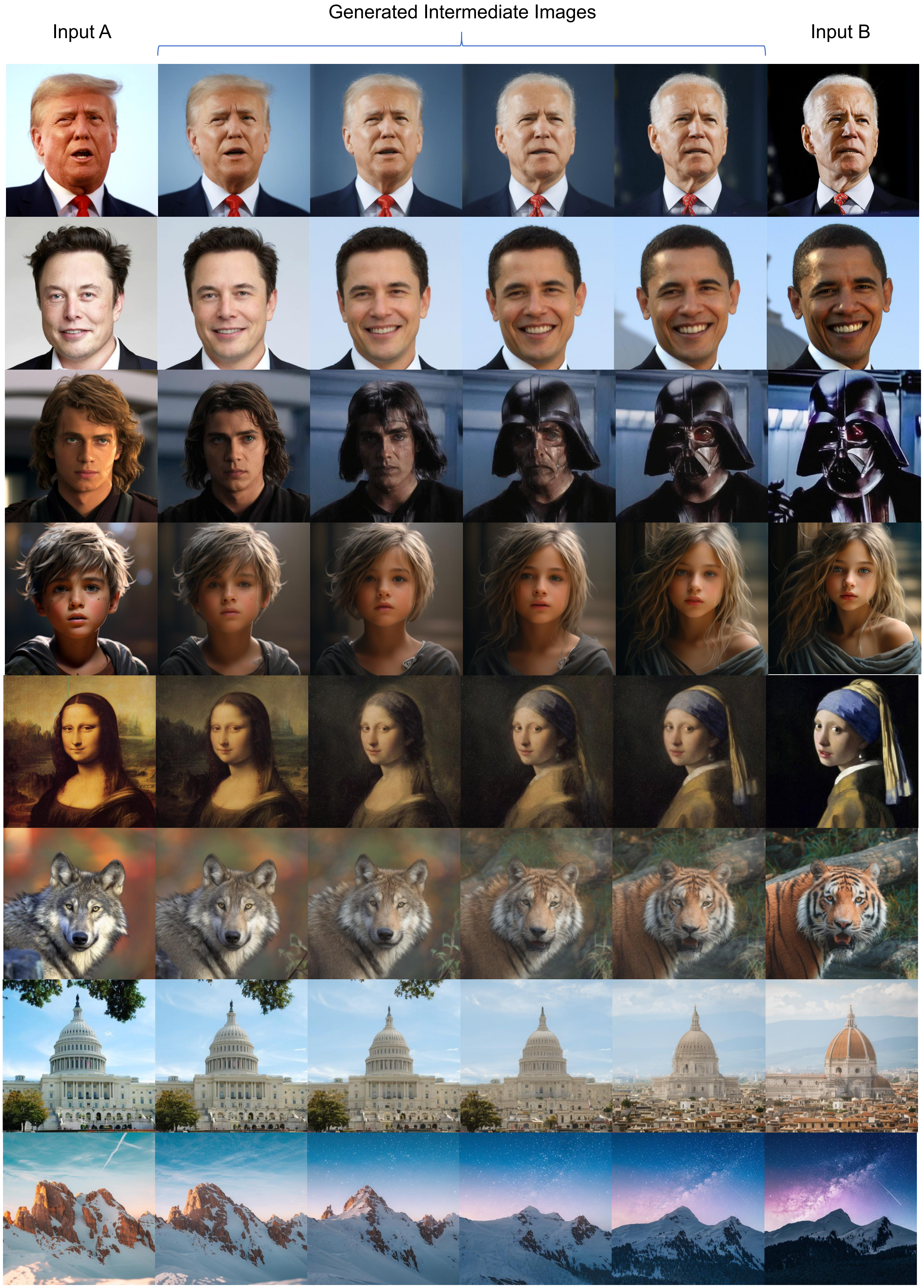

Figure 7. More qualitative results of our approach.

Figure 8. An example of the questionnaire we used in the user study. Note that all the results shown here are videos.

References

[1] Alyaa Aloraibi. Image morphing techniques: A review. *Technium: Romanian Journal of Applied Sciences and Technology*, 9:41–53, 2023. 2

[2] Bhumika G. Bhatt. Comparative study of triangulation based and feature based image morphing. *Signal & Image Processing : An International Journal*, 2:235–243, 2011. 2

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[4] Ming Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *ArXiv*, abs/2304.08465, 2023. 2

[5] Jonathan Ho. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022. 1

[6] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1510–1519, 2017. 1

[7] 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. In *CVPR*, pages 6007–6017, 2022. 2

[8] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101, 2017. 1

[9] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6038–6047, 2022. 1

[10] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. In *ECCV*, 2020. 2

[11] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM TOG*, 2021. 2

[12] Axel Sauer, Tero Karras, Samuli Laine, Andreas Geiger, and Timo Aila. StyleGAN-T: Unlocking the power of GANs for fast large-scale text-to-image synthesis. In *International Conference on Machine Learning*, 2023. 2

[13] Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets. 2022. 2

[14] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *ArXiv*, abs/2306.14435, 2023. 1, 2

[15] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. 2

[16] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, pages 10347–10357, July 2021. 2

[17] Clinton J. Wang and Polina Golland. Interpolating between images with diffusion models, 2023. 2

[18] George Wolberg. Image morphing: a survey. *The Visual Computer*, 14:360–372, 1998. 2

[19] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. Gan inversion: A survey. *IEEE TPAMI*, 45:3121–3138, 2021. 2

[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[21] Bhushan Zope and Soniya B. Zope. A survey of morphing techniques. *International Journal of Advanced engineering, Management and Science*, 3:81–87, 2017. 2