Local 3D Editing via 3D Distillation of CLIP Knowledge

Supplementary Material

Paper ID 5992

1. Implementation details

Zero-shot relevance mask generation 2D relevance mask $\mathbf{M}$ given t_{mask} is generated using the CLIP model and is used as a pseudo label for AFN training. Specifically, we denote a attention map of block b of the CLIP image encoder E_I as $\mathbf{A}^{(b)}$, and its gradients with respect to the model output y as $\nabla\mathbf{A}^{(b)} := \frac{\partial y}{\partial \nabla \mathbf{A}^{(b)}}$. Here y is the cosine similarity between the text embedding $E_T(t_{mask})$ and the image embedding $E_I(I)$. Then the aggregated relevance $\mathbf{N} \in \mathbb{R}^{s \times s}$ for the CLIP encoder consisting of B blocks is computed as:

$$\begin{aligned} \bar{\mathbf{A}}^{(b)} &= \mathbb{I}_s + \mathbb{E}_h(\nabla \mathbf{A}^{(b)} \odot \mathbf{A}^{(b)})^+ \\ \mathbf{N} &= \bar{\mathbf{A}}^{(1)} \cdot \bar{\mathbf{A}}^{(2)} \cdot \ldots \cdot \bar{\mathbf{A}}^{(B)}, \end{aligned} \tag{1}$$

where superscript x^+ denotes $\max(x, 0)$ operation, $\mathbb{E}_h$ is the mean operation along the transformer heads dimension, $\odot$ is the Hadamard product, and s is a sequence length of input tokens. Then we take the first row of $\mathbf{N}$ which corresponds to the relevance for [CLS] token $\mathbf{N}_{\text{[CLS]}} \in \mathbb{R}^s$, and reshape $\mathbf{N}_{\text{[CLS]}[2:s]}$ to $\sqrt{s-1} \times \sqrt{s-1}$ matrix. Finally, the matrix is upsampled to $\mathbf{M} \in \mathbb{R}^{H_V \times W_V}$ using bi-linear interpolation, where H_V and W_V are the height and the width of the neural rendering resolution before the super-resolution. Please refer to transformer visualization methods [1, 4] for theoretical background and additional details.

Network details 8-layer Multi-Layer Perceptron (MLP) with a width of 256 and LeakyReLU for nonlinear activation is used for all three modules: Latent Residual Mapper (LRM), Attention Field Network (AFN), and Deformation Network (DN). For DN and AFN, all the arguments are concatenated and used as input to the model.

Training details Our model utilizes pretrained EG3D [3] model with 128^2 neural rendering resolution for FFHQ [7] and AFHQv2 CATS [6], and 64^2 for ShapeNet Cars [5, 8]. We use the learning rate of 3×10^{-4}, and the lambda values used for the training is $\lambda_{L2} = \lambda_{mask} = 0.1$, $\lambda_{CLIP^+} = \lambda_{id} = 0.3$, and $\lambda_{tv} = 1$. As shown in Fig. 9 of the main paper, we use smaller values of $\lambda_{sparsity}$ for manipulations that require geometric changes, so that we obtain a smooth mask over a wide region, and use larger values otherwise. The training takes $4K$ iterations (about 2 hours) on a single NVIDIA A100 40GB GPU.

	FFHQ			
	Fidelity	Locality	ID	Text reflectance
CLIP-NeRF	2.75	4.50	2.43	4.33
FeNeRF + SC	2.98	4.13	3.18	5.48
IDE3D + SC	9.03	5.95	5.45	6.63
LeNeRF w/o AFN	8.53	8.38	9.40	9.55
LeNeRF (Ours)	**9.25**	**9.53**	**9.98**	**9.70**

Table 1. Results of a user study on four metrics: fidelity, locality, identity preservation (ID), and text reflectance. Scores are in the range of 1-10 and are averaged over 40 surveys. Best in **bold**.

Code release Please refer to the attached code.zip for more details.

2. Additional results

User study We requested 40 users to evaluate LENeRF along with various baselines in the range of 1 to 10 regarding 1) the fidelity, 2) the locality, 3) the identity preservation, and 4) how well the text prompt is reflected in the results. Table 1 shows that LENeRF outperforms all baselines by a large margin for each criterion.

Single-view 2D image editing We present results of single-view 2D image editing in Fig. 1. 2D image is inverted into a 3D model via pivotal tuning inversion (PTI) [9], and manipulated using LeNeRF.

ShapeNet Cars We demonstrate results for ShapeNet Cars in Fig. 2.

Lambda interpolation We can change the rate of manipulation by controlling the lambda value of residuals in Eq.

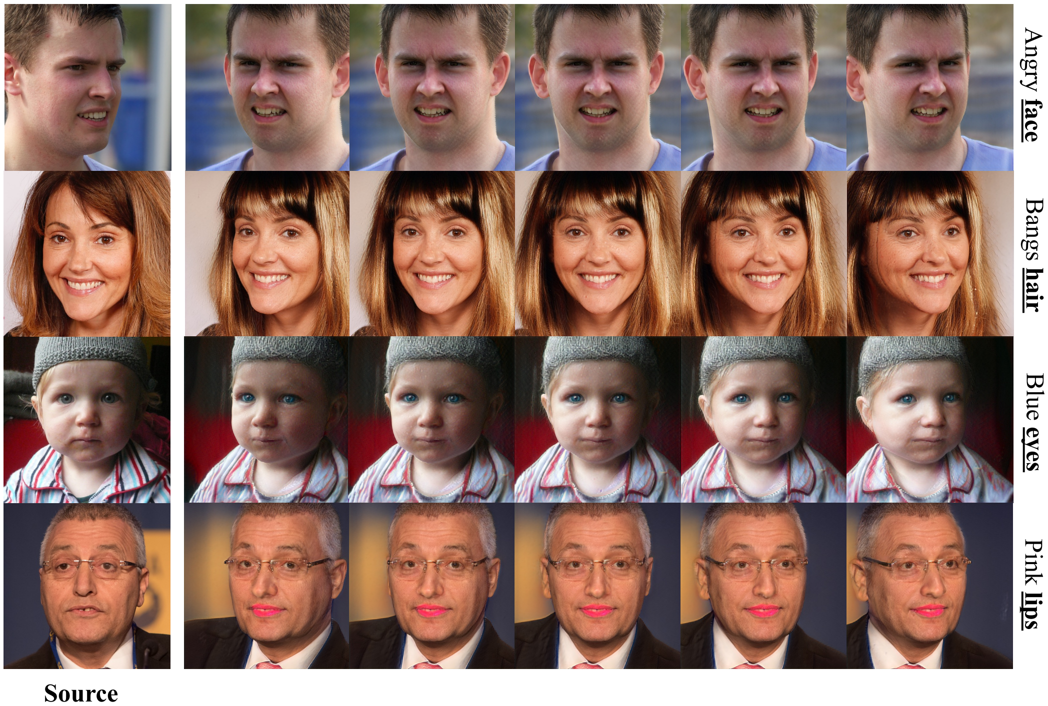


Figure 1. Results of single-view 2D image editing. The first column is the source image, the text on the right is t_{edit} (e.g., *Pink lips*), and the bold underlined word refers to t_{mask} (e.g., *lips*).

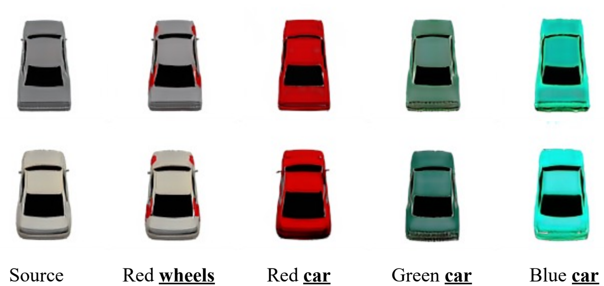


Figure 2. Results of ShapeNet Cars. The text written below is t_{edit} (e.g., *Red wheels*), and the bold underlined word refers to t_{mask} (e.g., *wheels*).

(4) of the main paper. That is, we can control λ of

$$\mathbf{w}_i = (\mathbf{w}_s^1 + \lambda\Delta\mathbf{w}^1, ..., \mathbf{w}_s^N + \lambda\Delta\mathbf{w}^N). \tag{2}$$

We demonstrate the results in Fig. 3.

Additional results Fig. 4 shows additional results of our method. We visualize the original images, manipulated images, predicted relevance mask $\mathbf{M}$, and the volume-rendered attention field $\hat{\mathbf{M}}_t$. Also, please refer to the videos for editing quality and multi-view consistency.

3. Limitation and future work

Our method depends on the capability of the pretrained 3D generator and the CLIP model. Therefore it struggles to generate content outside of the generator's latent space and results in degenerate solutions. CLIP is an encoder-only model that does not have an optimal embedding space for generation capabilities. Instead, we might seek to utilize recently proposed 2D text-to-image diffusion models [10, 11] which can provide stronger priors for manipulating 3D models. Also, our method cannot control the *degree* of manipulation, e.g., how much to open the mouth. Utilizing a deformable 3D generator along with control handles such as 3D Morphable Models (3DMM) [2] is a possible approach to overcome the such limitation.

References

[1] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015. 1

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In Warren N. Waggenspack, editor, *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1999, Los*

CVPR #5992

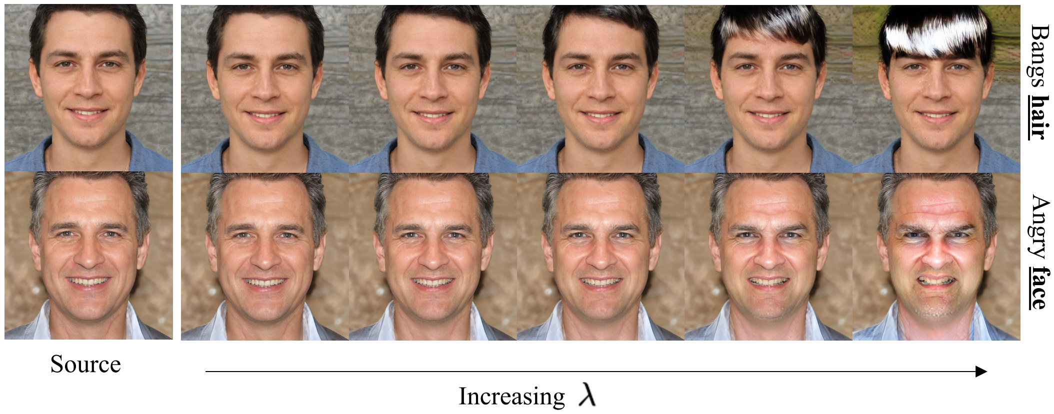


Figure 3. We can change the λ value that is multiplied to the delta latent code $\Delta\mathbf{w}$ estimated by Latent Residual Mapper (LRM) to control the manipulation strengths.

Angeles, CA, USA, August 8-13, 1999, pages 187–194. ACM, 1999. 2

[3] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16102–16112. IEEE, 2022. 1

[4] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 782–791. Computer Vision Foundation / IEEE, 2021. 1

[5] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5939–5948. Computer Vision Foundation / IEEE, 2019. 1

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8185–8194. Computer Vision Foundation / IEEE, 2020. 1

[7] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019. 1

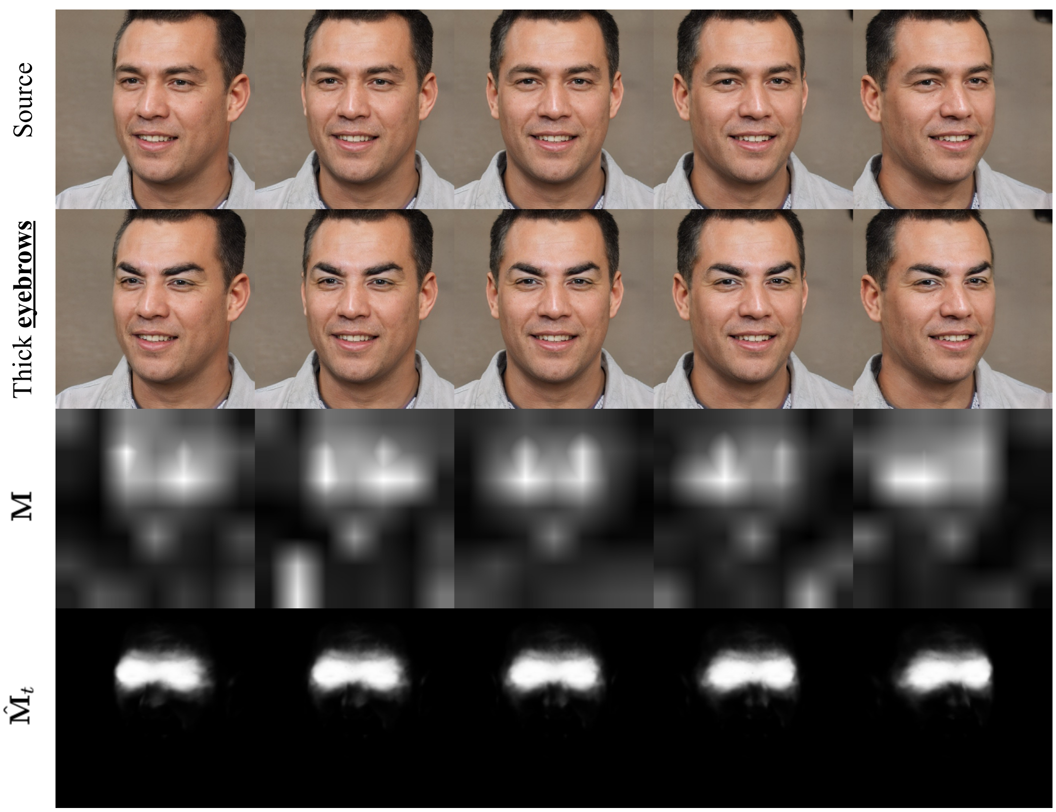
[8] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1

[9] Daniel Roich, Ron Mokady, Amit H. Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *CoRR*, abs/2106.05744, 2021. 1

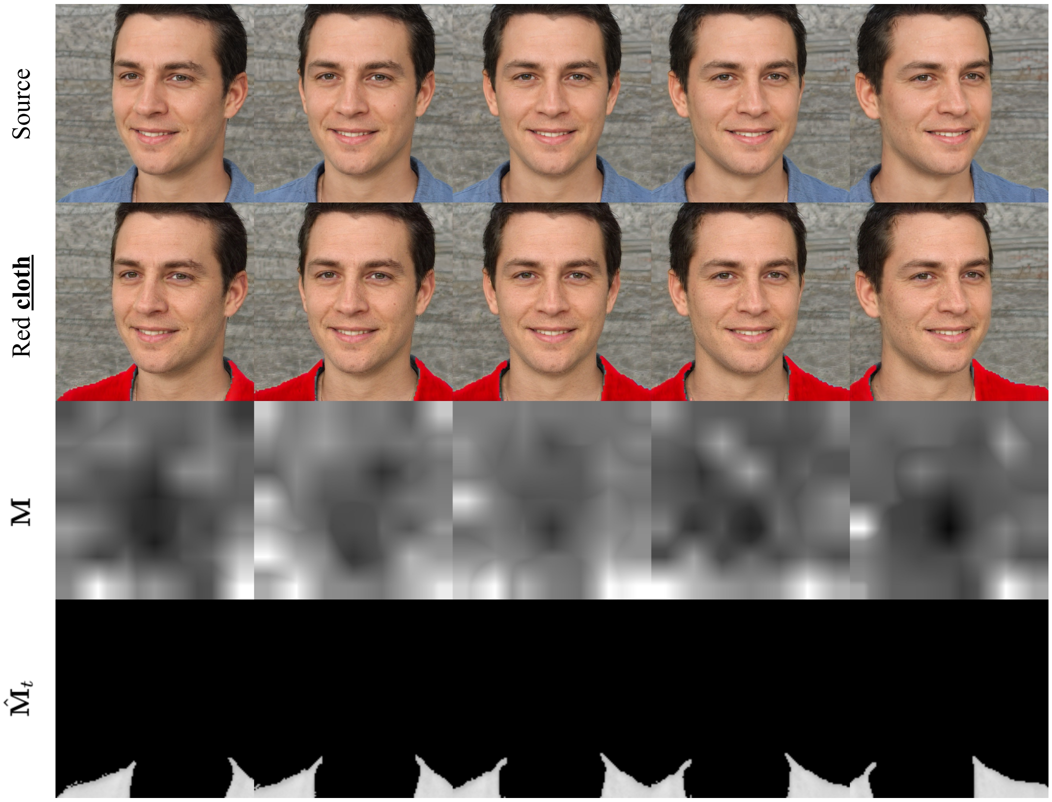
[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2

[11] 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, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *CoRR*, abs/2205.11487, 2022. 2

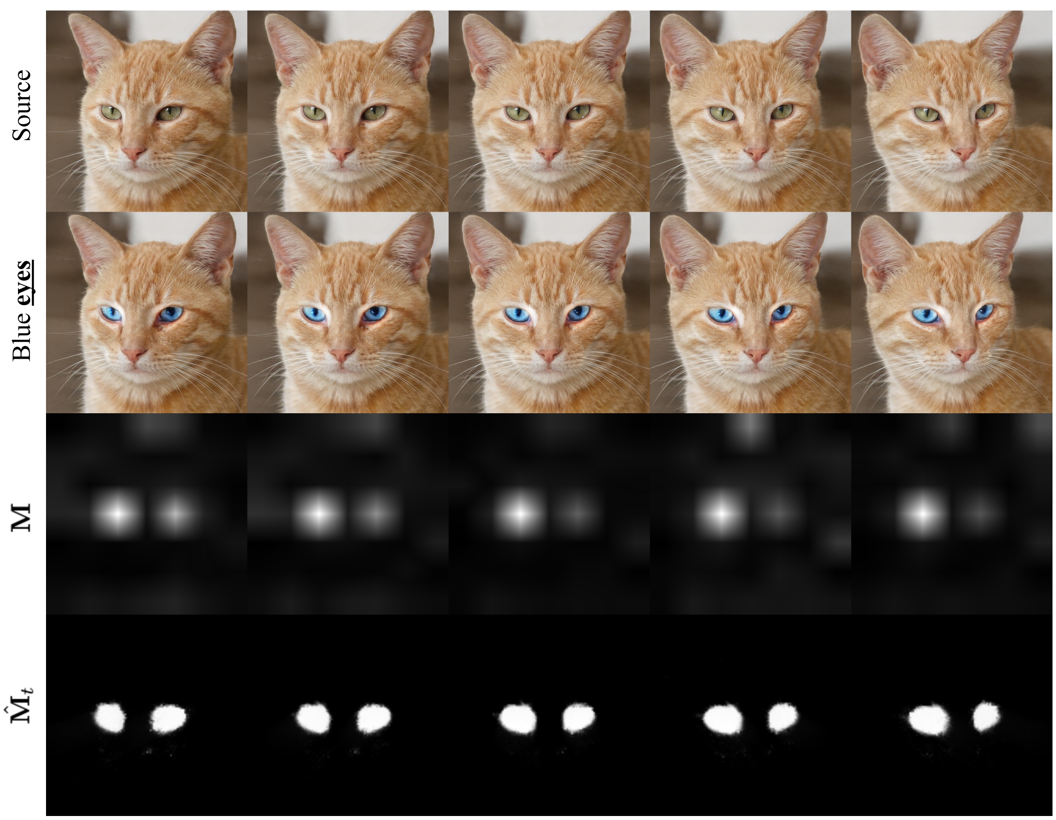
(a) Asian **face** (t_{edit} = Asian face, t_{mask} = face)



(b) Thick **eyebrows** (t_{edit} = Thick eyebrows, t_{mask} = eyebrows)



(c) Red **cloth** (t_{edit} = Red cloth, t_{mask} = cloth)



(d) Blue **eyes** (t_{edit} = Blue eyes, t_{mask} = eyes)

Figure 4. Additional results of LENeRF. We visualize the original images, manipulated images, predicted relevance mask $\mathbf{M}$, and the volume-rendered attention field $\hat{\mathbf{M}}_t$.