Lasagna: Layered Score Distillation for Disentangled Image Editing

Dina Bashkirova[1] **, Rupayan Mallick**[2] **, Arijit Ray**[1] **,**
Ranjay Krishna[3] , Sarah Adel Bargal[2] , Jianming Zhang[4] , Kate Saenko[1,5]
[1]Boston University, [2]Georgetown University, [3]University of Washington, [4]Adobe Research, [5]FAIR

Abstract

Recent text-guided image editing methods achieve great results on various edit types; however, they fail to perform edits that are underrepresented in the training data, such as relighting. Methods that involve finetuning on paired supervised data often fail to preserve the input semantics on out-of-distribution examples, especially if the amount of training data is scarce. In this paper, we propose ***Lasagna***, a method for disentangled image editing that distills the prior of a finetuned diffusion model in a separate visual layer. ***Lasagna*** uses score distillation to learn a plausible edit and preserves the semantics of the input by restricting the layer composition function. We show that ***Lasagna*** achieves superior shading quality compared to the state-of-the-art text-guided editing methods and can be used for other types of layers, *e.g.*, alpha-composition for colorization.

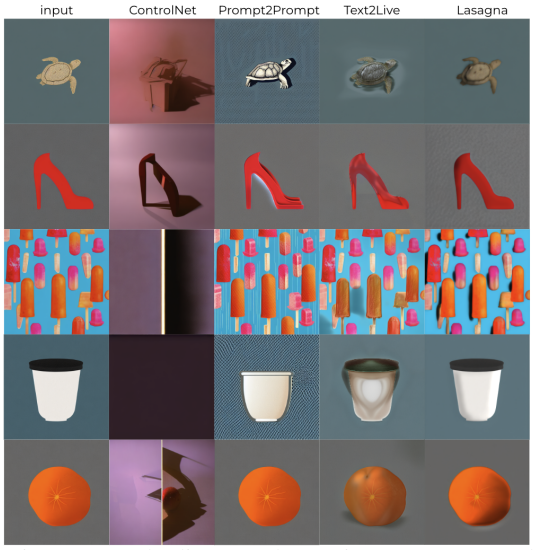

Figure 1: Shading results on images generated with Stable Diffusion v 2.1. ControlNet [13] trained on Rene fails to preserve the semantcs of the OOD examples, Prompt2Prompt achieves limited shading editing while changing the input semantics, Text2Live produces undesirable artifacts. ***Lasagna*** that distills ControlNet prior trained on Rene achieves promising shading capabilities while preserving all other aspects of an input fixed.

Layered Art Representation. State-of-the-art generative models have attained remarkable realism and expressive power, and have the potential to become an indispensable tool for artists of various skill levels. The majority of digital artists use specialized image editing applications, such as Adobe Photoshop [1] and Procreate [6], that allow a layered representation of images as illustrated in Figure 3 in the Appendix. Most artists prefer such layered representation since it disentangles different aspects of an image, *e.g.*, outlines, colorized foreground and background, shading, lighting, textures, *etc.*, which allows more flexible and fine-grained editing. Although recent works on text-guided image editing with diffusion models [8; 2; 5] and generative transformers [3] show impressive editing capabilities, they directly predict the editing result which often leads to changing other aspects of the image that should remain fixed. Moreover, these methods often fail when the edits are not well represented in the training data, such as lighting and shading, due to the lack of grounding prior in the language space. Directly training a model to make image edits is prone to overfitting if the concept is not represented well in the training data (illustrated in Figure 2). In this paper, we pro-

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

pose an image editing method, ***Lasagna***, that allows a disentangled layered composition of edits - which allows a model to apply edits just on a layer without changing any other semantics of the image - greatly improving control and accuracy of edits not well represented in the training data. We test our approach on applications such as colorization, relighting and shading.

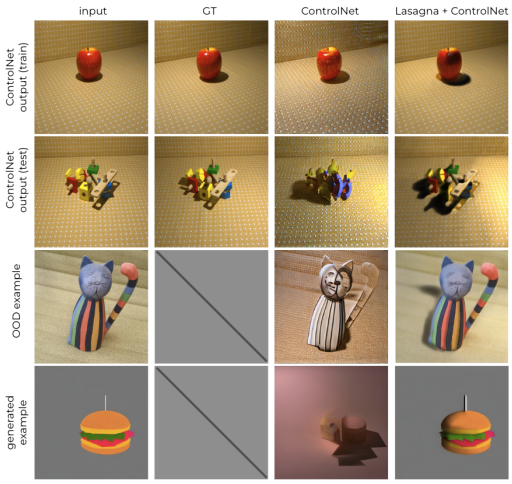

Figure 2: ControlNet [13] adapter trained on ReNe [12] dataset for relighting achieves satisfactory results on the images it was trained on (*first row*), but fails to preserve the content when given test examples (*second row*) or the real (*third row*) or generated (*last row*) examples from a different domain. In contrast, ***Lasagna*** that uses the same ControlNet adapter achieves higher quality shading while preserving the semantics of the input image on a wide range of input examples.

Preliminary: Score Distillation Sampling. Given an input image $x \in \mathcal{X}$, and the embeddings $\tau(c)$ of the conditioning text prompt c, a DDPM [4] model learns to predict the random Gaussian ϵ noise added to an image x (or its latent embedding in case of LDM [10]) at time step t:

$$\text{LDM} := \mathbb{E}_{x\sim\mathcal{X},\epsilon\sim\mathcal{N}(0,1)}||\epsilon - \epsilon_\theta(x+\epsilon, t, \tau(c))||_2^2 \quad (1)$$

DreamFusion [9] introduced Score Distillation Sampling (SDS) to train a NerF [7] MLP representing a 3D object via the 2D prior of a diffusion model. DreamFusion renders a 3D object into a 2D image, and a diffusion model is used to compute the denoising score of a version of that image distorted with random Gaussian noise. The denoising scores implicitly approximate realism of the rendered image, and is backpropagated to update the NerF parameters. More broadly, the DDPM objective can be efficiently used for differential image parametrization, in which a differentiable generator G parametrized by θ_G is trained to generate an image $\hat{x}$ from a learned data distribution of a frozen diffusion model ϵ_θ conditioned on $\tau(c)$. In this approach, the score matching error is backpropagated to the update the parameters θ_G to maximize the "realism" of the generated image.

Method and Comparison. ***Lasagna*** disentangles shading by leveraging the prior of a fine-tuned diffusion model via SDS and a layer composition function that restrict the edit to a certain functionality (e.g. luminosity multiplication for shading). In ***Lasagna***, we learn the editing layer by extracting the image prior via the SDS. Unlike DreamFusion, the trained generator is conditioned on the input image (base layer) and produces a 2D editing layer. Given an input base layer x (e.g. colors, or albedo), we train a generator G parametrized by θ_G that produces an editing layer $G(\theta_G) = y$, and compose the input and predicted layers with a fixed composition function $f(x, y)$ to a diffusion model to compute the score matching error (Eq. 1). This error implicitly optimizes the realism of the input composition of layers conditioned on the editing text prompt and constrained by the input image x and the composition function f. In our shading experiments, we used a luminosity multiplication function: $f(x, y) = xy$, where $x \in \mathbf{R}^{H\times W\times 3}$ is the base (input image or albedo) layer, and $y \in \mathbf{R}^{H\times W\times 1}$ is the luminosity multiplier layer, and a convolutional UNet [11] generator. While some edit types, such as colorization, can be achieved by directly using the off-the-shelf trained diffusion model (Figure 5 in Appendix), it does not extend to the edits for which language guidance is limited, such as shading (see 4 in the Appendix). To introduce the language-conditional prior necessary for shading, we finetune a ControlNet adaptor [13] on a real object relighting dataset ReNe [12]. Given an input image x, the lighting parameter r and the corresponding ground truth relighted image y, ControlNet adaptor minimizes the denoising DDPM error: $\text{DDPM}_c(x, y, \theta(c), \theta_c) := ||\epsilon - \epsilon_{\theta_c}(y+\epsilon, x, t, \tau(c))||_2^2$, where the conditioning prompt $c =$ *"A photo of an object with* $[r]$ *lighting"*. Figure 2 illustrates that while ControlNet learns to perform relighting, it fails to preserve the semantics of an input image (e.g. shapes and colors of the objects) when given examples outside the learned distribution. Figure 1 shows that the proposed disentangled editing approach, unlike ControlNet and text-conditional image editing baselines, achieves promising shading results while preserving content of the input base layer.

References

[1] Adobe. Photoshop. https://www.adobe.com/products/photoshop.html.

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[3] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[5] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.

[6] Savage Interactive Pty Ltd. Procreate App. https://procreate.com/.

[7] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[8] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

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

[10] 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.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[12] Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, and Samuele Salti. Relight my nerf: A dataset for novel view synthesis and relighting of real world objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20762–20772, 2023.

[13] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

1 Appendix

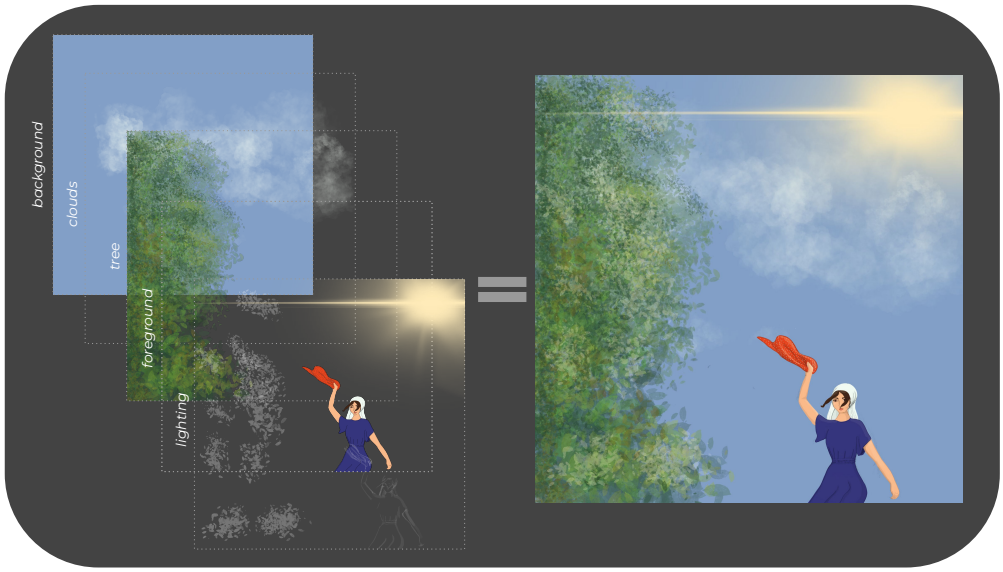

Figure 3: The most popular representation of digital art used in common image editing tools is the layered representation *(left)*. Different aspects of an image are separated into layers that are composed of predefined functions (*e.g.*, shading is composed via multiplicative luminosity adjustment, *etc.*).

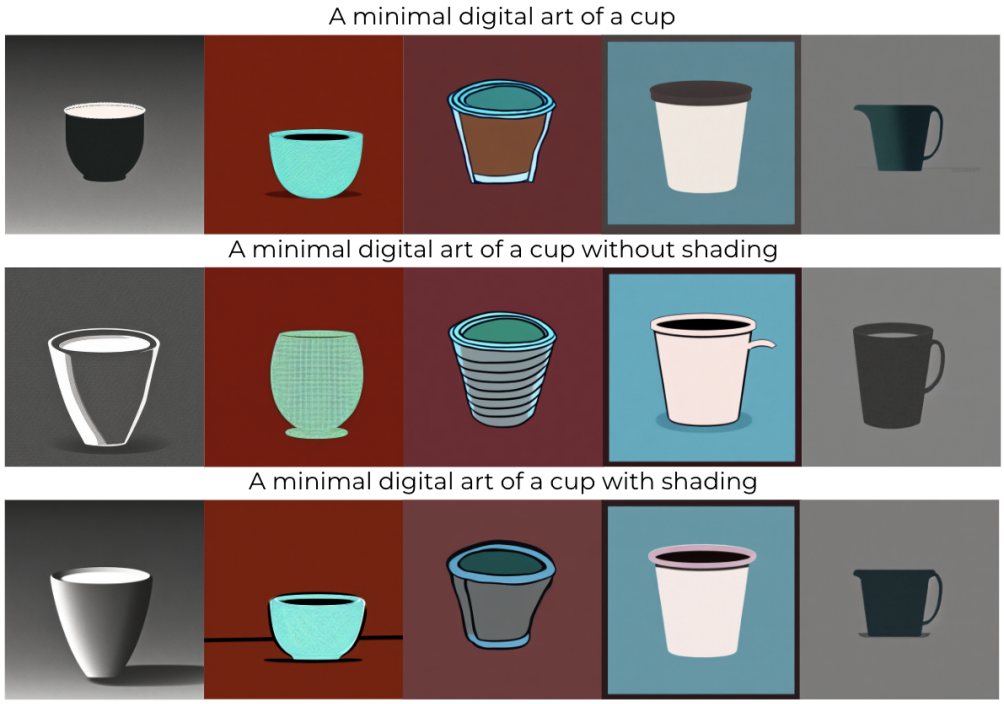

Figure 4: **Top:** Stable Diffusion, given the same random seed, produces images with shading inconsistent with the given text prompt due to the limitation in the language conditioning model prior and grounding.

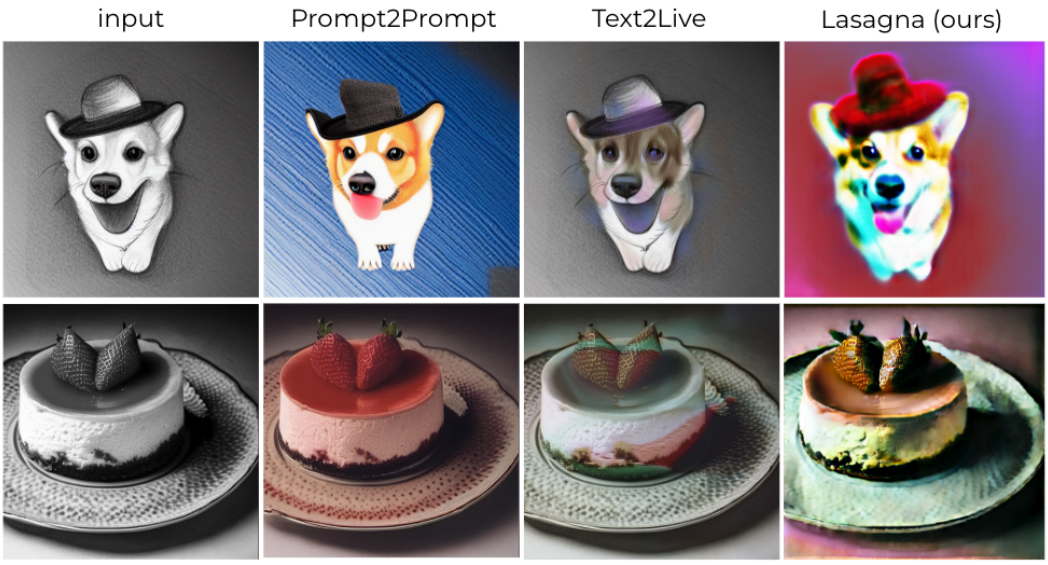

Figure 5: **Top:** Sketch-to-color translation results with Prompt2Prompt, Text2Live and ***Lasagna***on a sketch generated by StableDiffusion v2.1 with the prompt "A pencil sketch of a corgi wearing a hat", and with an editing prompt "A colored digital art of a corgi wearing a hat". Here, we used alpha weighted linear combination composition function for ***Lasagna***. **Bottom:** Colorization results on an image generated by StableDiffusion v2.1 with the prompt "A professional greyscale photo of a strawberry cheesecake" with an editing prompt "A professional colored photo of a strawberry cheesecake".