Latents2Semantics: Leveraging the Latent Space of Generative Models for Localized Style Manipulation of Face Images

Snehal Singh Tomar, A.N. Rajagopalan

Indian Institute of Technology Madras
snehal@smail.iitm.ac.in, raju@ee.iitm.ac.in

Abstract

With the metaverse slowly becoming a reality and given the rapid pace of developments toward the creation of digital humans, the need for a principled style editing pipeline for human faces is bound to increase manifold. We cater to this need by introducing the Latents2Semantics Autoencoder (L2SAE), a Generative Autoencoder model that facilitates highly localized editing of style attributes of several Regions of Interest (ROIs) in face images. The L2SAE learns separate latent representations for encoded images' structure and style information. Thus, allowing for structure-preserving style editing of the chosen ROIs. The encoded structure representation is a multichannel 2D tensor with reduced spatial dimensions, which captures both local and global structure properties. The style representation is a 1D tensor that captures global style attributes. In our framework, we slice the structure representation to build strong and disentangled correspondences with different ROIs. Consequentially, style editing of the chosen ROIs amounts to a simple combination of (a) the ROI-mask generated from the sliced structure representation and (b) the decoded image with global style changes, generated from the manipulated (using Gaussian noise) global style and unchanged structure tensor. Style editing sans additional human supervision is a significant win over SOTA style editing pipelines because most existing works require additional human effort (supervision) post-training for attributing semantic meaning to style edits. We also do away with iterative-optimization-based inversion or determining controllable latent directions post-training, which requires additional computationally expensive operations. We provide qualitative and quantitative results for the same over multiple applications, such as selective style editing and swapping using test images sampled from several datasets.

Introduction

With the rise of online photo-sharing applications, the demand for tools that enable automatic photorealistic edits on face images has seen exponential growth. Motivated by the capability of Generative Adversarial Networks (GANs) to generate high-resolution and photorealistic images and transfer the style attributes from an image to another (demonstrated by works like (Karras et al. 2017; Karras, Laine, and Aila 2018; Karras et al. 2020)),

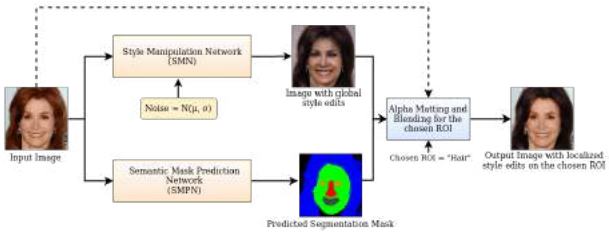

Figure 1: A high level representation of our ROI-specific style editing pipeline and its constituents.

several GAN-based approaches (Zhu et al. 2020; Alaluf, Patashnik, and Cohen-Or 2021a; Nizan and Tal 2020; Wang et al. 2021; d'Apolito et al. 2021; Nie et al. 2020; Shen et al. 2020; Patashnik et al. 2021; Härkönen et al. 2020; Niemeyer and Geiger 2021) have attempted this task. However, there exists a paucity of methods that can efficiently perform highly localized, structure-preserving, and photorealistic style edits on real images which are in accordance with the global style scheme. Very recently, a few methods like (Shi et al. 2022) addressed local style editing of real images, but their framework is dependent on the highly expensive iterative optimization-based image-to-latent inversion process (Alaluf, Patashnik, and Cohen-Or 2021b) and thus inconvenient for real-life applications.

We seek to diminish the aforesaid research gap in this work. The latent space of Generative Autoencoder (GA) models can be designed to encode the *structure* and *style* information present in the input image into separate representations. Our key insight is that achieving a strong correspondence between a part of the structure representation and semantic ROIs in a disentangled fashion is pivotal for solving the problem. To this end, we design a framework where we slice structure representation such that each slice represents (gets interpreted as) all the information required by the decoder to decode each ROI independently. The style representation can then be leveraged to affect global style changes onto the input image, and the sliced structure representation can be leveraged to produce semantic segmentation masks for each ROI. Consequentially, obtaining ROI-specific style edits shall amount to a simple alpha matting task with the

predicted semantic segmentation masks as the matte.

Furthermore, our design is based on Swapping Autoencoder (SAE) (Park et al. 2020) that learns a latent representation with a neat distinction between structure and style information of the input image. In contrast to iterative-optimization-based inversion approaches, our design is significantly faster, as it inverts using a single forward pass and converges to areas of the latent space, which are more suitable for editing. We demonstrate that our simple yet effective autoencoder-based approach achieves satisfactory visual quality while being extremely fast compared to prior art (Shi et al. 2022) for local style editing of face images. The contributions of this paper can be summarized as below:

- We employ the SAE's latent space consisting of the Structure Tensor and Texture (Style) Vector to perform localized, structure-preserving, and photorealistic style edits on real face images, which are all independent of one another and are in accordance with the global style scheme of the input image. Our approach does not require any priors or ground-truth (GT) supervision for the locally altered images, peripheral models, or additional data for editing an input image. To our knowledge, ours is the first work that does so for any Generative Autoencoder model; eliminating the need for expensive inversions. We provide qualitative results and quantitative metrics to support our claims.
- We infuse a strong disentanglement with respect to the structure of semantic ROIs in re-generated images in the latent space (Structure Tensor) of SAE (Park et al. 2020). Our model achieves satisfactory visual quality compared to SOTA while being extremely fast and efficient.
- Localized style editing for a given ROI from an input image amounts to a simple forward pass with the addition of noise to the texture (style) latents for achieving global style edits, followed by a forward pass with appropriate masking (retention of a single non-zero slice corresponding to the ROI) applied to the structure latents for predicting ROI specific segmentation masks. Finally, an alpha matting operation generates the output image. This eliminates the need for any additional human effort (supervision) post-training and underscores the applicability of our model to any generic automatic face editing application.

Method

An overview of our framework is shown in Fig. 1. Our ROI-specific style editing pipeline consists of two primary networks, viz., the Style Manipulation Network (SMN) and the Semantic Mask Prediction Network (SMPN). Given an image, the SMN is tasked with generating photorealistic global style modified images, and the SMPN is responsible for generating accurate segmentation maps. The Alpha matting block performs an alpha matting between the input image and the style modified image using the predicted ROI-specific segmentation mask as the matte. This results in the output image's style attributes being modified only for the chosen ROI. A key advantage of our method is that it does not require GT supervision for the locally-modified images and is prior agnostic. During training, it only needs

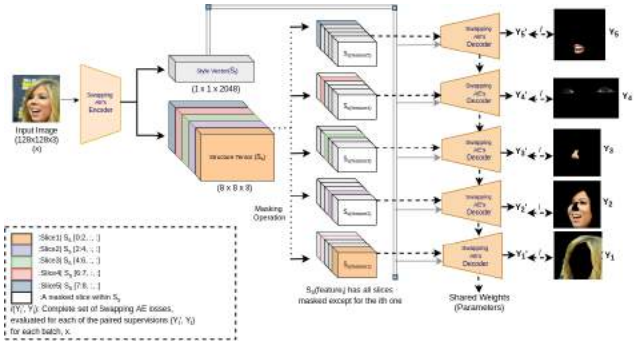

Figure 2: A schematic representation of the SAE (Park et al. 2020) inspired SMN and SMPN architectures and training methodology. The schema used for slicing S_s and for formation of masked $S_{s_{feature_i}}$ have been annotated in the legend. l refers to the operation defined by Eq. 1.

the weak supervision of the GT segmentation maps as guidance. The SMN and the SMPN share the same model architecture, derived from SAE (Park et al. 2020). The SAE (Park et al. 2020) is a generative Autoencoder model which embeds the *structure* and *style* information present in input images ($H \times H \times 3$) into a *structure tensor* (S_s, having dimensions $H/16 \times H/16 \times 8$) and *texture vector* ($S_t$, having dimensions $1 \times 1 \times 2048$). The latent space $S = \{S_s, S_t\}$ serves as the foundation of our work.

Our primary goal is to manipulate the style of a particular region in the image. To achieve that, we first aim to build a correspondence between the latent representation S_s and the structure of individual regions of interest, namely: hair, skin, nose, eyes, and (lips + mouth) in the reconstructed image. In our work, we follow a feature slicing scheme in the latent space and found it to perform reasonably well in practice. For each ROI, we first fix a particular set of feature maps of S_s and mask (setting to zero) all other feature slices. Next, we train the network to produce an image containing only the corresponding semantic ROI when the masked S_s is decoded together with S_t.

We use separate network instances for the SMN and the SMPN. The SMN is a pre-trained SAE (trained on the FFHQ dataset (Karras, Laine, and Aila 2018)). Whereas, the SMPN is a separate SAE instance, fine-tuned on the ROI-separated CelebAMask-HQ dataset comprising 27016 datapoints.

Network Architecture and Losses.

$$\begin{aligned} l(Y_0', Y_0) &= L_{\text{rec}}(Y_0', Y_0) + 0.5L_{\text{GAN, rec}}(Y_0', Y_i) \\ &+0.5L_{\text{GAN, swap}}(Y_0', Y_0) + 0.5L_{\text{CooccurGAN}}(Y_0', Y_0) \end{aligned} \quad (1)$$

The pretrained SAE (Park et al. 2020) is taken such that it id trained on the FFHQ dataset using the same loss functions as Eq. 1. Given an image, it first generates S_s and S_t.

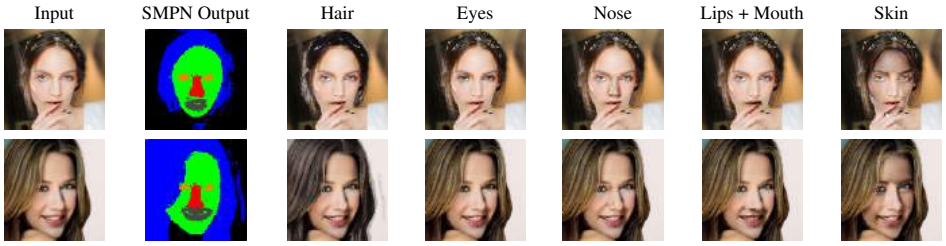

Figure 3: Qualitative results for the prediction of semantic segmentation maps (column 2) by the SMPN and selective style edits (columns 3 through 7) performed on several ROIs. All selective style editing results were obtained by giving noise vectors sampled from $\mathbb{N}(0, 1)$ as input to the SMN. The color coding used for semantic regions in the segmentation maps is given by; blue: *hair*, green: *skin*, red: *nose*, orange (approach 1): *eyes*, orange (approach 2): *lips* + *mouth*, grey (approach 1): *lips* + *mouth*, and grey (approach 2): *eyes* .

Algorithm 1: Inference algorithm

Input: x (input image), $Choice_{ROI}, \mu_{stlye-noise}$
Output: $\hat{x}$ (ROI-selective style edited image)

$SMN = SAE(), SMPN \leftarrow SAE_{sliced_{S_s}}()$
$noise \leftarrow \mu_{stlye-noise} \cdot \mathbb{N}(0, 1),$
$slice\_mask \leftarrow Mask_{Choice_{ROI}}$
$S_{s_1}, S_{t_1} = Encoder_{SMN}(x)$
$NoisyS_{t_1} = S_{t_1} + noise$
$x_{GloballyNoisy} = Decoder_{SMN}(S_{s_1}, NoisyS_{t_1})$
$S_{s_2}, S_{t_2} = Encoder_{SMPN}(x)$
$Sroi_{s_2} = slice\_mask \cdot S_{s_2}$
$ROI_{Mask} = Decoder_{SMPN}(Sroi_{s_2}, S_{t_2}) > 0$
$\hat{x} = AlphaMatting(x, x_{GloballyNoisy}, ROI_{SemanticMask})$
▷ Operation defined by Algorithm 2

Algorithm 2: The alpha matting and blending algorithm

Input: x (input image), m (ROI mask), y (global style edited image)
Output: $\hat{x}$ (ROI-selective style edited image)

$\alpha = m \circledast \sigma(3, 3)$ ▷ $\circledast$ denotes convolution, $\sigma(i, j) = (1/\sqrt{2\pi}) \cdot e^{-(i^2+j^2)/2}; i, j \in [0, 3)$
$\hat{x} = (1 - \alpha) \cdot x + \alpha \cdot y$
$\hat{x} = \hat{x} \circledast \sigma(3, 3)$ ▷ $\circledast$ denotes convolution

Any perturbation in the style tensor S_t results in global style manipulation in the reconstructed image.

The SMPN solely focuses on segmenting different ROIs given the input image. We follow the same architecture and slicing scheme of the structure latent S_s as elucidated in Fig. 2. For every batch of training data, parameters of the encoder and decoder (all Siamese decoders share the same parameters) are optimized using the following overall loss:

$$L_{\text{overall}} = \sum_{i=1}^{5} 0.2 \cdot l(Y_i', Y_i) \tag{2}$$

Overall, the SMN produces an image with a different global style. The ROI masks, predicted by the SMPN, are used to allow style changes only in certain semantic regions using alpha blending.

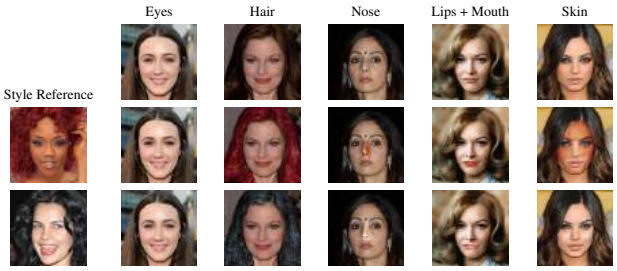

Figure 4: Selective style swapping across several ROIs. Column 1 depicts images used as a style reference. Columns 2 through 6 in the top row show input images (structre reference) for performing style edits. Images in the matrix denote the outputs for selective swapping (with respect to the ROI denoted by the column) of style attributes between the corresponding input and style reference image, respectively.

Alpha Matting and Blending Given the input image (x), the semantic mask for the chosen ROI (m), and the global style edited image (y). The $AlphaMatting()$ operation used in Algorithm 1 to obtain the ROI localized style edited image ($\hat{x}$) corresponds to the series of steps given by Algorithm 2, where $\sigma(3, 3)$ denotes a standard 3×3 Gaussian convolution kernel. The Gaussian blurring performed, ensures smoothing of edges in the matte and combined image, respectively.

Training

Training was initiated using the pre-trained weights provided by (Park et al. 2020), post training on the FFHQ dataset (Karras, Laine, and Aila 2018). The optimizer and training loop used were the same as used by (Park et al. 2020). As show in Figure 3, a batch of training data comprised of $\{X, Y_1, ..., Y_5\}$ where X denotes a batch of input images and Y_i denotes a batch of region specific images, R_i.

The SMPN model was trained on the CelebAMask-HQ dataset (Lee et al. 2020).

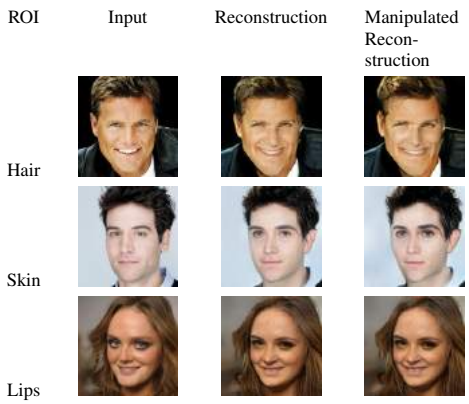

Figure 5: ROI-wise structure editing. Column 2 depicts the reconstructed image obtained sans any modification to encoded latent. Column 3 shows the structure modifications obtained upon reconstruction with noise added to ROI-specific-slices of the encoded latents.

Table 1: Quantitative results for our methods performance versus SemanticStyleGAN (Shi et al. 2022). We evaluate the perceptual similarity of edits obtained (FID, LPIPS) and the time taken to perform one texture edit per ROI.

Method	FID ↓	LPIPS ↓	Time Taken (s) ↓
SemanticStyleGAN	0.3072	**22.3771**	120.602
Approach 2 (Ours)	**0.2026**	26.6084	**0.07**

Inference Mechanism

Inferring segmentation maps from the model amounts to a simple forward pass over the network with appropriate masking applied to S_s in the latent space, depending upon the ROI for which the segmentation has to be performed. Similarly, performing global style edits entails adding noise to the encoded S_t followed by decoding it in tandem with unchanged S_s. The process of performing ROI specific style edits using our model is encapsulated by Algorithm 1.

Experiments

We present a detailed analysis of our model's efficacy towards the claims made in this paper. We provide qualitative and quantitative results for our method's performance with respect to degree of photorealism, ROI-wise localization of edits, and time of computation per edit. The SemanticStyleGAN (Shi et al. 2022) is one of the few recent works that claim to perform highly localized texture editing of real images with minimal human supervision post training. We present a detailed comparative study between our method and the SemanticStyleGAN (Shi et al. 2022). The subsequent subsections elucidate that our method is much faster while being comparably good at performing ROI-wise edits. We do not attempt to estimate the amount of additional human effort (supervision) required by competing methods for the sake of brevity. Figures 4 and 5 illustrate selective style swapping and structure editing which are promising applications of our method.

Dataset and Implementation Details

Our model was trained only on the CelebAMask-HQdataset (Karras et al. 2018) and was evaluated on 862 images sampled from it. The models were trained on 2 NVIDA GeForce RTX 3090 GPUs using a batch size of 4, respectively. We used the optimizer presented by the SAE (Park et al. 2020) to train our models.

L2SAE versus SOTA

Computation Time for obtaining edits Owing to the optimization-based pipelines for inverting real images to latents, most SOTA methods such as (Shi et al. 2022) and (Wu, Lischinski, and Shechtman 2021) require heavy computations for projecting images onto their latent space. These methods lack neat disentanglement with respect to semantic ROIs in their latent manifold. Thus, making editing via inverted latents time consuming. Most of this time is lost in additional human supervision required for attributing meaning to controllable directions and inferring from latent space classification models. We ignore the effects of additional human supervision in our study. Given its disentangled latent space, our method is much faster than the SOTA in producing ROI-specific edits. Table 1 shows that our approach is faster than SOTA by multiple orders of magnitude.

Quality of Style Edits Fig. 3 highlights the qualitative results in terms of segmentation maps predicted by the SMPN and the ROI-specific style edits obtained in tandem with the SMN. Fig. 6 (Appendix-A) shows ROI-wise style edits obtained by our mehtod in contrast with those obtained by the SemanticStyleGAN (Shi et al. 2022). It is evident that the SemanticStyleGAN compromises with structure retention from input images in its texture editing pipeline. Since, our method emaerges from the SAE (Park et al. 2020), it does not face any issues with structure retention. Moreover, the edits obtained by the SemanticStyleGAN aren't as localized as ours. Editing a region using SemanticStyleGAN affects multiple other regions as well. We produce more noticeable edits as well.

Conclusion

In conclusion, this work presents a framework for performing structure-preserving, localized, and photorealistic style edits on face images, which agree with the global style scheme of the input image. The presented method does not require any additional human supervision post training and also does away with the need for a computationally expensive iterative-optimization-based latent-inversion process. Performing localized style edits in the presence of occlusions over ROIs is a challenging test scenario for our method. Our method may be used for interesting applications in AR/VR (digital humans) and medicine (dermatology). However, it might find few potentially harmful applications in much the same manner as deepfakes and the likes.

Acknowledgement

Support from Institute of Eminence (IoE) project No. SB22231269EEETWO005001 for Research Centre in Computer Vision is gratefully acknowledged.

References

Alaluf, Y.; Patashnik, O.; and Cohen-Or, D. 2021a. Only a Matter of Style: Age Transformation Using a Style-Based Regression Model. *ACM Trans. Graph.*, 40(4).

Alaluf, Y.; Patashnik, O.; and Cohen-Or, D. 2021b. ReStyle: A Residual-Based StyleGAN Encoder via Iterative Refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

d'Apolito, S.; Paudel, D. P.; Huang, Z.; Romero, A.; and Van Gool, L. 2021. GANmut: Learning Interpretable Conditional Space for Gamut of Emotions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 568–577.

Härkönen, E.; Hertzmann, A.; Lehtinen, J.; and Paris, S. 2020. GANSpace: Discovering Interpretable GAN Controls. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 9841–9850. Curran Associates, Inc.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *CoRR*, abs/1710.10196.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive growing of GANs for improved quality, stability, and variation. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 1–26.

Karras, T.; Laine, S.; and Aila, T. 2018. A Style-Based Generator Architecture for Generative Adversarial Networks. *CoRR*, abs/1812.04948.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2020. MaskGAN: Towards Diverse and Interactive Facial Image Manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nie, W.; Karras, T.; Garg, A.; Debnath, S.; Patney, A.; Patel, A.; and Anandkumar, A. 2020. Semi-Supervised StyleGAN for Disentanglement Learning. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7360–7369. PMLR.

Niemeyer, M.; and Geiger, A. 2021. GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Nizan, O.; and Tal, A. 2020. Breaking the Cycle - Colleagues Are All You Need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Park, T.; Zhu, J.-Y.; Wang, O.; Lu, J.; Shechtman, E.; Efros, A. A.; and Zhang, R. 2020. Swapping Autoencoder for Deep Image Manipulation. In *Advances in Neural Information Processing Systems*.

Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; and Lischinski, D. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2085–2094.

Shen, Y.; Yang, C.; Tang, X.; and Zhou, B. 2020. InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs. *TPAMI*.

Shi, Y.; Yang, X.; Wan, Y.; and Shen, X. 2022. SemanticStyleGAN: Learning Compositional Generative Priors for Controllable Image Synthesis and Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11254–11264.

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2021. Cross-Domain and Disentangled Face Manipulation with 3D Guidance. *arXiv preprint arXiv:2104.11228*.

Wu, Z.; Lischinski, D.; and Shechtman, E. 2021. StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12863–12872.

Zhu, P.; Abdal, R.; Qin, Y.; and Wonka, P. 2020. SEAN: Image Synthesis With Semantic Region-Adaptive Normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Appendix

A. Qualitative Comparisons

Figure 6: A comparative analysis of our qualitative results for Localized Style Editing (Texture Manipulation), with respect to SemanticStyleGAN (Shi et al. 2022). Our method is better at preserving the input image structure. Moreover, it performs more localized and pronounced edits.