StyleYourSmile: Diffusion-Driven One Shot Cross-Domain Retargeting for Portraits

Avirup Dey and Vinay Namboodiri

University of Bath, Bath, United Kingdom

Figure 1: ***StyleYourSmile*** *can preserve fine-grained identity features as well as domain-specific attributes while retargeting facial expressions. Our model achieves disentanglement between identity and domain style* ***without*** *using any curated multi-style pairs.*

Abstract

Cross-domain portrait retargeting requires disentangled control over identity, expressions, and domain-specific stylistic attributes. Existing methods, typically trained on subjects in a single domain, either fail to generalize across image styles, need test-time optimizations, or require fine-tuning with curated multi-style data to achieve domain-invariant identity representations. In this work, we introduce StyleYourSmile, a novel one-shot cross-domain face retargeting method that eliminates these bottlenecks. We propose a dual-encoder architecture alongside an efficient data augmentation strategy for representing domain-invariant identity cues and capturing domain-specific stylistic variations. Leveraging these disentangled control signals, we condition a diffusion model to retarget facial expressions across domains. Extensive experiments demonstrate that StyleYourSmile achieves superior identity preservation and retargeting fidelity across a wide range of visual styles.

CCS Concepts

• ***Computing methodologies*** → ***Non-photorealistic rendering; Image representations;***

1. Introduction

Editing portraits across different visual domains, such as transferring expressions from a photograph to a pencil sketch, requires disentangling identity from transient attributes (expression, pose, lighting) and from domain-specific style cues. While modern portrait retargeting techniques achieve good identity preservation, they often assume a single real-image domain and break down when confronted with style shifts. This is a critical limitation, as domain-robust editing is essential for applications in digital art, animation pipelines, and virtual avatars.

Parametric face models and GAN-based reenactment methods [BTA*23, GGU*20] partially address disentanglement by mapping faces to structured spaces. However, trained primarily on real images, these models struggle to retain fine details in stylized domains, producing smoothed or inconsistent results. Diffusion-based approaches [DZX*23, PLM*24] inherit the same issue due to limited training data and their reliance on identity embeddings that neglect non-identity cues. Furthermore, the lack of curated multi-style datasets makes training extremely challenging.

Our key insight is that identity and style must be represented through complementary encoders, rather than compressed into a single discriminative space. We propose *StyleYourSmile*, a

DOI: 10.2312/egs.20261010

diffusion-based framework for one-shot cross-domain portrait retargeting. The system combines (i) a face-recognition encoder that isolates domain-invariant identity features, (ii) a style encoder capturing domain-specific cues, and (iii) 3DMM-based spatial conditioning module.To overcome the data bottleneck, we introduce a lightweight style-augmentation module that overlays diverse artistic styles onto standard video datasets (e.g., VoxCeleb) without requiring additional stylized footage. Despite its simplicity, StyleYourSmile achieves state-of-the-art cross-domain retargeting accuracy and preserves both identity and stylistic fidelity across varied artistic domains.

2. Method

We formalize cross-domain face retargeting as follows: Given a source portrait X with descriptors $(x_{id}, x_{pose}, x_{sty})$ and target portrait Y with descriptors $(y_{id}, y_{pose}, y_{sty})$, we construct a mapping $\mathcal{F} : (X,Y) \rightarrow X'$ such that $X' \approx (x_{id}, y_{pose}, x_{sty})$ i.e., it retains the identity and domain-specific visual style of the source while faithfully transferring the pose/expression encoded in the target. A key requirement of our formulation is that the model must learn orthogonal representations for identity and style: the identity encoder should be invariant to stylistic domain changes, while the style encoder must capture domain-specific structure without entangling identity cues. We address this challenge on two fronts. First, to ensure clean factorization of attributes, we implicitly disentangle x_{id} and x_{sty} using two complementary encoders (Section 2.1). Second, to overcome the data bottleneck arising from the lack of multi-style face datasets, we introduce a lightweight style-augmentation pipeline that synthesizes diverse stylized variants of standard video datasets (Section 2.2).

2.1. Architecture

Our key insight is that cross-domain retargeting requires disentanglement between source style and source identity and hence, we use two input channels:

Encoding Identity: Identity embeddings from face recognition models (e.g., ArcFace) offer a compact and discriminative representation, but integrating them into existing diffusion models poses a challenge. Arc2Face [PLM*24] addresses this by retraining the CLIP text encoder to interpret ID embeddings wrapped in a pseudo-token, effectively enforcing identity fidelity. However, this approach limits the influence of other control signals such as pose or style, leading to rigid retargeting behaviour and poor style transfer, as seen in our experiments. To balance identity fidelity with controllability, we propose a shallow transformer-based module $\mathcal{P}_{id}$ that maps identity embeddings $f_{id} \in \mathbf{R}^{512}$ into the CLIP [RKH*21] text space. This allows the model to preserve identity while remaining responsive to additional conditioning signals.

Encoding Style: While identity embeddings ensure discriminative control, they lack the rich perceptual detail required for realistic appearance transfer. CLIP [RKH*21] vision tokens, as observed in prior works like IP-Adapter [YZL*23], emphasize perceptual likeness- capturing cues such as lighting, hair, and local texture. To leverage this, we extract visual tokens from the source image using the CLIP encoder $\mathcal{E}_{sty}$, leveraging both patch-level and global

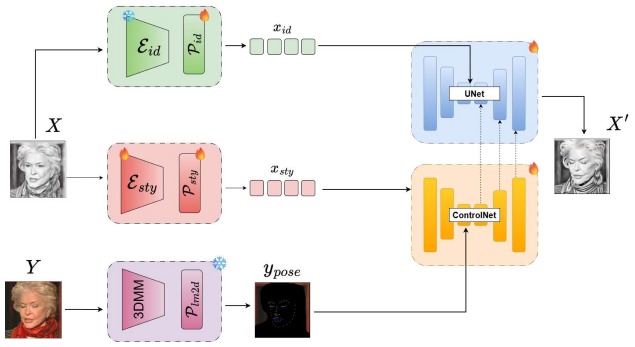

Figure 2: ***Model Overview****: First, the source image X is encoded as follows - (i) a face recognition encoder $\mathcal{E}_{id}$ extracts domain invariant identity features and they are projected into CLIP text space by a decoder $\mathcal{P}_{id}$ as identity tokens x_{id}. (ii) A style encoder (CLIP-V) $\mathcal{E}_{sty}$ extracts domain specific style features and they are projected into CLIP [RKH*21] text space by a decoder $\mathcal{P}_{sty}$ as style tokens x_{sty}. Simultaneously, a spatial conditioning image y_{pose} is which is a composite of 3DMM landmarks and foreground masks, is computed from the target image Y. Then, the denoising UNet [RFB15], containing trainable low rank matrices, is optimized to disentangle identity and domain style, conditioned with x_{id} and a ControlNet [ZRA23] signal which combines y_{pose} and x_{sty}.*

tokens. They are projected into the text space via a shallow transformer decoder $\mathcal{P}_{sty}$ with learnable queries, mirroring our identity encoder. The resulting tokens x_{sty} complement the identity tokens x_{id}, enabling more faithful preservation of domain-specific details during retargeting.

Training: As shown in Figure 2, the target 3DMM-based landmarks, estimated with D3DFR [DYX*19] is fed to a ControlNet [ZRA23] along with the style tokens. The aggregated control signal acts as the conditioning for the diffusion UNet [RFB15]. We train our model for 40,000 epochs with a constant learning rate of 1e-4 and a batch size of 4 on a 4x NVIDIA A5000 GPU setup.

2.2. Data Augmentation

We incorporate stylistic diversity without resorting to curated datasets of animations, artistic portraits, etc. Instead we use a novel attention-based style transfer method [CHH24] that augments real video frames, from VoxCeleb-1 with random styles. The key observation is that *queries* from content image can be matched with corresponding *keys* in the style image when they share semantic similarities, thus maintaining spatial coherence. For instance, when transferring 'sktech' style onto a portrait the strokes will occur only at the edges or high contrast areas. While DDIM inversion, the (Q,K,V) of both images are cached and while denoising the (K,V) pairs from the style image are coupled with the queries Q of output at every timestep. To enforce structural consistency of the output, its queries are linearly combined with the content queries. Formally, the whole mechanism is described as:

$$\widetilde{Q}_t^{out} = \gamma \times Q_t^c + (1-\gamma) \times Q_t^{out}, \quad (1)$$

$$\phi^{out} = \mathrm{Attn}\left(\widetilde{Q}_t^{out}, K_t^s, V_t^s\right) \quad (2)$$

Figure 3: *Visual comparison of various models on stylized VoxCeleb1 [NCXZ20] test set. Our model outperforms previous models in terms of identity retention and style preservation.*

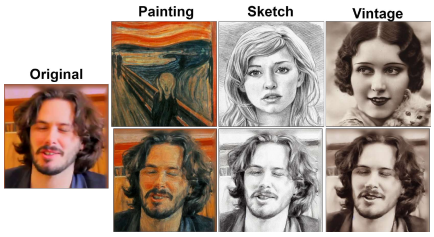

Figure 4: *We augment the training data with different styles, with varying degrees of abstraction. Training on such data incentivize the model to decouple identity from style.*

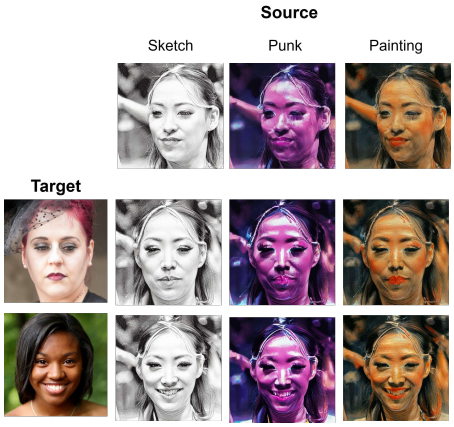

Figure 5: *Qualitative Results on FFHQ [KLA19] dataset. Our model can generalize across diverse domain styles.*

We visualize sample augmentations in Figure 4.

3. Results

3.1. Datasets and Metrics

For evaluation, we choose 20 subjects from VoxCeleb1 test split, each with 3 video sequences. We extract 10 frames from each, giving us a total of 600 frames. For augmentation we choose from a set of 5 domain styles, with varying degrees of abstraction, some of which are shown in Fig. 4. This gives us a total of 3000 augmented frames which are used for quantitative analysis.

We use cosine similarity between ID embeddings (dubbed as **CS-ID**) to measure identity retention, as previously done in [BTA*23, HZH*24]. For expression and pose, we measure **motion transfer error**, given by the Euclidean distances between the expression and pose coefficients of the generated and driving images. To measure stylistic similarity between the generated and ground truth images, we use **ArtFID** metric which evaluates both content and style preservation and strongly coincides with human judgement. We also use **PSNR** and **LPIPS** to measure reconstruction quality in self-retargeting setting.

3.2. Qualitative Results

Fig. 3 shows a visual comparison among various models for cross-domain face retargeting on the test set. We observe that ***StyleYourSmile*** outperforms existing ones in self and cross-identity settings. DiffusionRig [DZX*23] requires a personal album for inference-time fine-tuning. As our task is defined in a single image-to-image setting, it falls short of it's full potential. However, it is able to capture the high-level details of the source image like face orientation and overall colour tone. HyperReenact [BTA*23] shows good retargeting fidelity but is unable to capture domain-specific details and fine-grained facial features accurately. ROME [KSLZ22] shows competitive retargeting performance when the foreground is accurately segmented. However, this cannot be guaranteed in in-the-wild or stylized images. It also fails to capture texture details correctly. We also show our model's performance on in-the-wild faces and styles in Figure 5.

Methods	Self Retargeting						Cross ID Retargeting		
	PSNR↑	LPIPS↓	CS-ID↑	Exp↓	Pose↓	ArtFID↓	CS-ID↑	Exp↓	Pose↓
HyperReenact [BTA*23]	12.225	0.377	0.410	0.368	7.334	35.536	0.270	0.387	**6.344**
ROME [KSLZ22]	10.037	0.511	0.189	0.491	8.486	38.002	0.091	0.420	8.375
DiffusionRig [DZX*23]	13.650	0.402	0.324	0.273	7.034	35.392	0.221	0.414	9.111
Ours	**19.889**	**0.146**	**0.615**	**0.241**	**6.321**	**32.377**	**0.553**	**0.333**	8.072

Table 1: *Quantitative evaluations among various methods for cross-domain face retargeting on Voxceleb1 test set. Best figures are given in* ***bold*** *and second-best figures are underlined.*

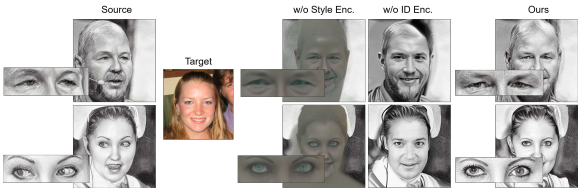

Figure 6: *Ablations: Effect of removing identity and style encoders. We use the sketch domain in this example.*

3.3. Quantitative Results

We report the performance of all models on different metrics in Table 1. In self retargeting setting, our model achieves highest reconstruction quality and style preservation. In cross-identity retargeting we achieve better identity retention and expression fidelity than the baselines.

3.4. Ablations

We disable the identity and style encoders to evaluate their effect on the output. In Figure 6, we note that without the style encoder the output loses the textural details of the sketch domain but captures the core identity markers (observe the inset image). Disabling the identity encoder causes identity drift but the model is able to represent the source style well. We also evaluate the effect of removing data augmentation and report the results in the Supplementary material.

3.5. Conclusions

In this work, we introduce ***StyleYourSmile***, a diffusion-based framework to solve cross-domain retargeting for portraits. We hypothesize that existing methods have entangled representations for identity and style since they are trained on a singular domain. To address this, we train two complementary encoders to learn disentangled representations for identity and domain style and we adopt a light training-free image stylization method that augments images with abstract styles from multiple domains. Through our proposed method, we avoid data curation and are able to train our model on a small consumer grade work station. Through extensive evaluations, we demonstrate that our model outperforms existing ones in terms of identity retention and style preservation.

References

[BTA*23] BOUNARELI S., TZELEPIS C., ARGYRIOU V., PATRAS I., TZIMIROPOULOS G.: Hyperreenact: One-shot reenactment via jointly learning to refine and retarget faces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (2023). 1, 3, 4

[CHH24] CHUNG J., HYUN S., HEO J.-P.: Style injection in diffusion: A training-free approach for adapting large-scale diffusion models for style transfer. In *Proceedings of the IEEE/CVF CVPR* (2024), pp. 8795–8805. 2

[DYX*19] DENG Y., YANG J., XU S., CHEN D., JIA Y., TONG X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF CVPR workshops* (2019), pp. 0–0. 2

[DZX*23] DING Z., ZHANG X., XIA Z., JEBE L., TU Z., ZHANG X.: Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF CVPR* (2023), pp. 12736–12746. 1, 3, 4

[GGU*20] GHOSH P., GUPTA P. S., UZIEL R., RANJAN A., BLACK M. J., BOLKART T.: GIF: Generative interpretable faces. In *International Conference on 3D Vision (3DV)* (2020), pp. 868–878. URL: http://gif.is.tue.mpg.de/. 1

[HZH*24] HAN Y., ZHU J., HE K., CHEN X., GE Y., LI W., LI X., ZHANG J., WANG C., LIU Y.: Face-adapter for pre-trained diffusion models with fine-grained id and attribute control. In *European Conference on Computer Vision* (2024), Springer, pp. 20–36. 3

[KLA19] KARRAS T., LAINE S., AILA T.: A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF CVPR* (2019), pp. 4401–4410. 3

[KSLZ22] KHAKHULIN T., SKLYAROVA V., LEMPITSKY V., ZAKHAROV E.: Realistic one-shot mesh-based head avatars. In *European Conference on Computer Vision* (2022), Springer, pp. 345–362. 3, 4

[NCXZ20] NAGRANI A., CHUNG J. S., XIE W., ZISSERMAN A.: Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language 60* (2020), 101027. 3

[PLM*24] PAPANTONIOU F. P., LATTAS A., MOSCHOGLOU S., DENG J., KAINZ B., ZAFEIRIOU S.: Arc2face: A foundation model for id-consistent human faces. In *European Conference on Computer Vision* (2024), Springer, pp. 241–261. 1, 2

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (2015), Springer, pp. 234–241. 2

[RKH*21] RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., ET AL.: Learning transferable visual models from natural language supervision. In *International conference on machine learning* (2021), PmLR, pp. 8748–8763. 2

[YZL*23] YE H., ZHANG J., LIU S., HAN X., YANG W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023). 2

[ZRA23] ZHANG L., RAO A., AGRAWALA M.: Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 3836–3847. 2