SMILE: Semantically-guided Multi-attribute Image and Layout Editing

Andrés Romero, Luc Van Gool, Radu Timofte

Computer Vision Lab, ETH, Zürich, Switzerland.
{roandres, vangool, radu.timofte}@vision.ee.ethz.ch

AIM 2021 | 2021 ICCV VIRTUAL OCTOBER 11-17

Readme

TL;DR: Our main purpose is to manipulate facial attributes whether exemplar images exist or not. We decoupled this problem into two sub-problems: semantic manipulation of the structure of the face using condition information (*e.g.*, one-hot labels), and semantically guided image synthesis.

Facial attribute manipulation is an active topic in the community. There are several ways to tackle this problem. We identified that current solutions either use one model per attribute (GeneGAN & alike), take both target and reference image as input to the model (MulGAN & alike), or does not include mutually inclusive domains (StarGANv2). Our model can successfully address all these issues using a single model.

Methods	Latent-guided synthesis	Image-guided synthesis	Mutually-inclusive domains	Fine-grained mapping
CycleGAN	✗	✗	✗	✗
StarGAN	✗	✗	✓	✓
MUNIT$_{\text{\& alike}}$	✓	✓	✗	✗
GeneGAN$_{\text{\& alike}}$	✗	✓	✗	✓
MulGAN$_{\text{\& alike}}$	✗	✓	✓	✓
StarGANv2	✓	✓	✗	✗
SMILE (ours)	✓	✓	✓	✓

Table 1: Feature comparison with state-of-the-art approaches in I2I translation. SMILE successfully performs both latent-guide and image-guide attribute transformations for fine-grained or more global mappings in a mutually inclusive domain manner.

Overview

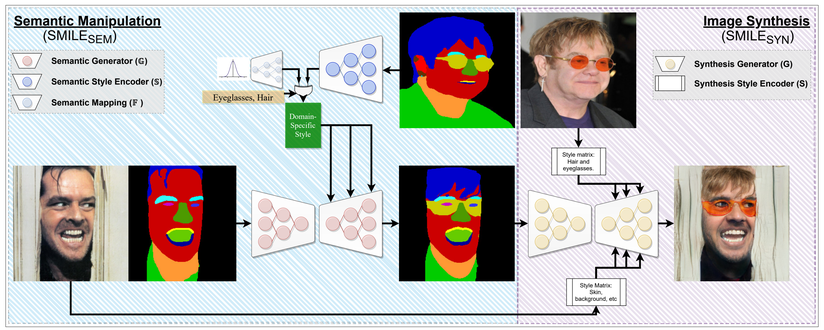

Figure 1: We translate an image by either taking as input a random style or target attributes into the generator (we use a reference image in this example). We first manipulate the shape of the attribute by using a semantic segmentation map (left), and then we synthesize the style of each semantic region by using both input and reference styles to produce a photo-realistic merge of the two images (right). Our proposed approach SMILE is an ensemble SMILE$_{\text{SYN}}$ ∘ SMILE$_{\text{SEM}}$.

Semantic Manipulation

Our selected attributes are: Gender, eyeglasses, baldness, bangs, hat, and earrings. The selection is done purely on how visible the attribute is with respect to others in the semantic segmentation map. Our baseline is StarGANv2 which has proven to be very good at manipulating images in both latent or exemplar synthesis. However, for mutually inclusive domains, it does not perform well. We train StarGAN on semantics and use it as our baseline (case A in Table 2).

	CelebA-HQ | **Latent Synthesis**								
	Pose			Attributes		Reconstruction	Perceptual		
	Roll↓	Pitch↓	Yaw↓	AP↑	F1↑	mIoU↑	FID↓	Diversity↑	
StarGANv2	2.952 ± 0.856	16.900 ± 6.264	29.331 ± 8.134	0.795 ± 0.092	0.797 ± 0.079	0.964 ± 0.012	81.945 ± 24.276	0.018 ± 0.008	
SMILE$_{\text{SEM}}$	2.589 ± 0.684	15.082 ± 4.097	11.286 ± 1.983	0.960 ± 0.031	0.946 ± 0.032	0.989 ± 0.002	43.151 ± 15.527	0.399 ± 0.020	
(*C*): SMILE$_{\text{SEM}}$ −WC	2.683 ± 0.792	18.628 ± 6.243	10.553 ± 2.560	0.965 ± 0.028	0.953 ± 0.027	0.986 ± 0.002	48.123 ± 14.759	0.390 ± 0.013	
(*B*): SMILE$_{\text{SEM}}$ −WC −MC	2.732 ± 0.681	18.172 ± 4.500	17.626 ± 7.250	0.940 ± 0.039	0.928 ± 0.038	0.987 ± 0.003	46.797 ± 14.204	0.395 ± 0.013	
(*A*): SMILE$_{\text{SEM}}$ −WC −MC −DS	2.359 ± 0.678	13.520 ± 4.476	15.424 ± 6.432	0.889 ± 0.062	0.884 ± 0.051	0.994 ± 0.001	61.015 ± 22.235	0.382 ± 0.014	
	CelebA-HQ | **Reference Synthesis**								
	Pose			Attributes		Reconstruction	Perceptual		
	Roll↓	Pitch↓	Yaw↓	AP↑	F1↑	mIoU↑	FID↓	Diversity↑	
StarGANv2	2.472 ± 0.726	14.691 ± 3.987	31.071 ± 15.769	0.811 ± 0.086	0.806 ± 0.077	0.971 ± 0.012	72.910 ± 18.961	0.214 ± 0.051	
SMILE$_{\text{SEM}}$	1.948 ± 0.450	13.225 ± 3.428	9.439 ± 1.826	0.942 ± 0.030	0.928 ± 0.031	0.989 ± 0.002	50.257 ± 24.735	0.129 ± 0.083	
(*C*): SMILE$_{\text{SEM}}$ −WC	2.182 ± 0.652	17.142 ± 6.113	9.117 ± 1.280	0.943 ± 0.031	0.930 ± 0.029	0.986 ± 0.002	52.327 ± 23.352	0.111 ± 0.064	
(*B*): SMILE$_{\text{SEM}}$ −WC −MC	2.277 ± 0.595	16.362 ± 4.304	14.952 ± 5.364	0.919 ± 0.047	0.909 ± 0.043	0.987 ± 0.003	53.298 ± 23.361	0.132 ± 0.057	
(*A*): SMILE$_{\text{SEM}}$ −WC −MC −DS	2.011 ± 0.698	10.811 ± 4.247	13.765 ± 7.567	0.899 ± 0.063	0.887 ± 0.060	0.994 ± 0.001	65.863 ± 26.084	0.136 ± 0.058	

Table 2: Quantitative contribution of each component of our system for Latent Synthesis manipulation (upper part) and Exemplar Image manipulation (lower part). ↓ and ↑ mean that lower is better and higher is better, respectively. Note that Diversity computes the LPIPS perceptual *dissimilarity* across different styles for a single input, therefore higher is better. WC, MC, and DS stand for Weighted Classes, Modulated Convolutions, and Detaching Style, respectively.

Semantic Image Synthesis

In order to StyleGAN to generate images from semantic maps, we replace StyleGAN2 modulated convolutions (ModConv) with our proposed improved semantic region-wise adaptive convolutions (SACs). Let w be the kernel weight, h the input features of the convolution, s the condition information, and σ_E the standard deviation also known as the demodulating factor, we define SACs in Equation 1.

$$\text{ModConv}_w(\mathbf{h}, s) = \frac{w * (s\mathbf{h})}{\sigma_E(w, s)} \Leftrightarrow s \in \mathbb{R}^{1\times C\times 1\times 1}$$
$$\text{SAC}_w(\mathbf{h}, \mathbf{s}) = \frac{w * (\mathbf{s} \odot \mathbf{h})}{\sigma_E(w, \mathbf{s})} \Leftrightarrow \mathbf{s} \in \mathbb{R}^{1\times C\times H\times W}, \quad (1)$$

where,

$$\mathbf{s} = \alpha_w SM + (1 - \alpha_w)M, \quad (2)$$

Inspired by SEAN, we couple an encoder to use a single model to handle both random generation and exemplar generation. Our ablation study includes the encoder and our SACs layers.

Experiment	FFHQ								
	Latent Synthesis			**Reference Synthesis**				Training	# Params
	FID↓	Diversity↑	Runtime [s/img]	LPIPS↓	PSNR↑	SSIM↑	RMSE↓	Runtime [s/img]	[days]	[millions]
StyleGAN2	15.15	-	0.03	0.14 ± 0.02	20.13 ± 1.14	0.66 ± 0.03	0.11 ± 0.02	120	2.5	30.0
SMILE$_{\text{SYN}}$	16.99	0.43 ± 0.03	0.13	0.21 ± 0.06	18.19 ± 2.84	0.54 ± 0.10	0.13 ± 0.03	0.13	17.5	42.2
(*B*): SMILE$_{\text{SYN}}$ −Encoder	13.08	0.42 ± 0.04	0.13	0.18 ± 0.06	17.86 ± 2.97	0.60 ± 0.09	0.13 ± 0.05	210	9.4	39.9
(*A*): SMILE$_{\text{SYN}}$ −Encoder −SAC	24.12	0.08 ± 0.03	0.60	0.42 ± 0.07	10.38 ± 2.07	0.34 ± 0.08	0.31 ± 0.07	180	3.8	36.1
SPADE	-	-	-	0.40 ± 0.02	12.33 ± 0.69	0.40 ± 0.03	0.25 ± 0.02	0.56	1	92.5
SEAN	-	-	-	0.24 ± 0.02	16.68 ± 0.82	0.52 ± 0.03	0.15 ± 0.01	0.28	4	266.9

Table 3: Image synthesis quantitative evaluation under different configurations, and in comparison with recent works.

Application to videos

Figure 2: Although our method is not trained using videos, it performs well on facial reenactment for videos in the will.

Code: https://github.com/affromero/SMILE | Project page: https://affromero.github.io/SMILE.