Supplementary Materiels: Controllable Image Synthesis via SegVAE

Yen-Chi Cheng[1,2], Hsin-Ying Lee[1], Min Sun[2], Ming-Hsuan Yang[1,3]

[1]University of California, Merced [2]National Tsing Hua University [3]Google

A Overview

We provide more qualitative and quantitative results in Section B and Section C. For a quick walkthrough of this paper, please also check the video provided in the zip file.

B Quantitative Results

We measure the "compatibility error" and "reconstruction error" on the ablated models in Table 1. Adopting LSTM in the iterative prediction process helps the model to generate more compatible and reasonable shapes, which greatly improves the performance on the compatibility and the reconstruction. Leveraging the learned shape prior further enhances the results since the distribution varies a lot when the input contains a different combination of classes and each shape of the class has variations. Table 1 shows the necessity of the proposed model's architecture design.

Label-set length analysis. We compute FID on HumanParsing with different length of label-sets to analyze its effect on the proposed method in Table 2. The results show that the larger the length, the larger the FID.

Randomized prediction order. To further analyze how the prediction order will affect the performance, we randomly shuffle the prediction order during training and evaluate the FID in Table 3.

C Qualitative Results

We present more generated results in this section. For multimodal generation, please see Figure 1 and Figure 2. For comparison with baselines and the existing method, please refer to Figure 3. Finally, more editing results are demonstrated in Figure 4.

Failure cases. The proposed method has the following limitations. First, SegVAE fails when the shape vector of a certain class is located in an under-sampled space. For example, *hat* in the CelebAMask-HQ as shown in Figure 5 (a). Second, the label-set combination is unseen. Figure 5 (b) shows the output of the proposed model when all classes are used in a label-set. SegVAE has difficulty to handle this input since HumanParsing do not contain this example.

Table 1: **Compatibility error and reconstruction error.** We train a shape predictor to measure the compatibility error (abbreviated as "Compat. err.") over the generated shapes for all methods. We also train an auto-encoder to measure the quality of our generated results by calculating the reconstruction error (denoted as "Recon. err.").

	HumanParsing		CelebAMask-HQ	
Method	Compat. err. ↓	Recon. err. ↓	Compat. err. ↓	Recon. err. ↓
Ours w/o LSTM	$.7534^{\pm.0249}$	$.6808^{\pm.009}$	$.0987^{\pm.0025}$	$.1144^{\pm.002}$
Ours w/o Learned Prior	$.6926^{\pm.0115}$	$.5856^{\pm.010}$	$.0839^{\pm.0029}$	$.1045^{\pm.003}$
Ours	$\mathbf{.6174^{\pm.0147}}$	$\mathbf{.5663^{\pm.011}}$	$\mathbf{.0754^{\pm.0013}}$	$\mathbf{.0840^{\pm.001}}$

Table 2: **Label-set length analysis**. We compute FID on HumanParsing with different length of label-sets. The results show that the larger the length, the larger the FID.

Length	FID↓
10	39.2692
11	39.5351
12	39.9063
13	41.0591
14	44.2884

Table 3: **Prediction order analysis**.

Order (HumanParsing)	FID↓	Diversity↑
1 *Body* → *Clothes* → *Accessories* (Ours)	$\mathbf{39.6496^{\pm.3543}}$	$\mathbf{.2072^{\pm.053}}$
2 *Clothes* → *Body* → *Accessories*	$39.9008^{\pm.5263}$	$.2062^{\pm.0494}$
3 *Accessories* → *Body* → *Clothes*	$40.2909^{\pm.2195}$	$.2043^{\pm.0521}$
Random order	$52.6728^{\pm.3865}$	$.1714^{\pm0.044}$
Order (CelebAMask-HQ)	**FID↓**	**Diversity↑**
1 *Face* → *Face features* → *Accessories* (Ours)	$\mathbf{28.8221^{\pm.2732}}$	$\mathbf{.1575^{\pm.043}}$
2 *Face features* → *Face* → *Accessories*	$30.6547^{\pm.1267}$	$.1517^{\pm.0376}$
3 *Accessories* → *Face* → *Face features*	$32.0325^{\pm.1294}$	$.1489^{\pm.0363}$
Random order	$31.1238^{\pm.2088}$	$.1529^{\pm0.042}$

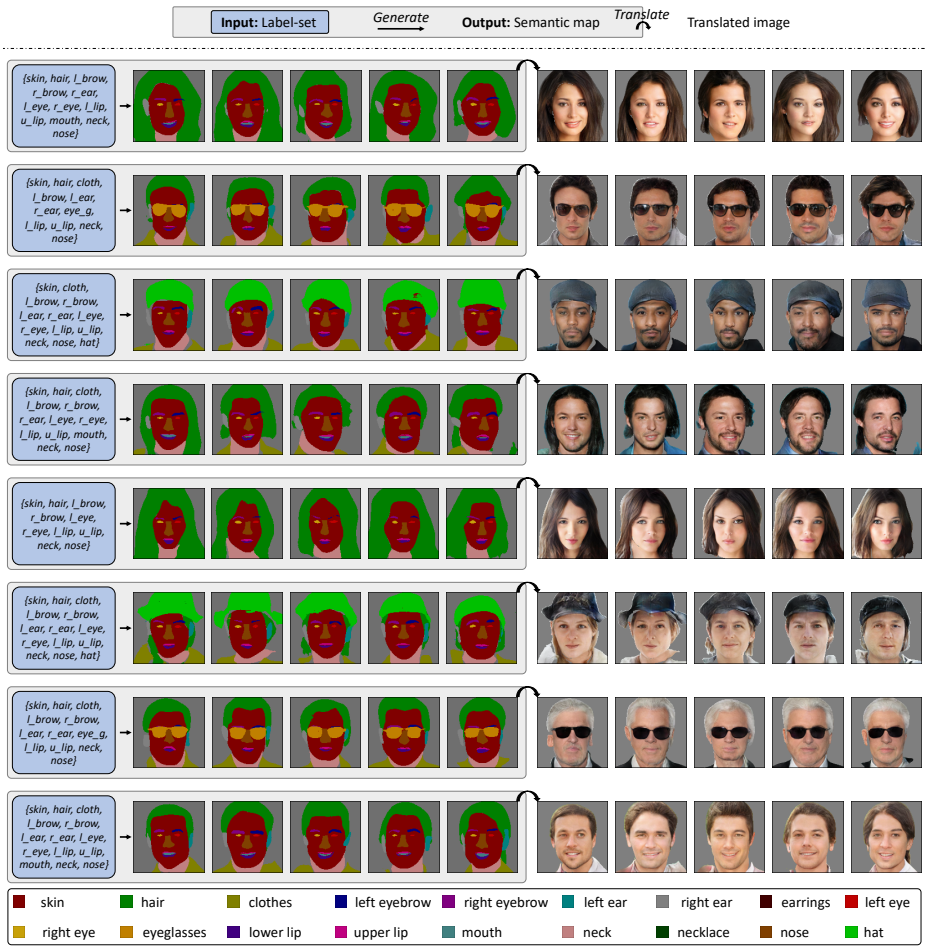

Fig. 1: **Multi-modality.** We demonstrate the ability of the proposed model to generate diverse results given a label-set on the CelebAMask-HQ dataset.

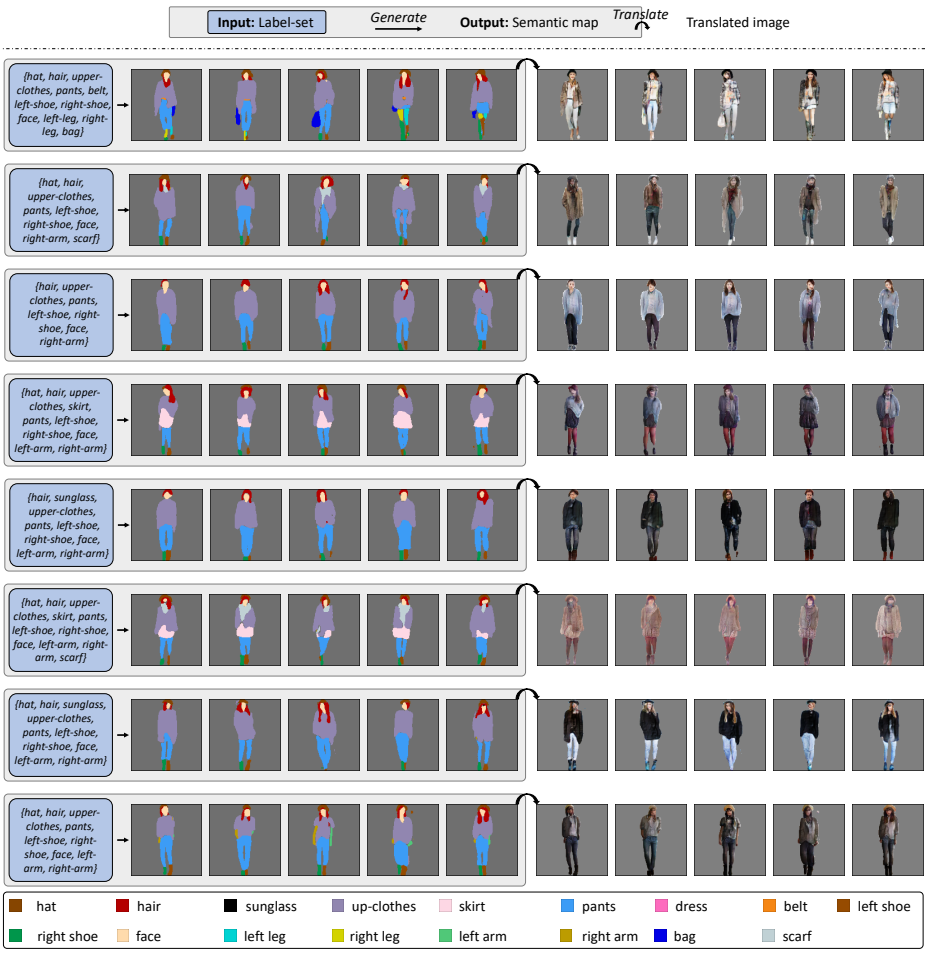

Fig. 2: **Multi-modality.** We demonstrate the diverse generation results given a label-set on the HumanParsing dataset.

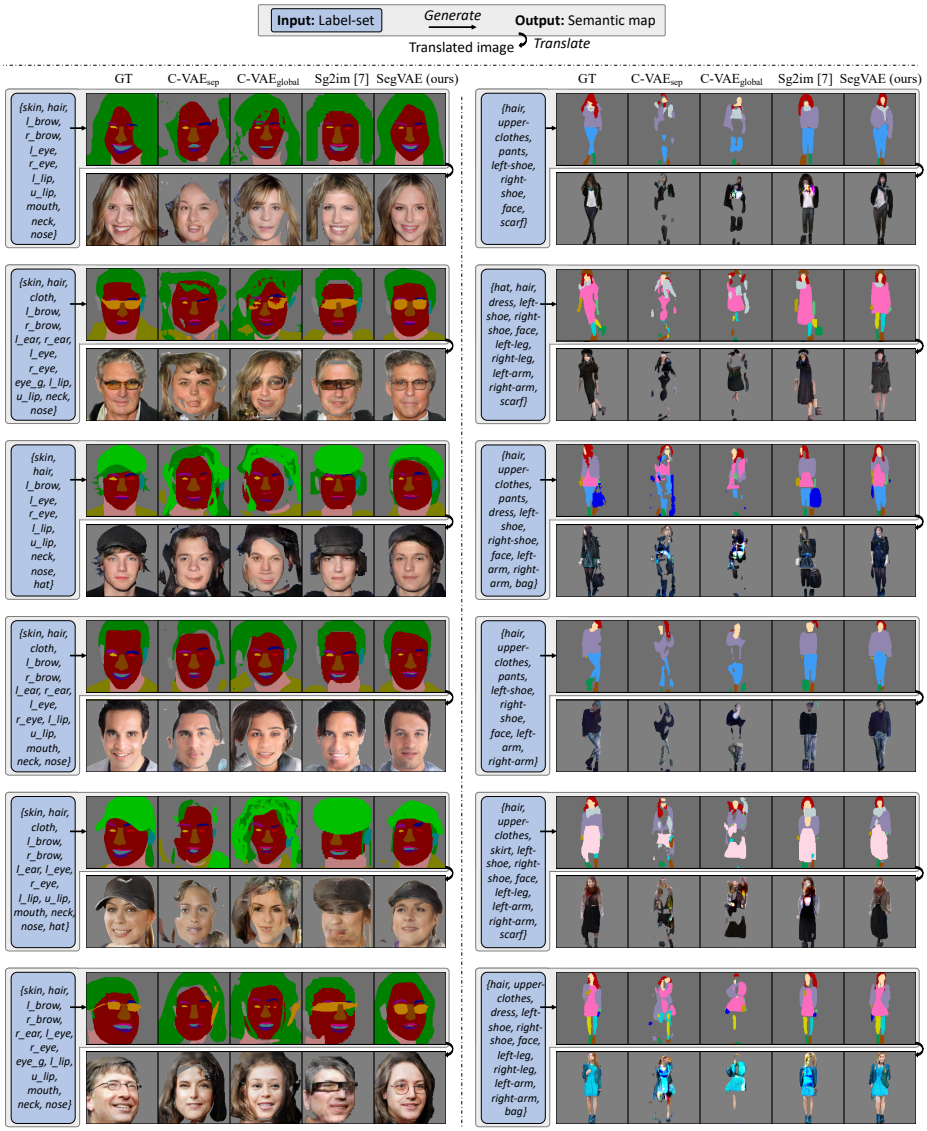

Fig. 3: **Qualitative comparison.** We present the generated semantic maps given a label-set on the CelebAMask-HQ (left) and the HumanParsing (right) datasets. The proposed model generates images with better visual quality compared to other methods. We also present the translated realistic images via SPADE [1]. Please refer to Figure 1 and Figure 2 for the color mapping for each category.

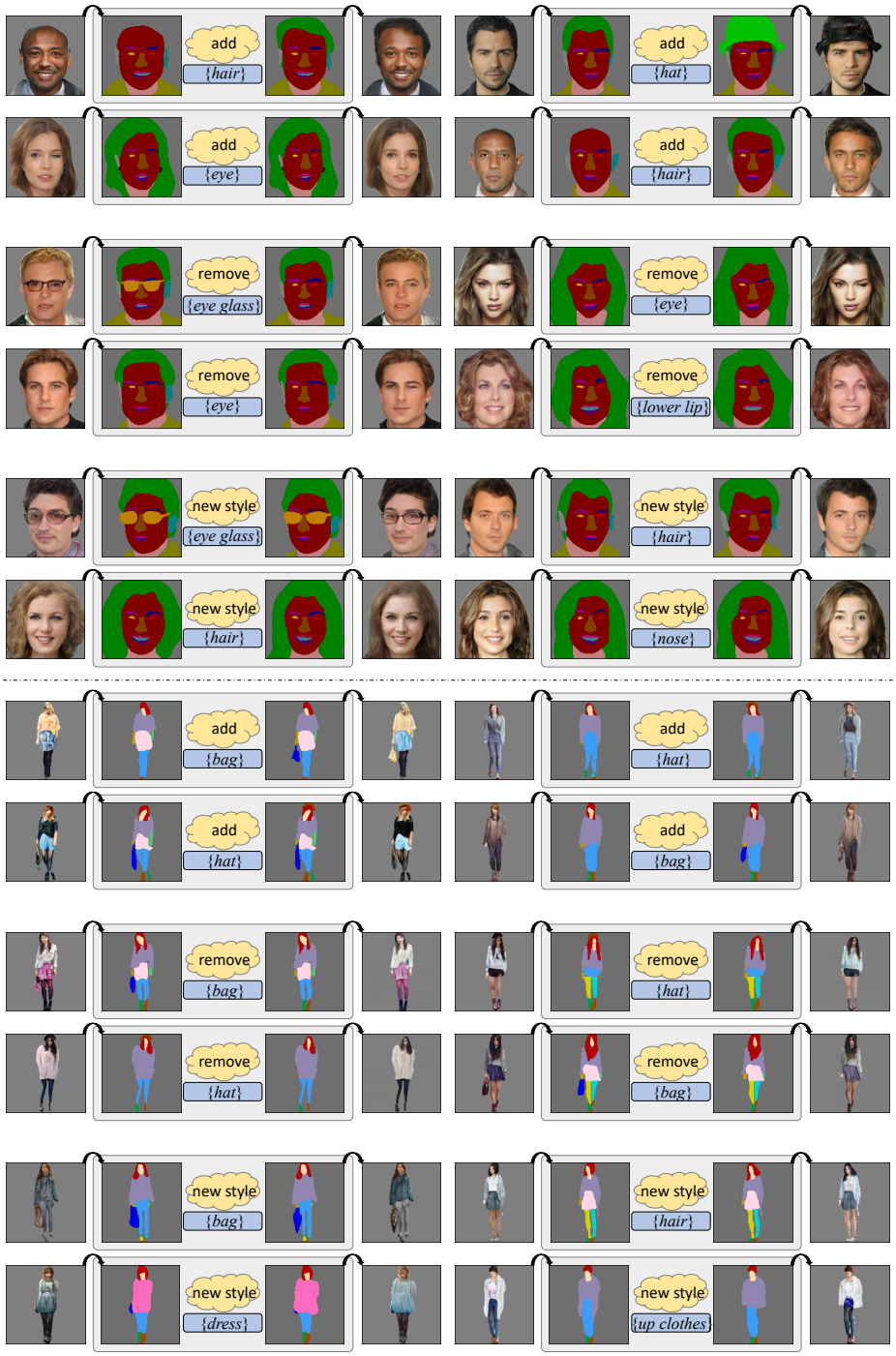

Fig. 4: **Editing.** We present three real-world image editing applications: *add*, *remove*, and *new style*, on the CelebAMask-HQ and HumanParsing datasets. SegVAE enables flexible and intuitive control over the generated outputs.

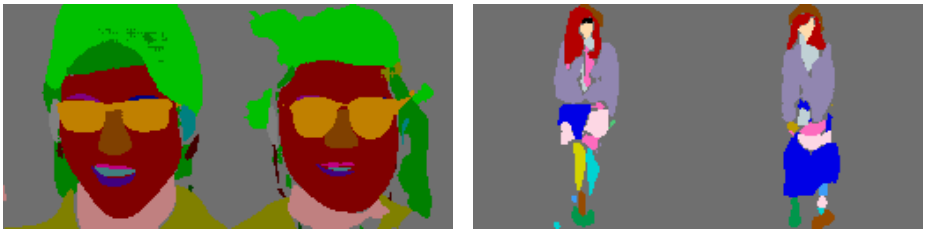

Fig. 5: **Failure cases.** We demonstrate two typical cases in two datasets: (a) the shape space is under-sampled, (b) the input label-set is unseen during training.

References

1. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019) 5