Controllable Face Synthesis with Semantic Latent Diffusion Models

Alex Ergasti[0009−0005−8110−9714], Claudio Ferrari[0000−0001−9465−6753], Tomaso Fontanini[0000−0001−6595−4874], Massimo Bertozzi[0000−0003−1463−5384], and Andrea Prati[0000−0002−1211−529X]

University of Parma, Department of Architecture and Engineering, Parma, Italy
https://github.com/ErgastiAlex/SCA-DM
{alex.ergasti,claudio.ferrari2, tomaso.fontanini, massimo.bertozzi,andrea.prati}@unipr.it

Abstract. Semantic Image Synthesis (SIS) is among the most popular and effective techniques in the field of face generation and editing, thanks to its good generation quality and the versatility is brings along. Recent works attempted to go beyond the standard GAN-based framework, and started to explore Diffusion Models (DMs) for this task as these stand out with respect to GANs in terms of both quality and diversity. On the other hand, DMs lack in fine-grained controllability and reproducibility. To address that, in this paper we propose a SIS framework based on a novel Latent Diffusion Model architecture for human face generation and editing that is both able to reproduce and manipulate a real reference image and generate diversity-driven results. The proposed system utilizes both SPADE normalization and cross-attention layers to merge shape and style information and, by doing so, allows for a precise control over each of the semantic parts of the human face. This was not possible with previous methods in the state of the art. Finally, we performed an extensive set of experiments to prove that our model surpasses current state of the art, both qualitatively and quantitatively.

Keywords: Semantic Image Synthesis · Diffusion Models · Face Editing

1 Introduction

Diffusion models (DMs) are the current state of the art in the majority of fields involving generative tasks. Since their introduction, they were applied to a large variety of scenarios, yet their most important application is in the field of image generation. In this domain, they were applied in several contexts, from text-to-image [19] to semantic segmentation [1] and many more [2], thanks to the excellent quality and variety of data that can be generated through them. However, even if results obtained with DMs are unquestionably impressive, they face two major limitations, that are *(i)* the lack of controllability and reproducibility of the generated samples and *(ii)* the missing ability of reproducing and editing real images. This means that it is both extremely difficult to have a precise and

Fig. 1: Our model can generate images in three ways: (a) Given a reference image, (b) mixing styles from a reference image and noise, (c) fully noise-based.

fine-grained control of what is generated, mostly at a local-level, and to manipulate only specific objects in a given image. In the attempt to overcome these problems, several solutions were proposed. Some methods focus on controlling the output of the generative process in terms of spatial arrangement or shapes, such as in ControlNet [28], Composer [11], T2I [16] or Semantic-Diffusion [27]. These architectures pursue the goal of controlling the layout and appearance of the generate images by using sketches, color palettes, depth or semantic maps as additional information. Other works instead aim to encode a set of reference images to reproduce an object and its appearance in order to generate new samples including such object, as in Textual Inversion (TI) [5] or Dreambooth [20].

Scenarios where generative models are applicable is vast. Nowadays, pretty much any data can be generated. We chose to explore the domain of human faces for several reasons: first, generative face models raise severe safety and privacy related concerns, as effective models can be maliciously used to alter or change the identity of a given individual *i.e.* DeepFakes. On the other hand, generative face models can be employed for biometric applications such as increasing the robustness of face recognition systems to adversarial attacks using synthetic data, achieving important privacy preserving properties. Finally, faces present a high degree of correlation across facial parts. A model is thus prone to reproduce biases that are commonly found in human face datasets *e.g.* people with light colored eyes usually have blonde hair, which makes disentangling, controlling and generating diverse appearances for local face parts more challenging. Nevertheless, in general, all the methods in the literature based on diffusion models are all designed to *generate* new data, while trying to control the generation in some way. To the best of our knowledge, there is yet no convincing solution to

exploit diffusion models for addressing the task of reconstructing and precisely manipulating *real images*. Hence, in this paper we present a solution to augment diffusion models with the above capability.

More specifically, we cast the problem of face editing, manipulation and generation in a Semantic Image Synthesis (SIS) framework. This choice is motivated by the observation that SIS approaches demonstrated the most effective for the task of precise image manipulation. SIS methods address the task of generating photo-realistic images using their semantic mask as a condition to control the spatial layout. A semantic mask is an image in which each pixel represents a semantic class, and is pixel-wise aligned with the RGB reference. In the literature, the vast majority of SIS architectures make use of custom normalization layers to modulate the features activation with the information contained in the semantic mask. In particular, the most common paradigm is represented by SPADE [17] which introduced spatially-adaptive normalization layers. There exist several SIS methods with different goals: some focus of generating a random style for each semantic parts (noise-based) [23,18], while others are trained to extract specific styles from a reference image and map them to each semantic region (reference-based) [30,23]. Until recently, the majority of SIS models were based on Generative Adversarial Networks. With the surge of diffusion models, the research efforts started moving towards them. In particular, the most relevant diffusion SIS architecture is Semantic Diffusion Model (SDM) [27] which fuses the powerful generation capability of DM with SPADE normalization layers in order to precisely control the shape of the generated samples. Still, SDM is fully noise-based and cannot reproduce a set of specific styles during generation.

To overcome this limitation, in this paper we propose a novel diffusion-based SIS model, named Semantic Class-Adaptive Diffusion Model (SCA-DM), that is able to both generate diverse samples conditioned on a semantic mask but, at the same time, can also be used to extract precise styles from any reference image with the specific goal of accurate human face editing. Thus, our goal is to design and investigate the first diffusion-based architecture that encapsulates both the above features. We do so by adding a style reference image in addition to the semantic mask as condition to guide the diffusion process. To achieve that, we trained from scratch a Latent Diffusion Model (LDM), conditioned with both the semantic mask and styles extracted from the reference image by means of a dedicated style encoder. Conditioning a diffusion model can be achieved in multiple ways, such as via Cross Attention, concatenation or, as in our specific case, with SPADE normalization layers. Since our proposed architecture needs to merge both the information provided by the mask and the reference image to be edited, we chose to use a variation of the Cross Attention to control the style condition, and SPADE to condition the spatial layout with the semantic mask.

We will show that the proposed solution enables both accurate face generation and editing of real images, a property that is still under-explored with diffusion models, and showcase its advantages with respect to prior GAN-based methods and the recent SDM [27]. In sum, the contributions of this work are:

- We developed a novel LDM architecture for SIS that is both able to generate diverse samples and exactly reproduce a given style extracted from a reference image, opening the way to precise human face editing.
- We propose and explore the combination of SPADE normalization layers and cross-attention to fuse layout and style information together, as opposed to prior works that only use SPADE. Additionally, we designed a modified mask-conditioned cross-attention layer in order to improve the disentanglement of different facial parts.
- We provide an extensive set of qualitative and quantitative evaluation to showcase the proposed model capability and its superiority w.r.t. to current state-of-the-art solutions.

2 Related Work

Diffusion Models. Diffusion models are currently the state of the art of generative models, firstly proposed by *Ho et Al.* [8] as Denoising Diffusion Probabilistic Models (DDPM). Numerous studies have focused on enhancing diffusion models to mitigate their primary drawbacks, namely prolonged training and inference time. To address the former, the Latent Diffusion Model (LDM) [19] was proposed introducing an encoder/decoder structure to reduce the dimensionality of the space in which the diffusion model operates. By enabling the diffusion model to function within a latent space, researchers have successfully reduced the required training time. To tackle the latter, a new sampling technique called DDIM [22] has been proposed. It allows to sample from a DM without any retraining requirements, reducing in this way the steps required by paying just a little of image quality, speeding up the process even up to a factor of 20. Moreover, the advancement of diffusion models has opened up for various works aimed at enhancing their capabilities. For instance, works like Control Net [28], Composer [11] or T2I-Adapter [16] represent notable expansions of LDM. All of these approaches were proposed with the common objective of enhancing the control and precision of image generation in diffusion models, incorporating different condition mechanism, such as semantic masks or scribbles.

Other works based on pretrained LDM focus on adding new information inside the model, by teaching a new word S^*, as in Textual Inversion [5] and Dreambooth [20]. Nevertheless, both methods require a fine-tuning without being one shot and are not tailored to exactly reproduce human faces but rather subjects for which inconsistencies are less spottable by a human eye. Other disadvantages of these techniques are that they do not allow a precise style mixing, delegating all the generation variety to the chosen text, thus without any possibility to a fine grained control over the the generated samples.

Semantic Image Synthesis. The current landscape of semantic image synthesis can be broadly divided into two main categories: GAN-based and Diffusion-based approaches. Within the GAN-based category, further subdivision is possible based on whether the models are reference-based or not. Under the former category, numerous GAN models exist, many of which share a similar structure.

In the current state of the art, SEAN [30] is a notable model that utilizes SPatial Adaptive DE-normalization layers (SPADE, [17]) to inject both semantic masks and style embeddings into the generator. However, it lacks the ability to generate diverse images; given the same input, it always produces the same image. Another noteworthy model is CLADE (CLass Adaptive DE-normalization) [25], which has the capability to generate images both with and without an RGB reference image. This flexibility enhances its applicability in various scenarios. Additionally, INADE [24] is another significant model in the landscape of semantic image synthesis. Unlike SEAN, INADE focuses on producing images using instance style rather than class style. This allows for conditioning each instance in the image individually, leading to more precise and nuanced results. Next, FrankenMask [4] focused on the automatic manipulation of the semantic mask. Another remarkable model is SemanticStyleGAN [21], which innovatively expands upon StyleGAN [13]. Unlike other cited approaches, SemanticStyleGAN does not directly utilize a given semantic mask during the generation process. Instead, it splits the style embedding with the semantic mask embedding information. This novel approach enables the model, with the aid of GAN inversion [29] techniques, to generate a face given a semantic mask. Finally, Tarollo *et al.* [26] employed a SIS model to perform adversarial attacks on face recognition systems, proving the broader range of applications of these models.

One of the most renowned and accessible diffusion models for semantic image synthesis is the Semantic Diffusion Model (SDM) [27]. Unlike other models, SDM is conditioned solely by the semantic mask without any reference image. Consequently, it excels in generating diverse images, but without the option to guide the generation through a reference image. Other notable models in this domain include Control Net [28], with every other similar works as Composer [11] or T2I [16], and Collaborative Diffusion [12]. Control Net (and similar) incorporates semantic masks using UNet injection, providing additional control over the synthesis process. On the other hand, Collaborative Diffusion utilizes multiple diffusion models that collaborate with each other to produce high-quality face images, showcasing the potential for collaborative approaches in SIS.

The goal of our research is to propose a Latent Diffusion Model (LDM) that is able to produce high quality and diverse samples, which shape is controlled by semantic masks while the styles can be both extracted from a reference image or generated from scratch, as shown by 1. Our system is capable of maintaining the overall coherence and allowing controllability of the generated images better than current state of the art in every of its different configurations.

3 Proposed Architecture

We propose SCA-DM (Semantic Class-Adaptive Diffusion Model), a Latent Diffusion Model (LDM [19]) for semantic face image synthesis. Our model is composed of: a pretrained VQGAN, a custom Diffusion Model which integrates SPADE and cross-attention layers, and a Multi-Resolution Style Encoder to extract the style given an RGB image and the corresponding semantic style mask.

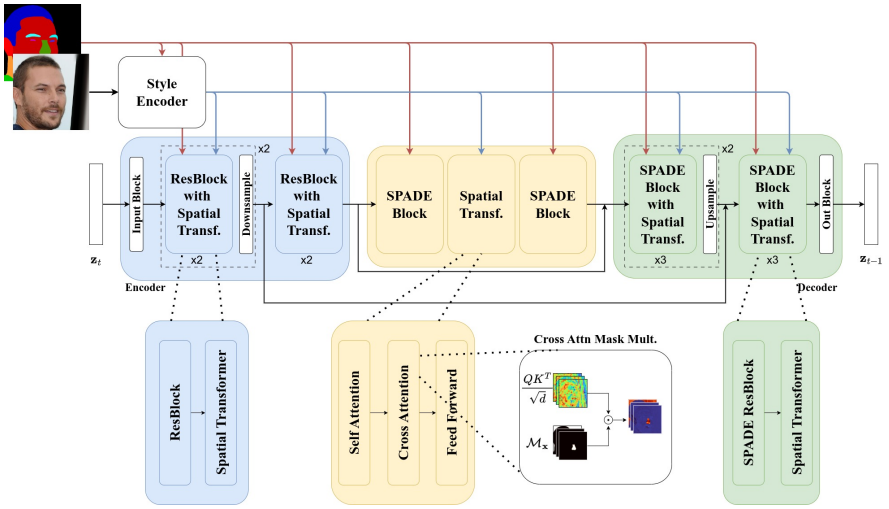


Fig. 2: Model architecture. The encoder part of the UNet uses only standard Resnet Block with SpatialTransformer to guide the diffusion process with the style embedding obtained from $\mathcal{E}_s$. The middle block and the decoder part use SPADEResBlock, as in SDM, to encapsulate the semantic mask info. The Mask attention is applied inside the SpatialTransformer on the Cross Attention Map.

Our custom diffusion model has SPADE ResBlock in the middle block and in the decoder block, inspired by SDM [27], and it uses Spatial Transformer to inject the Style Embedding inside the UNet architecture, as shown by Fig. 2.

During training, the model will learn to merge shape and style information together in order to faithfully reproduce any given subject.

3.1 Latent Diffusion Model

Diffusion models are probabilistic models designed to gradually remove noise from a normally distributed sample $\mathbf{z}$. The process that gradually adds Gaussian noise with fixed and scheduled variance $\beta_1, \ldots, \beta_T$ to the data $\mathbf{x}_0$ is called forward process and it is described by:

$$\begin{aligned} q\left(\mathbf{x}_T|\mathbf{x}_0\right) &= \prod_{t=1}^{T} q\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right) \\ q\left(\mathbf{x}_t|\mathbf{x}_{t-1}\right) &= \mathcal{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}; \beta_t\mathbf{I}\right) \end{aligned} \tag{1}$$

The model learns a data distribution $p(\mathbf{x})$, which corresponds to learning the reverse process of a fixed Markov chain of length T. The most successful models [3,9] rely on a reconstruction loss of the image $\mathbf{x}_{t-1}$ given $\mathbf{x}_t$, which is indeed a reweighted variant of the variation lower bound on $p(\mathbf{x})$. In our case, the training loss is defined as follows:

$$\mathcal{L}_{\mathrm{DM}} = \mathbb{E}_{\mathbf{x},\epsilon\sim\mathcal{N}(0,1),t}\left[\left\|\epsilon - \epsilon_\theta\left(\mathbf{x_t}, \mathbf{t}, \mathcal{E}_\mathbf{s}\left(\mathbf{x}\right), \mathcal{M}_\mathbf{x}\right)\right\|_2^2\right] \tag{2}$$

The model architecture ϵ_θ is depicted in 2. It is conditioned with time t, semantic mask $\mathcal{M}_\mathbf{x}$ and the style embedding $\mathcal{E}_s(x)$ of the given style image x. By doing so, the proposed system will link the shape information of the generated sample to $\mathcal{M}_\mathbf{x}$ and map the extracted semantic styles to each semantic region. Since, as said before, the forward process is fixed, during training it is possible to obtain $\mathbf{x}_t$ directly for each uniform sampled t via:

$$\mathbf{x}_t = \sqrt{\overline{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\overline{\alpha}_t}\epsilon, \quad \text{where } \overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i,\ \alpha_i = 1 - \beta_i. \tag{3}$$

Our proposed model is a Latent Diffusion Model, which means it works on latent space and not directly on the image space. Because of this, we use a pretrained VQGAN to encode and decode the image from and to the latent space. Apart from the pretrained VQGAN, both the model ϵ_θ and the style encoder $\mathcal{E}_s$ are trained together using the aforementioned loss. It is also possible to encourage the model to place greater reliance on conditioning, as demonstrated by Ho et al. [10]. This can be achieved by computing and combining the estimated noises (ϵ_θ) with and without conditioning, obtaining a new noise $\overline{\epsilon}$. This approach involves passing an empty conditioning through the network to compute the estimated noise, thereby obtaining a refined estimation that incorporates the influence of conditioning terms $\mathcal{E}_s(x)$ and $\mathcal{M}_\mathbf{x}$:

$$\begin{aligned}\overline{\epsilon}(\mathbf{x_t}, \mathbf{t}, \mathcal{E}_\mathbf{s}(\mathbf{x}), \mathcal{M}_\mathbf{x}) = \epsilon_\theta(\mathbf{x_t}, \mathbf{t}, \mathcal{E}_\mathbf{s}(\mathbf{x}), \mathcal{M}_\mathbf{x}) + \\ + s(\epsilon_\theta(\mathbf{x_t}, \mathbf{t}, \mathcal{E}_\mathbf{s}(\mathbf{x}), \mathcal{M}_\mathbf{x}) - \epsilon_\theta(\mathbf{x_t}, \mathbf{t}, \mathcal{E}_\mathbf{s}(\emptyset), \emptyset))\end{aligned} \tag{4}$$

where s controls the intensity of the combination.

3.2 Multi-Resolution Style Encoder

The Multi-Resolution Style Encoder $\mathcal{E}_s$ takes as input an RGB image $x \in \mathbb{R}^{3\times H\times W}$ and a corresponding semantic mask image $\mathcal{M}_x \in \mathbb{N}^{C\times H\times W}$. For each of the L convolutional layers in the style encoder, we extract style features specific to each of the C semantic classes. Specifically, at the i-th layer, with feature maps of size $\mathbb{R}^{D_i\times H_i\times W_i}$, we split the D_i channels into C groups, yielding a feature map $F_{i,j} \in \mathbb{R}^{D_i/C\times H_i\times W_i}$ for each semantic class j. We then apply element-wise multiplication with the corresponding mask channel, followed by average pooling (AP), to obtain the style feature $S_{i,j}$:

$$S_{i,j} = AP\left(F_{i,j} \cdot \mathcal{M}_j\right) \tag{5}$$

Subsequently, we reshape each $S_{i,j}$ to have D channels using a 1×1 convolution. This process is repeated for each layer $i = 1, \ldots, L$, resulting in a multiscale style embedding $S \in \mathbb{R}^{C\times L\cdot D}$, where D is the size of the embedding at any layer.

3.3 Mask-conditioned Cross-Attention

During training, our model makes use of cross-attention layers in order to map the style codes in the corresponding semantic regions of the images. More in detail, cross-attention layers are defined as follows:

$$\mathcal{L}_{CA}(Q, K, V) = \mathcal{S}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{6}$$

where $\mathcal{S}$ is the Softmax function. Additionally, differently than self-attention that combines embeddings from a single source, in cross-attention embeddings come from two separate sources. In particular, in our case, Q is obtained from the projection of the features of previous convolutional layers $\mathcal{F}_i$, while K and V are obtained from the projection of the style codes $\mathcal{E}_s(x)$. More in detail:

$$Q = W_Q^{(i)} \cdot \mathcal{F}_i, K = W_K^{(i)} \cdot \mathcal{E}_s(x, \mathcal{M}_x), V = W_V^{(i)} \cdot \mathcal{E}_s(x, \mathcal{M}_x) \tag{7}$$

In the proposed architecture we modify the cross-attention layers in order to force disentanglement between each class embedding extracted by the Style Encoder. More in detail, we multiply in each Cross Attention the given attention map with the semantic mask (as shown in 2). By doing this, a single style embedding can only modify a very specific part of the image, therefore encouraging the style encoder to extract only local information, delegating to self-attention the image coherency. A cross-attention layer then becomes:

$$\mathcal{L}_{CA}(Q, K, V) = \mathcal{S}\left(\frac{QK^T}{\sqrt{d}}\mathcal{M}_x\right)V \tag{8}$$

By multiplying the mask *before* applying the Softmax function we make sure that the numerical coherence is maintained.

3.4 SPADE ResBlock

In our architecture, the Multi Style Encoder $\mathcal{E}_s$ is responsible for the style embedding and cross attention layers inject the extracted style in the diffusion process. In order to inject also the semantic mask $\mathcal{M}_x$, we substitute the basic ResBlock of U-Net backbone with SPADE ResBlock [17,27] (Fig. 2). The SPADE ResBlock, an extension of the ResNet block [6], is equipped with SPADE layers that take the semantic image mask ($\mathcal{M}_x$) as a conditioning input. SPADE layers work by generating two feature maps, γ and β, which are subsequently used to scale and shift the features inside the ResBlock. By incorporating these blocks, we enable the diffusion model to be conditioned on the semantic mask.

4 Experiments

Training Details. We train our model on a NVIDIA 4090 gpu with a learning rate of $2.0e-06$. During training, 50% of the images to the style encoder were

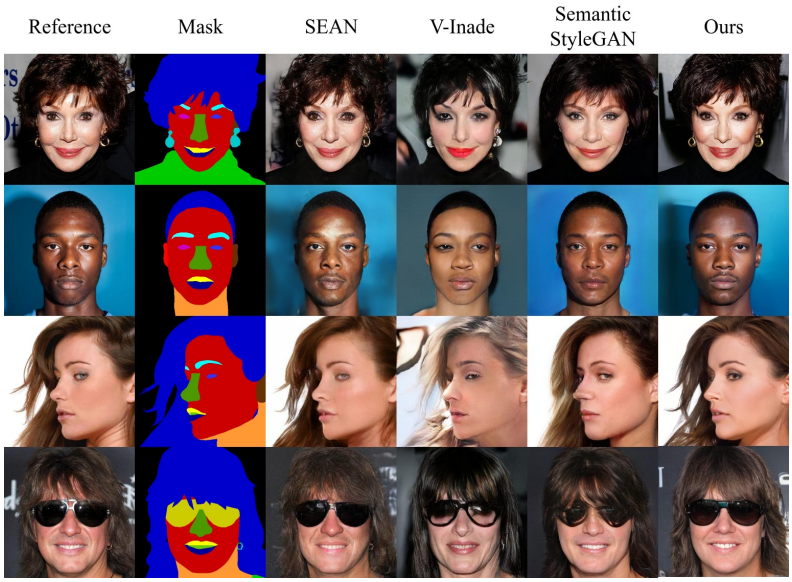


Fig. 3: Reconstruction comparison between the state of the art and our model.

set to zeros; this lets the model learn to produce random images without any reference, and improved the capability of generating conditioned images manipulating the scale s in 4. For all the tests, we empirically set $s = 1.2$.

Dataset. we train our model on CelebAMask-HQ [14] which is composed by 30k human face images paired with the corresponding 19-channel semantic mask. The dataset is splitted as follows: 28k images for training and 2k for testing.

Metrics. To evaluate our method, we employ the current state-of-the-art evaluation metrics of SIS methods. In particular, we use Frechet Inception Distance [7] (FID) to estimate the generation quality compared with the test dataset. Furthermore, we evaluate the similarity between the given semantic mask and the semantic mask parsed from the generated images using mean Intersection-over-Union (mIOU) and pixel accuracy. We use a pretrained FaceParsing model [15] to generate the masks from the fake samples. Finally, to validate how closely the model is able to reconstruct the reference human face, we calculate Structural Similarity Index Measure (SSIM). We test our model in the following experimental settings: reconstruction (reference-based), total style swap, partial style swap and diversity (noise-based).

4.1 Results on reconstruction

As a first experiment, we tested the capability of our model to faithfully reconstruct any human face starting from the corresponding semantic mask and styles. A qualitative evaluation of this task can be seen in Fig. 3 where different reconstruction results produced by several state-of-the-art methods are compared to

Architecture	CelebA-HQ			
	SSIM ↑	FID ↓	mIOU ↑	Acc. ↑
MaskGAN [14]	0.51	59.91	76.3	87.8
SEAN [30]	0.49	20.82	**82.4**	**95.0**
V-INADE [24]	0.29	17.49	78.0	93.5
SemanticStyleGAN [21]	0.52	26.83	69.8	90.2
Ours	**0.54**	**16.85**	81.78	94.67

Table 1: Comparison with the state of the art of reconstruction models in terms of SSIM, FID, mIOU and segmentation accuracy.

the results generated by our model. Our model can precisely reconstruct the input image even in challenging setting (Fig. 3, bottom rows).

Then, we performed an extensive quantitative evaluation, as shown in 1. Our model achieves better FID compared to state-of-the-art models. On the other side, we have slightly worst performance in terms of mIOU and Accuracy even if we still position as second-best. This is likely caused by the fact that the proposed diffusion model works in the latent space, using an VQGAN encoder/decoder architecture to go from and to pixel space leading to some misalignment between mask and generated image. Finally, we calculate SSIM to verify the similarity between the generated samples and the reference image and we obtained the best result overall. In this section, a comparison with SDM was not possible given its inability to perform reference-based inference.

4.2 Results on total style swap

We conducted a comparative analysis to assess our model's ability to generate images using the style of another image (total style swap). The quality of the generated images is shown in 4 where all models capable of style transfer were presented. Our system demonstrates superior overall consistency and produces images of higher quality compared to the state of the art. Additionally, thanks to our training approach, the model can generate a random style for elements not present in the reference image, as showcased in the second column where the hat is accurately generated despite its absence in the reference. Similarly, in the first column, the model generates meaningful eyes without explicit reference, highlighting its versatility.

Furthermore, to provide additional evidence of our model capabilities both in terms of flexibility and robustness, we showcase in 6 (bottom two rows) the interpolation of a full style embedding.

4.3 Results on partial style swap

In this section, our objective is to demonstrate our model's ability to selectively alter only a portion of the style in the target image, as illustrated in 5a. Our

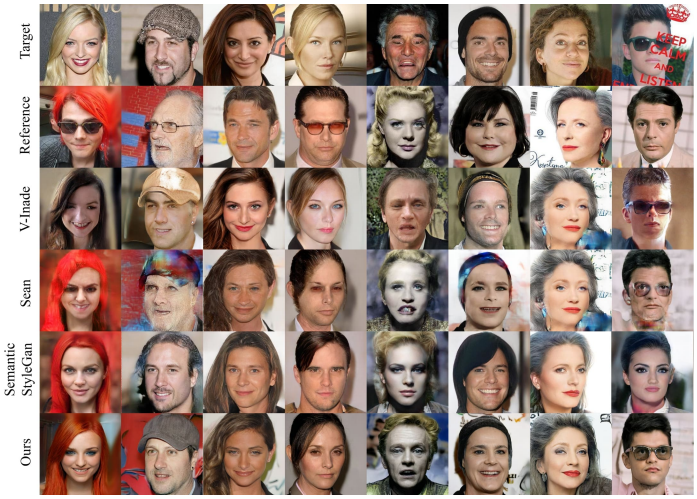


Fig. 4: Style transfer comparison between different methods and our model. The style of the reference image is applied to the target image. The overall consistency in style swap is far better compared to state-of-the-art methods.

model successfully swaps a subset of styles in the generated image, preserving the integrity of the other elements.

This test also serves the purpose of demonstrating the efficacy of the masked attention that is crucial for this task. Indeed, without it, there would be too much entanglement between each style part and the model would not be able to generate images with mixed style, as evidenced in 5b. As expected, the model without the masked attention is not able to partially edit the style, *e.g.* the skin tone does not change in the first row. In the last two rows, we tried to change the style of a single eye. Because the eyes style is entangled (the eyes color is the same for both eyes), the model without any mask attention is not able to perform this local editing, but the style is applied to both.

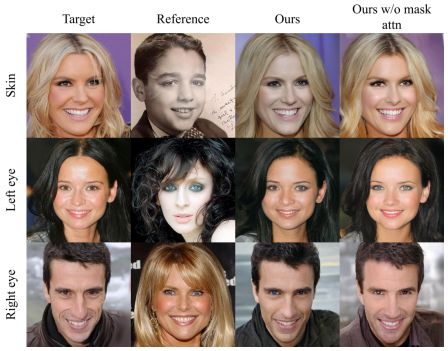
Further evidence of our model proficiency in partial style swapping can be found in 6, where we showcase its ability to interpolate partial feature styles from full target to full reference.

4.4 Results on noise-based Generation

Our model is also able to generate images without any reference, by passing an all zero image to the style encoder. This allows our model to be compared also to solutions, like Semantic Diffusion Model (SDM), that can only generate diverse results but are unable to exactly reproduce a specific human face.

We reported both a quantitative and a qualitative evaluation in 7. Starting from the former (on the right), a comparison between our model and SDM when using different diffusion steps for generation is presented. In particular, FID and inference time (in seconds) were calculated for 50, 100, 200 and 1000

(a) Style transfer of single face features. Our model can successfully swap single style part with high image coherence.



(b) Ablation study which compares the model with and without mask attention. The model without mask attention has more entanglement between each style feature.

diffusion steps. Indeed, the FID score of our model outperforms SDM in every configuration except the last one. We argue our model performs better with fewer diffusion steps due to the fact that it works in the latent space and, for this reason, can recover, thanks to the VQGAN, details that would have been lost with a small number of diffusion steps. This is crucial since, looking at the inference time, performing inference with 1000 steps is often unpractical, especially for SDM which overall takes $5\times$ more time than our method to perform a single inference step. Finally, we calculated LPIPS to measure the diversity in the generated results, with an higher value corresponding to better diversity. Our model scores an LPIPS score of 0.33 against 0.42 of SDM. This was expected as adding reconstruction capability to a model can slightly hinder its generation of diverse sample due to some overfitting, *e.g.* the model learns to associate semantic mask shapes to specific styles.

Finally, we argue that our results are visually on par or even better than those produced by SDM expecially when a small number of diffusion steps is employed during inference. Indeed, this can be seen in 7 on the left.

5 Conclusion

In this paper, we proposed a novel SIS model for precise semantic editing and human faces synthesis. Our system is based on LDM [19] and introduces in the decoder of the U-Net architecture a series of SPADE layers to condition the generation with the semantic masks. Additionally, masked cross-attention layers are employed to map styles extracted from a reference image to the semantic region of the masks without causing entanglement. With the combination of SPADE and cross-attention layers our model is able to both generate diverse results and to faithfully reproduce any given face. Regarding diversity, our model

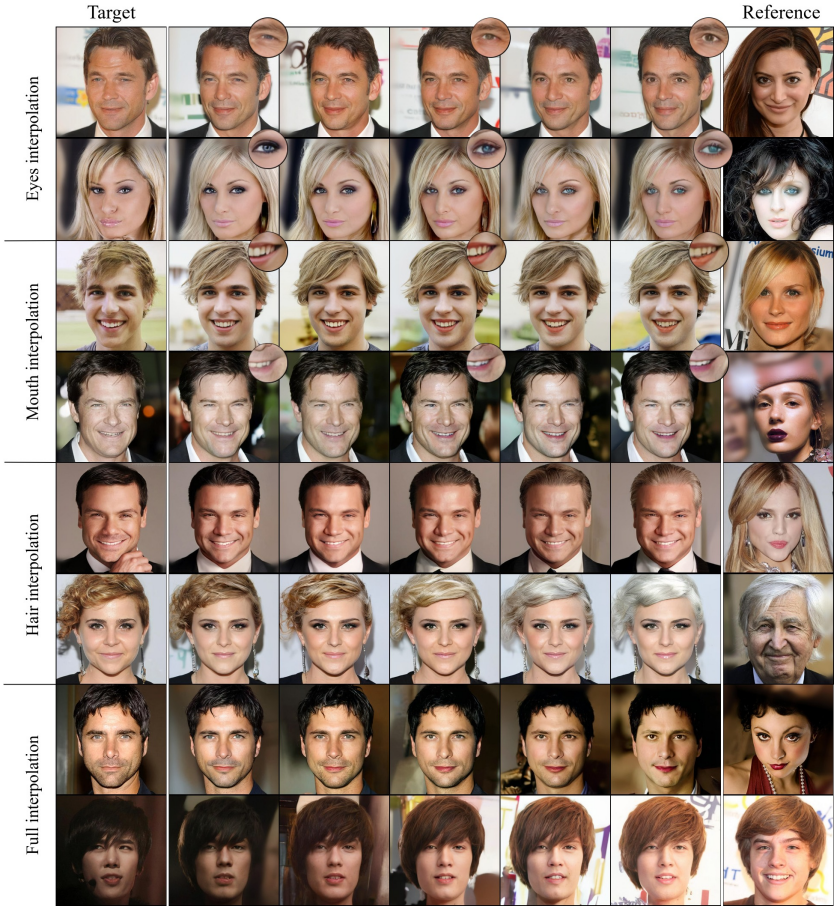


Fig. 6: Interpolation of eyes, mouth, hair style and full style going from full target (left) to full reference (right). Some details are highlighted for a clear observation of changes.

suffers from some overfitting during training and therefore sometimes struggle to generate samples completely different than the original RGB image. Additionally, the system allows for local and global style transfer, as well as style interpolation. Indeed, with current diffusion models conditioned with semantic information like SDM [27] this was not possible, making our system more flexible and suited for several human face editing applications.

Acknowledgements This work was partially funded by "Partenariato FAIR (Future Artificial Intelligence Research) - PE00000013, CUP J33C22002830006" funded by the European Union - NextGenerationEU through the italian MUR within NRRP, project DL-MIG.

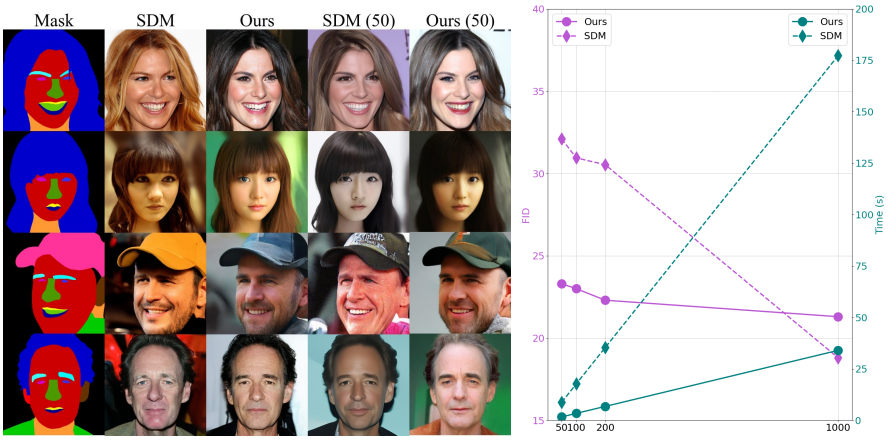


Fig. 7: Left: comparison of capability of generating images with random style between SDM and our model with 1000 or 50 steps. Right: quantitative evaluation of FID and inference time for different numbers of diffusion steps.

References

1. Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. arXiv preprint arXiv:2112.03126 (2021)
2. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
3. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. CoRR **abs/2105.05233** (2021), `https://arxiv.org/abs/2105.05233`
4. Fontanini, T., Ferrari, C., Lisanti, G., Galteri, L., Berretti, S., Bertozzi, M., Prati, A.: Frankenmask: Manipulating semantic masks with transformers for face parts editing. Pattern Recognition Letters **176**, 14–20 (2023)
5. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion (2022)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. CoRR **abs/2006.11239** (2020), `https://arxiv.org/abs/2006.11239`
10. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021), `https://openreview.net/forum?id=qw8AKxfYbI`
11. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions (2023)

12. Huang, Z., Chan, K.C.K., Jiang, Y., Liu, Z.: Collaborative diffusion for multi-modal face generation and editing (2023)
13. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks (2019)
14. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation (2020)
15. liuziwei7: Celebamask-hq. https://github.com/switchablenorms/CelebAMask-HQ/tree/master/face_parsing
16. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453 (2023)
17. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization (2019)
18. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2287–2296 (2021)
19. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models (2022)
20. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation (2023)
21. Shi, Y., Yang, X., Wan, Y., Shen, X.: Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing (2022)
22. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models (2022)
23. Tan, Z., Chai, M., Chen, D., Liao, J., Chu, Q., Liu, B., Hua, G., Yu, N.: Diverse semantic image synthesis via probability distribution modeling. In: IEEE/CVF Conf. on Computer Vision and Pattern Recognition. pp. 7962–7971 (2021)
24. Tan, Z., Chai, M., Chen, D., Liao, J., Chu, Q., Liu, B., Hua, G., Yu, N.: Diverse semantic image synthesis via probability distribution modeling (2021)
25. Tan, Z., Chen, D., Chu, Q., Chai, M., Liao, J., He, M., Yuan, L., Hua, G., Yu, N.: Efficient semantic image synthesis via class-adaptive normalization (2021)
26. Tarollo, G., Fontanini, T., Ferrari, C., Borghi, G., Prati, A.: Adversarial identity injection for semantic face image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1471–1480 (2024)
27. Wang, W., Bao, J., Zhou, W., Chen, D., Chen, D., Yuan, L., Li, H.: Semantic image synthesis via diffusion models (2022)
28. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
29. Zhu, J., Shen, Y., Zhao, D., Zhou, B.: In-domain gan inversion for real image editing. In: European conference on computer vision. pp. 592–608. Springer (2020)
30. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (Jun 2020). https://doi.org/10.1109/cvpr42600.2020.00515, http://dx.doi.org/10.1109/CVPR42600.2020.00515