IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 December 2024, accepted 5 January 2025, date of publication 14 January 2025, date of current version 17 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3529216

RESEARCH ARTICLE

Semantic Image Synthesis via Class-Adaptive Cross-Attention

TOMASO FONTANINI[1], CLAUDIO FERRARI[1], GIUSEPPE LISANTI[2], MASSIMO BERTOZZI[1], AND ANDREA PRATI[1], (Senior Member, IEEE)

[1]Department of Architecture and Engineering, University of Parma, 43124 Parma, Italy
[2]Department of Computer Science and Engineering, University of Bologna, 40126 Bologna, Italy

Corresponding author: Tomaso Fontanini (tomaso.fontanini@unipr.it)

This work was supported in part by the Progetti di Rilevante Interesse Nazionale 2020 (PRIN 2020) "LEGO.AI: LEarning the Geometry of knOwledge in AI systems" funded by Italian Ministero dell'istruzione, dell'università e della ricerca (MIUR) under Grant 2020TA3K9N, and in part by the "Partenariato Future Artificial Intelligence Research (FAIR)" funded by European Union (EU)—NextGenerationEU through Italian Ministero dell' università e della ricerca (MUR) within National Recovery and Resilience Plan (NRRP) through the Project Data-agnostic Learning for Multimodal Integration and Generation (DL-MIG) under Grant PE00000013 and Grant CUP J33C22002830006.

ABSTRACT In semantic image synthesis the state of the art is dominated by methods that use customized variants of the SPatially-Adaptive DE-normalization (SPADE) layers, which allow for good visual generation quality and editing versatility. By design, such layers learn pixel-wise modulation parameters to de-normalize the generator activations based on the semantic class each pixel belongs to. Thus, they tend to overlook global image statistics, ultimately leading to unconvincing local style editing and causing global inconsistencies such as color or illumination distribution shifts. Also, SPADE layers require the semantic segmentation mask for mapping styles in the generator, preventing shape manipulations without manual intervention. In response, we designed a novel architecture where cross-attention layers are used in place of SPADE for learning shape-style correlations and so conditioning the image generation process. Our model inherits the versatility of SPADE, at the same time obtaining state-of-the-art generation quality improving FID score by 5.6%, 1.4% and 3.4% on CelebMask-HQ, Ade20k and DeepFashion datasets respectively, as well as improved global and local style transfer. Code and models available at https://github.com/TFonta/CA2SIS.

INDEX TERMS Semantic image synthesis, cross-attention, image editing.

I. INTRODUCTION

Semantic image synthesis is the task of generating realistic images conditioned on a semantic mask, *i.e.* a pixel-wise annotation of the semantic classes defining the spatial layout. Since the initial stages of exploration in this domain, the predominant trend consisted in using spatially-adaptive normalization layers, firstly proposed by Park et al. and known as SPADE [1]. Those are designed to modulate the activations in the generator layers to propagate the semantic information, and so condition the generated samples in a spatially-adaptive manner. This mechanism proved effective in terms of generation quality and semantic control. Under this paradigm, images can be generated either by encoding the style, *i.e.* texture, from a reference image (*reference-based*), or in a fully generative setting (*diversity-based*). Most methods belong to the latter set, where the objective is to produce realistic yet diverse outputs given the same semantic mask [1], [2], [3], [4]. Whereas reference images are usually employed during training and can be used to guide the generation, such methods focus primarily on multi-modality and diversity. Oppositely, reference-based approaches tackle the problem of encoding the style of a specific image, with the primary goal of accurately reconstructing and editing *real images* [5], [6]. The two settings, despite similar in their goals and often sharing architectural design, demand for dedicated solutions to deal with different challenges. In diversity-driven approaches, some essential objectives are being able to generate multi-modal outputs given the same semantic layout, or avoiding overfitting. It is quite common that, for example, if a semantic mask of a human face has long hair, then the model will output an image of a woman, even if that information is completely unknown. On the opposite, reference-based methods deal with totally different issues.

The associate editor coordinating the review of this manuscript and approving it for publication was Szidonia Lefkovits.

For instance, a good reference-based method should be able to retain, in the case of human faces, the perceived identity of the portrayed individual. In this paper, we are specifically interested in the reference-based scenario. Most parts of the paper will hence be focused on this aspect, yet examples in the diversity-driven setting will be provided as well.

Despite alternatives have been explored in the literature, such as using layout-to-image conditional convolutions [3], [4], [5], [6], [7] SPADE and its variants still represent the standard choice in the field of semantic image synthesis. Even very recent approaches such as Semantic Diffusion [8] employ SPADE layers to condition the generation with the semantic layout. However, a drawback of all SPADE-based approaches (both diversity- and reference-based) is the tendency to introduce overall inconsistencies in the generated images. Indeed, they apply the feature modulation through learning class-wise normalization parameters, independently for each semantic class. Thus, global statistics such as illumination or color distribution, as well as long-range dependencies, are neglected, ultimately leading to inconsistencies in the generated images.

To address the above, in this paper we propose a significant paradigm change for semantic image synthesis, to answer the question on how to simultaneously improve the generation quality while maintaining a fine-grained control over the semantic classes. Specifically, we explore the use of cross-attention [9], [10] as an alternative to SPADE for conditioning the image generation in a generative adversarial setting. Cross-attention layers proved mostly effective when used with diffusion models to condition the image generation via text embeddings [11], [12]. They allow the conditioning mechanism to be: *(a)* very flexible, since cross-attention lets any intermediate representation to be mapped inside the network; *(b)* much more consistent, since attention accounts for long-range dependencies in the input data. Our proposed model naturally blends the capability of class-level style control as in previous GAN methods based on SPADE, with the versatility, improved quality and consistency of cross-attention.

In our framework, the input to the generator is the semantic mask, while the style features extracted from a reference image are the condition to the cross-attention layers. In SPADE layers, the spatial layout of the semantic mask can be directly used to apply class-specific feature normalization at precise spatial locations defined by the mask pixels, which forces a strong spatial consistency in the output image. On the other hand, it represents a noticeable constraint as inter-class dependencies or global statistics cannot be captured. Cross-attention layers allow instead to learn shape-style correlations by injecting style information into the model while the semantic mask can be encoded into a latent code and used as input to the whole system. We will show that this is advantageous in many ways: it leads to increased robustness to inaccuracies in the semantic masks, while also improving the overall image consistency. Being the mask encoded into a latent feature, it also enables to perform latent manipulation so to edit *the shape* other than the style. Further collateral advantage of this solution is the prospect of conditioning the generator with arbitrary style features. This versatility allowed us to design a specific style encoder that extracts multi-scale features and is equipped with class-adaptive grouped convolutions to optimize the representation of each class. In sum, the main contributions of this work are:

- We explore the use of cross-attention in place of SPADE layers in reference-based semantic image synthesis, and design a novel architecture to blend such mechanism into a GAN framework, inheriting the advantages of both.
- To show the advantage of using cross-attentions, we designed a new style encoder optimized for extracting multi-scale, class-level style features that can effectively mitigate the loss of details resulting from style pooling;
- We introduce a novel attention loss to force the learned attention maps to match the shape of the semantic mask, improving style mapping and controllability;
- We extensively show that our solution improves upon the state of the art, and provides an alternative paradigm that brings several advantages over prior works.

II. RELATED WORK

From a technical standpoint, all semantic image synthesis methods in the literature share similar frameworks. The two most common modules are the Style Encoder and the Generator network. The former is responsible for encoding the style (either from a reference image or noise), while the generator progressively up-samples the semantic layout provided as input. The output image is generated by injecting the style at specific spatial locations defined by the semantic mask.

A. SEMANTIC SYNTHESIS VIA SPADE LAYERS

Notwithstanding the different architectural choices, the major challenge is to properly encode and inject the style information into the generator. In this regard, Park et al. [1] first noted that in conventional synthesis architectures [13], [14], the normalization layers tend to remove the information contained in the input semantic masks. They thus proposed the SPatially-Adaptive (DE)normalization (SPADE) method to overcome this problem. The same model, also known as GauGAN [15], was deployed in a GAN-based image synthesis application. Since the introduction of SPADE, a lot of effort has been put into investigating its limitations and finding solutions for improvement. For example, Zhue et al. [6] noted that in SPADE only a single style code is used to control the style of the whole image. To gain more fine-grained generation control, they designed a SEmantic region-Adaptive Normalization block (SEAN), allowing to control the style of each semantic class individually. Simultaneously, Lee et al. [5] proposed a similar framework

specific for human faces, named MaskGAN, to enable diverse and interactive face manipulation. Nevertheless, SEAN outperforms MaskGAN in almost every aspect. Similar to SEAN, Tan et al. [16] proposed a CLass-Adaptive (DE)normalization layer (CLADE) that uses the input semantic mask to modulate the normalized activation in a class-adaptive manner. Later, to further push the style control beyond the class level, the same authors proposed the INstance-Adaptive DEnormalization (INADE) [4] approach that is capable of producing diverse results even at the instance level. Both CLADE and INADE are also equipped with an extra style encoder trained with KL loss to provide quality-driven results (V-CLADE, V-INADE). Along this line, several other works were proposed [7], [17], [18], [19], [20]. From a different perspective, Sushko et al. [21] proposed a SPADE-based architecture where an alternative training paradigm was designed that relies solely on adversarial supervision. Among all the reference works, one method that stands out in terms of similarity with ours is SEAN [16], which is the only one purely reference-based.

All the above mentioned papers employ SPADE or its variants. One limitation of SPADE layers is their dependency on spatially invariant normalization techniques, which can be less effective in capturing global dependencies or ensuring long-range consistency in generated images. Other works attempted to define alternative solutions such as SC-GAN [3], which introduced a novel semantic encoding and stylization methods via spatially-variant and appearance-correlated operations, inspired by the layout-to-image conditional convolutions earlier proposed by Liu et al. [7]. Other alternatives take inspiration from StyleGAN [22] and adapt it for the semantic synthesis task. Specifically, we mention Semantic-StyleGAN [23], which is a StyleGAN-based architecture that models local semantic parts separately via learning structural priors through the semantic mask. Differently from all the above, in this work we revisit the way in which style is injected into the generator by exploring the use of self- and cross-attention in place of SPADE. Among all the above, our results show that cross-attentions are the most promising.

B. DIFFUSION MODELS FOR SEMANTIC SYNTHESIS

All the previously referenced works are based on GAN frameworks, while Diffusion Models (DM) have now emerged as the new state of the art in the generative field [24]. After the introduction of Stable Diffusion [12], they became much less cumbersome to train, and quickly saw a widespread usage, even if their application in pure semantic image synthesis is still limited. The main reason is the difficult local and fine-grained controllability of the diffusion process. External modalities, *e.g.* text or semantic masks, typically used to condition the generative process, indeed do not provide locally-detailed information, making generation and editing vague and imprecise. We instead propose an opposite paradigm where encoded *style features* are used as condition, and the semantic layout is the input to the generative process. Collaborative Diffusion [25] or ControlNet [26] are two notorious examples where additional modalities (including semantic masks) are used in support of text-to-image generation. One example of mask-based solution is Semantic Diffusion (SDM) [8], which uses only semantic layouts to condition the diffusion process. Nevertheless, SDM is a diversity-based approach. Reference-based DMs are still under-explored, with the recent exception of [27] which proposed the first partial solution to this problem, but is still outperformed by our model. Indeed, convincing solutions that allow faithfully reconstructing and manipulating a real image are still lacking. Recent methods such as Dreambooth [28] or Textual Inversion [29] attempted to do so, yet they do not really manipulate a real image, but rather try to reproduce specific objects in different scenarios but cannot edit them, as opposed to our method. By borrowing the attention mechanism typical of DMs and adapting it to a GAN framework, we take advantage of the versatility of such layers while simultaneously avoiding the lack of controllability precision typical of diffusion models. Moreover, our solution enables local style control over the image generation, which is quite hard to obtain with DMs, even if semantic masks are used as condition as in SDM [8].

III. ARCHITECTURE

The proposed architecture (see Fig. 1) is called Class-Adaptive Cross-Attention, $(CA)^2$-*SIS*, and is composed by three main modules: (a) a Multi-Resolution Style Encoder $\mathcal{E}_s$, with Grouped Convolutions, Group Normalization layers and skip connections, that is used to extract style features from an RGB reference, (b) a Mask Embedder $\mathcal{E}_m$ that extracts a latent representation separately for each semantic class and, finally, (c) a Cross-Attention Generator $\mathcal{G}$ that exploits the attention mechanism to inject the multi-resolution style codes inside the network in order to control the generated image. Additionally, a Discriminator $\mathcal{D}$ is employed during training to enforce the adversarial loss.

A. MULTI-RESOLUTION GROUPED STYLE ENCODER

The purpose of the style encoder is to extract style features from an RGB image $y \in \mathbb{R}^{H \times W \times 3}$ to condition the generator. In order to independently control the style of different semantic classes, we improve the SEAN [6] style encoder, which uses the semantic mask to adaptively pool style features from specific areas of the RGB image. However, SEAN and other methods based on SPADE [1] extract style features only from the last layer of the encoder. We observed that this led to sub-optimal results. Specifically, style vectors are obtained by performing an average pooling on the masked feature maps. On the one hand, this operation removes spatial information, thus preventing the network from learning a trivial mapping (copying) between the input masks and the generated image. This is fundamental to later apply specific styles to arbitrary shapes, or manipulating them. On the other hand, pooling features necessarily leads to a loss of structural details of the texture, *e.g.* skin spots, cast shadows.

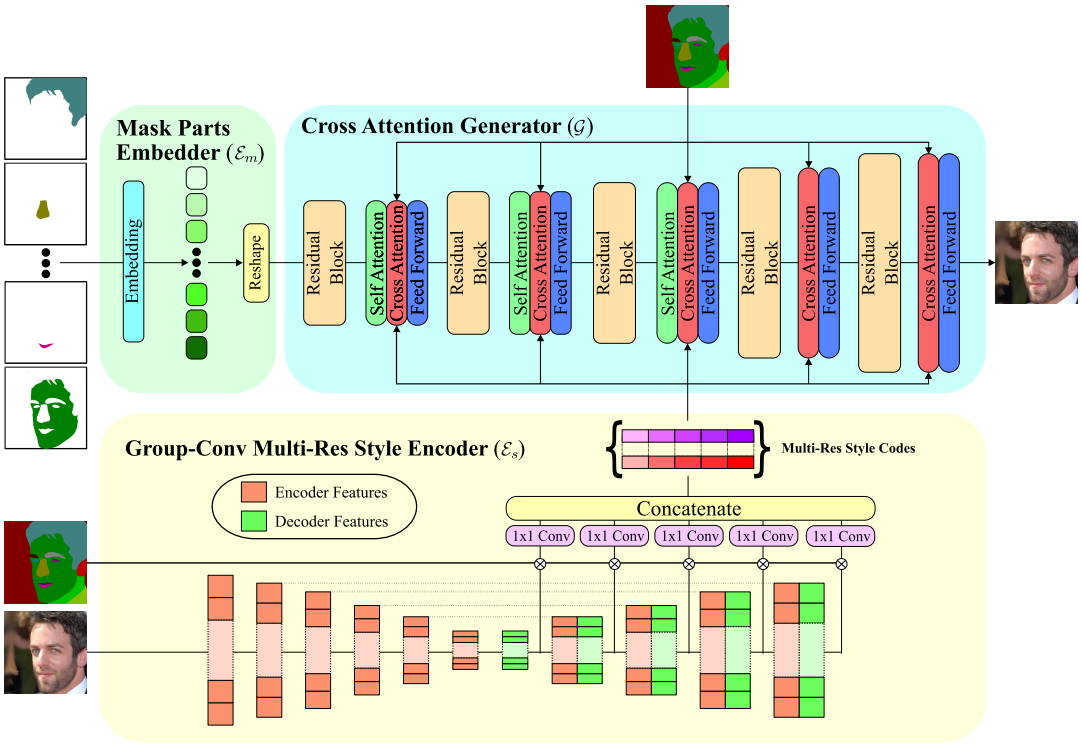

FIGURE 1. ***(CA)^2-SIS*** **architecture: style codes are extracted using a Multi-Resolution Style Encoder $\mathcal{E}_s$ equipped with grouped convolutions; the Mask Embedder $\mathcal{E}_m$ embeds each of the semantic mask parts into a set of latent codes; finally, these codes are fed to the Cross-Attention Generator $\mathcal{G}$ that is conditioned with the style codes thanks to the cross-attention mechanism. Additionally, the semantic mask is also used in the cross-attention layer to calculate the attention loss $\mathcal{L}_{att}$ in order to push each attention map to follow the mask shape.**

We identified most of these problems being due to using single-resolution features, which need to encode all the necessary information to reconstruct the image in a single set of feature maps. To mitigate this effect and address the loss of details induced by average pooling, we defined a deeper encoder to extract features at multiple scales. Although using multi-scale features was explored for mapping style features to a pre-trained StyleGAN generator [2], here we build further improvements upon the versatility brought by cross-attention. Specifically, we enhance the feature representation of each class by also employing grouped convolutions that allow capturing details at different granularity levels, mitigating the loss caused by pooling, while retaining its advantages in preventing trivial input-output mapping.

Grouped convolutions: Let a semantic mask be a C-channel image $\mathcal{M} \in \mathbb{N}^{C\times H\times W}$, where each channel $\mathcal{M}_j$ is a binary image encoding the spatial location of a specific class. Then, we define the i-th convolutional layer of the style encoder to have $(C \times f)$ filters *i.e.* one group for each semantic class, each group having f filters with their own learnable weights. Differently from previous solutions [6] which process all features together for each layer with classic convolutions, this design allows us to sample styles from different feature groups, each relative to a specific semantic class. Each layer is followed by group normalization to learns specific statistics of the features of a specific mask channel, and a ReLU.

Multi-resolution feature pooling: Let the style features of group j resulting from the i-th layer of the encoder be $\mathcal{F}_{i,j} \in \mathbb{R}^{f_j\times H_i\times W_i}$, then the class-wise style codes are extracted with an average pooling (AP) of the masked features:

$$S_{i,j} = AP\left(\mathcal{F}_{i,j} \cdot \mathcal{M}_j\right) \tag{1}$$

Note that, depending on the resolution of the i-th layer, the mask $\mathcal{M}$ is resized accordingly. This is done for each upsampling layer $i = 1, \cdots, L_{up}$ of the encoder and for each semantic class *i.e.* mask channel, j. Pooled features $S_{i,j} \in \mathbb{R}^{f_j}$ are then processed by a 1×1 convolutional layer so to reduce their dimensionality, and make all codes from each layer of the style encoder of the same size *i.e.* 256. They are then concatenated to form a code of size $C \times (L_{up} \cdot 256)$.

The overall architecture of our grouped multi-resolution style encoder ($\mathcal{E}_s$) is composed of 6 down-sampling layers and 5 up-sampling layers (i.e., L_{up}), linked together using skip connections. So, the final style codes have size $C \times 1280$.

B. MASK PARTS EMBEDDER

All SPADE-based semantic image synthesis methods use the raw semantic masks as input to the de-normalization layers since they require spatial information to apply feature maps modulations. With cross-attentions we get rid of this constraint. We hence pass the semantic mask through a Mask Embedder ($\mathcal{E}_m$) to obtain a latent representation. We will show (Table 4, Fig. 10) that this leads to both quantitative and

qualitative improvements. The design of $\mathcal{E}_m$ is very simple yet effective: we flatten each mask channel, and pass them *separately* through a linear layer, producing a latent code m_j of dimension 256 for each of the C channels. These C codes are then reshaped to form a $C \times 16 \times 16$ representation of the mask, that can be finally used as input to the generator.

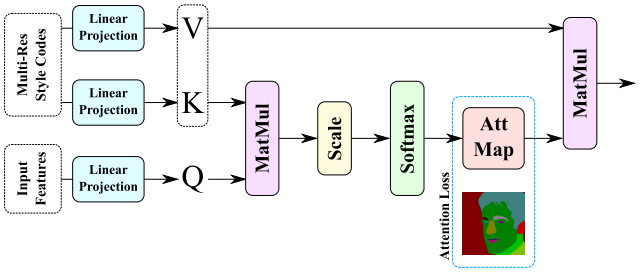

FIGURE 2. Cross-Attention layer: The Query Q is derived from the features of the previous residual block, while Key K and Value V are calculated starting from the multi-resolution style codes. Additionaly, an attention loss between the output of the Softmax and the semantic map $\mathcal{L}_{att}$ is calculated.

C. CROSS ATTENTION GENERATOR

The generator network is the main contribution of this work, as it significantly differentiates from the previous literature where custom normalization layers were used. Our generator is made up of 5 up-sampling blocks, each composed by a convolutional residual block and a spatial transformer block. More in detail, the spatial transformer blocks include a self-attention layer, a cross-attention layer and a feed-forward layer, following the structure of [12]. Given that the memory footprint of self-attention layers is quadratic with respect to the input, we use self-attention up to feature maps of size 64×64, therefore the last two attention blocks only employ cross-attention. Attention is defined as follows:

$$\mathcal{A}(Q, K, V) = \mathcal{M}_A(Q, K) \cdot V \tag{2}$$

where $\mathcal{M}_A(Q, K)$ is the attention map, defined as:

$$\mathcal{M}_A(Q, K) = \mathcal{S}\left(\frac{QK^T}{\sqrt{d}}\right) \tag{3}$$

d is the output dimension of each attention head in a multi-head configuration ($d = 64$, as in [9]), and $\mathcal{S}()$ is the Softmax activation function. In self-attention, Q, K and V are all obtained from the projection of the same embeddings, that is the flattened features $\phi^{(i)}$ of the generator at layer i:

$$Q = W_Q^{(i)} \cdot \phi^{(i)}, K \quad = W_K^{(i)} \cdot \phi^{(i)}, V = W_V^{(i)} \cdot \phi^{(i)} \tag{4}$$

Cross-attentions are meant to map the styles extracted by the encoder into the generator, in order to condition the generated samples. Hence, in cross-attention, K, V are computed from the style codes $\mathcal{E}_s(y)$ as follows:

$$Q = W_Q^{(i)} \cdot \phi^{(i)}, K \quad = W_K^{(i)} \cdot \mathcal{E}_s(y), V = W_V^{(i)} \cdot \mathcal{E}_s(y) \tag{5}$$

Attention Loss. By mixing style features $\mathcal{E}_s(y)$ with the flattened generator features $\phi^{(i)}$ which encode layout information from the semantic mask, cross-attentions implicitly learn how to map the style of each class into the relative spatial locations. This mapping is learned and encoded in the attention map $\mathcal{M}_A(Q, K)$. However, both self- and cross-attention compute long-range global dependencies, meaning that they capture correlations across different classes. This results in the attention maps $\mathcal{M}_A(Q, K)$ to spread beyond the class boundaries defined by the segmentation mask $\mathcal{M}$. In fact, certain styles might be class-wise correlated, *e.g.* hair color and skin tone, which is captured by the attention maps. On the one hand, this helps preserving the overall image quality and realism by forcing consistency across styles; on the other, it limits local style controllability (see Fig. 14).

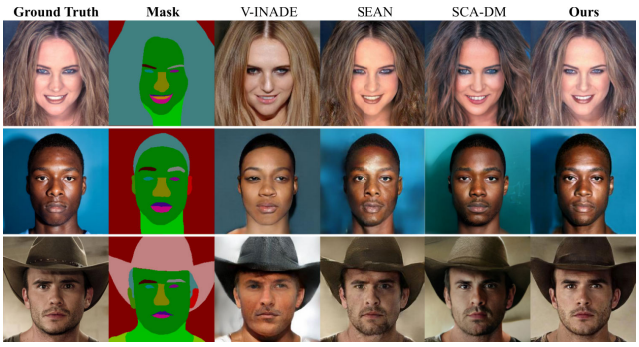

FIGURE 3. Qualitative comparison between state-of-the-art methods and our architecture on CelebMask-HQ. Our approach better preserves the color distribution (top row) and illumination coherence (second row).

In order to maximize both quality and controllability, during training we impose a loss term to push the learned attention maps $\mathcal{M}_A(Q, K)$ to follow the shape of the semantic mask $\mathcal{M}$ (Fig. 2). This allows to increase the style controllability without sacrificing the generation quality. The proposed *attention loss* is a binary cross-entropy (BCE) computed between the attention map and the semantic mask. Since, in cross-attention, the key values are the projected style codes, the attention map $\mathcal{M}_A(Q, K)$ has size $h \times C \times H \times W$, where h is the number of attention heads, C is the number of semantic classes and H, W the height and width of the feature maps. Thus, it is possible to impose that all the h attention maps of the c-th semantic class match the c-th input mask channel pixel-wise. The loss for a semantic class c is computed as:

$$\mathcal{L}_{att}^c = \frac{1}{HWh} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{h} y_{cij}^k \log(x_{cij}^k) + (1 - y_{cij}^k) \log(1 - x_{cij}^k) \tag{6}$$

where y_{cij}^k is the class of pixels at channel c, row i, column j, for the k-th head, and x_{cij}^k the corresponding attention map value. Fig. 14 shows how, by imposing this loss when swapping styles, the entanglement is greatly reduced.

IV. EXPERIMENTAL RESULTS

Implementation details The model is trained to reconstruct a target image from its semantic mask. This is enforced by means of an adversarial loss, a feature matching loss [14]

TABLE 1. Comparison with the state-of-the-art in terms of FID, mIoU and segmentation accuracy (Acc.). Best results in bold. Highlight indicates the best result among our variants and the competitors. "*" indicates MaskGan is trainable only for CelebMask-HQ. Results marked with "-" are not available since training SCA-DM on Ade20K requires more than one month.

	CelebMask-HQ			Ade20k			DeepFashion		
Method	**FID ↓**	**mIoU ↑**	**Acc. (%) ↑**	**FID ↓**	**mIoU ↑**	**Acc. (%) ↑**	**FID ↓**	**mIoU ↑**	**Acc. (%) ↑**
Pix2PixHD [14]	22.26	78.40	92.88	66.65	28.47	63.78	15.33	89.52	98.47
SPADE [1]	21.08	78.32	92.76	53.70	44.21	69.05	11.18	**92.87**	**99.11**
MaskGAN [5]	59.91	76.34	87.89	*	*	*	*	*	*
SEAN [6]	18.72	**78.62**	93.54	38.63	43.82	67.42	10.70	92.19	98.72
V-INADE [4]	17.49	78.04	93.50	39.87	45.93	**70.58**	10.33	92.40	98.85
SCA-DM [27]	16.74	77.49	93.65	-	-	-	10.93	89.94	98.46
Ours - w/o AttLoss	**15.80**	78.01	93.42	**38.08**	45.72	69.04	**9.97**	90.95	98.56
Ours - Full	15.84	78.54	**93.78**	38.34	**46.05**	69.27	10.63	91.24	98.61

and a perceptual loss [30]. An attention loss is also used, as described in Sec. III-C. We employ a multi-scale discriminator [1] and train the model for 100 epochs on a NVIDIA A100 GPU, using the Adam optimizer with a learning rate of 0.0002. Generated images have size 256×256.

Datasets. Experiments are conducted on three datasets: CelebAMask-HQ [5], Ade20k [31] and DeepFashion [32].

Metrics. We employ the following metrics for evaluation: the Frechet Inception Distance (FID), to estimate the generation quality; a semantic-segmentation-based metric is used to compare the ground-truth layouts against those obtained by running a pre-trained segmentation method on the generated images. It evaluates the mean Intersection-over-Union (mIoU) and pixel accuracy. We employ *FaceParsing*[1] for CelebAMask-HQ and DeepFashion, and *SceneSegmetantion*[2] for Ade20k, respectively.

A. RECONSTRUCTION

We first report a quantitative comparison against state-of-the-art works in terms of semantic synthesis quality (Table 1) of two versions of the proposed method: with and without attention loss. As discussed in Sect. II, most methods are diversity-driven, with the exception of SEAN [6] which is purely reference-based. However, some of them are equipped with a style encoder that allows them to use a reference image. Thus, here we provide a comparison with such, in particular Pix2PixHD [14], SPADE [1], MaskGAN [5], SEAN [6], INADE [4] when equipped with its style Variational autoencoder (V-INADE) and SCA-DM [27]. A comparison against diversity-driven approaches is in Sect. IV-G.

Table 1 shows that our novel solution based on shape-style attention blocks performs competitively with respect to state-of-the-art methods, obtaining the best FID score in all datasets. Interestingly, when adding the attention loss, FID score gets slightly worse; this can be explained by the fact that imposing stronger disentanglement in the attention maps means less freedom in the generation, which is always reflected in a slightly degraded FID score. Unfortunately

[1]https://github.com/switchablenorms/CelebAMask-HQ/tree/master/face_parsing

[2]https://github.com/CSAILVision/semantic-segmentation-pytorch

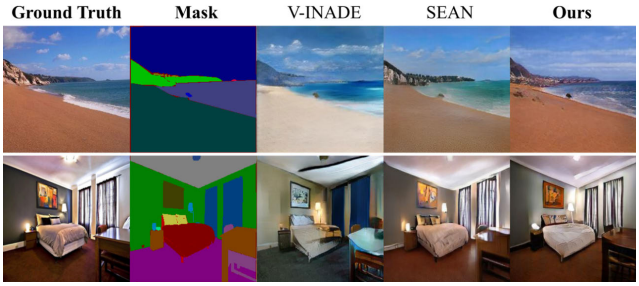

FIGURE 4. Qualitative comparison between state-of-the-art methods and our architecture on Ade20k. Our approach better preserves the color distribution (top row) and illumination coherence (second row).

though, the FID score does not fully reflect the quality improvement obtained by using cross-attentions in place of SPADE, which can be better appreciated in the qualitative examples in Fig. 3, 4 and 5. The reader can appreciate how our method better preserves the overall color distribution such as the skin tone (Fig. 3 top row) or sand color (Fig. 4 top row), and more complex details such as the cast shadow (Fig. 3, second row), or the clothing folds (Fig. 4, bottom row).

FIGURE 5. Qualitative comparison between state-of-the-art methods and our architecture on Deepfashion.

On the other hand, we obtain slightly lower performance for segmentation accuracy measures (mIoU and accuracy). A reasonable explanation is that SPADE provides hard pixel-wise guidance by using the semantic mask (resized according to the size of each layer) in *all* the generator layers to modulate the activations at specific spatial locations. This is done both during training and inference. Instead, we *do not directly inject* layout information in the generator, but only use the (embedded) semantic mask as input, and during

training for computing the attention loss $\mathcal{L}_{att}$, providing weaker supervision. Nonetheless, performance are totally comparable. This is a clear hint that, even without the need of an explicit spatial mapping, cross-attention do automatically learn meaningful shape-style relationships. This downside is partially avoided when employing the attention loss; imposing a constraint to localize the attention has a positive effect over the shape related measures.

It is worth noting that for CelebAMask-HQ our method performs significantly better compared to all methods in terms of FID score. A reason is that in CelebAMask-HQ there is a higher level of style correlation compared to other datasets. In human faces, eyes, hair and skin tone are highly correlated and this correlation is captured by attention blocks. The same applies to a lesser extent for Ade20k, mostly for outdoor scenes, where grass color is likely correlated with that of the trees, or sea color with that of the sand.

In terms of running time, our method is comparable to the other approaches (0.07 seconds per image), but significantly faster than SEAN (0.16 seconds per image) and diffusion-based approaches such as SCA-DM (5 seconds per image).

B. STYLE TRANSFER

In Fig. 6 we report some examples of style transfer of all face parts from one reference face to another one. Note that we did not put effort into picking particular examples, yet we chose some that best highlight the limitations of previous solutions. Our method achieves significantly more realistic transfer results even in very challenging cases. Our model is able to correctly integrate texture even when it is not present in the reference image but its semantic class exists in the target image (*e.g.* opposite ear in the top row, eyeglasses in the middle row, teeth in the bottom row). On the opposite, SEAN cannot handle such cases. This is possible thanks to the cross-attention layers which learn shape-style correlations.

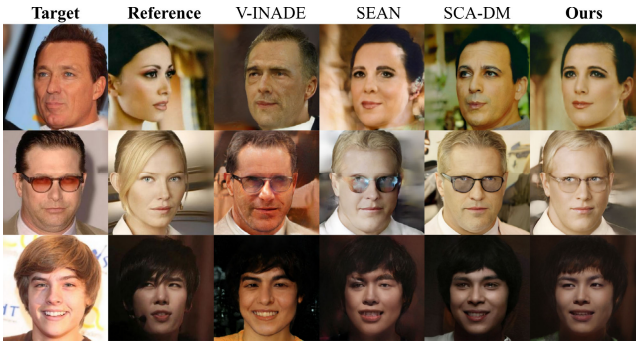

FIGURE 6. Comparison between the state-of-the-art and *(CA)²-SIS*.Our method convincingly transfers the style of all face parts even if a specific style *i.e.* opposite ear in the top row, eyeglasses in the middle row, and teeth in the bottom row, are absent in the reference image.

In Fig. 7 and 8 we also report some examples of style transfer at the class level. Our model can accurately apply local styles and generate realistic results even for challenging cases (*e.g.* different poses) without sacrificing image consistency, and maintaining complex details such as global illumination coherence. We note this is a remarkable result given that we *do not use any explicit spatial information* in the intermediate layers to guide the style mapping into the generator features.

FIGURE 7. Style transfer of face parts. *(CA)²-SIS* can perform local style manipulation even without spatial information in the generator.

FIGURE 8. Style transfer of single parts in ADE20k and DeepFashion.

FIGURE 9. Transferring the style of local parts can lead to overall image inconsistencies, for example, if the target and reference image differ in illumination, or if changing highly correlated parts. Methods using spatially-adaptive normalization layers (i.e. SEAN) cannot handle such cases. Our solution fixes this effect to a large extent.

Fixing Image Consistency. Performing style transfer – or editing – of local features is a tricky task, as it can cause global inconsistencies to arise in the manipulated image. For example, this can happen when trying to mix the style of images with different illumination, or if trying to transfer the style of strongly correlated classes. In Fig. 9 some examples are depicted where we compare the transfer of two parts, the nose and left eye. SEAN is able to precisely swap the style of the single parts yet at the cost of making the final image resulting overall unrealistic and inconsistent. Our method, instead, tends to fix this lack of consistency; in particular, it does so by, in one case (Fig. 9, third column) removing the unnatural cast shadow by adjusting the overall illumination and shades, and, in the other case (Fig. 9, fourth column) by adjusting the style of the symmetric eye, as eyes color is normally the same in human faces. This behavior is two-faced: on the one hand, it restricts the ability of the model to precisely apply local styles; on the other, it maintains a high level of overall visual quality and realism. Note that all the results in Fig. 9 are obtained using attention loss, which retains a good trade-off between generation quality and controllability.

C. SHAPE TRANSFER AND MANIPULATION

Unlike previous methods, in our framework the semantic mask is embedded into a set of class-wise latent codes. We briefly anticipated that in certain circumstances, other than resulting in improved reconstruction, this allows for additional advantages such as the possibility of globally or even locally manipulating the shape. In Fig. 10 we report some qualitative examples of *shape transfer*, *i.e.* changing the embedding of a semantic class from a reference mask to a target one. Again, we compare our solution with SEAN as representative of previous SPADE-based methods. Using the raw mask induces several artifacts (holes) in the generated images even for slight misalignment of face parts. This spatial inconsistency cannot be handled by methods that use SPADE to explicitly inject the style into specific spatial locations. On the other hand, our solution allows for fixing such minor inconsistencies, opening the way to fully automatic control of both style and shape. This advantage is confirmed by comparing the columns FID and FID-sh in Table 4, which reports the FID score obtained on reconstructed images versus shape-manipulated ones.

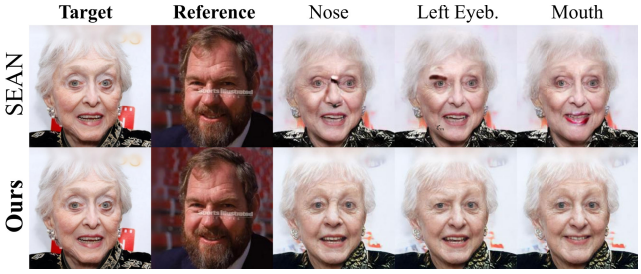

FIGURE 10. Shape transfer comparison between SEAN and our final architecture. SEAN cannot handle automatic shape changes, and require manual intervention to fix the mask. Our method equipped with the mask embedder $\mathcal{E}_m$ can instead automatically do so.

Shape interpolation Other than transferring the shape of face parts from one image to another, we can extend this ability to perform geometry *interpolation*. The only other method able to do so is MaskGAN [5]. However, it can only perform global mask interpolations and needs an additional alpha-blender to refine the generated images. Differently, we are able to linearly interpolate any arbitrary channel j from two masks $\mathcal{M}^1$ and $\mathcal{M}^2$, and generate an interpolated mask embedding as $m_j^{int} = \alpha \cdot \mathcal{E}_m(\mathcal{M}_j^1) + (1-\alpha) \cdot \mathcal{E}_m(\mathcal{M}_j^2)$. Fig. 11 reports some examples of face images generated in this way. Interestingly, when interpolating the whole mask, the 3D head pose is also changed. On the one hand, this is an intriguing effect that could open the way to novel applications; on the other, it can be problematic when manipulating local parts, in which case inconsistencies still occur if not properly handled.

FIGURE 11. Interpolation between two mask embeddings. The generated images naturally transition between two shapes.

D. COMPARISON WITH STYLEGAN-BASED METHODS

In the field of human face generation, StyleGAN [22] and all its subsequent improvements are among the most effective unconditional face generation models to date. Indeed, almost all recent methods for face generation and

manipulation are based on a pre-trained StyleGAN model. Being purely generative though, they cannot be directly used for reconstructing and editing *real faces*. However, a workaround has been found, which technique is referred to as *GAN-Inversion* [33]. It is based on the idea of applying some optimization for finding the embedding in the StyleGAN latent space that best approximates a given real face image. Once this latent is found, it is possible to edit the reconstructed face by applying latent manipulation techniques.

Using GAN-Inversion, remarkable results have been achieved. Yet, a question is whether the inversion process is sufficiently accurate to *(i)* produce an image where the identity of the subject is sufficiently preserved, *(ii)* maintain the level of quality and realism achieved in a purely generative setting and *(iii)* be applied to out-of-domain data *i.e.* faces not included in CelebA-HQ. In this section, we aim at comparing the results of our method against the most recent alternative based on StyleGAN, which is SemanticStyleGAN [34]. This specific approach also performs local face editing based on a semantic mask. Results are reported in Table 2. The performance of SemanticStyleGAN drops dramatically when applied in a GAN-inversion setting. The quality of the generated images decreases significantly as evidenced by the FID score which increases to 26.83 from 6.42 (value taken from [34]) of the fully generative setting. Also, the ability of retaining the identity of the subject is compromised. We measure the latter using the Face Recognition Similarity (FRS) score: it is computed by embedding both real and reconstructed (inverted) images using an InceptionV3 pre-trained on VGG-Face2 dataset [35], and then computing the cosine similarity between embedding pairs. Our *$(CA)^2$-SIS* scores a way higher similarity, indicating the identity is more effectively preserved. This is an essential feature for methods that aim at accurate editing of real images.

TABLE 2. Comparison against Semantic StyleGAN on CelebMask-HQ.

Method	FID ↓	FRS ↑
Semantic StyleGAN [34]	26.83	0.68
Ours	**15.82**	**0.81**

Methods based on StyleGAN are extremely promising in the generation task, yet these results evidence that they still fall short when manipulating a real image is the goal as they fail to apply the inversion process in a sufficiently accurate way. Another drawback is related to the worse generalization to out-of-distribution data. In Figure 12, we qualitatively compare *$(CA)^2$-SIS*, SemanticStyleGAN and SEAN on some images from FFHQ [22], a commonly used dataset for training generative models. Semantic masks are not provided in FFHQ, so we employed the same face parser of Section IV to generate them. All three models were trained on CelebA-HQ. SemanticStyleGAN fails dramatically in maintaining the identity traits of the subjects, mostly for severely under-represented classes such as the eyeglasses

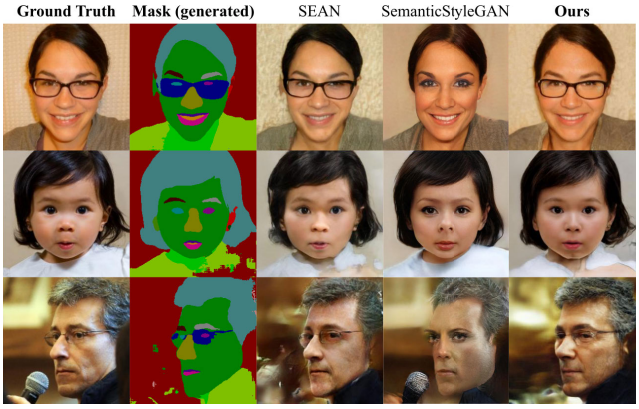

FIGURE 12. Results on FFHQ with a model trained on CelebA-HQ.

(top row), babies (middle row) or non-frontal faces (bottom row). SEAN instead can better preserve the identity and handle unseen samples, but results less robust to noise in the masks. Our model more robustly handles detection noise as compared to SEAN. This is another advantageous result of embedding the mask into a latent vector and relying on attention mechanism.

E. USER STUDY

In this section, we report some results of a user study conducted on CelebMask-HQ. Following the protocol of [4] or [36] (20 images are shown to 20 participants), we recruited 30 volunteers and asked them to select the most realistic result among those generated by different methods given the same input on 30 randomly selected images. The average percentage of times that users selected a specific method is reported in Fig. 13 (Reconstruction). Participants picked our results as the most realistic in the majority of cases. Using attention mechanisms to condition the image generation has the advantage that global style correlations are preserved more faithfully than using adaptive-normalization layers, which treat each class (or instance) independently, eventually making the generated image perceived as less natural.

In addition, we also asked the participants to evaluate the results for the task of style transfer. Here we compared only against SEAN. V-INADE was excluded; even if equipped with the variational style encoder, its goal is to approximate the class distribution so as to generate diverse outputs. As shown also in Fig. 6, it does not perform well on style transfer as it was not designed for that goal. We applied the style of a reference image A to the semantic mask of another image B, for 10 random image-mask pairs. Results in Fig. 13 (Style Transfer) set off our method as the most realistic.

F. ABLATION STUDY

To make the contribution of the attention loss explicit, in Figure 14 we provide some qualitative examples. We take one target image and apply the style of a face part taken from four different reference images. Then, we compute

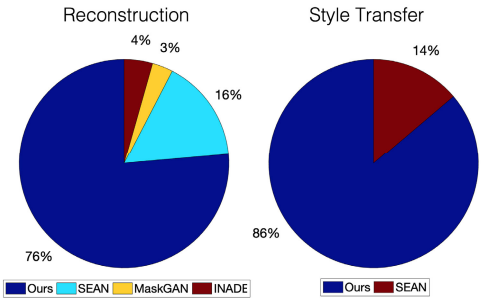

FIGURE 13. User study. Both charts show *(CA)²-SIS* is preferred by users, who selected our results as the most realistic in the majority of cases.

the average $L1$ difference and convert it to a heatmap. The latter clearly evidences that the attention loss leads to a more localized editing. Without it, the cross-attention tends to expand its influence beyond the part of interest. For example, in the top row, eyes are edited, but some manipulation to the hair is applied as well. The attention loss corrects this (second row).

Finally, in Table 3 and Table 4, we report a detailed ablation study on different architectural designs. We performed three sets of experiments, exploring several configurations for the style encoder and the generator, respectively, and for the task of shape transfer, as detailed below.

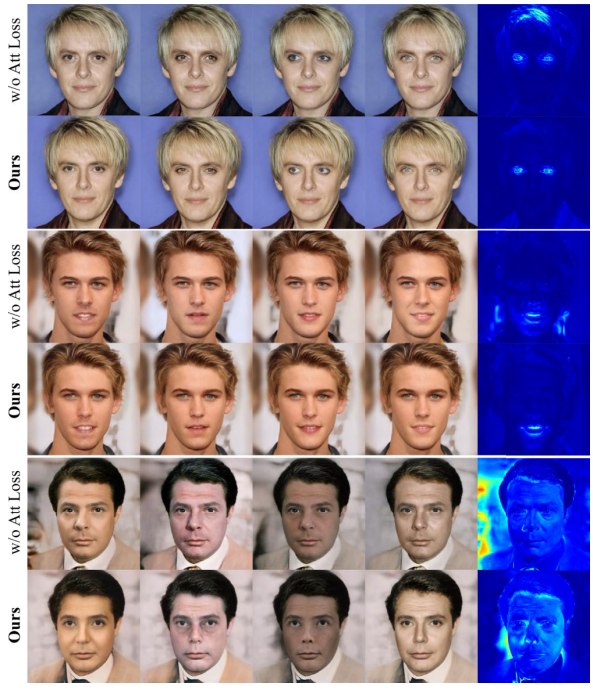

FIGURE 14. Local style transfer with and without attention loss. The same target image is edited by transferring the style of a face part (top: eyes, middle: mouth, bottom: skin) from 4 different images. The heatmpas show the average $L1$ difference across the edited images and the original one. The attention loss leads to more localized editing.

1) STYLE ENCODER

We first trained a version of our architecture equipped with the style encoder of SEAN [6] (Ours - SEAN $\mathcal{E}_s$). Compared to pure SEAN (FID 18.72), our FID score is lower (FID 17.81), indicating the cross-attention-based generator leads to better results given the same style features. Then, we tested a version of our multi-scale style encoder without group convolutions (Ours w/o GC). Compared to our full model, gathering features with group convolutions enables capturing class-specific texture details at different levels of granularity.

2) GENERATOR

Regarding the generator, removing the mask embedder (Ours w/o $\mathcal{E}_m$) worsens the quality. Indeed, embedding the mask to remove spatial information demonstrated beneficial to better capture shape-style correlations with cross-attention modules. Finally, we replaced cross-attention layers with SPADE layers (Ours w/o CA). These do not exploit well the multi-resolution styles as compared to cross-attention.

TABLE 3. Ablation study on CelebMask-HQ. "*CA*" stands for cross-attention. "*GC*" stands for group convolutions.

Style Encoder	FID ↓
Ours - SEAN $\mathcal{E}_s$	17.81
Ours w/o GC	16.42
(CA)²-SIS	**15.84**

Generator	FID ↓
Ours w/o $\mathcal{E}_m$	18.65
Ours w/o CA	18.14
(CA)²-SIS	**15.84**

3) SHAPE TRANSFER

In Table 4, we report a quantitative comparison for the task of shape transfer. Those are computed by swapping a random part (random seed is fixed) across pairs of images, without imposing any constraints for choosing the pairs. Results are crystal clear; the FID in case of shape transfer (FID-Sh) of both SEAN and V-INADE increases significantly, while it remains stable for our method, supporting the qualitative results shown in Figure 10.

TABLE 4. Results for shape (Sh) transfer on CelebMask-HQ.

Method	FID ↓	FID-Sh ↓
SEAN	18.72	20.04
V-INADE	17.49	31.16
(CA)²-SIS	**15.84**	**15.90**

The above experiments highlight how the proposed model is a promising alternative with respect to SPADE and variants as it allows for *(i)* more versatility in designing proper style features, *(ii)* embedding the semantic mask into latent vectors, leading to more freedom in geometry manipulation and *(iii)* improved image quality without sacrificing editing control.

G. DIVERSITY-DRIVEN GENERATION

The main goal of this work was to explore an alternative architecture for semantic image synthesis guided by a reference style image. However, the proposed generator architecture based on spatial transformer layers in place of SPADE is general in its purpose and can be potentially used

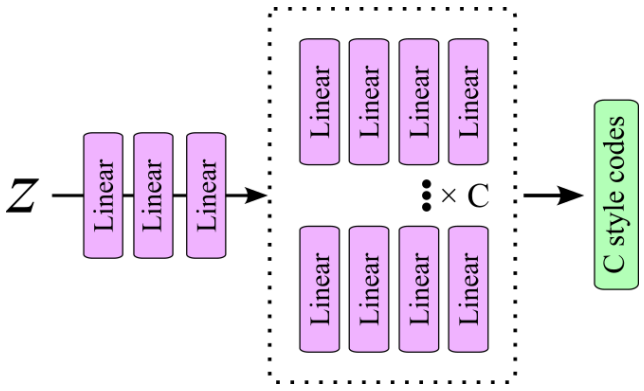

FIGURE 15. Mapping network $\mathcal{M}_s$ architecture: a noise vector $z \sim \mathcal{N}(0, 1)$ is processed to obtain a set of C style codes.

also in a fully-generative setting. Clearly, the latter requires specific training strategies to force both quality and diversity, which is beyond the scope of this work. Nevertheless, in this section we report the results of an additional set of experiments to evaluate *(CA)*^{2}*-SIS* adapted for the diversity-driven setting.

To achieve this, we exploit the pre-trained *(CA)*^{2}*-SIS* architecture as described in Section III, but we substitute the style encoder $\mathcal{E}_s$ with a mapping network $\mathcal{M}_s$ designed to map a noise vector into the latent style codes (see Fig. 15).

The design of the mapping network $\mathcal{M}_s$ is quite simple: it takes a noise vector $z \sim \mathcal{N}(0, 1)$ as input, which passes through 3 linear layers, each outputting a latent vector of size 512. Then, the output of the third layer branches into C paths, where C is the number of semantic classes. Each branch is composed of 4 linear layers. The first three output a latent vector of size 512, while the last one's output has size $1,280$, which is needed to match the size of the multi-scale style codes resulting from $\mathcal{E}_s i.e. 256 \times 5$. The output of the mapping network $\mathcal{M}_s$ becomes the *Key* and *Value* for the cross-attention layers; recalling Eq. (2), that is:

$$Q = W_Q^{(i)} \cdot \phi^{(i)}, K = W_K^{(i)} \cdot \mathcal{M}_s(z), V = W_V^{(i)} \cdot \mathcal{M}_s(z) \quad (7)$$

TABLE 5. Comparison with GAN and DM based state-of-the-art method in terms of FID, and LPIPS. Best results in bold, second best underlined.

Type	Method	FID ↓	LPIPS ↑
GAN	SC-GAN [3]	20.85	0.170
	INADE [4]	**18.31**	0.365
DM	SDM [8]	32.96	0.391
	ControlNet [26]	41.90	**0.606**
	CoDiff [25]	22.69	<u>0.382</u>
	Ours	<u>19.79</u>	0.367

To train the mapping network, we exclude the style encoder $\mathcal{E}_s$, keep the generator G frozen and re-initialize the discriminator. The training is guided by only two losses: the discriminator loss $\mathcal{L}_D$ and a diversity loss $\mathcal{L}_{dv}$. The latter maximizes the $L1$ discrepancy between two images I_1, I_2 generated from two noise vectors z_1, z_2 given the same semantic mask m, that is $\mathcal{L}_{dv} = -\|G(\mathcal{M}_s(z_1), m) - G(\mathcal{M}_s(z_2), m)\|_1$. We train it for 100 epochs, using the same parameters as in Sect. IV.

Results on CelebMask-HQ are shown in Table 5, in comparison with both GAN and diffusion-based state-of-the-art approaches, namely INADE [16], SC-GAN [3], SemanticDiffusion [8], ControlNet [26] and Collaborative Diffusion [25]. Note that, differently from the methods in Table 1, these are purely diversity-driven and do not use any reference style images. In terms of image quality (FID) and diversity (LPIPS [37]), our method performance are comparable to previous GANs equipped with SPADE layers (INADE) or other custom modules such as conditional convolutions (SC-GAN). We remark that our method was not designed for a diversity-driven setting, evidencing that cross-attentions are a promising replacement for previous alternatives.

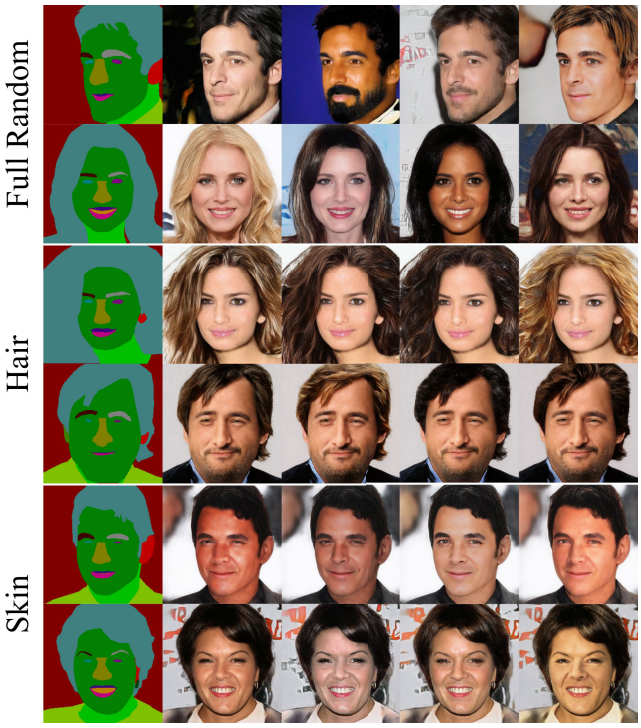

FIGURE 16. Images generated with our model trained for diversity. In the first two rows all the 19 style codes corresponding to the CelebMask-HQ semantic classes are generated by the mapping network $\mathcal{M}_s$. In the remaining rows, we kept all the style codes fixed except hair and skin to demonstrate the disentanglement capability of the model.

To compare against diffusion-based methods, we tested them using the DDIM [38] technique to reduce the diffusion steps and make the computational times comparable. With $1,000$ steps, for instance, SDM [8] takes 8 minutes per image on a NVIDIA A100 GPU, which makes generating the whole test set (2K images) impractical. By using DDIM with 50 steps, times reduce to 10 seconds per image. Despite reduced, such is still way larger compared to *(CA)*^{2}*-SIS*, which generates an image in 0.07 seconds. In general,

compared to diffusion-based approaches, our method reports a better FID score but lower LPIPS. Whereas DMs are known to have an excellent ability of generating diverse samples, we again remark that our method was not specifically tailored for such task. To compare against ControlNet [26] in a fair setting, we excluded the text condition, and re-trained it using only semantic masks and the recommended hyper-parameters. Without textual clues, the generation ability is severely compromised, with a FID score that is largely worse than our approach, even though it stands out in terms of diversity. The same holds, although to a lesser extent, for Collaborative Diffusion [25]. In this case, we used the single modality pre-trained model that was publicly released.

Finally, our solution allows for a nice feature not shown by previous methods, that is generating diverse and disentangled styles for different semantic classes. For example, one can keep the style of some parts fixed while only generating that of a specific class, such as skin or hair, allowing for a detailed generation control over local classes. Some qualitative results are shown in Figure 16.

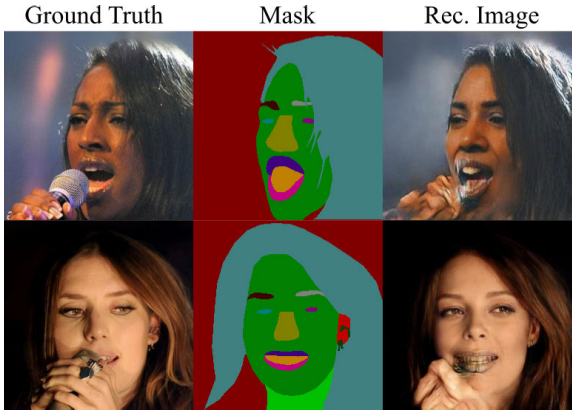

FIGURE 17. Some examples of edge cases in which the semantic mask does not faithfully reflect the original image resulting in artifacts.

V. LIMITATIONS, CONCLUSION, AND FUTURE WORKS

In this section, we discuss some issues, limitations, and peculiar behaviors of our solutions. First, although shape transfer results clearly show that our model can manipulate the shape significantly better than previous works, in case of strong misalignment or large translation, gaps and inconsistencies still occur. This prevented us from successfully applying shape transfer on other datasets (CelebMask-HQ is particularly suitable as face images are inherently quite well aligned). Nevertheless, the ability of our model to learn to apply shape changes without specific supervision, achieved thanks to the proposed paradigm change, is valuable and can open the way to further improvements. From a different perspective, as can be noted from Fig. 6, using cross-attention layers in place of adaptive normalization ones is advantageous as they can fix the generation process for an overall increased image consistency. On the one hand, in cases where parts styles are strongly correlated, cross-attention can prevent full-style control over single classes (eyes color in Fig. 9). On the other hand, style transfer outcomes look significantly more realistic and consistent. Finally, in Fig. 17 we show some examples of imperfect reconstruction due to missing semantic parts in the input mask. This issue is common in semantic image synthesis models and still affects our model that, when generating the reconstructed samples, it fails to produce a coherent result.

In conclusion, in this paper we proposed a novel architecture that uses attention mechanisms as an alternative to SPADE-like normalization layers to perform semantic image synthesis. In addition, we employ an attention loss to force the attention in the model to match the semantic mask. This has the effect of reducing entanglement between different styles and maintaining the semantic mask shape in the generated samples. We provided a detailed analysis of its values and limitations w.r.t. prior art, and a preliminary solution to the new task of automatic manipulation of local geometry.

As another valuable trait of our proposed solution, we mention the possibility of extending this framework to include additional conditioning information other than styles. This is made possible thanks to the high flexibility of cross-attention layers. This is particularly helpful for face editing where pose, expression and identity could be used to control the generated samples. In a recent preliminary investigation, we experimented with injecting identity information in the form of embedding extracted by a face recognition model into the generator [39]. The versatility of cross-attentions allowed to seamlessly fuse low-level (style and texture) and high-level (identity) information to further push the realism of the generated images. In [39] we also showed how this feature could be used to perform effective adversarial attacks. Extending this strategy to other characteristics such as age, expression or pose could lead to a general, fully controllable framework for face editing that can be also extended to different contexts.

ACKNOWLEDGMENT

The authors would like to thank the CINECA Award under the ISCRA Initiative (Project: IsCa5 X-SYS), for the availability of computing resources and also would like to thank Andrea Pilzer and Giuseppe Fiameni of NVIDIA for the support.

REFERENCES

[1] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2337–2346.

[2] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2287–2296.

[3] Y. Wang, L. Qi, Y.-C. Chen, X. Zhang, and J. Jia, "Image synthesis via semantic composition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13749–13758.

[4] Z. Tan, M. Chai, D. Chen, J. Liao, Q. Chu, B. Liu, G. Hua, and N. Yu, "Diverse semantic image synthesis via probability distribution modeling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7962–7971.

[5] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5549–5558.

[6] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "SEAN: Image synthesis with semantic region-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5104–5113.

[7] X. Liu, G. Yin, J. Shao, X. Wang, and H. Li, "Learning to predict layout-to-image conditional convolutions for semantic image synthesis," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Jan. 2019, pp. 1–9.

[8] W. Wang, J. Bao, W. Zhou, D. Chen, D. Chen, L. Yuan, and H. Li, "Semantic image synthesis via diffusion models," 2022, *arXiv:2207.00050*.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Jun. 2017, pp. 5998–6008.

[10] C. R. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 357–366.

[11] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Jan. 2022, pp. 36479–36494.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10684–10695.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[14] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[15] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "GauGAN: Semantic image synthesis with spatially adaptive normalization," in *Proc. ACM SIGGRAPH Real-Time Live!*, Jul. 2019, pp. 1–2.

[16] Z. Tan, D. Chen, Q. Chu, M. Chai, J. Liao, M. He, L. Yuan, G. Hua, and N. Yu, "Efficient semantic image synthesis via class-adaptive normalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4852–4866, Sep. 2022.

[17] T. Fontanini, C. Ferrari, M. Bertozzi, and A. Prati, "Automatic generation of semantic parts for face image synthesis," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, Jan. 2023, pp. 209–221.

[18] T. Fontanini, C. Ferrari, G. Lisanti, L. Galteri, S. Berretti, M. Bertozzi, and A. Prati, "FrankenMask: Manipulating semantic masks with transformers for face parts editing," *Pattern Recognit. Lett.*, vol. 176, pp. 14–20, Dec. 2023.

[19] Y. Li, Y. Li, J. Lu, E. Shechtman, Y. J. Lee, and K. K. Singh, "Collaging class-specific GANs for semantic image synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14418–14427.

[20] Y. Shi, X. Liu, Y. Wei, Z. Wu, and W. Zuo, "Retrieval-based spatially adaptive normalization for semantic image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11224–11233.

[21] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele, and A. Khoreva, "OASIS: Only adversarial supervision for semantic image synthesis," *Int. J. Comput. Vis.*, vol. 130, no. 12, pp. 2903–2923, Dec. 2022.

[22] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[23] Y. Shi, X. Yang, Y. Wan, and X. Shen, "SemanticStyleGAN: Learning compositional generative priors for controllable image synthesis and editing," 2021, *arXiv:2112.02236*.

[24] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, Jan. 2021, pp. 8780–8794.

[25] Z. Huang, K. C. K. Chan, Y. Jiang, and Z. Liu, "Collaborative diffusion for multi-modal face generation and editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 6080–6090.

[26] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," 2023, *arXiv:2302.05543*.

[27] A. Ergasti, C. Ferrari, T. Fontanini, M. Bertozzi, and A. Prati, "Controllable face synthesis with semantic latent diffusion models," 2024, *arXiv:2403.12743*.

[28] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22500–22510.

[29] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," 2022, *arXiv:2208.01618*.

[30] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, Jan. 2016, pp. 694–711.

[31] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 633–641.

[32] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.

[33] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain GAN inversion for real image editing," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2020, pp. 592–608.

[34] Y. Shi, X. Yang, Y. Wan, and X. Shen, "SemanticStyleGAN: Learning compositional generative priors for controllable image synthesis and editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 11254–11264.

[35] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[36] Z. Zhu, Z. Xu, A. You, and X. Bai, "Semantically multi-modal image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5467–5476.

[37] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[38] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent.*, Jan. 2020, pp. 1–12.

[39] G. Tarollo, T. Fontanini, C. Ferrari, G. Borghi, and A. Prati, "Adversarial identity injection for semantic face image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 1471–1480.

TOMASO FONTANINI received the Ph.D. degree in information engineering from the University of Parma, in 2021, under the supervision of Prof. Andrea Prati. In 2020, he was a Visiting Student with the VisLAB Laboratory, University of California at Riverside (UCR), under the supervision of Prof. Bir Bhanu. He is currently an Assistant Professor with the Department of Engineering and Architecture, University of Parma. His current research interests include image generation and manipulation of facial attributes and styles in an unsupervised setting. In the past, he worked on applying meta-learning techniques to generative adversarial networks and image retrieval. He has co-authored several publications about these topics in journals and international conferences.

CLAUDIO FERRARI received the M.Sc. degree (cum laude) in computer engineering and the Ph.D. degree in information engineering from the University of Florence, in 2014 and 2018, respectively. He is currently an Assistant Professor with the University of Parma. In 2014, he spent six months as a Visiting Research Scholar with the IRIS Computer Vision Laboratory, University of Southern California (USC), working on the problem of pose invariant face recognition, under the supervision of Prof. Gerard Medioni. His research interests include the analysis of humans, including face recognition, emotions and expressions analysis, 3D face and body modeling and reconstruction, and deep generative models for 2D/3D/4D face generation. On these topics, he has authored or co-authored more than 40 publications in international journals and conference proceedings.

GIUSEPPE LISANTI is currently an Associate Professor with the Department of Computer Science and Engineering, University of Bologna. He has co-authored more than 40 papers in the most prestigious journals and conferences in computer vision, multimedia analysis, pattern recognition, and machine learning. He is involved in research collaborations with other research centers, both national and international. He has actively participated in various roles in several research projects and a number of industry-funded projects. He received the Best Paper Award from the IEEE Computer Society Workshop on Biometrics, in 2017.

MASSIMO BERTOZZI received the Dr. (Eng.) degree in electronic engineering from the University of Parma and the Ph.D. degree in information technology from the Dipartimento di Ingegneria dell'Informazione, Universitã di Parma, in October 1997. He chaired the local IEEE Student Branch with the Universitã di Parma. He discussing his master's thesis about the implementation of Simulation of Petri Nets on the CM-2 Massive Parallel Architecture. Since November 1997, he has been holding a permanent position. He is currently an Associate Professor with the Department of Engineering and Architecture. His research interests include the application of image processing to real-time systems and specifically to autonomous driving, on the optimization of machine code at assembly level, and on parallel and distributed computing, use of CNN for autonomous driving and industrial inspection. From 2015 to 2020, he was with VisLab on the development of SW and system integration for the Ambarella CVflow chip architecture. Since 2020, he has been focusing his research interests to CNN in cooperation with the IMP Laboratory, research group.

ANDREA PRATI (Senior Member, IEEE) received the Graduate degree in computer engineering and the Ph.D. degree from the University of Modena and Reggio Emilia, in 1998 and 2002, respectively. He was an Assistant Professor, from 2005 to 2011, and has been an Associate Professor with the Iuav University of Venice, since 2015. In December 2015, he moved to the University of Parma and got promoted to full professorship, in 2019. He is currently the Head of the IMP Laboratory, research group. His research interests include computer vision and image processing, deep learning, and generative models. He is the author of nine book chapters, more than 40 articles in international refereed journals, and more than 100 papers in proceedings of international conferences and workshops. To date, his H-index on Google Scholar is 44, with a total of 10 231 citations. On Scopus, his H-index is 31, with a total of 5610 citations. He is a fellow of IAPR and a member of CVPL.