HAL open science

SCAM! Transferring humans between images with Semantic Cross Attention Modulation

Nicolas Dufour, David Picard, Vicky Kalogeiton

▶ **To cite this version:**

Nicolas Dufour, David Picard, Vicky Kalogeiton. SCAM! Transferring humans between images with Semantic Cross Attention Modulation. European Conference on Computer Vision (ECCV 2022), Oct 2022, Tel Aviv, Israel. pp.713-729, 10.1007/978-3-031-19781-9_41 . hal-03840802

HAL Id: hal-03840802

https://hal.science/hal-03840802v1

Submitted on 12 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

SCAM! Transferring humans between images with Semantic Cross Attention Modulation

Nicolas Dufour[1,2][0000−0002−1903−5110], David Picard[1][0000−0002−6296−4222], and Vicky Kalogeiton[2][0000−0002−7368−6993]

[1] LIGM, Ecole des Ponts, Univ Gustave Eiffel, CNRS, Marne-la-Vallée, France
{nicolas.dufour, david.picard}@enpc.fr
[2] LIX, CNRS, Ecole Polytechnique, IP Paris
vicky.kalogeiton@lix.polytechnique.fr
Project page: https://imagine.enpc.fr/~dufourn/scam

Abstract. A large body of recent work targets semantically conditioned image generation. Most such methods focus on the narrower task of pose transfer and ignore the more challenging task of subject transfer that consists in not only transferring the pose but also the appearance and background. In this work, we introduce SCAM (Semantic Cross Attention Modulation), a system that encodes rich and diverse information in each semantic region of the image (including foreground and background), thus achieving precise generation with emphasis on fine details. This is enabled by the Semantic Attention Transformer Encoder that extracts multiple latent vectors for each semantic region, and the corresponding generator that exploits these multiple latents by using semantic cross attention modulation. It is trained only using a reconstruction setup, while subject transfer is performed at test time. Our analysis shows that our proposed architecture is successful at encoding the diversity of appearance in each semantic region. Extensive experiments on the iDesigner, CelebAMask-HD and ADE20K datasets show that SCAM outperforms competing approaches; moreover, it sets the new state of the art on subject transfer.

Keywords: Semantic Generation, Semantic Editing, Generative Adversarial Networks, Subject Transfer

1 Introduction

Being able to perform subject transfer between two images is a key challenge for many applications, from post-processing in game or art industries to software addressing the needs of the public. For instance, in film industries, one could replace a stunt performer by the main actors, thus alleviating the need of finding a look-alike performer; hence, increasing the filmmakers freedom. Similarly, it could enable finishing a film when an actor is indisposed.

Given a source and a target subject, the idea of subject transfer is for the source subject to replace the target subject in the target image *seamlessly*. The target image should keep the same background, the same interactions between

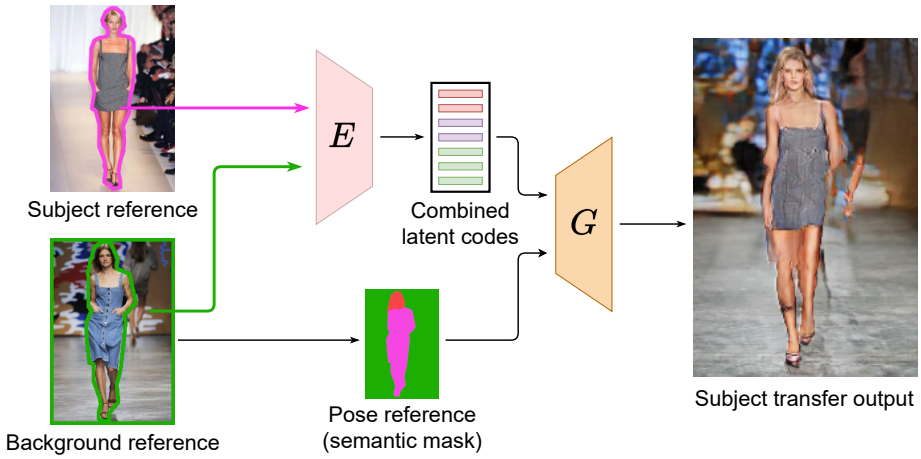

Fig. 1: **Subject transfer with the proposed SCAM.** We first encode the desired subject with the encoder E and get the subject latent codes. Then, we encode the background and the semantic mask for the pose and background reference. Finally, the generator G synthesizes an image, where the subject is transferred with the desired background and pose. Pictures taken from the Internet[3].

subject and objects, and the same spatial configuration, to account for possible occlusions. Figure 1 illustrates this. Note, in contrast to faces, buildings, or landscapes, human bodies are malleable with high morphological diversity, thus casting the task hard to model.

Most methods focus either on pose transfer [49,2,44], where the pose changes, or on style transfer [54,31], where the pose remains fixed but the subject's styling changes. These are limited and cannot be used out of the box for our task, as they are: (1) restrictive; they only work on uniform backgrounds, failing in complex ones (PISE [49], SEAN [54], [44]), and (2) expensive; they require hard training [44] or training one model per human (Everybody Dance Now [2]). Instead, subject transfer changes both the pose and the style/identity of the subject. Thus, a successful system is decoupled in both pose and style transfer and performs both tasks simultaneously.

Semantic editing is a related task, consisting in controlling the output of a generative network by a segmentation mask. Indeed, subject transfer can be performed with semantic editing by using the mask of the target subject with the style of the source. However, modern methods cannot handle complex layout and rich structure (like full-bodies) with in the wild backgrounds. For instance, SPADE [31] fails to control each region style independently, while SEAN [54] fails to handle complex, detailed scenes, such as multiple background objects.

[3]Kate Moss picture by JB Villareal/Shoot Digital, Natalia Vodianova picture by Karl Prouse/Catwalking.

To this end, we propose the **S**emantic **C**ross **A**ttention **M**odulation system (**SCAM**), a semantic editing model that accounts for all aforementioned challenges relevant to subject transfer. SCAM captures fine image details inside the semantic region by having multiple latents per semantic regions. This enables capturing unsupervised semantic information inside the semantic labels, which allows for better handling coarse semantic labels, such as background. Our model can generate more complex backgrounds than previous methods and outperforms SEAN [54] both on the subject transfer and semantic reconstruction tasks.

We propose three architectural contributions: First, we propose the Semantic Cross Attention (**SCA**) that performs attention between a set of latents (each linked to a semantic region) and an image feature map. SCA constrains the attention, such as the latents only attend the regions on the image feature map that correspond to the relevant semantic label. Secondly, we introduce the **SAT** operation and encoder (**S**emantic **A**ttention **T**ransformer) that relies on cross attention to decide which information to gather in the image and for which latent, thus allowing for richer information to be encoded. Third, we propose the **SCAM**-Generator (after which SCAM is named) that modulates the feature maps using the SCAM-Operation, which allows every pixel to attend to the semantically meaningful latents. Note, the whole architecture is trained using a reconstruction setup only, and subject transfer is performed at test time.

2 Related Work

Image to Image synthesis with GANs. GANs [10] generate images by processing a random vector sampled from a predefined distribution with a dedicated network [3,32,18]. A major improvement is StyleGAN [19,20] that allows to modulate the feature map at each resolution according to a given style vector. Typically, unconditional GANs allow for minimal control over the generator's output. For more flexibility, Pix2Pix [14,45] trains a generator coupled to an encoder, allowing for output control with multiple modalities (sketches, keypoints). One of its drawbacks is the need for data pairs, which can be hard to collect (drawings). To tackle this, CycleGAN [53] uses unpaired data by enforcing cycle consistency across domains. However, acquiring paired data is feasible when leveraging external models, such as semantic masks from segmentation models. In our case, we do not have access to ground-truth images where the subject has been transferred as it would require both the subject and the reference to have exactly the same pose and occlusion. We circumvent this by training on a reconstruction proxy task and performing subject transfer at test time.

Semantic Image Generation. Even if Pix2Pix [14,45] manage to control the output image with a segmentation mask, it suffers from a semantic information washing-up. SPADE [31] propose to fix this problem by introducing layer-wise semantic conditioning. CLADE[38] propose a more efficient version of SPADE to reduce runtime complexity. Other approaches such as [46,27,25,36,40,38,7,11] propose improvements over spade. However, these approaches work well generating images from semantic information, but they do not focus on the case

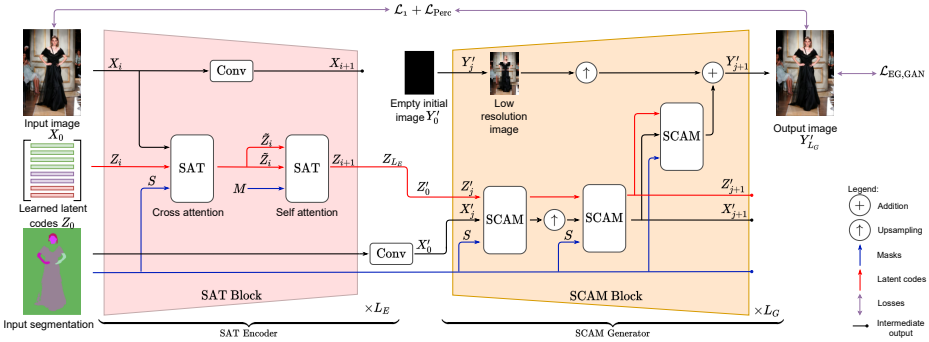

Fig. 2: **Training setup of the proposed SCAM architecture**. It consists of the SAT-Encoder (pink) and the SCAM-Generator (yellow). The SAT-Encoder allows the latents to retrieve information from an image, exploiting both the raw image and the convolution feature maps. Once the image is encoded, the latents are fed to the SCAM-Generator, which captures top-down and bottom-up interactions with a semantic constraint, allowing to easily alter the desired regions thanks to the latents that are dedicated to a given region.

where we want to re-generate a given image to then be able to edit it. To do so, the style of the image must be carefully encoded and move from the single style vector used in SPADE and its variants. SEAN[54] propose to introduce a single style vector per semantic label by performing average pooling on the encoder CNN features. GroupDNet [56] propose to solve this by encoding each semantic region separately using grouped convolution. INADE [37] propose to use instance conditioned convolution to extract a single style code per instance. Although this allows better control of the output and richer representation, it still has two problems: (1) it is limited when handling coarse labels with diverse objects, (2) it creates a single vector per semantic region. In our approach, SCAM, we solve this by introducing the SAT-Encoder which can extract rich representation from images and is able to output multiple and diverse latents per semantic region. In turn, the SEAN-Generator modulates the output by both the semantic mask and the extracted style. This, however, modulates each pixel of a semantic region with the same style vector. Instead, our SCAM-Generator, uses attention to leverage different tokens to interact with, leading to different modulations per pixel, and hence enabling the emergence of unsupervised semantic structure in the semantic regions. Other approaches propose to use diffusion process approaches [30] to perform this editing process. However, these diffusion approaches are very expensive to sample from.

Attention in Computer Vision. Despite their remarkable success [4,33,34,1], transformers suffer from a quadratic complexity problem, which makes it hard to use. To tackle this, most vision methods [6,41,28] subdivide images into patches, resulting in losing information. Instead, the recent Perceiver [16,15] tackles this complexity issue by replacing self attention by cross attention. The image pixels are attended by a significatively smaller set of learned tokens.

Attention in GANS. It [48,8,17,51,22,47] has shown great progress over the past years. GANsformer [13] leverages cross attention to exploit multiple style codes between style vectors and feature maps. These architectures use attention for unconditional generation; however, they do not focus on subject transfer. Instead, our proposed SCA improves upon GANsformer's duplex attention for semantically constrained generation by assigning latents to semantic regions.
Pose Transfer. Using keypoints for pose transfer [29,55,39,23] typically results in coarse representation of bodies. To tackle this, some methods use semantic masks [12,5,50,49]. These, however, focus only on pose transfer, which does not alter the background. Instead, we aim at subject transfer, where preserving the background is crucial. Most methods are limited to simple backgrounds. [2] overfit a GAN to a video and regenerate the background; hence, it cannot be used for dynamic scenes. [44] address this by adapting the weights of the generator, but this ties the subject to the background, not allowing for subject transfer. Here, we focus on subject transfer that changes both pose and background.

3 Method

Our goal is to perform semantic editing with a focus on subject transfer. We propose the **SCAM** method (Semantic Cross Attention Modulation, Figure 2). It relies on **SCA** (Semantic Cross Attention), i.e. a novel mechanism that masks the attention according to segmentation masks, thus encoding semantically meaningful latent variables (Section 3.1). SCAM consists of: (a) **SAT-Encoder** (Semantic Attention Transformer) that relies on cross attention to decide which information to gather in the image and for which latent (Section 3.2); and (b) **SCAM-Generator** (Semantic Cross Attention Modulation) that captures rich semantic information in an unsupervised way (Section 3.3).
Notation. Let $X \in \mathbb{R}^{n \times C}$ be the feature map with n the number of pixels, and C the number of channels. Let $Z \in \mathbb{R}^{m \times d}$ be a set of m latents of dimension d and s the number of semantic labels. Each semantic label is attributed k latents, such that $m = k \times s$. Each semantic label mask is assigned k copies in $S \in \{0;1\}^{n \times m}$. $\sigma(.)$ is the softmax operation.
Motivation. Since many regions of the image have visually diverse content (e.g., the background), we propose to encode this varied information in several complementary latents. Motivated by the findings of Gansformers[13], we use attention to introduce both a constraint on which part of the image a latent can get information from and a competing mechanism between latents attending the same region so as to specialize them. Reciprocally, using duplex attention, we introduce the same strategy by limiting the latents that the feature map can attend to and introduce a competition between parts of a semantic region that can attend the same latent.

3.1 Semantic Cross Attention (SCA)

Definition. The goal of SCA is two-fold depending on what is the query and what is the key. Either it allows to give the feature map information from a

semantically restricted set of latents or, respectively, it allows a set of latents to retrieve information in a semantically restricted region of the feature map. It is defined as:

$$\mathrm{SCA}(I_1, I_2, I_3) = \sigma\left(\frac{QK^T \odot I_3 + \tau\,(1 - I_3)}{\sqrt{d_{in}}}\right) V \quad , \tag{1}$$

where I_1, I_2, I_3 the inputs, with I_1 attending I_2, and I_3 the mask that forces tokens from I_1 to attend only specific tokens from I_2[4], $Q{=}W_Q I_1$, $K{=}W_K I_2$ and $V{=}W_V I_2$ the queries, keys and values, and d_{in} the internal attention dimension.

We use three types of SCA. *(a) SCA with pixels X attending latents Z*: $\mathrm{SCA}(X, Z, S)$, where $W_Q{\in}\mathbb{R}^{n\times d_{in}}$ and $W_K, W_V{\in}\mathbb{R}^{m\times d_{in}}$. The idea is to force the pixels from a semantic region to attend latents that are associated with the same label. *(b) SCA with latents Z attending pixels X*: $\mathrm{SCA}(Z, X, S)$, where $W_Q{\in}\mathbb{R}^{m\times d_{in}}$, $W_K, W_V{\in}\mathbb{R}^{n\times d_{in}}$. The idea is to semantically mask attention values to enforce latents to attend semantically corresponding pixels. *(c) SCA with latents Z attending themselves*: $\mathrm{SCA}(Z, Z, M)$, where $W_Q, W_K, W_V{\in}\mathbb{R}^{n\times d_{in}}$. We denote $M \in \mathbb{N}^{m\times m}$ this mask, with $M_{\text{latents}}(i, j){=}1$ if the semantic label of latent i is the same as the one of latent j; 0 otherwise. The idea is to let the latents only attend latents that share the same semantic label.

3.2 SAT-Encoder

Following [16], our SAT-Encoder relies on cross attention. It consists of L_E consecutive layers of **SAT-Blocks**, where the input of the $i{+}1$-th one is the output of the i-th one (Figure 2 (left)). Given a set of learned queries Z_0 (i.e., parameters updated with gradient descent using back-propagation), it outputs latents Z_{L_E} that have encoded the input image. This allows to create multiple latents per semantic regions resulting in specialized latents for different part of each semantic region of the image. The encoder is also flexible enough to easily assign a different number of latent for each semantic region, allowing to optimize the representation power given to each semantic region. At each layer, the latent code retrieves information from the image feature maps at different scales.
The **SAT-Block** is composed of three components: two **SAT-Operations** and a strided convolution. SAT-Operations are transformer-like [43] operations, replacing self-attention by our proposed SCA. They are defined as:

$$\mathrm{SAT}(I_1, I_2, S) = \mathrm{LN}(f(\mathrm{LN}(\mathrm{SCA}(I_1, I_2, S) + I_1)) + I_1), \tag{2}$$

with LN the layer norm and f a simple 2-layer feed forward network. The first SAT-Operation, $\mathrm{SAT}(X, Z, S)$, let the latents retrieve information from the image feature map (case *(b)* from 3.1). The second SAT-Operation, $\mathrm{SAT}(Z, Z, M)$, is refining the latents using SCA in a self attention setup where the latents attend themselves, keeping the semantic restriction (case *(c)* from 3.1). The strided convolution encode the previous layer image feature map, reducing its spatial

[4]The attention values requiring masking are filled with $-\infty$ before the softmax. (In practice $\tau{=}-10^9$)

dimension. Implementation details and reference code about SAT-Blocks and SAT-Operation are in the supplementary material

3.3 SCAM-Generator

Definition and Architecture. SCAM-Generator takes as an input the latent codes Z'_0 ($= Z_{L_E}$ encoder's output) and the segmentation mask S and outputs the generated image Y'_{L_G}. It consists of L_G **SCAM-Blocks** (Figure 2 (right)). The input latent of the generator is given by the encoder's output, whereas the input latents of each block within the generator are the output latents of previous blocks. Similarly, the input feature maps of each block are the feature map outputs at each resolution, while for the first features, we encode the segmentation with a convolutional layer following [31,54].

The **SCAM-Block** has a progressive growing architecture similar to the one of StyleGAN2 [20] and consists of 3 SCAM-Operations (See Figure 2 (right)). 2 SCAM-Operations process the generator feature-map, with an upscaling operation between the two, while a parallel SCAM-operation retrieves information from the feature maps and generate the image in the RGB space. Implementation details and reference code are in the supplementary.

SCAM-Operation. It aims at exchanging information between pixels/features and latents of the same semantic label and is depicted in Figure 3. Each SCAM-Operation has inputs: (1) the set of input latents Z'_{in},(2) the input feature map X'_{in} , and (3) the segmentation mask S. Its outputs are: (1) the output latents Z'_{out}, and (2) the output feature map X'_{out}. SCAM-Operation consists of three parts: (a) the latent SAT, (b) the feature SCA, (c) the Modulation operation.

(a) The latent SAT. It uses a SAT operation to update the current latents Z'_{in} based on the current feature map: Z'_{out}:=SAT$(Z'_{\text{in}}, X'_{\text{in}}, S)$. This allows for latent refinement while enforcing the latents semantic constraint thanks to the SCA operation inside SAT.

(b) The feature SCA performs latent to image attention: it incorporates the latent information to the pixels/features using SCA. Given X'_{in}, the output la-

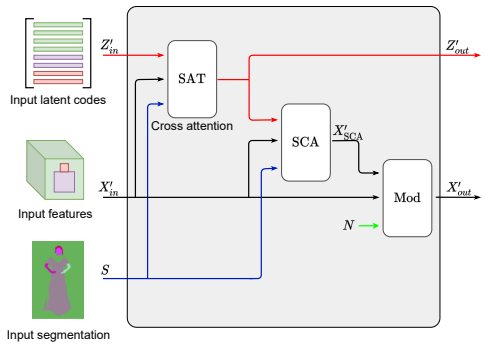

Fig. 3: **SCAM-Operation**. It modulates a feature map according to a segmentation map, allowing each pixel to retrieve information from a semantically restricted set of latents. It enables both top-bottom (latents retrieve information from the feature map) and bottom-top interactions (the map gets information from latents).

tents from the SAT-Operation Z'_{out} and the mask S, it outputs X'_{SCA}:

$$X'_{\text{SCA}} = \text{SCA}(X'_{\text{in}}, Z'_{\text{out}}, S) \quad . \tag{3}$$

(c) The Modulation operation takes as input the X'_{in} maps and the X'_{SCA} from SCA, and outputs features that are passed through a convolution layer $g(.)$ to produce the final output image features X'_{out}. It is defined as:

$$X'_{out} = g\left(\gamma(X'_{\text{SCA}}) \odot \text{IN}(X'_{\text{in}}) + \mu(X'_{\text{SCA}}) + N\right) \quad . \tag{4}$$

It consists of the $\gamma(.)$ and $\mu(.)$ operators that determine the scale and bias factor for the modulation operation. We use Instance Normalization (IN) [42] to perform the feature map normalization. Following [19,20], we also add noise $N \in \mathbb{R}^{n \times C}$, with $N \sim \mathcal{N}(0, \sigma^2 I_{n \times C})$, and σ learned parameter. This encourages the modulation to account for stochastic variation.

Discussion. SCAM-Operation outputs both the modulated feature map and the updated latents. Thus, the information from the previous maps is propagated to the latents and to the feature maps. This brings several advantages: first, it preserves the semantic constraint; second, it provides finer refinement within the semantic mask by attending to multiple latents; and third it allows each pixel to choose which latent to use for modulation.

3.4 Training losses

We train SCAM with GAN and reconstruction losses. We denote by D the discriminator, G the generator and E the encoder. For the **discriminator**, we follow [14,45,45,31] and use PatchGAN, as it discriminates for patches of the given image instead of the global image. For the **GAN loss**, we use Hinge GAN loss [26]: $\mathcal{L}_{\text{EG,GAN}}$, $\mathcal{L}_{\text{D,GAN}}$. For the **reconstruction loss**, we use the perceptual loss $\mathcal{L}_{\text{Perc}}$ as in [54] and the $\mathcal{L}_1$ between the input and the reconstructed input. The **final losses** are $\mathcal{L}_{\text{EG}} = \mathcal{L}_{\text{EG,GAN}} + \lambda_{\text{perc}}\mathcal{L}_{\text{Perc}} + \lambda_1\mathcal{L}_1$ and $\mathcal{L}_{\text{D}} = \mathcal{L}_{\text{D,GAN}}$ with λ_{perc} and λ_1 hyperparameters. We use $\lambda_{\text{perc}} = \lambda_1 = 10$ in our experiments. Training details are in the supplementary.

3.5 Subject transfer

Once SCAM is trained for reconstruction, at test time we perform subject transfer. Given two images X_A, X_B with their respective segmentation masks (S_A, S_B), we retrieve their latent codes as Z_A and Z_B using the SAT-Encoder. To transfer the subject from X_B to the context of X_A, we create Z_{mix}, where the style codes related to the background come from Z_A and the remainder codes come from Z_B. Then, we retrieve $Y'_{\text{mix}} = G(Z_{\text{mix}}, S_B)$. See Figure 1.

4 Experiments

We now present experimental results for SCAM. More results and ethical impacts are discussed in supplementary.

Method	iDesigner					CelebAMask-HQ			ADE20K	
	PSNR ↑	R-FID ↓	S-FID ↓	REIDSim ↑	REIDAcc ↑	PSNR ↑	R-FID ↓	S-FID ↓	PSNR ↑	R-FID ↓
SPADE [31] [CVPR19]	10.4	66.7	67.5	0.67	0.26	10.9	38.2	38.3	10.7	59.7
CLADE[38][TPAMI21]	11.3	45.4	46.1	0.68	0.29	10.8	41.8	42.0	10.4	53.7
SEAN-CLADE [38] [TPAMI21]	15.3	48.4	56.1	0.75	0.31	16.2	19.8	24.3	14.0	38.7
INADE[37] [CVPR21]	12.0	33.0	33.9	0.72	0.34	12.24	22.7	23.4	11.3	48.6
SEAN [54] [CVPR20]	14.9	53.5	58.7	0.74	0.30	16.2	18.9	22.8	14.6	47.6
SCAM (Ours)	**21.4**	**13.2**	**26.9**	**0.81**	**0.56**	**21.9**	**15.5**	**19.8**	**20.0**	**27.5**

Table 1: **Comparison** on iDesigner [35] and CelebAMask-HQ [21] and ADE20K[52].

Implementation details. We train all models for 50k steps, with batch size of 32 on 4 Nvidia V100 GPUs. We set k=8 latents per label of dim d=256. We generate images of resolution 256px.

4.1 Datasets and metrics

iDesigner [35] contains images from designer fashion shows, including 50 different designers with 50k train and 10k test samples. We segment human parts with [24] and then merge the labels to end up with: face, body and background labels.
CelebAMask-HQ [21] contains celebrity faces from CelebA-HQ [18] with 28k train and 2k test images labelled with 19 semantic labels of high quality.
ADE20K [52] contains diverse scenes images with 20k train and 2k test images. The images are labelled with 150 semantic labels of high quality.
Metrics. We use PSNR, reconstruction FID (R-FID), and swap FID (S-FID). R-FID is computed as the FID between the train set and the reconstructed test set, while S-FID is between the test set and a set of subject transfer images computed on the test set. We introduce REIDAcc and REIDSim, computed in the latent space of a re-identification network [9]. REIDSim computes the average cosine similarity between the subject image and the subject transfer image. REIDAcc accounts for the proportion of images where the cosine similarity of the transferred subject is higher with the subject than with the background.

4.2 Comparison to the state of the art

We first compare our proposed SCAM to INADE [37], SEAN [54], CLADE [38], SEAN-CLADE [38] and SPADE [31]. We reproduce all the methods and provide code for our implementations in the supplementary material.
Results on iDesigner are shown in Table 1 (left). Overall, SCAM outperforms competing approaches for all metrics. Specifically, for PSNR it outperforms SEAN-CLADE by approximately +5.9dB, whereas it reaches 13.2 R-FID vs 33.0 for INADE. These major boosts show that our reconstructed images better preserve the details of the initial images, meaning the representation power of SCAM is higher than that of other approaches. The difference is also notable for S-FID, with INADE reaching 33.9 vs 26.9 for SCAM, showing that our method

	Encoder		Generator	Losses			iDesigner						CelebAMask-HQ					
	Conv	SA	SAT	$\mathcal{L}_1$	$\mathcal{L}_{\text{Perc}}$	$\mathcal{L}_{\text{GAN}}$	PSNR↑	$+\Delta$	R-FID ↓	$-\Delta$	S-FID ↓	$-\Delta$	PSNR↑	$+\Delta$	R-FID ↓	$-\Delta$	S-FID ↓	$-\Delta$
i	✓	✓	✓	✓	✓	✓	21.0		13.2		27.1		22.0		15.7		20.7	
ii	✗	✓	✓	✓	✓	✓	19.2	−1.8	26.3	−13.1	34.1	−7.0	**22.1**	+0.1	**15.5**	+0.2	20.2	+0.5
iii	✓	✗	✓	✓	✓	✓	21.3	+0.3	**12.7**	+0.5	**26.1**	+1.0	22.0	0.0	16.5	−0.8	**20.0**	+0.7
iv	✗	✗	✓	✓	✓	✓	18.5	−2.5	24.8	−11.6	29.8	−2.7	19.8	−2.2	19.0	−3.3	21.6	−0.9
v	✓	✓	✗	✓	✓	✓	21.1	+0.1	15.6	−2.4	27.7	−0.6	21.8	−0.2	15.6	+0.1	21.7	−1.0
vi	✗	✓	✗	✓	✓	✓	17.7	−3.3	49.8	−36.6	55.9	−28.8	20.2	−1.8	21.7	−6.0	25.3	−4.6
vii	✓	✗	✗	✓	✓	✓	21.0	0.0	16.7	−3.5	32.3	−5.2	21.4	−0.6	16.7	−1.0	21.8	−1.1
viii	✓	✓	✓	✗	✓	✓	19.5	−1.5	16.1	−2.9	30.7	−3.6	-	-	-	-	-	-
ix	✓	✓	✓	✓	✗	✓	**22.9**	+1.9	43.2	−30.0	91.3	−64.2	-	-	-	-	-	-
x	SEAN		SCAM	✓	✓	✓	17.5	−3.5	27.6	−14.4	32.7	−5.6	17.6	−2.4	20.4	−5.3	24.2	−3.5

Table 2: **Ablation** study on iDesigner [35] and CelebAMask-HQ[21]

perform better subject transfer on datasets with coarse semantic labels than other approaches. We also observe a superiority of SCAM on REIDSim (+0.06 compared to INADE) and REIDAcc (+0.22 compared to INADE). Overall REIDAcc is a hard metric. This can be explained by the fact that the subject transfer image shares the same semantic information with the background image.

Results on CelebAMask-HQ are shown in Table 1 (center), where we observe that SCAM outperforms all methods. For instance, for PSNR, SCAM outperforms SEAN by +5.7dB. SCAM also improves over SEAN by 3.4 R-FID points (15.5 vs 19.8). For subject transfer, SCAM outperforms SEAN by a S-FID decrease of almost 3 points (19.8 vs 22.8), clearly indicating that our method is also better at transferring. We observe that even for a dataset that has precise labelling, our approach still outperforms competing approaches.

Results on ADE20K are shown in Table 1 (right), where we observe that SCAM outperforms all methods. SCAM has the best PSNR of 20.0 whereas second to best SEAN has a PSNR of 14.6. SCAM also beats SEAN-CLADE by 11.2 R-FID points (27.5 vs 38.7). We cannot evaluate S-FID on this dataset since it is hard to select what is the main subject in the image, and not all images share the same semantic labels.

4.3 Ablations

Here, we perform several ablations to validate the effectiveness of all components of SCAM. We denote by − experiments that do not converge.

Ablations of SAT-Encoder and SCAM-Generator. We examine the effectiveness of our encoder and generator by modifying either the SAT-Encoder or the SCAM-Generator with baseline versions and report the results on iDesigner in Table 2. The first row (i) corresponds to our SCAM. We benchmark some variants of our SAT-Encoder: Conv denotes whether we use convolutions in SAT or not. SA denotes whether we use self-attention SAT block or not. The results show that overall convolutions in the SAT-Encoder provide a big encoding advantage. This is especially true for the complex iDesigner dataset. Indeed (i) outperforms (ii, iv, vi) by a high margin on iDesigner e.g., by -13.1 R-FID and -7.0 S-FID for (ii), which validates our use of multiple resolutions in the encoder. We also examine a variation of the generator by removing the SAT block in the

SCAM block. Having SAT in SCAM leads to better results; such as -2.4 in R-FID, and -0.6 in S-FID for (v) on iDesigner. Similar results can be observed in (vi and vii).

In (x), we use the same encoder as in SEAN and the SCAM generator. To manage to have multiple latents per semantic latents, we split the SEAN encoding in 8 smaller latents for each semantic latent. We observe here that our SAT-Encoder is better than the SEAN encoder at extracting information from images. Indeed, we obtain better R-FID and S-FID (-14,4/-5.6) with our encoder than the SEAN encoder.

Ablation of Losses. Table 2 (i,viii,ix) ablates the three $\mathcal{L}_1, \mathcal{L}_{\text{Perc}}, \mathcal{L}_{\text{GAN}}$ losses used in SCAM. The full combination (i) reaches the best results. Interestingly, removing $\mathcal{L}_{\text{Perc}}$ (ix) results in the best PSNR point (22.9db) while having among the worst R-FID and S-FID (43.2, 91.3, respectively). This is expected, as removing the perceptual loss makes the generator rely only on the L_1 loss, and may artificially increase PSNR at the cost of realism.

Number k of latents per semantic label. We've studied the impact of the number of latents k per semantic label on the performance of the model. We tested $k \in [4, 8, 16, 32]$ on iDesigner [35]. We find that the R-FID increase with the number of latents where the S-FID decreases. Indeed for $k = 4$ we have R-FID of 15.9 and a S-FID of 27.3. For $k = 32$, we have R-FID of 9.7 and a S-FID of 30.3. We settle for $k = 8$ in our experiments since it offers the best trade-off between S-FID and R-FID. We also privilege a smaller k since the time complexity of our model is $O(k)$.

Visualization of the attention matrix. To investigate how using multiple latents per region is handled by SCAM, we visualize in Figure 4 the last SCAM layer attention matrix. We colour each pixel according to the corresponding latent with the highest attention value. Overall, we observe that for each semantic region, the latents attend to different subregions, capturing semantic information without supervision. The first example (a) shows that even without specialized segmentation labels, SCAM specializes some latents to reconstruct the complex face pattern (eyes, mouth, and hair) and others for the different body parts (dress and shoes). The second row displays an interesting case: SCAM is capable of assigning different latents to the humans in the background even if they are not labelled as such.

Fig. 4: (a) SCAM specializes latents on complex patterns (shirt, eyes, shoes); (b) it learns semantic information *on its own* inside the semantic labels (background people).

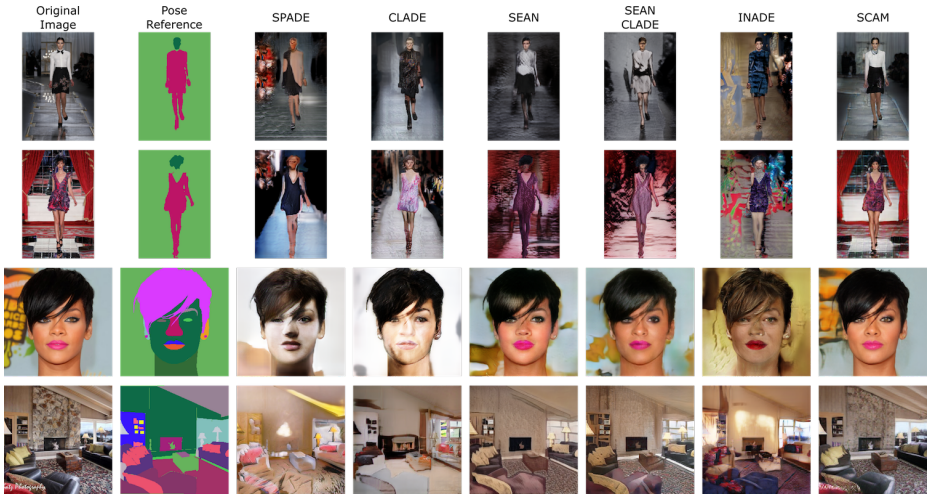

Fig. 5: **Reconstructions** on iDesigner [35], CelebAMask-HQ[21] and ADE20K[52].

4.4 Qualitative results on reconstruction

Overall, we observe that the reconstruction quality of SCAM is superior to competing approaches. Figure 5 displays the input images, their masks (first two columns) and the reconstructions by SPADE [31], CLADE [38], SEAN [54], SEAN-CLADE [38], INADE [37] and SCAM (last six columns). The first two rows are samples from iDesigner, the third is from CelebAMask-HQ and the lastrow from ADE20K.

Subject reconstruction on iDesigner. SCAM reconstructs more structure than SEAN, both in the background and the human. INADE, CLADE and SPADE tend to generate images that doesn't match the style of the original image. For the *background*, we observe that the curtains and window frame of the second row are well-reconstructed by SCAM, in contrast to SEAN that includes colors but no other frame-cues. This highlights the rich generation capabilities of SCAM-Generator, which manage to generate complex backgrounds where competing approaches fail. For the *subject*, SCAM results in finer reconstructions compared to other approaches. For instance, in the first row SCAM reconstructs coherent clothes, while SEAN generate a blurred out version of the clothes.

Reconstruction on CelebAMask-HQ. Overall, SCAM generates crisper and more realistic results than competing approaches. For instance, in the third row, SEAN fails entirely to reconstruct the background by producing an averaged texture. Our method does a better job at this, figuring out a better positioning for the logo and capturing better colors and shapes. On the subject, we also observe that SEAN fails to capture small details such as eye colors.

Reconstruction on ADE20K. SCAM has a more reliable reconstruction than competing approaches. In the fourth row, we observe that SCAM is the only approach that manage to reconstruct some texture on the chimney wall. We

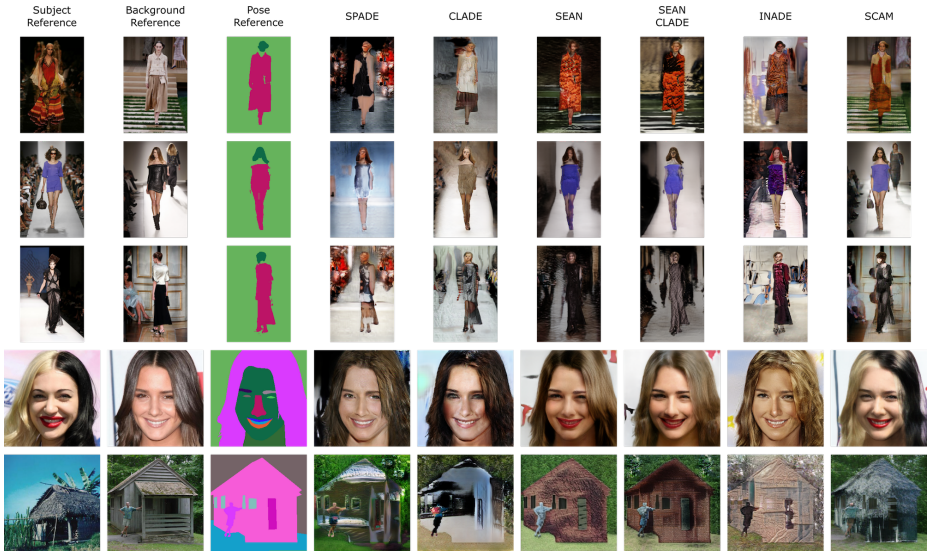

Fig. 6: **Subject Transfer** on the test set of iDesigner [35], CelebAMask-HQ[21] and ADE20K[52]. Note the hard case in row 3, where only SCAM rotates the subject. For ADE20K, we consider the house as the subject in the 5th row.

also observe that overall, SCAM has more detail in the reconstructed object whereas approaches such as SEAN tend to have more averaged textures.

4.5 Qualitative results on subject transfer

Figure 5 displays the subject and background images and the segmentation mask of the pose (first three columns). Then we show the subject transfer for SPADE [31], CLADE [38], SEAN [54], SEAN-CLADE [38], INADE [37] and SCAM (last six columns). The first three rows are samples from iDesigner, the fourth is from CelebAMask-HQ and the last one is on ADE20K. Overall, amongst all methods, SCAM is the one to successfully preserve all components of subject transfer: subject appearance, background appearance and pose.

Subject transfer on iDesigner. SCAM leads to superior subjects (e.g., washed out colors in SEAN vs coherent structured clothing in SCAM in the first row) and background reconstruction (i.e., global background structure, positioning of people on the sides and colours) than SEAN. For instance, in the first row, SEAN completely fails to reconstruct the background, while SEAN-CLADE reconstructs some texture but lacks detail. Similarly, in the second row, SCAM captures the model in the background whereas other approaches miss it; note also the precise people reconstruction in the left part of the catwalk for SCAM compared to competing approaches. In the third row, we have a very hard case, where the subject has a completely different pose than the pose reference. While other approaches fail to rotate the subject, SCAM does succeed in doing, even if

the quality of the generated image is low. Overall, subjects have better appearance in SCAM than in other methods, like details in clothes, shoes, or faces.
Subject transfer on CelebAMask-HQ. The fourth row shows that SCAM recovers more details in the transferred image, such as the colour of the skin or facial expression. Notably, SCAM does capture the bicolor separation of the hair, while SEAN, SEAN-CLADE and INADE display an averaged hair color.
Subject transfer on ADE20K. In the fifth row, we consider the house as the subject we want to transfer. We can see that SCAM does transfer the hut like appearance of the subject, whereas competing approaches fail to do so. We also observe that most approaches have difficulties with the person generation, only SCAM generates a human that is coherent with the background reference.

4.6 User Study on iDesigner

We perform an user study on iDesigner. We compare 3 models one against each other: INADE, SEAN and SCAM. We have asked 38 different people to select among 20 reconstruction images which method in their opinion had the best reconstruction. We denote the percentage of best pick for each method as R-UP (Reconstruction User Preference). Similarly, we have the same people asked to select among 20 subject transfer images which method performed the best subject transfer in their opinion. Participant were asked to take into account image quality and quality of transfer. We denote the percentage of best pick for each method as ST-UP (Subject Transfer User Preference).
We observe that our method, SCAM, outperforms competing methods by a high margin. Indeed, 98.4% of users chose SCAM as the best image reconstruction technique, 1.5% picked SEAN and 0.1% for INADE. As for subject transfer 92.8% of users preferred SCAM, 5.0% chose SCAM and 2.2% INADE.

5 Conclusion

We introduced SCAM that performs semantic editing and in particular subject transfer in images. The architecture contributions of SCAM are: first, the semantic cross attention (**SCA**) mechanism performing attention between features and a set of latents under the constraint that they only attend to semantically meaningful regions; second, the Semantic Attention Transformer Encoder (**SAT**) retrieving information based on a semantic attention mask; third, the Semantic Cross Attention Modulation Generator **(SCAM)** performing semantic-based generation. SCAM sets the new state of the art by leveraging multiple latents per semantic region and by providing a finer encoding of the latent vectors both at encoding and decoding stages.
Acknowledgments: We would like to thank Dimitrios Papadopoulos, Monika Wysoczanska, Philippe Chiberre and Thibaut Issenhuth for proofreading. This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012630 made by GENCI and was supported by a DIM RFSI grant and ANR project TOSAI ANR-20-IADJ-0009.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: NeurIPS (2020)
2. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: ICCV (2019)
3. Denton, E., Chintala, S., Szlam, A., Fergus, R.: Deep generative image models using a laplacian pyramid of adversarial networks. In: NeurIPS (2015)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
5. Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., Yin, J.: Soft-gated warping-gan for pose-guided person image synthesis. In: NeurIPS (2018)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
7. Endo, Y., Kanamori, Y.: Diversifying semantic image synthesis and editing via class-and layer-wise vaes. In: Computer Graphics Forum (2020)
8. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021)
9. Fu, D., Chen, D., Bao, J., Yang, H., Yuan, L., Zhang, L., Li, H., Chen, D.: Unsupervised pre-training for person re-identification. In: CVPR (2021)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
11. Gu, S., Bao, J., Yang, H., Chen, D., Wen, F., Yuan, L.: Mask-guided portrait editing with conditional gans. In: CVPR (2019)
12. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: ICCV (2019)
13. Hudson, D.A., Zitnick, C.L.: Generative adversarial transformers. In: Proc. ICML (2021)
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
15. Jaegle, A., Borgeaud, S., Alayrac, J.B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Hénaff, O., Botvinick, M.M., Zisserman, A., Vinyals, O., Carreira, J.: Perceiver io: A general architecture for structured inputs & outputs. In: arXiv (2021)
16. Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., Carreira, J.: Perceiver: General perception with iterative attention. In: Proc. ICML (2021)
17. Jiang, Y., Chang, S., Wang, Z.: Transgan: Two transformers can make one strong gan. In: CoRR (2021)
18. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: ICLR (2018)
19. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019)
20. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: CVPR (2020)

21. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: Towards diverse and interactive facial image manipulation. In: CVPR (2020)
22. Lee, K., Chang, H., Jiang, L., Zhang, H., Tu, Z., Liu, C.: Vitgan: Training gans with vision transformers. ArXiv (2021)
23. Li, K., Zhang, J., Liu, Y., Lai, Y.K., Dai, Q.: Pona: Pose-guided non-local attention for human pose transfer. In: IEEE Transactions on Image Processing (2020)
24. Li, P., Xu, Y., Wei, Y., Yang, Y.: Self-correction for human parsing. In: IEEE TPAMI (2020)
25. Li, Y., Li, Y., Lu, J., Shechtman, E., Lee, Y.J., Singh, K.K.: Collaging class-specific gans for semantic image synthesis. In: ICCV (2021)
26. Lim, J.H., Ye, J.C.: Geometric gan. In: arXiv (2017)
27. Liu, X., Yin, G., Shao, J., Wang, X., Li, h.: Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In: NeurIPS (2019)
28. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
29. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Gool, L.V.: Pose guided person image generation. In: NeurIPS (2017)
30. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: ICLR (2022)
31. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: CVPR (2019)
32. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR (2016)
33. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training. In: ArXiv (2018)
34. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. In: ArXiv (2019)
35. R.J. Lehman, H.H.: idesigner 2019. In: FGVC6 (2019)
36. Schönfeld, E., Sushko, V., Zhang, D., Gall, J., Schiele, B., Khoreva, A.: You only need adversarial supervision for semantic image synthesis. In: ICLR (2021)
37. 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: CVPR (2021)
38. 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. In: IEEE TPAMI (2021)
39. Tang, H., Bai, S., Zhang, L., Torr, P.H., Sebe, N.: Xinggan for person image generation. In: ECCV (2020)
40. Tang, H., Xu, D., Yan, Y., Torr, P.H., Sebe, N.: Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In: CVPR (2020)
41. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: Proc. ICML (2021)
42. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. In: CVPR (2020)
43. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
44. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. In: NeurIPS (2019)

45. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: CVPR (2018)
46. Wang, Y., Qi, L., Chen, Y.C., Zhang, X., Jia, J.: Image synthesis via semantic composition. In: ICCV (2021)
47. Zhang, B., Gu, S., Zhang, B., Bao, J., Chen, D., Wen, F., Wang, Y., Guo, B.: Styleswin: Transformer-based gan for high-resolution image generation. ArXiv (2021)
48. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: Proc. ICML (2019)
49. Zhang, J., Li, K., Lai, Y.K., Yang, J.: Pise: Person image synthesis and editing with decoupled gan. In: CVPR (2021)
50. Zhang, J., Liu, X., Li, K.: Human pose transfer by adaptive hierarchical deformation. In: CGF (2020)
51. Zhao, L., Zhang, Z., Chen, T., Metaxas, D., Zhang, H.: Improved transformer for high-resolution gans. In: NeurIPS. vol. 34 (2021)
52. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR (2017)
53. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
54. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: Image synthesis with semantic region-adaptive normalization. In: CVPR (2020)
55. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: CVPR (2019)
56. Zhu, Z., Xu, Z., You, A., Bai, X.: Semantically multi-modal image synthesis. In: CVPR (2020)