# Discovering Class-Specific GAN Controls for Semantic Image Synthesis

Edgar Schönfeld[1] Julio Borges[1] Vadim Sushko[1] Bernt Schiele[2] Anna Khoreva[1,3]

[1]Bosch Center for AI [2]MPI for Informatics [3]University of Tübingen

## Abstract

*Prior work has extensively studied the latent space structure of GANs for unconditional image synthesis, enabling global editing of generated images by the unsupervised discovery of interpretable latent directions. However, the discovery of latent directions for conditional GANs for semantic image synthesis (SIS) has remained unexplored. In this work, we specifically focus on addressing this gap. We propose a novel optimization method for finding spatially disentangled class-specific directions in the latent space of pretrained SIS models. We show that the latent directions found by our method can effectively control the local appearance of semantic classes, e.g., changing their internal structure, texture or color independently from each other. Visual inspection and quantitative evaluation of the discovered GAN controls on various datasets demonstrate that our method discovers a diverse set of unique and semantically meaningful latent directions for class-specific edits.*

## 1. Introduction

Semantic image synthesis (SIS) transforms user-specified semantic layouts to realistic images. Its applications range widely from image editing and content creation to synthetic data augmentation, where training data is generated to fulfill specific semantic requirements. For SIS, GANs [9] have demonstrated their superiority in terms of the visual quality of synthesised images and their alignment to input semantic label maps [17, 24, 26, 32, 38]. Although some of GAN-based SIS models allow local appearance editing of single classes or regions in an image – either by style transfer from a reference image [16, 32, 46] or by sampling noise independently for specific image regions [26, 47], they are lacking the technique of enabling interpretable semantic changes for the specific class without reference image and user-in-the-loop supervision.

On the other hand, prior work has extensively studied the latent space of unconditional GANs [8, 10, 25, 29, 35, 41], finding interpretable latent directions which activate distinctive factors of variations in the generation process in an unsupervised fashion, without exploiting reference images. Moving latent code(s) along a certain direction can result in domain-agnostic transformations, e.g. rotation or zooming [13,25,36], or domain-specific alterations, e.g. age or nose length of a person [6, 7, 18, 28, 39]. Despite their progress, it remains a challenge to find interpretable latent directions to control interactively the synthesis of specific semantic classes in the image without changing other image regions. Since the above methods were designed specifically for unconditional GANs, they are not well suited to discover class-specific latent directions in the presence of semantic label maps, inherently given for SIS.

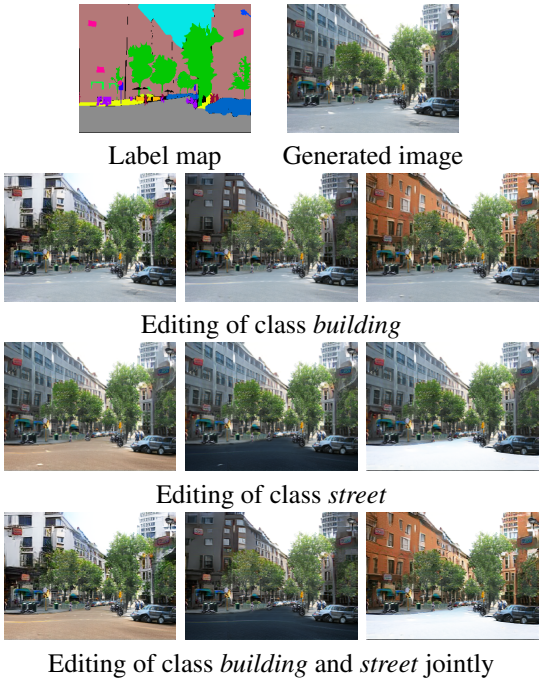


Figure 1. Our Ctrl-SIS method learns class-specific directions in the latent space of a SIS model, which can be applied jointly for different semantic classes for local editing of the generated image.

In this work, we address this limitation and study the latent space of conditional GANs designed for SIS, which to the best of our knowledge has not been explored previously. In particular, making use of the label maps we devise a method to discover meaningful latent directions that only change a specific semantic class in the image. These directions can, for example, encode different designs of the facade for the building class or surfaces for the street class (Fig. 1), enabling the user to perform local semantic edits independently from the rest of the image. Note that in recent state-of-the-art SIS GANs, the generator is already

designed to be sensitive to spatial information [26, 47], allowing region-based noise sampling of specific classes.

On this basis, we introduce a simple, efficient optimisation method to discover class-specific controls in pretrained SIS GANs, which we call *Ctrl-SIS* (see Fig. 2). Our optimization objective is designed to ensure that the learnt latent directions are 1) diverse and different from each other (*diversity loss*); 2) only affect the image area of the selected class, preserving the appearance of other areas (*disentanglement loss*); and 3) induce the same semantic edits consistently across different initial latent codes and label maps containing the class (*consistency loss*). See Sec. 3 for more details. We demonstrate that GAN controls discovered automatically by Ctrl-SIS can effectively manipulate the appearance of the selected semantic class in specific ways, without affecting other classes in the image. For example, we can change the house facade (see Fig. 1), remove leaves from trees or cover mountains in snow (see Fig. 4). Moreover, we can edit different classes jointly, e.g., alter both the building and the road in the street scene (see Fig. 1). Since we use only train-time optimization, instead of exhaustive search as in [39] or test-time optimization as in [18,22,44,45], our training time stays relatively fast compared to the former, while also allowing interactive image editing compared to the latter.

The evaluation of GAN control discovery methods is commonly left to subjective visual inspection. To address this, we introduce new metrics to quantitatively assess diversity, spatial disentanglement and consistency properties of learnt latent directions (see Sec. 4.2). We compare Ctrl-SIS with other GAN control methods on different SIS GANs [24, 26, 38] on two datasets [4, 43]. Our experiments show that latent directions found by prior methods adapted to SIS [10, 29] lead to weaker class edits, comparable to random directions (see Sec. 4.3). In contrast, Ctrl-SIS finds directions that enable diverse and semantically meaningful class edits while maintaining high image quality.

In summary, our contributions are as follows: 1) We propose Ctrl-SIS – a method to discover interpretable latent controls for individual semantic classes in pretrained SIS GANs. To the best of our knowledge, the discovery of class-specific latent direction has not yet been addressed in the SIS literature. 2) We define diversity, consistency, and spatial disentanglement as desirable properties of class-specific latent controls and propose new metrics to quantify them.

## 2. Related work

**GAN models for SIS.** SIS GANs attracted a lot of attention for their application in controllable image synthesis [23] and editing [21, 32]. To achieve photorealism, Pix2Pix [12] used an encoder-decoder generator and a patch-based discriminator, providing label maps as input to the first layers of both networks. SPADE [24] demonstrated that using label maps as input only to the first layer tends to weaken semantic conditioning and proposed to modulate intermediate generator layers via a spatially-adaptive normalization. Follow-up works improved image quality through different ways of using label maps in the generator, e.g., via other conditional normalizations [21, 33, 38, 46], conditional convolutions [19], a label map encoder [17], or by learning class-specific sub-generators [34].

Many SIS models struggled to achieve diversity, as the generator tended to ignore the input latent code [12, 37]. To address this issue, SPADE [24] and SC-GAN [38] utilized an image encoder to extract a global style vector from a reference image. By changing a reference image, these models can generate different images from the same label map. [16,46] further enabled class-wise style transfers from the reference image. [47] allowed local editing by controlling the appearance of different classes via separate latent codes using group convolutions. OASIS [26] improved image diversity by feeding a 3D latent code tensor jointly with the label map into the conditional batch normalization layers, thus, enabling both global image editing as well as local region-based editing. For this reason, in this work we discover class-specific directions in the 3D latent spaces of already pre-trained generators of the state-of-the-art SIS models [24, 26, 38]. Since latent direction discovery is not applicable to the exemplar-based approaches, such as [16,46], we focus on non-exemplar-based models, that rely only on the input latent code to generate diverse images.

**GAN control discovery.** It has been shown that the latent space of GANs frequently exhibits semantically relevant vector space arithmetic [2, 8, 13, 27, 36]. However, finding steerable directions in the latent space is challenging due to its high dimensionality and the large variety of image semantics. Consequently, some works use human supervision [27], attribute predictors [28,39] or predetermined visual transformations such as zooming and rotation [13,25] to identify interpretable latent directions. As the dependence on supervision limits the practical use of these methods, [10, 29, 30, 35, 36, 41] investigated unsupervised discovery of GAN controls. GANSpace [10] performed PCA on the intermediate generator features, discovering useful directions in the latent space resulting from layerwise perturbations along the principal directions. SeFa [29] identified semantically meaningful directions in closed-form, through eigendecomposition of the generator weights. In contrast, [35, 41] relied on gradient-based optimization. WarpedGanSpace [35] used an image classifier to discriminate among a fixed set of directions, while LatentCLR [41] employed a contrastive loss optimizing directions to have orthogonal effects on the generator features. A common limitation of unsupervised methods is that the obtained latent directions are left to subjective visual inspection and

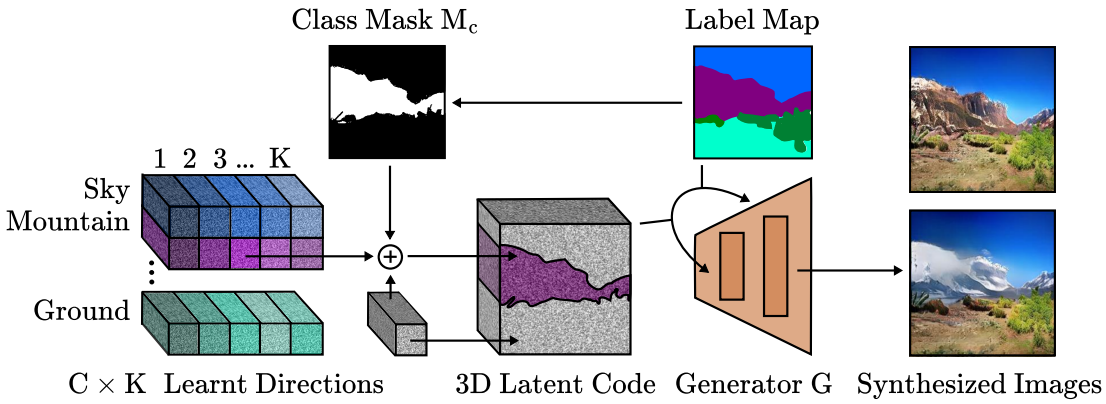


Figure 2. Ctrl-SIS provides a set of $K$ class-specific latent directions which control the appearance of $C$ semantic classes (shown left). To alter the appearance of class $c$, a class-specific latent direction is added to the input 3D latent code $z$ of the pretrained generator $G$ in the label map area corresponding to class $c$ ($M_c$).

manual identification of significant controls.

While the above work focused on finding latent directions for global image manipulation of unconditional GANs, our method finds class-specific latent directions for conditional SIS models. We pick state-of-the-art GANSpace and SeFa for comparison due to their simplicity, easy adaptation to SIS GAN models, and code availability.

**Local editing with GANs.** Recent work enabled local image editing through optimization in the latent space on specific image regions [18,22,31,39,44,45]. EditGAN [18] modelled images and label maps jointly, requiring the user to modify the label map in order to perform the edit via latent space optimization. Image2StyleGAN++ [1] changed images locally via a masked style transfer or by re-encoding images edited with provided scribbles. LELSD [22] proposed an area loss that, given a binary mask, optimizes changes only within the specified area. Yet, the found directions are still applied globally. ReSeFa [45] proposed to optimize the change of pixel values with respect to the latent code, identifying latent variations in a user-specified image region. The main limitation of the above work is that it requires test-time optimization, preventing the user from interactive image editing. In contrast, Ctrl-SIS is optimized end-to-end once to provide latent directions for interactive editing in the spirit of GANSpace and SeFa. It enables the user to perform image editing interactively, without the need for further supervision or mask area definitions.

Ctrl-SIS differs from the aforementioned latent space optimization methods in two ways. First, Ctrl-SIS is specifically designed for SIS, making use of semantic label maps inherent in this task. Second, the discovered latent directions are class-specific. Several SIS models allow class-specific noise sampling [17, 26, 32, 47], but do not provide a method to discover interpretable directions in their latent space. Lastly, existing methods for class-specific editing in SIS models manipulate classes using predetermined concepts [16,46], requiring a source image to extract a style. In contrast, our unsupervised approach allows users to browse through a set of discovered semantic concepts that the pretrained SIS GAN has learned.

## 3. Ctrl-SIS method

The goal of this work is to discover steerable latent directions for GAN-based SIS models. Enabled by the given semantic label maps, we aim to find GAN controls specific to semantic classes, e.g. a set of latent directions for controlling the appearance of the street and another set of directions for the appearance of house facades, see Fig.1. However, this task presents two major challenges.

The first challenge is that SIS GANs commonly do not provide the same image diversity as unconditional models [3, 14], nor have region-specific latent codes. We alleviate both problems by applying a 3D latent code injection [26], which we describe in Sec. 3.1. The second challenge is that prior GAN control discovery methods are not designed to consider label maps, nor to find *class-specific* directions, as they are devised for unconditional GANs. We address both aspects in Sec. 3.2 with a simple, efficient optimisation method which we call Ctrl-SIS.

### 3.1. GAN controls for SIS models

Current SIS models employ different ways to inject latent code into the generator, which affects its ability to perform class-specific edits. The default approach is to feed a one-dimensional latent vector as input to the generator [19,24,38], resulting in no direct opportunity to perform local region-based edits of the image. Thus, in order to enable local editing in SIS, we employ the 3D latent code injection scheme from [26], adopting it to all SIS models considered in this work. The 3D latent codes $z \in \mathbb{R}^{H \times W \times D}$ are created by replicating the original noise vector along

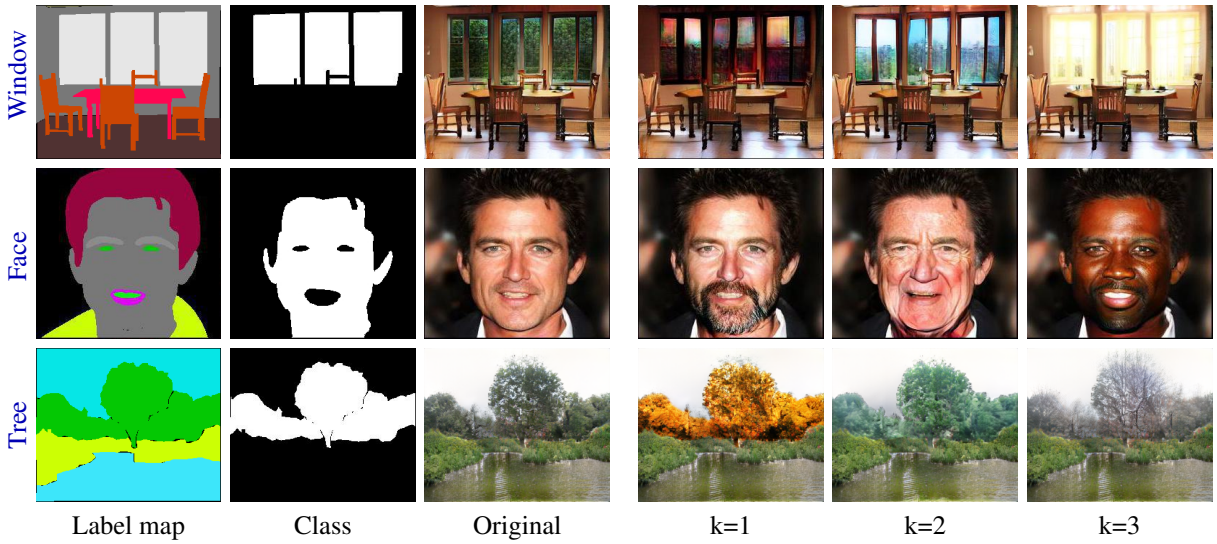


Figure 3. Examples of directions discovered by Ctrl-SIS for various classes, such as different views of a window, face appearances or tree leafage for different seasons of the year. The directions give insight into the concepts that the pretrained SIS model is able to represent.

the height $H$ and width $W$ of the label map. The 3D latent code allows to apply different latent vectors to different image regions (see Fig. 2). In practice, altering the 3D latent code only for a specific image region can still affect other image areas, due to spatial correlations learnt by the generator during training. Nevertheless, the 3D latent space provides better spatial disentanglement and, thus, improves image manipulation control for local edits compared to 1D latent codes. In the remainder of this paper, we assume a 3D latent space for the discovery of SIS GAN controls.

Let $G$ be a well-trained GAN generator of a SIS model. The generator $G(z, y)$ synthesises an image given a 3D latent code $z$ and label map $y$, i.e. $x = G(z, y) = F(h(z, y))$, where $h = \{G_l(z, y)\}_{l \in L}$ is a chosen subset of features from intermediate layers $l \in L$ in the network $G$, and $C$ is the total number of semantic classes. The latent code $z$ controls the appearance of the synthetic image, while the label map $y$ specifies the scene layout. Then an image $x$ can be globally edited by moving $z$ along a specific direction $v_k$:

$$x(v_k) = F(h(z, v_k, y)) = G(z + \alpha v_k, y), \qquad (1)$$

where $\alpha$ controls the intensity of the change, and the latent direction $v_k$ determines the semantics of the image transformation. Local editing of class $c$ in $x$ is done by moving $z$ along a class-specific direction $v_k^c$ only in the area of class $c$ in the label map $y$:

$$x(v_k^c) = F(h(z, v_k^c, y)) = G(z + \alpha M_c \odot v_k^c, y), \qquad (2)$$

where $M_c = \mathbb{1}_{[y=c]}$ is a binary mask indicating pixels in the image belonging to $c$ (see Fig. 2). We next define the task of class-specific GAN control discovery and introduce an optimization objective to find $v_k^c$ directions for any pretrained SIS model with a spatially-aware generation process induced by 3D latent codes.

### 3.2. Discovery of class-specific GAN controls

For the class of interest $c \in C$ we aim to find a diverse set of class-specific directions $V^c = \{v_0^c, v_1^c, ..., v_K^c\}$, $K > 1$, that can meaningfully edit the appearance of class $c$ in the synthetic image $x$, such that image $x(v_k^c)$ has a visually distinct appearance of class $c$ compared to $x$, but all other classes have the same appearance as in $x$. Based on this logic, we form an optimization objective, which consists of diversity, disentanglement and consistency loss terms:

$$\min_{V^c} \mathcal{L}_{div} + \mathcal{L}_{dis} + \mathcal{L}_{const} \, . \qquad (3)$$

The diversity loss $\mathcal{L}_{div}$ encourages a set of class-specific GAN controls $V^c$ to be diverse and introduce different semantic changes to class $c$, the disentanglement loss $\mathcal{L}_{dis}$ prevents changes outside the class area, and the consistency loss $\mathcal{L}_{const}$ ensures that the semantics of an edit are consistent between different initial latent codes $z$. We next provide the mathematical formulation of these loss terms.

**Diversity loss.** Given a label map $y$ and a class of interest $c$, the diversity loss aims to ensure that the set of found latent directions $V^c$ applied to identical input latent code $z$ yields maximally different semantic visual effects, i.e. change an appearance of class $c$ in a different way. It is formulated as

$$\mathcal{L}_{div} = -\mathbb{E}_{(z,y)} \Big[ \sum_{\substack{k_{1,2}=1 \\ k_1 \neq k_2}}^{K} M_c \cdot ||h(z, v_{k_1}^c, y) - h(z, v_{k_2}^c, y)||_2 \Big], \qquad (4)$$

where $\| \cdot \|$ is the $L_2$ norm, and for the class-specific area $M_c$ the distance between the two resulting images $x(v_{k_1}^c)$ and $x(v_{k_2}^c)$ is maximised in the generator feature space $h$, ensuring semantically different directions for class $c$. Depending on the selected feature space in $G$, i.e. the subset

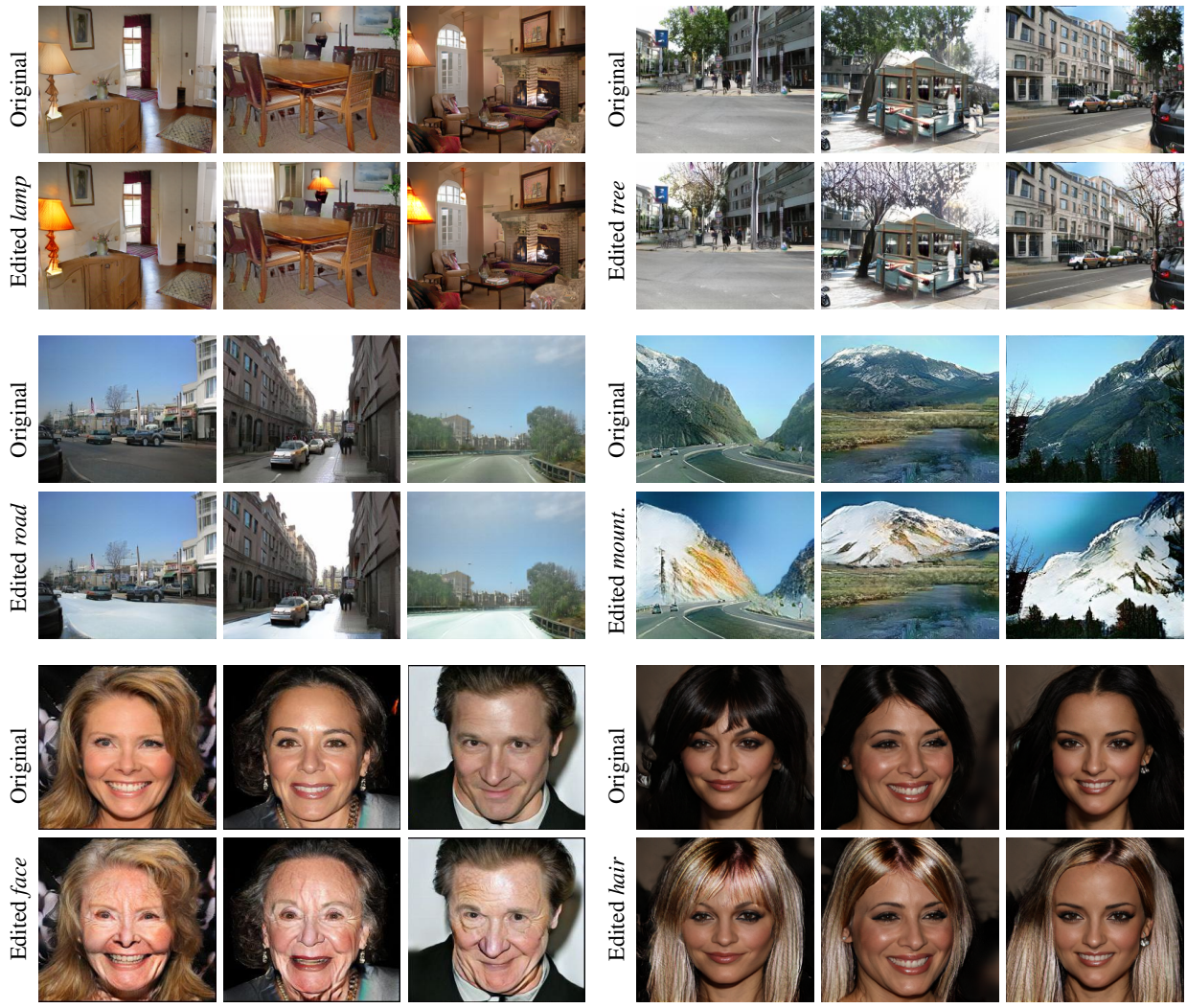


Figure 4. Interpretable latent directions learnt by Ctrl-SIS for various classes. Each triplet is edited with an identical direction. Class-specific edits, such as aging, snowy streets or bald trees, are highly consistent across different label maps and initial latent codes.

of intermediate layers $L$ in $h = \{G_l(z, y)\}_{l \in L}$, we can find various GAN control directions which correspond to different semantics encoded in the selected feature space of $G$.

**Disentanglement loss.** The discovered latent direction $v_k^c$ for class $c$ should only affect the image area belonging to $c$ in the label map $y$ and leave the rest of the image unaffected. Thus, we also minimize the change for images $x(v_{k_1}^c)$ and $x(v_{k_2}^c)$ in the feature space $h$ in the area outside of $M_c$:

$$\mathcal{L}_{dis}=\mathbb{E}_{(z,y)}\Big[\sum_{\substack{k_{1,2}=1 \\ k_1 \neq k_2}}^{K}(1-M_c)\cdot||h(z,v_{k_1}^c,y)-h(z,v_{k_2}^c,y)||_2\Big]. \quad (5)$$

**Consistency loss.** Identical GAN control directions should cause consistent semantic edits of class $c$ for different input latent codes and the same label map $y$ given to the generator. Therefore, for every found direction $v_k^c$ we minimize the feature space distance between two images generated with $z_1$ and $z_2$ in the class-specific area $M_c$ :

$$\mathcal{L}_{const}=\mathbb{E}_{(z,y)}\Big[\sum_{k=1}^{K} M_c \cdot ||h(z_1,v_k^c,y)-h(z_2,v_k^c,y)||_2\Big]. \quad (6)$$

Note that the directions in $V^c$ are the only parameters to be optimized; the weights of the pre-trained image generator $G(z, y)$ are kept frozen. The parameters are optimized by iterating over batches of label maps in the training set and minimizing the objective for selected classes at every step. During optimization, the directions $v_k^c$ are normalized along the channel dimension to unit length 1 and subsequently scaled by $\alpha$, sampled from the interval $[-n; n]$, where $n = \mathbb{E}[||z||_2]$ is the average norm of the latent code along the channel dimension. This ensures the latent edits are neither too small nor too extreme.

| Method | ADE20K | | | | | COCO-Stuff | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mCD ↑ | mCC ↓ | mOD ↓ | FID↓ | mIoU↑ | mCD ↑ | mCC ↓ | mOD ↓ | FID↓ | mIoU↑ |
| Baseline | - | - | - | 28.6 | 52.2 | - | - | - | **17.1** | 42.4 |
| Random | 0.11 | 0.30 | **0.01** | 31.3 | 49.4 | 0.16 | 0.07 | **0.00** | 17.6 | 42.3 |
| GANSpace | 0.09 | 0.29 | **0.01** | **28.1** | **53.3** | 0.15 | **0.06** | **0.00** | 17.2 | 42.1 |
| SeFa | 0.12 | **0.28** | **0.01** | **28.1** | 53.2 | 0.15 | **0.06** | **0.00** | **17.1** | **43.8** |
| **Ctrl-SIS** | **0.26** | **0.28** | **0.01** | 30.9 | 48.9 | **0.30** | 0.07 | 0.01 | 21.1 | 43.6 |

Table 1. Evaluation of OASIS GAN controls on ADE20K and COCO-Stuff. Red numbers indicate lower-than-random performance. Ctrl-SIS discovers significantly more diverse directions (mCD) without sacrificing consistency (mCC) or spatial disentanglement (mOD).

# 4. Experiments

## 4.1. Experimental setup

**Datasets.** We use three challenging datasets: CelebAMask-HQ [16], ADE20K [43], and COCO-Stuff [4]. CelebAMask-HQ consists of 30k face images. ADE20K and COCO-Stuff contain 20k and 164k images of indoor and outdoor scenes, and are used for the main experiments.

**SIS models.** We consider three pre-trained GANs for SIS: SC-GAN [38], SPADE [24] and OASIS [26], using the code provided by the authors [1]. We additionally implement 3D latent codes for SPADE and SC-GAN, which do not originally support it, enabling local image editing for them.

**GAN control methods.** Ctrl-SIS is compared against the two related latent discovery methods GANSpace [10] and SeFa [29], using the authors' code [2]. Following GANSpace-StyleGAN2 [10] and SeFA-StyleGAN2 [29], we train all latent direction methods on features extracted from the normalization layers of each ResNet block in the generator.

**Training details.** Ctrl-SIS is trained with a batch size of 16 on a NVIDIAv100 GPU, using the AdamW optimizer [20] and a learning rate of 1e-3. We train for 20 epochs on ADE20K, and 5 epochs on COCO-Stuff and CelebAMask-HQ, using $K = 5$. Finding class-specific directions with Ctrl-SIS takes $\sim$1h. For evaluation we scale the directions with $\alpha$ sampled in $[-n; n]$ (see Sec. 3.2), to ensure that the direction magnitude is in the same range as the average latent code. By scaling the magnitudes of latent directions from all methods in the same way, we ensure that the effect of the edit only depends on the learnt direction. For GANSpace and SeFa, we pick the directions belonging to the first $K$ components, as they cause the largest variations.

## 4.2. Evaluation

**Image quality evaluation.** Following [12,24,26], we monitor the visual quality of images generated with class-specific edits using FID [11] and mIoU metrics. FID is known to be well aligned with human judgement of image quality. mIoU assesses the alignment of images with ground truth label maps, calculated via a pre-trained semantic segmentation network. We use UperNet101 [40] for ADE20K and DeepLabV2 [5] for COCO-Stuff. In addition, we employ the precision and recall metrics of [15], which correlate with image quality and diversity, respectively.

**Evaluation of class-specific GAN controls.** Prior GAN control methods were mostly evaluated by subjective visual inspection [10, 29, 41]. Consequently, it was challenging to assess the important properties of GAN control methods. First, how many unique and visually distinct latent directions are found. Second, how general are the directions, i.e. if they consistently invoke the same semantic change in all images. Third, in the case of SIS, local class-specific edits must not affect the rest of the image. To quantitatively assess these properties, we introduce the following metrics.

To measure how unique and distinct the found directions are, we introduce the *mean control diversity* (mCD). mCD measures the mean pairwise LPIPS [42] distance between edited images generated with the same label map. We generate 10 global edits of each synthetic test set image by applying a randomly selected class-specific direction to each class area in the label map. The pairwise LPIPS distance is then averaged over all images and 5 initial input latent codes. We also measure a local version $\text{mCD}_l$, where only a single class is locally edited at a time. In this case, we compute the average masked LPIPS distance inside the class-specific area of a single class for all pairs of the $K$ class-specific directions. The final $\text{mCD}_l$ score is the average of the mean per-class scores. A high $\text{mCD}_l$ score implies that each discovered latent direction changes the class appearance in a uniquely different way.

To evaluate how much class-specific edits affect the areas outside of the target class area, we introduce the *mean outside class diversity* (mOD). mOD is computed similarly to $\text{mCD}_l$, except that we use the masked LPIPS distance in the image area outside of the target class area. Ideally, mOD is very low, as we want the latent direction to alter only the class-specific region. Both mCD and mOD scores are inspired by the metrics proposed by [47], see Sec. B in the supplementary for more details.

We introduce the *mean control consistency* (mCC) score to measure to which extent latent directions invoke the same interpretable edit in all images. The mCC is computed by editing each class in an image with a randomly picked class-specific direction. In this case, the score of an image is the

[1] Code for SIS models: SPADE, SC-GAN, OASIS
[2] Code for GAN control methods: GANSpace, SeFa

| | $\mathrm{mCD}_l$ ↑ | Precision↑ | Recall↑ | Human eval. SHE↑ | HDR↓ |
|---|---|---|---|---|---|
| Baseline | - | 0.84 | 0.63 | - | - |
| Random | 0.04 | 0.82 | 0.62 | 32.9 | 2.62 |
| GANSpace | 0.03 | **0.87** | 0.61 | 29.1 | 3.43 |
| SeFa | 0.05 | **0.87** | 0.62 | 30.3 | 2.88 |
| **Ctrl-SIS** | **0.12** | 0.85 | **0.64** | **60.7** | **1.07** |

Table 2. Evaluation of local class-specific image edits on ADE20K with OASIS. Ctrl-SIS shows higher diversity of latent directions ($\mathrm{mCD}_l$, SHE, HDR, Recall) while retaining good image quality (Precision).

mean pairwise LPIPS distance when different initial latent codes are applied while the class-specific latent direction is kept the same. The final mCC score is averaged over all label maps in the test set. Same as for $\mathrm{mCD}_l$, for the local $\mathrm{mCC}_l$ score, the edits are only applied to one class and the masked LPIPS distance is used. The lower $\mathrm{mCC}_l$ the more consistent are the class-specific edits for different initial latent codes. A detailed further description of the metrics is given in Sec. B in the supplementary. With the introduced metrics, a more unbiased and systematic comparison of GAN control discovery methods for SIS is possible.

**Human evaluation.** We also conduct a human evaluation of the learnt latent directions. To this end, we employ the SHE score from [47] and introduce a Human Diversity Rank (HDR) metric. For SHE, participants are shown two images edited only in the corresponding class area by applying the learnt class-specific latent direction. The final SHE score is the percentage of image pairs that the participants judge to be semantically different in the area of only one class. For HDR, participants are shown rows of locally edited images from four different methods (Random, SeFa, GANSpace, Ctrl-SIS) , as in Fig. C in the supplementary material but in a randomized order. The task is to rank the methods by their diversity. The final HDR score is an average rank (range 1 to 4) assigned to a GAN control discovery method. Each participant is provided with 50 questions and unlimited answering time for both scores.

### 4.3. Main results

We compare Ctrl-SIS, GANSpace and SeFa on global and local image editing. While local edits target a single class per image, global edits combine all class-specific edits within an image. The local edits show that the found directions encode semantic meaning, such aging faces, covering mountains in snow or turning on lamps (see Fig. 3 and 4). Global edits, change the whole image globally and are the result of combining all class edits in one image. In addition, we compare all methods to the performance of randomly sampled directions ("Random") as well as to the performance on unedited images ("Baseline").

As seen from Table 1, Ctrl-SIS achieves improved diversity by at least a factor of two, e.g., mCD of 0.26 vs. 0.12 of SeFa on ADE20K. The diversity of GANSpace and SeFa is lower (see numbers in red in Table 1) or close to random directions. Neither of these methods are designed to find class-specific directions. Yet, they still capture class-agnostic variations in the data, leading to directions that are closer to the mean of the image distribution and thus slightly better FID and mIoU. All methods exhibit similar consistency (mCC), with Ctrl-SIS and SeFa performing best on ADE20K, and SeFa and GANSpace on COCO-Stuff. The consistency of Ctrl-SIS is demonstrated in Fig. 4, e.g, where the learnt latent direction consistently defoliates trees or covers streets in snow. Similar to the consistency, the disentanglement (mOD) is strong for all methods, due to the spatially disentangled 3D latent space of the OASIS model.

| Model | Random | GANSpace | SeFa | Ctrl-SIS |
|---|---|---|---|---|
| OASIS | 0.04 | 0.03 | 0.05 | **0.12** |
| SC-GAN | 0.05 | 0.06 | 0.06 | **0.18** |
| SPADE | 0.05 | 0.08 | 0.06 | **0.09** |

Table 3. Comparison of $\mathrm{mCD}_l$ for GAN control methods across SIS models on ADE20K. Ctrl-SIS yields more diverse latent directions independent of the pretrained SIS model, while other methods find directions comparable to random sampling.

Due to the higher diversity of edited images, FID increases slightly for Ctrl-SIS compared to the baseline of unedited images (see Table 1). Since FID measures the overlap between the real and synthetic image distributions, images with weaker edits are closer to the original data. This can be visually confirmed in Fig. C in the supplementary material, where edits are shown side-by-side for all methods. Since SeFa and GANSpace only minimally change the class, their FID is close to the FID of unedited images (see Baseline in Table 1). Likewise, mIoU of images edited with Ctrl-SIS decreases, as the edited images move away from the mean mode of the synthetic image distribution. In contrast, for SeFa and GANSpace FID and mIoU are slightly better with respect to the baseline, while their diversity (mCC) is comparable to random directions. This observation suggests that SeFa and GANSpace images are closer to typical samples of the test set, while Ctrl-SIS learns more distinct directions.

We perform alternative evaluations of local class-specific edits in Table 2. Ctrl-SIS shows the highest recall and diversity ($\mathrm{mCD}_l$), and is the only method to improve both precision and recall over the OASIS baseline. Due to the precision-recall trade-off, SeFa and GANSpace have higher precision at the loss of recall, which is also reflected in their low $\mathrm{mCD}_l$ score and better FID over Baseline in Table 1. Moreover, both human diversity evaluation scores (SHE and HDR) are well aligned with the diversity metric $\mathrm{mCD}_l$ and recall, confirming the highest diversity of Ctrl-SIS.

Next, we compare Ctrl-SIS on different SIS models. Ta-

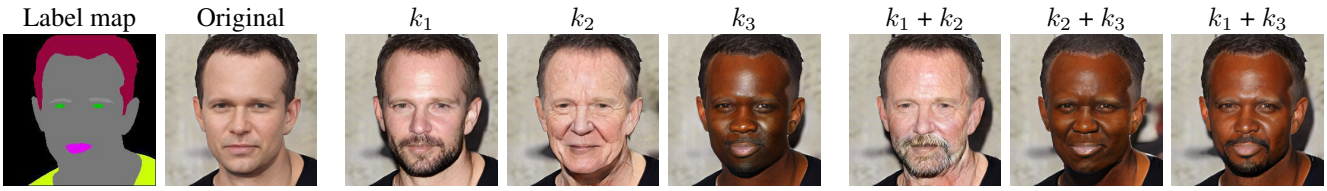


Figure 5. Combinations of directions $k_1$, $k_2$ and $k_3$ found for the *face* class (skin, neck, nose, ears) in the CelebAMask-HQ dataset. Different directions can be added in the latent space to combine their semantics, e.g., *k1 (beard)* and *k2 (age)* yield an old bearded person.


Figure 6. Optimizing Ctrl-SIS on the output features of ResNet blocks leads to an emphasis on structure (Res 1 and 2), while late normalization layers focus more on color (Norm 3 and 4). Directions from different layers can be combined (see last column).

ble 3 shows that Ctrl-SIS improves diversity strongly for local edits across all tested SIS models. SPADE naturally suffers from lower sensitivity to input latent code due to the strong regularization effect of its perceptual loss, as shown in [26]. While OASIS is trained without a perceptual loss, and SC-GAN uses a more powerful layer-wise conditioning strategy, leading to more diversity. Similar to Table 1, the diversity of GANSpace and SeFa is comparable to random directions. In other words, the directions that SeFa and GANSpace find differ just as much from each other as a set of randomly chosen directions. In contrast, the directions of Ctrl-SIS embody distinct appearances that are unlikely to appear in a set of random directions. An extended version of Table 3 with global metrics, FID and mIoU can be found in Table C in the supplementary material.

**Compositionality.** Individual class-specific latent directions can be combined. For example, Fig. 5 shows that directions corresponding to "age" and "beard" can be combined into "old and bearded". Further, the latent directions found by Ctrl-SIS depend on the subset of feature layers $G_l(z, y)$ of the SIS generator $G$ chosen for optimization (see Sec. 3). Fig. 6 highlights latent directions that were discovered by optimizing over layers from different blocks of the generator. For example, the set Norm 4 in Fig. 6 minimizes the loss over the first convolution in all conditional normalization layers [24] within the fourth ResNet block. While for the set Res 1 we minimized the loss for the final output features of the first ResNet block. We observe that the directions for Res 1 and 2 differ strongly in the internal structure of semantic classes, while Norm 3 and 4 encode changes in color. Interestingly, latent directions can be combined when synthesising images, by injecting different directions in different layers. In the last column of Fig. 6, a direction of an early ResNet block is injected into the first four blocks of the SIS model, while directions from the conditional normalization layers of a late ResNet blocks are injected from block five onwards. As the former directions encode structure, and the latter encodes colors, the resulting image combines both aspects.

| Method | mCD ↑ | mCC ↓ | mOD ↓ | FID ↓ | mIoU ↑ |
|---|---|---|---|---|---|
| **Ctrl-SIS** | 0.26 | **0.28** | **0.01** | 30.9 | 48.9 |
| No $\mathcal{L}_{div}$ | 0.24 | **0.28** | **0.01** | **30.5** | **49.4** |
| No $\mathcal{L}_{const}$ | 0.26 | 0.29 | **0.01** | 30.9 | 48.7 |
| No $\mathcal{L}_{dis}$ | **0.27** | **0.28** | 0.02 | 31.6 | 48.3 |

Table 4. Loss ablation of Ctrl-SIS on ADE20K.

**Ablation.** Table 4 presents an ablation on the proposed objective, using OASIS on the ADE20K dataset. Without the diversity term in our objective, the diversity decreases. Likewise, without the consistency or disentanglement term, consistency and disentanglement numbers worsen, respectively (see numbers in red). Further, the disentanglement term helps to improve synthesis and segmentation quality (see FID and mIoU in red), by helping to restrict the area affected by the edit only to the selected class area.

## 5. Conclusion

We propose Ctrl-SIS, which to our knowledge, is the first method for discovering class-specific interpretable GAN controls of SIS models. This is achieved by optimizing a set of class-specific latent directions via proposed diversity, consistency and disentanglement loss terms, making use of semantic label maps provided as part of the SIS task. The learnt latent directions can locally change the appearance of targeted semantic classes without affecting other classes in the image, and can be combined to sequentially change the image. Quantitative and qualitative analysis shows that Ctrl-SIS results in image edits of high quality, that are significantly more diverse than prior methods adapted to SIS.

## References


[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8296–8305, 2020. 3

[2] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Bolei Zhou, Joshua B. Tenenbaum, William T. Freeman, and Antonio Torralba. Gan dissection: Visualizing and understanding generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3

[4] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 6

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *International Conference on Learning Representations (ICLR)*, 2015. 6

[6] A. V. Cherepkov, Andrey Voynov, and Artem Babenko. Navigating the gan parameter space for semantic image editing. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[7] Edo Collins, Raja Bala, Bob Price, and Sabine Süsstrunk. Editing in style: Uncovering the local semantics of gans. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[8] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 1

[10] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2, 6

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. 2017. 6

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6

[13] Ali Jahanian, Lucy Chai, and Phillip Isola. On the" steerability" of generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[15] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *NeurIPS*, 2019. 6

[16] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3, 6

[17] Yuheng Li, Yijun Li, Jingwan Lu, Eli Shechtman, Yong Jae Lee, and Krishna Kumar Singh. Collaging class-specific gans for semantic image synthesis. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3

[18] Huan Ling, Karsten Kreis, Daiqing Li, Seung Wook Kim, Antonio Torralba, and Sanja Fidler. Editgan: High-precision semantic image editing. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 3

[19] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2, 3

[20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[21] Evangelos Ntavelis, Andrés Romero, Iason Kastanis, Luc Van Gool, and Radu Timofte. Sesame: Semantic editing of scenes by adding, manipulating or erasing objects. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[22] Ehsan Pajouheshgar, Tong Zhang, and Sabine Süsstrunk. Optimizing latent space directions for gan-based local image editing. *arXiv preprint arXiv:2111.12583*, 2021. 2, 3

[23] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Gaugan: semantic image synthesis with spatially adaptive normalization. In *ACM SIGGRAPH*. 2019. 2

[24] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 6, 8

[25] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 2

[26] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 3, 6, 8

[27] Sarah Schwettmann, Evan Hernandez, David Bau, Samuel Klein, Jacob Andreas, and Antonio Torralba. Toward a visual concept vocabulary for gan latent space. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[28] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[29] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 6

[30] Nurit Spingarn, Ron Banner, and Tomer Michaeli. Gan "steerability" without optimization. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[31] Ryohei Suzuki, Masanori Koyama, Takeru Miyato, Taizan Yonetsuji, and Huachun Zhu. Spatially controllable image synthesis with internal representation collaging. *arXiv preprint arXiv: 1811.10153*, 2018. 3

[32] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. Diverse semantic image synthesis via probability distribution modeling. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3

[33] Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, Gang Hua, and Nenghai Yu. Semantic image synthesis via efficient class-adaptive normalization. *arXiv preprint arXiv: 2012.04644*, 2020. 2

[34] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[35] Christos Tzelepis, Georgios Tzimiropoulos, and Ioannis Patras. Warpedganspace: Finding non-linear rbf paths in gan latent space. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[36] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning (ICML)*, 2020. 1, 2

[37] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[38] Yi Wang, Lu Qi, Ying-Cong Chen, Xiangyu Zhang, and Jiaya Jia. Image synthesis via semantic composition. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 6

[39] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3

[40] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision (ECCV)*, 2018. 6

[41] Oğuz Kaan Yüksel, Enis Simsar, Ezgi Gülperi Er, and Pinar Yanardag. Latentclr: A contrastive learning approach for unsupervised discovery of interpretable directions. In *International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6

[42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[43] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6

[44] Jiapeng Zhu, Ruili Feng, Yujun Shen, Deli Zhao, Zhengjun Zha, Jingren Zhou, and Qifeng Chen. Low-rank subspaces in gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3

[45] Jiapeng Zhu, Yujun Shen, Yinghao Xu, Deli Zhao, and Qifeng Chen. Region-based semantic factorization in gans. *arXiv preprint arXiv:2202.09649*, 2022. 2, 3

[46] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 3

[47] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. Semantically multi-modal image synthesis. In *CVPR*, 2020. 1, 2, 3, 6, 7