Discovering Class-Specific GAN Controls for Semantic Image Synthesis
Supplementary

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

April 6, 2023

This supplementary material is structured as follows:

A Qualitative results

A.1 Visual examples of joint class editing

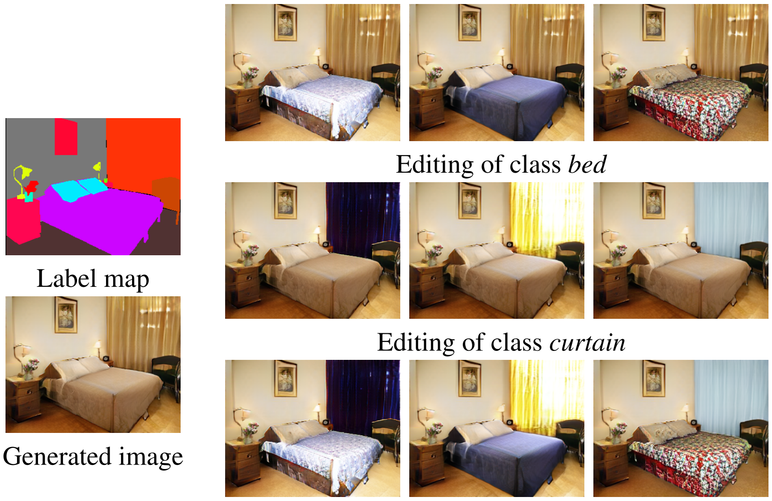
In Fig. A we show examples of images in which two classes are edited jointly with Ctrl-SIS. For example, Fig. A shows how bed and curtain, as well as window and wall can be edited separately or jointly. This property is enabled by the use of 3D latent codes, which are spatially aware and can vary image regions independently.

A.2 Visual examples of discovered directions

In this section we show latent directions learnt by Ctrl-SIS for the semantic image synthesis model OASIS. We pick OASIS, as it provides the best image quality and diversity (see Table 3 in the main paper). Results on ADE20K and COCO-Stuff are shown in Fig. B We observe that the directions are consistent across different label maps and change only the image area corresponding to the class of interest. The directions carry different semantics, such as the color of the bus, clouds in the sky, different kinds of house facades, various bed covers, different types of snow, and the lighting of the lamp.

A.3 Qualitative comparison to related work

Here, we visually compare the diversity between Ctrl-SIS, SeFa and GANSpace. These comparisons are presented in Fig. C. For all methods, the directions are applied to the 3D latent code within the image area corresponding to the selected class. We observe stronger diversity for Ctrl-SIS, which discovers meaningful class-specific directions. For example, in Fig. C Ctrl-SIS provides unique views from a window, tree leafage and street surfaces. The stronger diversity is explained by the fact that in contrast to SeFa and GANSpace, Ctrl-SIS is capable of leveraging the label maps that are already available for the task of semantic image synthesis during optimization to learn



Label map

Generated image

Editing of class *bed*

Editing of class *curtain*

Editing of class *bed* and *curtain* jointly

class-specific directions.

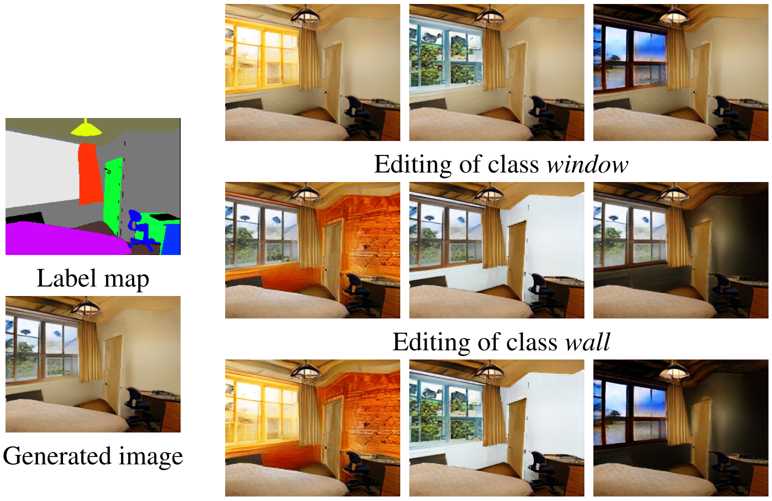
B Quantitative evaluation of class-specific GAN controls

In this section we provide details on our proposed metrics for evaluating GAN controls discovery methods and relate them to prior work [6]. A method for discovering semantically meaningful class-specific directions in the latent space of SIS GANs should exhibit the following three properties: First, the found directions should be as unique and different as possible. We assess this property via the *mean control diversity* - mCD. Second, a latent direction should invoke the same semantic edit independent of the initial latent code, which we assess via the *mean control consistency* - mCC. Third, class-specific edits should not affect image areas outside of the target class area. We verify this requirement via the *mean outside class diversity* - mOD. The scores are based on computing the LPIPS distance between pairs of images with different edits and the same initial latent code (mCD and mOD), or the same edits but different initial latent codes (mCC). For the global mCD and mCC scores the edits are applied to all classes simultaneously with latent directions that are randomly picked from the set of discovered class-specific directions. On the other hand, the local scores mCD_l, mCC_l and mOD rely on pairwise distances between images where only one class is edited at a time. To compute the pairwise distance between images where only one class is edited, we use the *masked* LPIPS distance. In the following, we explain the masked LPIPS distance and provide the formulations of the local scores mCD_l, mCC_l and mOD, as well as the global scores mCD and mCC.

The masked LPIPS distance. The default LPIPS distance between two images is based on extracting deep features from both images using a VGG network pretrained on ImageNet classification [4]. The features of all layers are normalized and re-scaled along the channel dimension. The final LPIPS distance is the L2 distance between these features. To compute the *masked* LPIPS distance, we multiply the deep features with a binary mask before computing the L2 distance. We distinguish between LPIPS^{M_c} and LPIPS^{1-M_c}. The former uses the binary mask M_c, which is 1 where the label map contains class c and 0 everywhere else. The latter applies the inverted mask $1 - M_c$.

Mean control diversity. The mean control diversity is computed for global edits (mCD) and local edits (mCD_l). The mCD_l is computed via:

$$\mathrm{mCD}_l = \frac{1}{C}\sum_{c=1}^{C}\mathbb{E}_c[\mathcal{P}_{CD}], \qquad (1)$$



Editing of class *window*

Editing of class *wall*

Editing of class *window* and *wall* jointly

Figure A: Joint editing of semantic classes using latent directions learnt by Ctrl-SIS.

where C is the total number of classes and $\mathcal{P}_{CD}$ denotes the control diversity measured for a label map containing class c. To compute $\mathcal{P}_{CD}$, a fixed initial latent code is sampled for each label map containing class c. Given a label map and its initial latent code, one locally edited image is created for each of the K latent directions specific to class c. Next, the average locally masked LPIPS distance is computed between all pairs of the K edited images. This score is averaged over Z initial latent codes, which can be formulated as follows:

$$\mathcal{P}_{CD} = \frac{1}{ZK} \sum_{z}^{Z} \sum_{\substack{k_{1,2}=1 \\ k_1 \neq k_2}}^{K} \mathrm{LPIPS}_{z,k_1,k_2}^{M_c}. \tag{2}$$

Here, $\mathrm{LPIPS}_{z,k_1,k_2}^{M_c}$ denotes the LPIPS distance masked with M_c between two images created with the same initial latent code z, where class c is edited with latent direction k_1 and k_2, respectively.

The mCD for global edits is computed as the average distance between globally edited images on the same label map. For each label map, we create pairs of images with different global edits, changing all classes at once. The class-specific latent directions are randomly chosen for each class. We compute the mean of the default LPIPS distance over all pairs and different initial latent codes. The score is averaged over all label maps in the test set. Higher mCD and mCD$_l$ scores indicate better diversity.

Mean outside class diversity. The spatial disentanglement metric mOD is computed for local edits via

$$\mathrm{mOD} = \frac{1}{C} \sum_{c=1}^{C} \mathbb{E}_c \left[\mathcal{P}_{OD} \right], \tag{3}$$

where $\mathcal{P}_{OD}$ is the outside class diversity measured for a label map containing class c. In contrast to mCD$_l$, the masked LPIPS is computed for the area outside the target class:

$$\mathcal{P}_{OD} = \frac{1}{ZK} \sum_{z}^{Z} \sum_{\substack{k_{1,2}=1 \\ k_1 \neq k_2}}^{K} \mathrm{LPIPS}_{z,k_1,k_2}^{1-M_c}. \tag{4}$$

$\mathrm{LPIPS}_{z,k_1,k_2}^{1-M_c}$ denotes the LPIPS distance masked with $1 - M_c$ between two images created with the same initial latent code z, where class c is edited locally with the latent direction k_1 and k_2, respectively. A lower mOD indicates better spatial disentanglement.

Mean control consistency. Lastly, to measure the consistency of an edit under different initial latent codes, we compute the mean control consistency for global edits

Method	ADE20K					COCO-Stuff				
	$\text{mCD}_l \uparrow$	$\text{mCC}_l \downarrow$	$\text{mOD} \downarrow$	FID↓	mIoU↑	$\text{mCD}_l \uparrow$	$\text{mCC}_l \downarrow$	$\text{mOD} \downarrow$	FID↓	mIoU↑
Baseline	-	-	-	28.6	52.2	-	-	-	17.1	42.4
Random	0.04	0.17	**0.01**	30.6	50.1	0.02	0.07	**0.00**	17.2	44.0
GANSpace	0.03	**0.15**	**0.01**	**28.3**	**53.9**	0.02	**0.06**	**0.00**	**16.7**	43.6
SeFa	0.05	**0.15**	**0.01**	**28.3**	53.7	0.02	**0.06**	**0.00**	16.9	44.2
Ctrl-SIS	**0.12**	0.16	**0.01**	28.8	51.6	**0.04**	0.07	0.01	17.5	**44.4**

Table A: Evaluation of OASIS GAN controls on ADE20K and COCO-Stuff on local class-specific edits.

	LPIPS			MS-SSIM		
	mCD ↑	mCC ↓	mOD ↓	mCD ↓	mCC ↑	mOD ↑
Random	0.11	0.30	**0.01**	0.98	0.76	**1.0**
GANSpace	0.09	0.29	**0.01**	0.94	**0.78**	**1.0**
SeFa	0.12	**0.28**	**0.01**	0.92	**0.78**	**1.0**
Ctrl-SIS	**0.26**	**0.28**	**0.01**	**0.74**	0.77	**1.0**

Table B: Evaluation of GAN controls with LPIPS and MS-SSIM using OASIS on ADE20K.

(mCC) and local edits (mCC_l). The mCC_l is

$$\text{mCC}_l = \frac{1}{C}\sum_{c=1}^{C}\mathbb{E}_c\left[\mathcal{P}_{CC}\right], \tag{5}$$

where $\mathcal{P}_{CC}$ is the control consistency of a label map containing class c. We compute the pairwise distances between images with different initial latent codes and the same local edit:

$$\mathcal{P}_{CC} = \frac{1}{ZK}\sum_{k}^{K}\sum_{\substack{z_{1,2}=1 \\ z_1 \neq z_2}}^{Z}\text{LPIPS}_{k,z_1,z_2}^{M_c}. \tag{6}$$

Here, $\text{LPIPS}_{k,z_1,z_2}^{M_c}$ denotes the LPIPS distance masked with M_c between two images created with different initial latent codes z_1 and z_2, where class c is edited locally with latent direction k for both images.

The global mCC score is computed as the average distance between images with the same global edit but different initial latent codes. For each label map, we create pairs of images with different initial latent codes, but a shared global edit. We compute the mean of the default LPIPS distance over all pairs and across different shared global edits. The score is averaged over all label maps in the test set. Ideally, the mCC and mCC_l are low, indicating high consistency under different initial latent codes.

Relation to prior diversity and disentanglement scores. The mCD_l and mCC_l are related to the *mean class diversity* (mCSD) and *mean other class* (mOCD) proposed by [6]. These two metrics evaluate diversity and spatial disentanglement for SIS models that allow class-specific manipulations [6, 2]. Note that mCSD and mOCD measure the class-specific diversity and disentanglement of a *SIS model*, while our metrics evaluate the class-specific diversity and disentanglement of *a set of discovered latent directions*, allowing us to compare different control discovery methods on the same SIS model. The mCSD measures intra-class diversity as a property of the SIS model itself. In contrast, mCD_l measures the diversity of a set of latent directions, which is a property of the GAN control discovery method. The same relationship holds between mOCD and mOD. We next present an extended evaluation using our proposed local metrics mCD_l, mCC_l and mOD.

C Extended quantitative evaluation

C.1 Evaluation on local class-specific edits

In this section we present an additional comparison between Ctrl-SIS and the related work using OASIS on the ADE20K and COCO-Stuff datasets. For evaluation we employ image quality metrics (FID and mIoU) as well as our proposed diversity (mCD), consistency (mCC) and disentanglement (mOD) scores. In contrast to Table 1 in the main paper, Table A presents this comparison for *local* edits. While the related work SeFa and GANSpace are designed for global edits, local edits are achieved by adding the learnt global directions to the 3D latent code only in a class-specific image area, as demonstrated in Fig. C. As Table A shows, Ctrl-SIS achieves at least twice the diversity score with respect to SeFa and GANSpace, while the consistency and disentanglement scores stay similar between all methods. The red numbers mark scores which are equal or worse than the ones originating from random directions. For SeFa and GANSpace, the FID and mIoU are slightly improved compared to unedited images (see Baseline in Table A), due to generating more ”typical” images (see Sec. 4.3 in the main paper). In summary, the results from Table A are in alignment with Table 1 (main paper), suggesting that the editing properties of Ctrl-SIS and related works are similar between local and global edits.

C.2 Evaluation with alternative distance measure

Our proposed scores are based on computing the mean LPIPS distance between pairs of images. Here, we also present our metrics computed with the multi-scale structural similarity distance (MS-SSIM) [3] as an alternative to LPIPS. The main differences between LPIPS and MS-SSIM are as follows. LPIPS computes the L2 distance between image features extracted with a network pre-trained on ImageNet classification. MS-SSIM is not neural network-based and instead computes the similarity between images based on the mean, variance and covariance of two images. A high similarity between two images results in high MS-SSIM but low LPIPS, since LPIPS measures dissimilarity. This means MS-SSIM-based mCD, mCC and mOD scores rise when the LPIPS-based scores fall, and vice versa. In Table B, we compare GAN control discovery methods with our metrics based on LPIPS and MS-SSIM. We note the same trends between the MS-SSIM-based metrics and the LPIPS-based metrics. In particular, Ctrl-SIS also sees a strong increase in diversity under the MS-SSSIM-based mCD metric. The results show that the evaluation metrics are not strictly dependent on the distance measure, and that other ways of estimating image (dis-) similarity may be used.

Model	Method	Global edits			Local edits		
		mCD ↑	FID ↓	mIoU ↑	mCD$_l$ ↑	FID ↓	mIoU ↑
OASIS	Random	0.11	31.3	49.4	0.04	30.6	50.1
	GANSpace	0.09	**28.1**	**53.3**	0.03	**28.3**	**53.9**
	SeFa	0.12	**28.1**	53.2	0.05	**28.3**	53.7
	Ctrl-SIS	**0.26**	30.9	48.9	**0.12**	28.8	51.6
SC-GAN	Random	0.08	34.3	38.1	0.05	**34.2**	38.6
	GANSpace	0.11	**34.2**	**38.3**	0.06	34.3	38.8
	SeFa	0.10	34.4	37.8	0.06	34.4	**38.9**
	Ctrl-SIS	**0.25**	36.4	34.7	**0.18**	**34.2**	38.4
SPADE	Random	0.08	**34.6**	**39.4**	0.05	34.6	39.6
	GANSpace	0.12	35.1	39.3	0.08	**34.6**	**39.7**
	SeFa	0.09	34.7	**39.4**	0.06	34.8	**39.7**
	Ctrl-SIS	**0.14**	35.4	38.6	**0.09**	**34.6**	39.4

Table C: Comparison of GAN control methods across SIS models on ADE20K.

C.3 Comparison of GAN control methods across SIS models

In this section, we compare Ctrl-SIS on different SIS models. Table C shows that Ctrl-SIS exhibits stronger diversity for local and global edits across all tested SIS models. The diversity of GANSpace and SeFa is comparable to the diversity measured for random directions (see red numbers in Table C). In other words, the directions that SeFa and GANSpace find differ just as much from each other, as a set of randomly chosen directions.

A visual comparison of the diversity of Ctrl-SIS, SeFa and GANSpace is shown in figure B: In contrast to SeFa and GANSpace, Ctrl-SIS yields latent directions with distinct appearances. The directions of GANSpace and SeFa all look very similar. Note that this is comparable to a set of *random* directions. In contrast to regular unconditional or class-conditional GANs, random directions in SIS yield images with low diversity. The low diversity of random directions is a well-known issue for SIS models [1, 5, 2].

D Limitations

There are two main limitations to what a method for class-specific latent direction discovery can do. First, class-specific directions do not encode shape-based semantics. For example, there cannot be a class-specific direction encoding a ”smile” in a face dataset if the shape of the mouth is already hard-coded by the label map. Second, the diversity of Ctrl-SIS is limited by the diversity of the SIS model to which it is applied. Notably, the diversity of SIS models is far lower than the diversity of regular unconditional or class-conditional GANs. While a standard unconditional GAN produces seemingly infinitely many different images, the diversity of SIS models like OASIS [2] was limited to a manageable number of distinct appearances, based on our experience. The problem of diversity in SIS models is a well-known problem [1, 5, 2]. Consequently, more diverse SIS models will lead to more diverse sets of discovered latent directions.

References

[1] 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. 5, 7

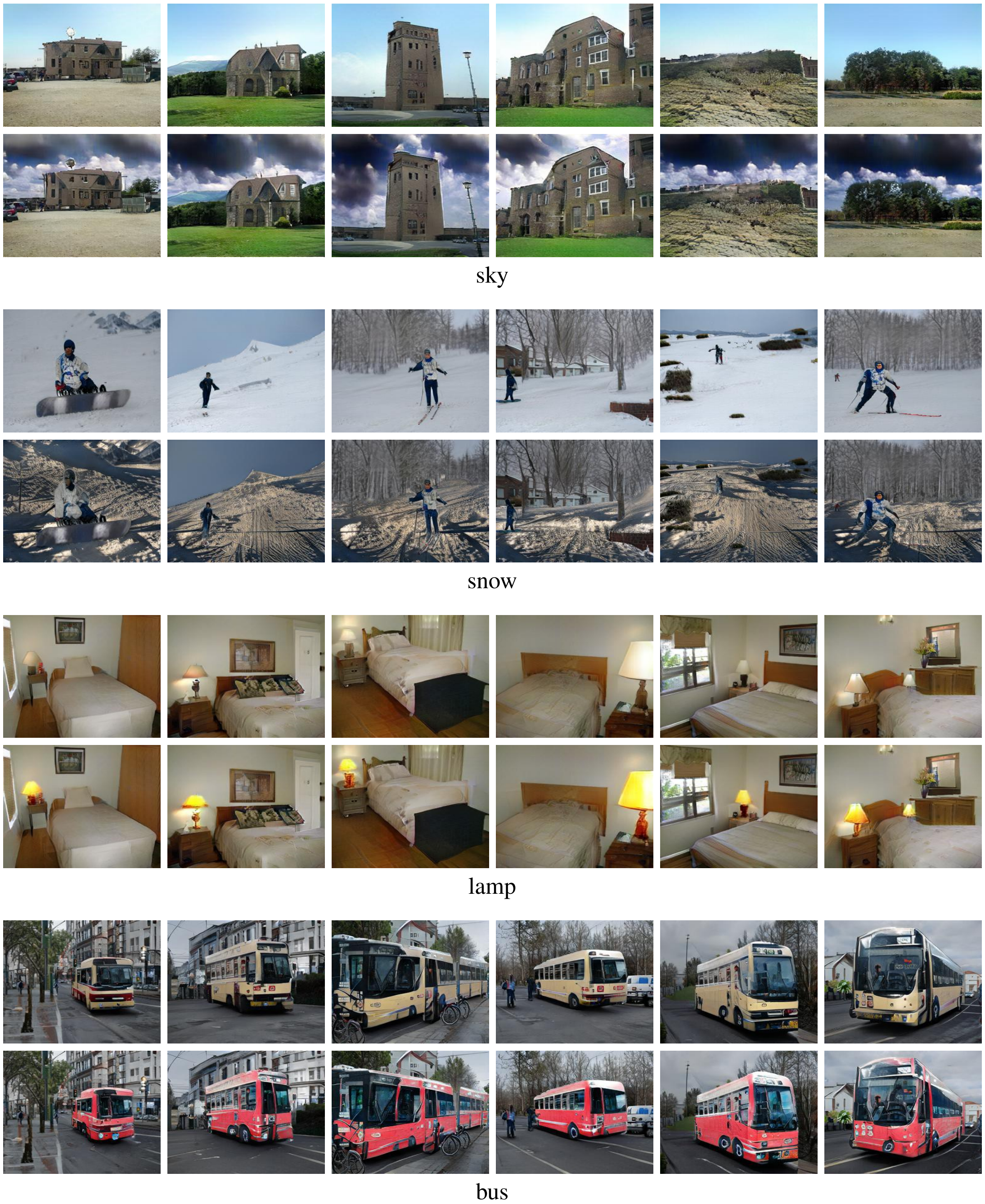
[2] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *International Conference on Learning Representations (ICLR)*, 2021. 4, 5, 7

[3] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers*, 2003. 5

[4] 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. 2

[5] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017. 5, 7

[6] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. Semantically multi-modal image synthesis. In *CVPR*, 2020. 2, 4


sky

snow

lamp

bus

Figure B: Latent directions learnt by Ctrl-SIS on ADE20K and COCO-Stuff.

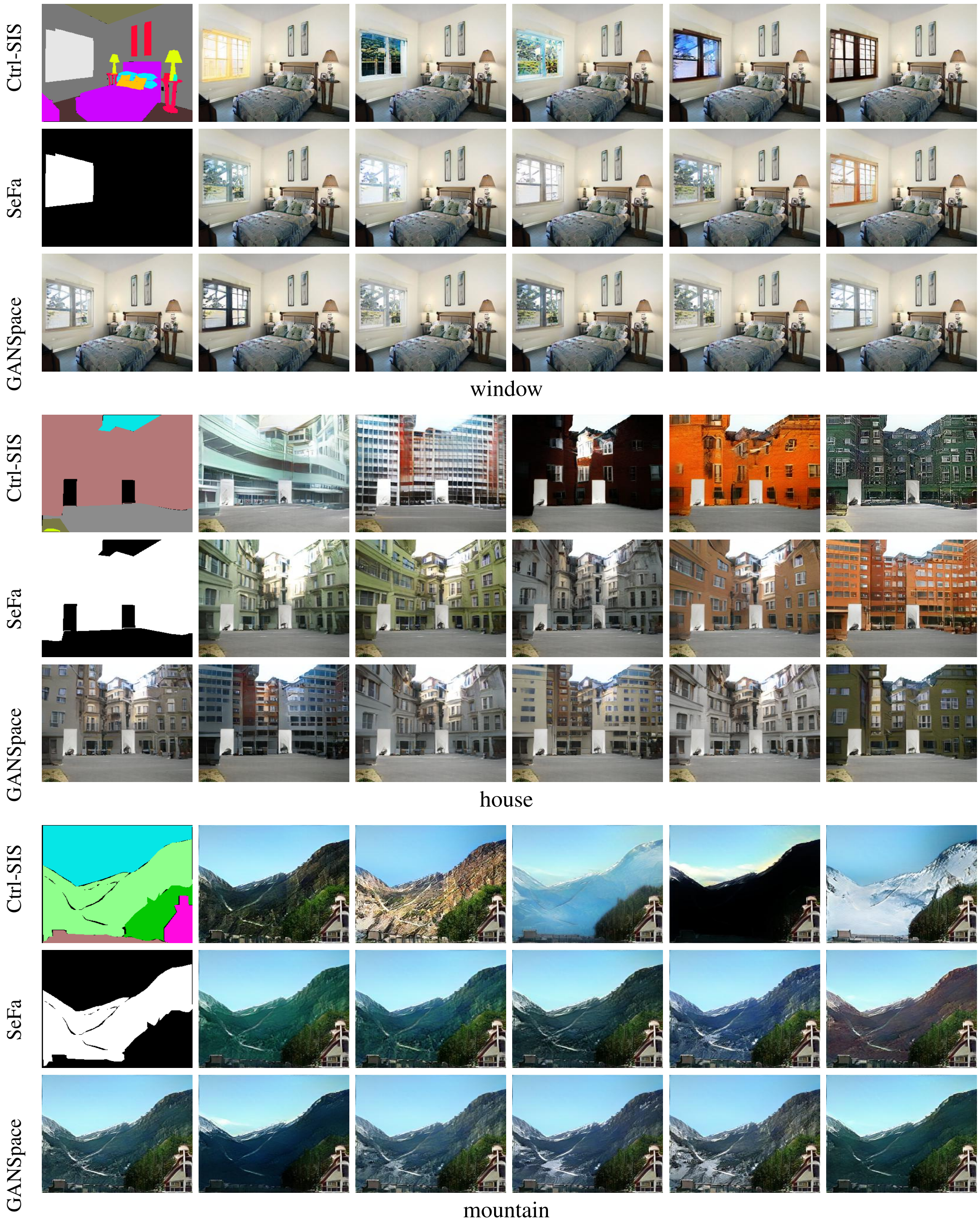


Figure C: Qualitative comparison of Ctrl-SIS against SeFa and GANSpace.