Few-shot Semantic Image Synthesis with Class Affinity Transfer

Marlène Careil[1,2] Jakob Verbeek[2] Stéphane Lathuilière[1]
[1]LTCI, Télécom Paris, IP Paris [2]Meta AI

Supplementary material

In this supplementary material, we provide additional details regarding the code and datasets used in our experiments in Section A. In Section B, we detail the proposed used architectures. In Section C, we provide additional implementation details. In Section D we list the class correspondences between source and target found by our method. We report extended quantitative results in Section E, and results, and provide additional qualitative results in Section F.

A. Assets and licensing information

In Table 1, we provide the links to the datasets and code repositories we used, and list their licenses in Table 2.

To avoid training our generative models on sensitive personal data, we train our models on filtered data. For Cityscapes, we use the version of the dataset in which human faces and license plates have been blurred, as provided through the dataset distribution website. For ADE20K and COCO, we processed the data ourselves to detect human faces, segment them, and blur them. We find that overall this processing has limited impact on performance. For the OASIS source models trained on COCO and ADE20K: we find that the FID changes from 17.0 to 18.7 for COCO, and from 28.3 to 29.8 for ADE20K. The FID metric is still computed w.r.t. the original non-filtered datasets.

B. Architectural designs

To complement the description of our method provided in the main paper, we report additional technical details.

Architectural design to adversarial model. In Figure 1 (top panel), we illustrate the modifications to SPADE blocks used in the OASIS generator to include our class affinity matrix. In red, we display original blocks from OASIS generator which are initialized with pretrained weights from the source model during finetuning on a target dataset. We prepend a 1×1 convolution indicated in orange blocks initialized with our affinity class matrix **A**. The residual layers are shown in yellow and are initialized with zeros. They take as input the target segmentation maps. During the first finetuning stage, only red blocks with open padlock in the drawing as well as all the orange and yellow blocks are finetuned, while in a second stage we train all layers in the spade blocks.

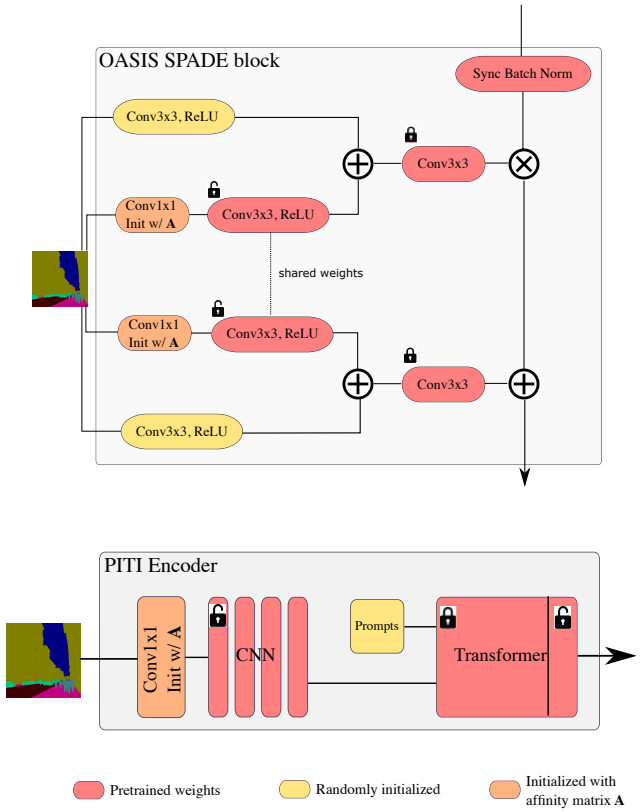

Figure 1. Overview of our modifications of the OASIS and PITI architectures for semantic image synthesis. The class affinity matrix is used to align the source model with the target label space, and then models are further finetuned using the target data.

We provide an ablation in Table 3 on whether using shared or separate paths to compute scale and shift parameters. When shared, the modules in the SPADE block are not duplicated, and the last 3×3 convolutional layer outputs both the scale and shift parameters. When separate, all blocks are duplicated as in Figure 1 (top panel), but weights

Dataset	
COCO-Stuff [1]	https://cocodataset.org/
Cityscapes [2]	https://www.cityscapes-dataset.com/
Ade20K [8]	http://groups.csail.mit.edu/vision/datasets/ADE20K/
Model	
OASIS [4]	https://github.com/boschresearch/OASIS
PITI [6]	https://github.com/PITI-Synthesis/PITI

Table 1. Links to the assets used in the paper.

Dataset	
COCO-Stuff [1]	https://www.flickr.com/creativecommons
Cityscapes [2]	https://www.cityscapes-dataset.com/license
Ade20K [8]	MIT License
Model	
OASIS [4]	https://github.com/boschresearch/OASIS/blob/master/LICENSE
PITI [6]	https://github.com/PITI-Synthesis/PITI/blob/main/LICENSE

Table 2. Assets licensing information.

Paths	**COCO→ADE** ↓**FID**	↑**mIoU**
Shared	41.4	30.1
Separate	**40.9**	**31.4**

Table 3. Ablation on using shared or separate paths in the SPADE blocks of the modified OASIS architecture.

Model	**Aff. mat.**	**COCO → ADE** ↓**FID**	↑**mIoU**
OASIS	Hard	**40.9**	**31.4**
	Soft	44.7	30.1
PITI	Hard	**40.7**	**22.3**
	Soft	43.6	19.8

Table 4. Ablation on the use of hard *vs*. soft affinity matrices. Affinities computed with supervised segmentation network.

are shared on the first 3×3 convolutional layer. The results indicate benefit in mIoU and FID when using separate paths, and we use this option in our main experiments.

Architectural changes to diffusion model. In Figure 1 (bottom panel), we display our changes to the PITI encoder in orange and yellow blocks. Similarly as for OASIS, we prepend a 1×1 convolution initialized by the class affinity matrix. Then, as specified in the method section from the main paper, we add trainable "prompt" tokens randomly initialized at the input of Transformer block. During training, we finetune these prompts, the 1×1 convolution, the CNN, as well as the last residual block of the Transformer.

Hard or soft affinity matrices. In Table 4, we conduct experiments by using hard or soft affinity matrices. More precisely, we compute the soft affinity matrices as described in Section 3.1 of the main paper. For the "hard" version, for each target class we binarize the affinities by setting the largest value to one, and the rest to zeros. The results show that the hard version yields consistent gains in FID and mIoU for both for PITI and OASIS, and we retain it in our main experiments.

C. Complementary training details

When using the GAN-based OASIS architecture [4], we use the publicly available checkpoints of as our source models. For the diffusion-based PITI architecture [6], we use the released checkpoint for COCO-Stuff. The checkpoint for ADE20K is not released, and we therefore trained the model ourselves using the public code, obtaining an FID of 27.8, comparable to the 27.3 reported by the authors.

We monitor finetuning on target datasets by computing FID on the validation set and employ early stopping using this FID criterion. We optimize learning rates on both OASIS and PITI. For PITI, we try learning rates $lr = \{3.5e-5, 7e-5, 1e-4, 2e-4\}$, and $lr = \{1e-4, 4e-4, 8e-4\}$ for OASIS. For PITI, we choose a learning rate of $7e-5$ when transferring to Cityscapes, while we set $lr = 2e-4$ when doing transfer to ADE20K and COCO-Stuff. For OASIS, we use $lr = 4e-4$ when transferring to Cityscapes while we set $lr = 1e-4$ for ADE20K and COCO-Stuff.

D. Estimated class affinities

In Table 5 and Table 6 we show the class mappings obtained for Cityscapes and ADE20K target classes with COCO-Stuff classes as source using the combination method. We see that most of the class correspondences are coherent, but there are a few aberrations. For instance, the "truck" class from Cityscapes is associated with

Cityscapes	COCO	Cityscapes	COCO
unlabeled	sky-other	ego vehicle	road
rectification border	unlabeled	out of roi	unlabeled
static	building-other	dynamic	building-other
guard rail	railing	bus	bus
ground	pavement	bridge	building
road	road	pole	building-other
sidewalk	pavement	polegroup	fence
parking	road	traffic light	traffic light
rail track	road	traffic sign	builidng-other
building	building-other	vegetation	tree
wall	building-other	terrain	grass
fence	fence	sky	sky-other
person	person	rider	person
car	car	truck	truck
caravan	truck	trailer	truck
train	bus	motorcycle	motorcycle
bicycle	bicycle	license plate	unlabeled
tunnel	building		

Table 5. Affinity class mappings for Cityscapes target classes with COCO-Stuff source classes using combination method.

"building" class from ADE20K. This could be explained by image style discrepancy existing between ADE20K and Cityscapes, as samples from Cityscapes contain darker images and less colorful than ADE20K, which is misleading image-based methods for class similarity estimation.

E. Additional quantitative results

In tables 8, 9, 10, 11, we complement the quantitative results in the main paper by showing results when using Cityscapes as target dataset.

We provide results after finetuning and in training-free mode in Table 8 and Table 9 respectively. Similar to what was observed when transferring to ADE20K and COCO-Stuff, we see that in general the supervised and text-based methods perform better than the self-supervised one. Besides, the combination method obtains results that are among the best with respect to the different methods.

We add ablations on PITI and OASIS in Table **??** and Table **??**. The results are consistent, and show the benefit of each component added in the OASIS and PITI pipelines.

Hand-designed affinity matrix estimation. We did the ablation of manually assigning target classes to their most similar source classes in Table 7. In COCO→ADE, it requires manually checking 151×183 pairs, while CAT automates this. With OASIS, we obtained performance similar to CAT: FID of 41.3 *vs*. 40.9 and a mIoU of 31.6 *vs*. 31.4.

KID comparisons. In Table 12, we add KID metrics to further evaluate the benefit of finetuning compared to training-free mode on OASIS and PITI with ADE20K and COCO-

ADE	COCO	ADE	COCO
wall	wall-concrete	building'	building-other
sky	sky-other	floor	floor-tile
tree	tree	ceiling	ceiling-other
road	road	bed	bed
windowpane	window-other	grass	grass
cabinet	cabinet	sidewalk	pavement
person	person	earth	dirt
door	door-stuff	table	table
mountain	mountain	plant	potted plant
curtain	curtain	chair	chair
car	car	water	water-other
painting	building-other	sofa	couch
shelf	shelf	house	house'
sea	river	mirror	mirror-stuff
rug	rug	field	grass
armchair	couch	seat	chair
fence	fence	desk	desk
rock	rock	wardrobe	cabinet
lamp	light	bathtub	toilet
railing	clock	cushion	couch
base	house	box	plastic
signboard	street sign	chest	cabinet
counter	counter	sand	sand
sink	sink	skyscraper	skyscraper
fireplace	wall-stone	refrigerator	refrigerator
grandstand	platform	path	road
stairs	stairs	runway	road
case	counter	pool	boat
pillow	bed	screen	wall-tile
stairway	stairs	river	river
bridge	bridge	bookcase	shelf
blind	window-blind	coffee	table
toilet	toilet	flower	flower
book	book	hill	hill
bench	bench	countertop	counter
stove	oven	palm	tree
kitchen	dining table	computer	tv
swivel	chair	boat	boat
bar	desk	arcade	platform
hovel	wall-concrete	bus	bus
towel	towel	light	light
truck	truck	tower	building-other
chandelier	light	awning	tent
streetlight	building-other	booth	desk
television	tv	airplane	airplane
dirt	dirt	apparel	clothes
pole	metal	land	hill
bannister	railing	escalator	stairs
ottoman	furniture-other	bottle	bottle
buffet	counter	poster	mirror-stuff
stage	counter	van	car
ship	boat	fountain	sink
conveyer	oven	canopy	ceiling-other
washer	oven	plaything	teddy bear

Table 6. Affinity class mappings for ADE20K target classes with COCO-Stuff source classes using combination method.

	COCO →ADE	
	↓FID	↑mIoU
Computed	**40.9**	31.4
Hand-designed	41.3	**31.6**

Table 7. Ablation on the use of hand-designed affinity matrices *vs*. ones computed with our Combination method.

Stuff dataset. We report improvements in all the settings after finetuning.

F. Additional qualitative results

Comparison with the state of the art. In Figure 2 and Figure 3, we compare our CAT approach to the best state-of-the-art models in Table 2 of the main paper. For OASIS, we compare to cGANTransfer [5], TransferGAN [7], FreezeD [3] as they get the best FID/mIoU scores for one of the four source-target dataset pairs. We show samples for the four source-target dataset pairs and observe images with superior quality using our approach. It is particularly striking for PITI when comparing to finetuning all layers from a pretrained PITI model, or finetuning from a GLIDE checkpoint. In this case, generated images do not adhere well to label maps and are of poor quality.

Qualitatives with different target dataset sizes. In Figure 4, we complete Figure 4 of main paper by showing qualitative samples from OASIS with COCO-Stuff as source and using ADE20K as target, with target dataset of sizes 25, 100, 400, and 20k, where 20k corresponds to the full dataset. We notice how image quality gradually improves with the number of images in the target dataset.

Training-free samples. In Figure 5, we show samples obtained without finetuning the model, by using the pretrained weights and only adding our affinity matrix mapping to the model. We compare them to synthesized images conditioned on the same segmentation map and noise input after finetuning on three pairs of source-target datasets: COCO-Stuff→ADE20K and vice-versa, as well as COCO-Stuff→Cityscapes.

We observe that target classes which do not exist in source dataset are poorly rendered in the training-free mode. For instance, the closest COCO-Stuff class to "painting" in ADE20K is "building-other" according to our combination method, while "cradle" is associated to COCO-Stuff class "bed". We can see in the top right block of the Figure 5 in rows 2 and 4 that these objects are better recognizable after finetuning. It is also worth noticing that both PITI and OASIS generators adapt well to the style of the target dataset when transferring from COCO-Stuff or ADE20K to Cityscapes. While training-free samples look colorful and bright, images after finetuning are darker, more in line with the style of the original Cityscapes images.

	Affinity matrix initialization	COCO → ADE ↓FID	COCO → ADE ↑mIoU	ADE → COCO ↓FID	ADE → COCO ↑mIoU	ADE → Cityscapes ↓FID	ADE → Cityscapes ↑mIoU	COCO → Cityscapes ↓FID	COCO → Cityscapes ↑mIoU
OASIS	From scratch	55.0	29.8	79.8	17.8	55.3	64.0	47.6	66.0
	Text-based	41.1	30.4	55.2	17.3	51.6	66.0	47.7	67.3
	Supervised	42.0	30.8	58.2	12.9	51.8	65.7	**46.7**	68.2
	Self-supervised	41.3	29.8	57.9	15.4	53.0	65.1	47.5	67.6
	Combination	**40.9**	**31.4**	**53.7**	**17.4**	51.3	**66.4**	47.0	**68.1**
PITI	Random	57.1	11.6	83.7	0.8	65.4	20.3	58.5	33.1
	Text-based	40.9	20.2	47.4	7.1	**61.1**	**28.5**	57.1	38.0
	Supervised	41.1	22.0	52.5	5.3	63.0	26.7	57.3	**40.8**
	Self-supervised	41.9	21.2	50.9	5.6	63.8	27.5	57.7	39.4
	Combination	**40.7**	**22.3**	**46.8**	**7.5**	62.7	27.3	**54.7**	39.9

Table 8. Comparison of different class affinity estimation methods with target datasets of 100 images. Results after finetuning.

	Affinity matrix initialization	COCO→Cityscapes ↓FID	COCO→Cityscapes ↑mIoU	COCO→ADE ↓FID	COCO→ADE ↑mIoU	ADE→COCO ↓FID	ADE→COCO ↑mIoU	ADE→Cityscapes ↓FID	ADE→Cityscapes ↑mIoU
OASIS	Random	216.4	2.0	216.0	0.5	270.8	0.1	326.9	2.4
	Text-based	130.5	26.2	44.6	23.9	57.8	15.2	138.8	35.6
	Supervised	**82.1**	32.1	46.9	47.0	22.8	64.5	12.1	36.3
	Self-supervised	88.7	31.0	45.5	22.7	68.1	10.2	94.2	43.3
	Combination	82.4	33.4	**43.1**	**25.1**	**56.3**	**13.8**	**79.2**	36.9
PITI	Random	254.7	3.8	96.3	5.7	98.3	0.1	287.4	2.0
	Text-based	103.6	20.6	51.6	19.2	50.8	7.1	75.3	22.3
	Supervised	**78.1**	23.6	48.7	20.2	59.2	5.2	**72.3**	22.9
	Self-supervised	88.7	**24.0**	49.5	19.6	53.9	5.0	92.9	22.1
	Combination	78.7	23.8	**48.5**	**20.9**	**49.3**	**7.4**	72.6	**22.9**

Table 9. Transfer with a target dataset of size 100 using training-free approach.

FreezeD	2 stage	Resid.	CAT	COCO→ADE ↓FID	COCO→ADE ↑mIoU	ADE→COCO ↓FID	ADE→COCO ↑mIoU	ADE→Cityscapes ↓FID	ADE→Cityscapes ↑mIoU	COCO→Cityscapes ↓FID	COCO→Cityscapes ↑mIoU
✗	✗	✗	✗	87.2	20.7	117.3	11.0	56.0	61.6	51.3	63.7
✓	✗	✗	✗	65.7	25.8	98.6	14.8	56.9	62.9	49.7	66.6
✓	✓	✗	✗	55.9	28.6	83.4	15.2	55.2	63.1	50.0	66.2
✓	✓	✓	✗	55.2	29.4	79.9	15.7	52.7	65.5	47.6	66.0
✓	✓	✓	✓	**40.9**	**31.4**	**53.7**	**17.4**	**51.3**	**66.4**	**46.9**	**68.3**

Table 10. Ablation on OASIS-based architecture.

FixDec	Prompts	CAT	COCO → ADE ↓FID	COCO → ADE ↑mIoU	ADE → COCO ↓FID	ADE → COCO ↑mIoU	ADE → Cityscapes ↓FID	ADE → Cityscapes ↑mIoU	COCO → Cityscapes ↓FID	COCO → Cityscapes ↑mIoU
✗	✗	✗	84.0	0.1	85.9	17.4	70.5	36.9	56.5	14.4
✓	✗	✗	52.4	13.6	79.0	1.4	66.3	20.0	62.2	33.5
✓	✓	✗	51.1	14.1	78.9	1.3	65.4	20.3	58.5	33.1
✓	✓	✓	**40.7**	**22.3**	**46.8**	**7.5**	**62.5**	**27.2**	**57.3**	**40.8**

Table 11. Ablation on PITI-based architecture.

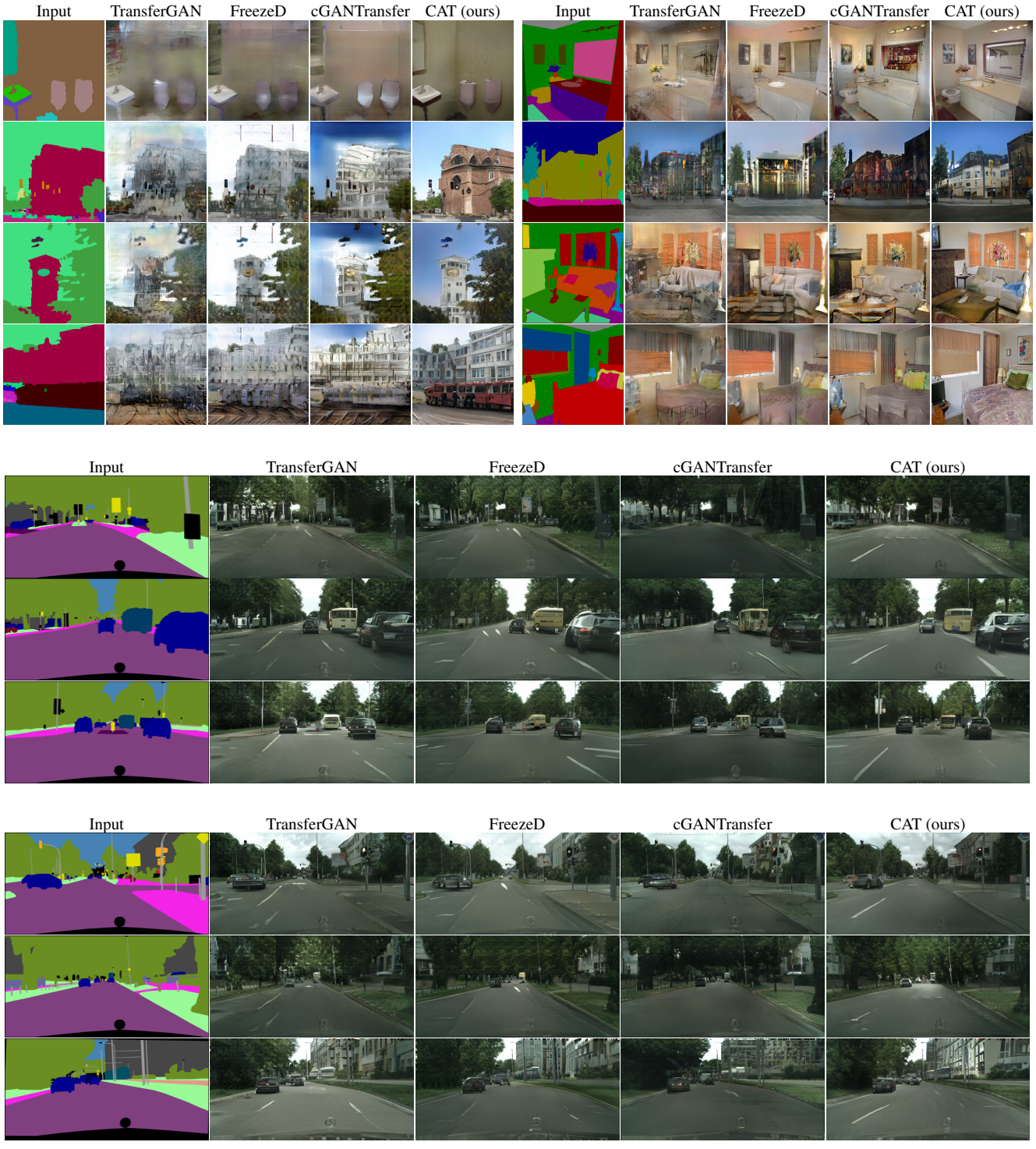

Figure 2. Samples from OASIS finetuned with 100 target images. Transfer from ADE to COCO (top left), COCO to ADE (top right), from COCO to Cityscapes (middle) and from ADE to Cityscapes (bottom).

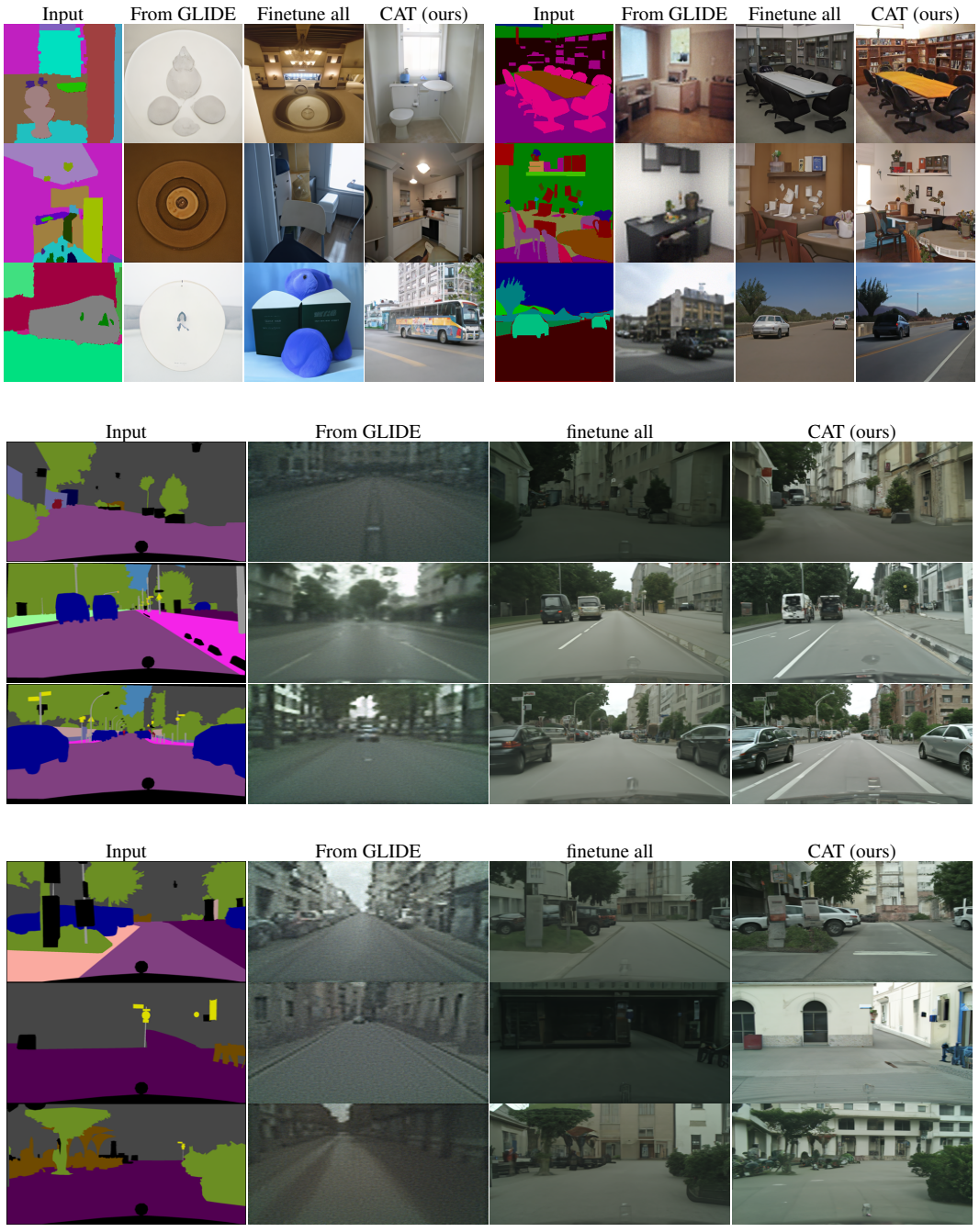

Figure 3. Samples from PITI finetuned with 100 target images. Transfer from ADE to COCO (top left), COCO to ADE (top right), from COCO to Cityscapes (middle) and from ADE to Cityscapes (bottom).

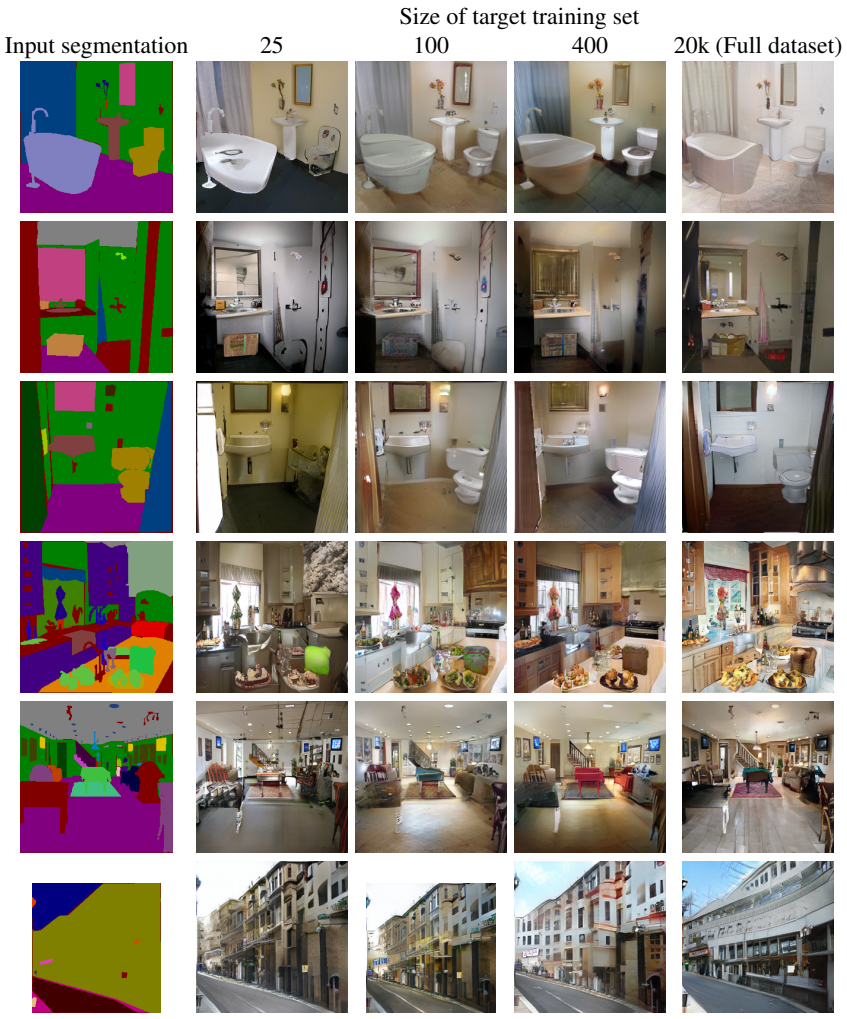

Figure 4. Samples from OASIS finetuned with target datasets from size in {25,100,400,20k}. Transfer from COCO to ADE.

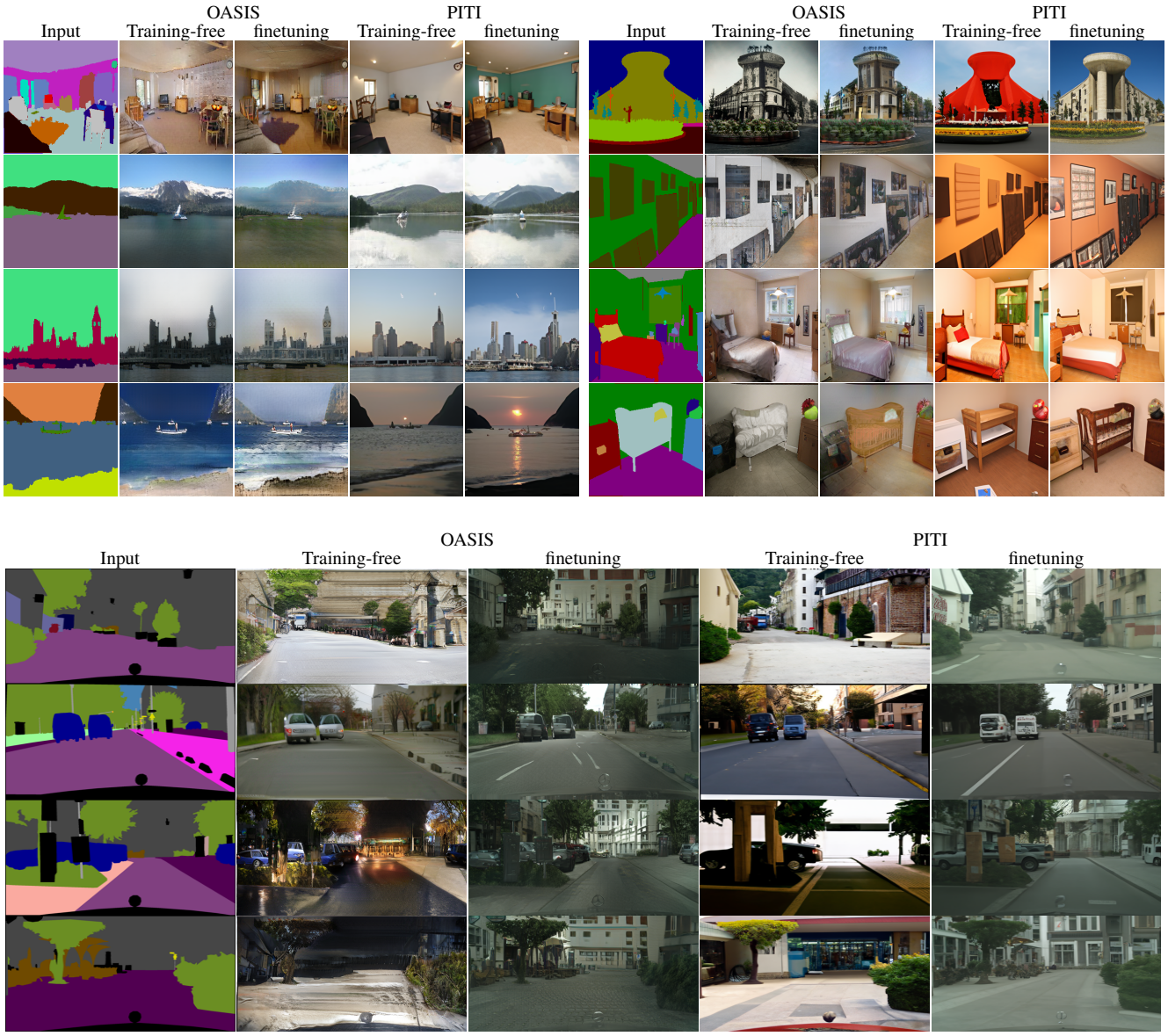

Figure 5. Samples from models trained with 100 target images. Transfer from ADE to COCO (top left), COCO to ADE (top right), and from COCO to Cityscapes (bottom). Samples obtained in training-free mode and after finetuning. Samples conditioned on the same noise.

		COCO→ADE	ADE→COCO
OASIS	Training-free	0.011	0.025
	After Finetuning	**0.008**	**0.024**
PITI	Training-free	0.015	0.023
	After Finetuning	**0.010**	**0.020**

Table 12. KID comparing training-free and finetuning mode

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. COCO-Stuff: Thing and stuff classes in context. In *CVPR*, 2018. 2

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[3] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning GANs. In *CVPR AI for Content Creation Workshop*, 2020. 4

[4] Edgar Schönfeld, Vadim Sushko, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You only need adversarial supervision for semantic image synthesis. In *ICLR*, 2021. 2

[5] Mohamad Shahbazi, Zhiwu Huang, Danda Pani Paudel, Ajad Chhatkuli, and Luc Van Gool. Efficient conditional GAN transfer with knowledge propagation across classes. In *CVPR*, 2021. 4

[6] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is all you need for image-to-image translation. *arXiv preprint*, arXiv:2205.12952, 2022. 2

[7] Yaxing Wang, Chenshen Wu, Luis Herranz, Joost van de Weijer, Abel Gonzalez-Garcia, and Bogdan Raducanu. Transferring GANs: generating images from limited data. In *ECCV*, 2018. 4

[8] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. 2