Supplementary Material
Diverse Semantic Image Synthesis via Probability Distribution Modeling

Zhentao Tan[1], Menglei Chai[2], Dongdong Chen[3], Jing Liao[4], Qi Chu[1],
Bin Liu[1], Gang Hua[5], Nenghai Yu[1]
[1]University of Science and Technology of China [2]Snap Inc. [3]Microsoft Cloud AI
[4]City University of Hong Kong [5]Wormpex AI Research LLC

1. Additional Implementation details

Network architectures. Here we give detailed network designs for each part. Figure 1 shows the architecture of our encoder. We use instance partial convolution and instance average pooling to get the parameters of each instance independently. The architecture of generator network is shown in Figure 2. The synthesis process starts with a random noise and goes through a series of the proposed INADE ResBIKs. Since the training is carried out on multiple GPUs, the batch normalization layer in INADE adopts the synchronous version. We use a multi-scale PathGAN [2] based discriminator whose architecture is shown in Figure 3.

Loss function. The loss function we adopted consists of four components:

Conditional adversarial loss. Let $\mathcal{E}$ be the prior noise remapping, G be the INADE generator, D be the discriminator, $\boldsymbol{m}$ be a given semantic mask, $\boldsymbol{o}$ and $\boldsymbol{p}$ be the corresponding image and instance map. The conditional adversarial loss built with hinge loss is formulated as:

$$\begin{aligned}\mathcal{L}_{GAN}(\mathcal{E},G,D)=\mathbb{E}[max(0,1-D(\boldsymbol{o},\boldsymbol{m},\boldsymbol{p}))]\\ +\mathbb{E}[max(0,1+D(G(\mathcal{E}(\boldsymbol{o},\boldsymbol{p}),\boldsymbol{m},\boldsymbol{p}),\boldsymbol{m},\boldsymbol{p}))].\end{aligned} \tag{1}$$

Feature matching loss. Let D_i and N_i be the output feature maps and the number of elements of the i-the layer of D respectively, S_D and E_D be the start number of layer for loss calculation and total number layers in D respectively. The feature matching loss is denoted as:

$$\begin{aligned}\mathcal{L}_F=\mathbb{E}\sum_{i=S_D}^{E_D}\frac{1}{N_i}[\|D_i(\boldsymbol{o},\boldsymbol{m},\boldsymbol{p})-\\ D_i(G(\mathcal{E}(\boldsymbol{o},\boldsymbol{p}),\boldsymbol{m},\boldsymbol{p}),\boldsymbol{m},\boldsymbol{p}))\|_1].\end{aligned} \tag{2}$$

To reduce the ambiguity, we only use high-level features and set S_D to 3.

Perceptual loss. Let V_i and M_i be the output feature maps and the number of elements of the i-the layer of VGG network respectively, S_V and E_V be the start number of layer for loss calculation and total number layers in VGG network respectively. The perceptual loss is denoted as:

$$\mathcal{L}_P=\mathbb{E}\sum_{i=S_V}^{E_V}\frac{1}{M_i}[\|V_i(\boldsymbol{o})-V_i(G(\mathcal{E}(\boldsymbol{o},\boldsymbol{p}),\boldsymbol{m},\boldsymbol{p}))\|_1]. \tag{3}$$

Similar to feature matching loss, we only use high-level features and set S_D to 3.

KL-Divergence loss. Let $q_\beta(z|\boldsymbol{o},\boldsymbol{p})$ and $q_\gamma(z|\boldsymbol{o},\boldsymbol{p})$ be the variational distribution of $\boldsymbol{N}_\gamma$ and $\boldsymbol{N}_\beta$ respectively. $p(z)$ be a standard Gaussian distribution. The KL-Divergence loss is denoted as:

$$\mathcal{L}_{KL}=0.5\times(\mathcal{D}(q_\beta(z|\boldsymbol{o},\boldsymbol{p})\|p(z))+\mathcal{D}(q_\gamma(z|\boldsymbol{o},\boldsymbol{p})\|p(z))). \tag{4}$$

The overall loss is made up of the above-mentioned loss terms as:

$$\min_{\mathcal{E},G}(\max_{D}(\mathcal{L}_{GAN})+\lambda_1\mathcal{L}_F+\lambda_2\mathcal{L}_P+\lambda_3\mathcal{L}_{KL}), \tag{5}$$

Following SPADE, We set $\lambda_1=10,\lambda_2=10,\lambda_3=0.05$.

2. Details of Datasets

The details about each dataset are described as follows:

- *Cityscapes* dataset [1] is a widely used dataset for semantic image synthesis [10, 7, 11]. The high-resolution images with fine semantic and instance annotations are taken from street scenes of German cities. There are 2,975 training images and 500 validation images. The number of annotated semantic classes is 35.
- *ADE20K* dataset [13] consists of 25,210 images (20,210 for training, 2,000 for validation and 3,000 for testing). The images in *ADE20K* dataset cover a wide range of scenes and object categories, including a total of 150 object and stuff classes.

- *CelebAMask-HQ* dataset [4, 3, 6] is based on CelebAHQ face imgae dataset. It contains of 28,000 training images and 2,000 validation images with 19 different semantic classes.

- *DeepFashion* dataset [5] contains 52,712 person images with fashion clothes. We use the processed dataset provided by GroupDNet [14] which consists of 30,000 training images and 2,247 validation images. There are 8 different semantic classes.

- *DeepFashion2* dataset is built from DeepFashion. We combine two adjacent images to generate the images containing two persons. The new semantic mask and the instance map are also derived from the corresponding two semantic masks. This dataset is only used to evaluate the performance of models trained on DeepFashion dataset in terms of instance level diversity.

In these datasets, *Cityscapes* and *DeepFashion2* have semantic and instance annotations, while the rest have only semantic annotations. In our experiment, the resolution of images is 256×256 except that Cityscapes dataset is 256×512.

3. Details of Diversity Metrics

We adopt the LPIPS [12, 8] to evaluate the overall diversity of the results. Specifically, we generate 10 groups of images or evaluation with randomly sampled noise, and calculate the diversity score between 2 random groups at a time. A total of 10 scores are calculated, and we measure the mean of these scores to reduce the potential fluctuation caused by random sampling.

To evaluate the instance-level diversity, we expand the metrics proposed by [14], called mean *Instance-Specific Diversity* (mISD) and mean *Other-Instances Diversity* (mOID), which represent the degree of change inside and outside the instance region when being manipulated. Specially, we generate several images by changing sampled noise for specified instance while keeping the noise for others unchanged. Then, the similarity inside and outside the instance region between these images makes up the mISD and mOID metrics. For datasets which have no instance annotations, these metrics degenerate to semantic level (mean *Class-Specific Diversity* (mCSD) and mean *Other-Classes Diversity* (mOCD)) which are the same with [14]. A high diversity inside the instance area (high mISD), as well as a low outside diversity (low mOID), are desired.

4. Additional ablation study

Here we give the additional ablation study for C^0 which represents the length of the initial sampling. Intuitively, the longer the sampling length is, the higher the diversity of the synthesized image will be. We conduct experiments on the *Cityscapes* and *CelebAMask-HQ* datasets, which include complex street scenes and delicate facial images. As summarized in Table 1, we compare the default setting ($C^o = 64$) with two variant settings: a shorter sampling length ($C^o = 8$, INADE-8) and a longer sampling length ($C^o = 128$, INADE-128). We find that INADE-8 shows the lower LPIPS score than INADE, while IANDE-128 correspondingly gets the highest score in this metric. And the model with the default setting (INADE) gets the best scores in terms of quality metrics. In our understanding, a short sampling length (e.g. 8) may limit the information capacity, thus reducing the generation quality (low scores of mIoU, acc and FID) and diversity (low score of LPIPS). In contrast, a longer sampling length (e.g. 128) can increase the diversity of the synthesized image (high score of LPIPS), but also increases the difficulty of high-quality image generation (low scores of mIoU, acc and FID).

In terms of model parameters, FLOPs and run time, INADE-8 is best, but the advantage is not obvious compared with INADE and INADE-128. Based on the above results, we set $C^o = 64$ on different datasets by default.

5. Additional results

In Figure 4, we show more multi-modal qualitative results on different datasets that only change one specified class or instance. The conclusions are basically the same as we mention in the main submission. BicycleGAN, DSC-GAN and VSPADE show the global style controllabitity, GroupDNet expands it to semantic level, while the synthesis results of our method can be controlled at both the semantic level and instance level. We notice that in some results, when we change one part, other parts slightly change as well, which is also mentioned in GroupDNet [14]. In fact, this is reasonable in some cases to increase the generation fidelity. For example, as shown in Figure 4 (h), the lighting often changes with the sky, if the appearance of the grass is totally unchanged, the final generated image will look unnatural to some extent. Therefore, though the metric mOCD (or mOID) may be a good indication of semantic/instance-level controllability, a slightly high mOCD or mOID do not represent worse quality. In other words, we do not expect them to be zero in real applications.

In Figure 5, Figure 6, Figure 7, Figure 8, we further show additional qualitative comparison results between the proposed INADE and other methods on the *DeepFashion*, *Cityscapes*, *ADE20K* and *CelebAMask-HQ* datasets. These results show that the images quality of INADE is better than or at least comparable to existing methods.

References

[1] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe

Table 1. Comparison of INADE with different C^o on the Cityscapes and CelebAMask-HQ daasets. $\mathbf{P}$, $\mathbf{F}$ and $\mathbf{T}$ represent the generator parameters, FLOPs and run time respectively.

Methods	Cityscapes							CelebAMask-HQ						
	mIoU	acc	FID	LPIPS	**P** (M)	**F** (G)	**T** (s)	mIoU	acc	FID	LPIPS	**P** (M)	**F** (G)	**T** (s)
INADE-64 (default)	**61.02**	**93.16**	**38.04**	0.248	77.39	75.26	0.0486	**74.08**	**94.31**	**22.55**	0.365	85.12	42.18	0.0298
INADE-8	60.25	93.07	38.68	0.220	**76.78**	**75.23**	**0.0482**	73.26	**94.31**	24.58	0.350	**84.50**	**42.16**	**0.0295**
INADE-128	59.57	92.68	39.30	**0.315**	78.10	75.28	0.0497	73.48	94.28	24.88	**0.366**	85.82	42.20	0.0306

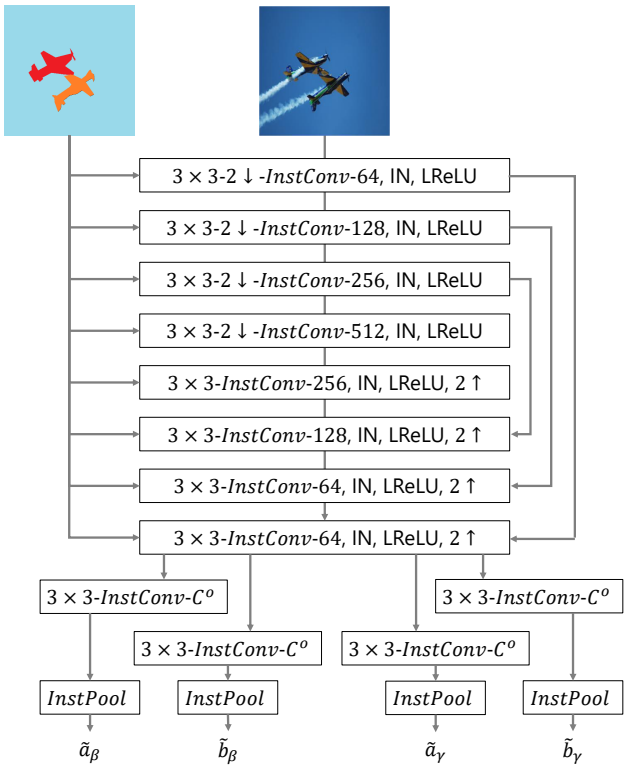

Figure 1. Architecture of our encoder network. We use UNet [9] based network to extract the features with the same resolution of input image, and then obtain the $\{\tilde{\boldsymbol{a}}_\gamma, \tilde{\boldsymbol{b}}_\gamma, \tilde{\boldsymbol{a}}_\beta, \tilde{\boldsymbol{b}}_\beta\}$ through independent instance partial convolution (InstConv) and instance average pooling (InstPool).

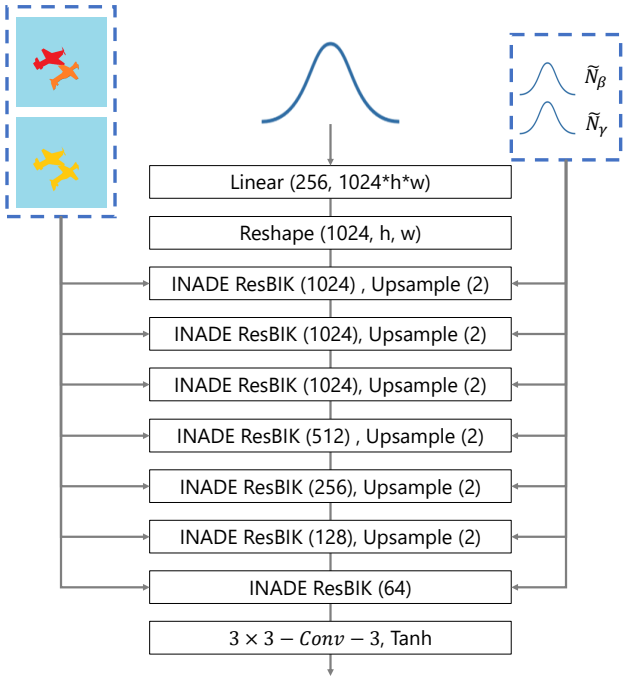

Figure 2. Architecture of our generator network. It consists of a linear transform layer, six INADE ResBIKs with upsampling and a final classification convolution layer. The upsampling operation on the second INADE ResBIK is removed if the resolution of generated images is 256×512. The initial noise $(\tilde{\boldsymbol{N}}_\gamma, \tilde{\boldsymbol{N}}_\beta)$ will be translated through a linear transformation mapping before fed to INADE ResBIKs.

Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 1

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 4

[3] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2

[4] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 2

[5] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2

[7] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8808–8816, 2018. 1

[8] richzhang. Perceptualsimilarity. https://github.com/richzhang/PerceptualSimilarity.git, 2020. 2

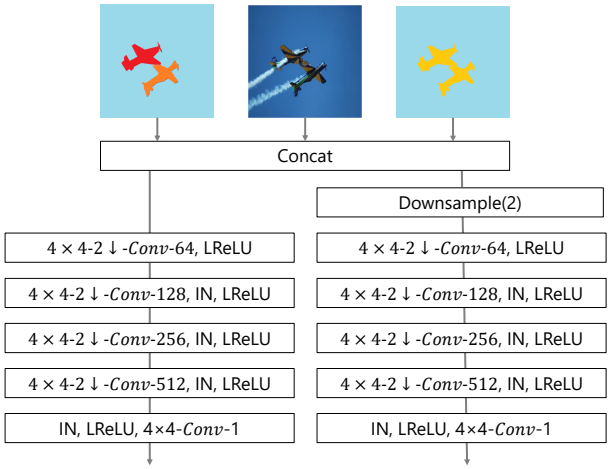

Figure 3. The discriminator of our method is based on the PatchGAN [2]. It takes the concatenation the segmentation map, instance map and the image as input.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[10] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018. 1

[11] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1

[12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 2

[13] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 1

[14] Zhen Zhu, Zhiliang Xu, Ansheng You, and Xiang Bai. Semantically multi-modal image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5467–5476, 2020. 2

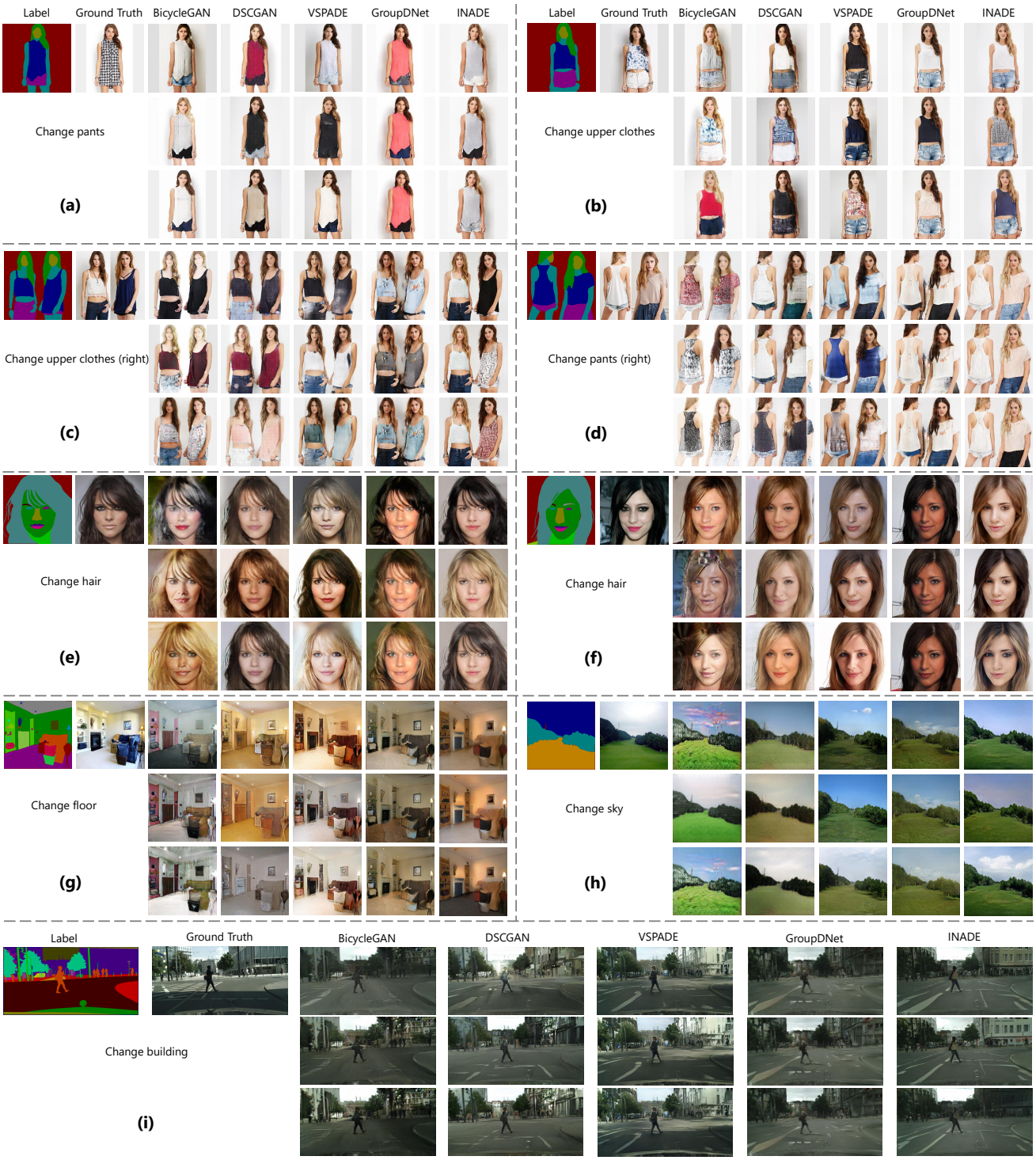

Figure 4. Multi-modal comparison of our INADE with previous state-of-the-art methods on DeepFashion (a-b), DeepFashion2 (c-d), CelebAMask-HQ (e-f), ADE20K (g-h) and Cityscapes (i) datasets.

Figure 5. Qualitative comparison of our INADE with previous state-of-the-art methods on the DeepFashion and DeepFashion2 datasets.

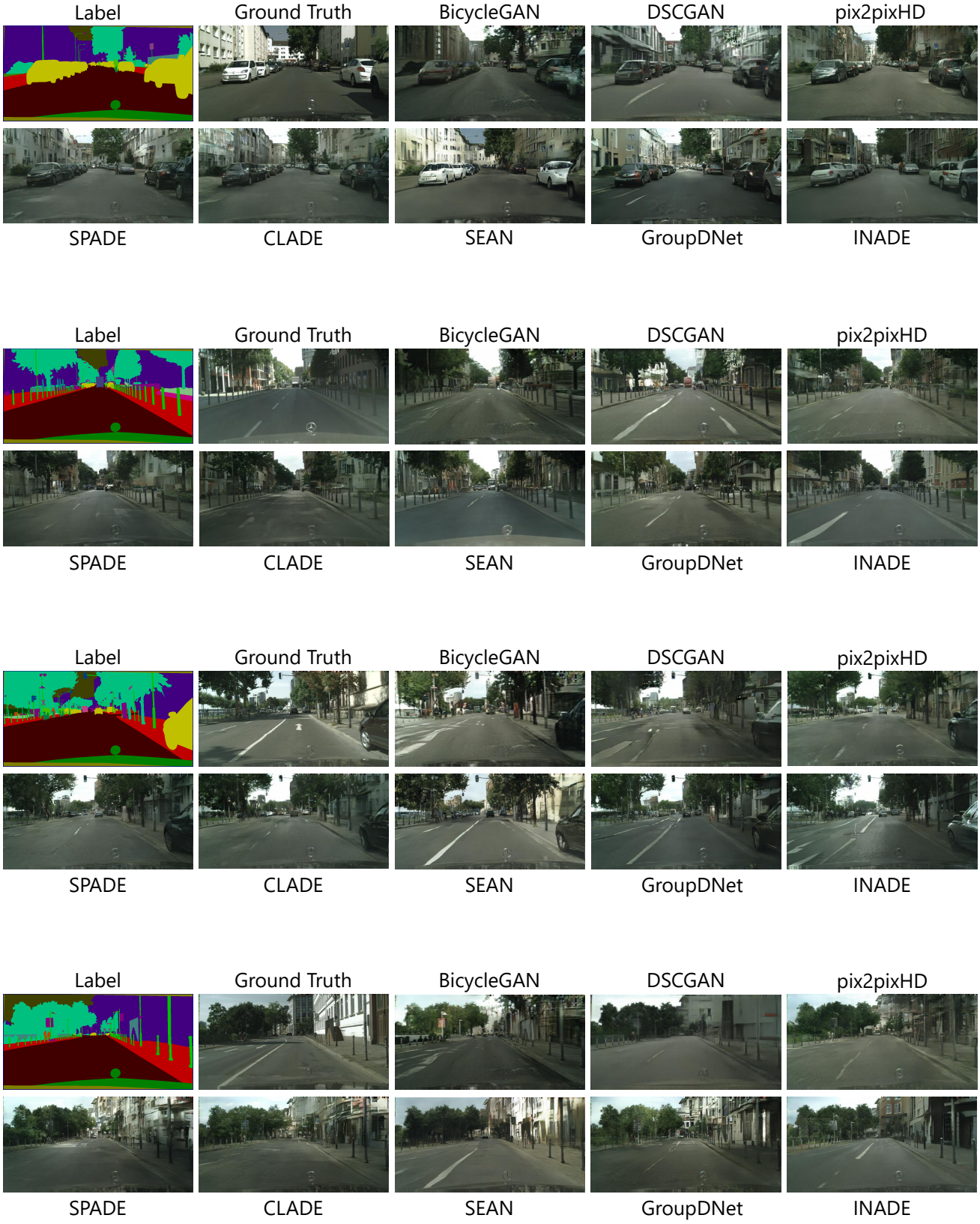

Figure 6. Qualitative comparison of our INADE with previous state-of-the-art methods on the Cityscapes dataset.

Figure 7. Qualitative comparison of our INADE with previous state-of-the-art methods on the ADE20K dataset.

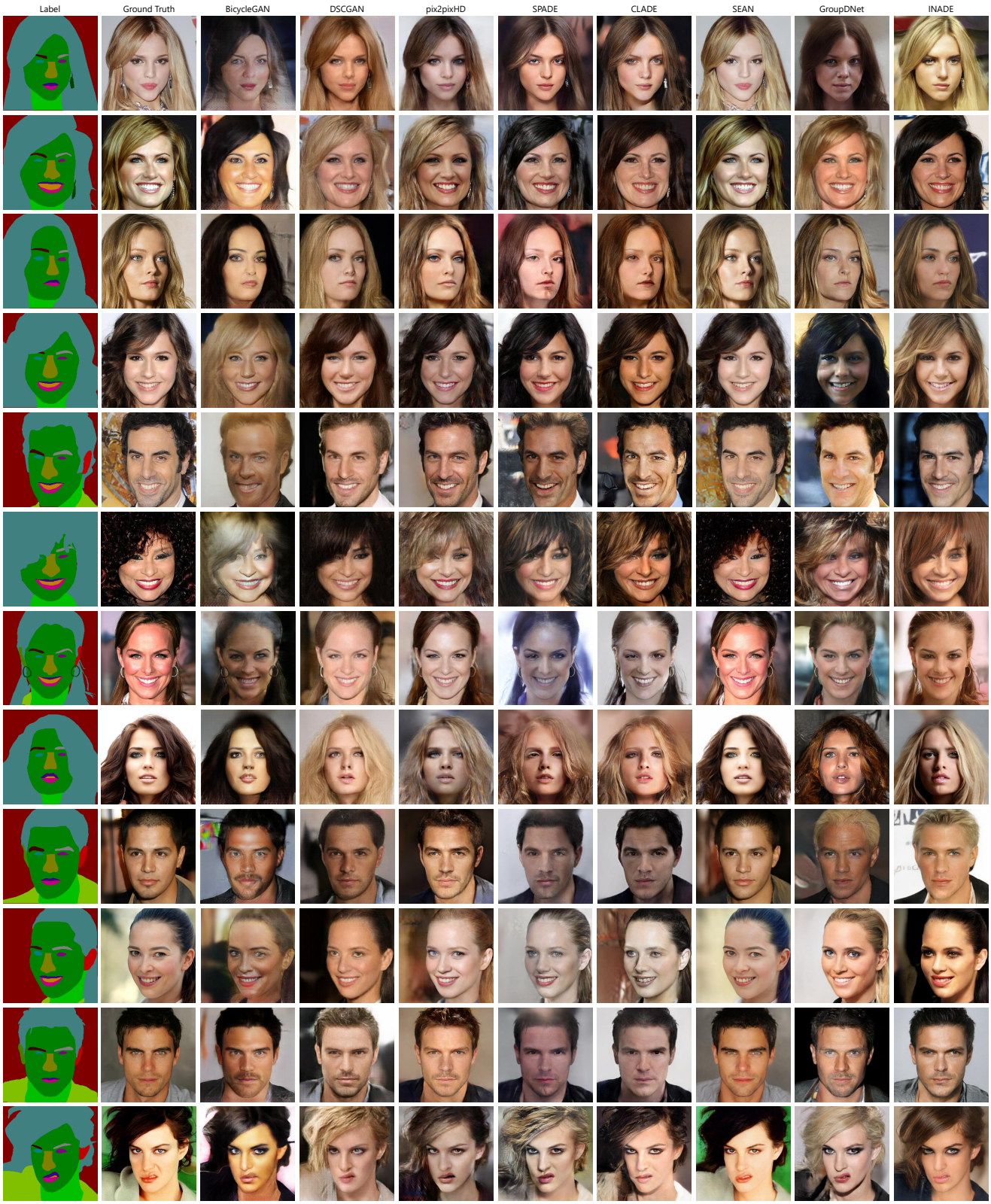

Figure 8. Qualitative comparison of our INADE with previous state-of-the-art methods on the CelebAMask-HQ dataset.