Supplementary Material for Neurips2022 Submission – Density Changing Regularized Image-to-Image Translation

Anonymous Author(s)
Affiliation
Address
email

1 More

Another perspective of density changing constraint. From the definition, we can see that we are enforcing constraint on the probability density ratios. If the density estimators are accurate, i.e., $f_{\mathcal{X}}(x) = f(x), f_{\mathcal{Y}}(G(x)) = f(G(x))$ and the mapping function G is invertible, then we are having $|det(J_{G(x)})| = \frac{f(x)}{f(G(x))}$ according to the change of variable formula, where $|det(J_{G(x)})|$ is the absolute value of determinant of the Jacobian matrix of the mapping function $G(x)$. This term can be viewed as a measure of volume changing caused by the function G in the space. For example, $X \in U[0,1]$, then we apply a transformation g such that $Y = 2X \in U[0,2]$, the volume changes from 1 to 2. As a consequence, the density decreases from 1 to $1/2$. In this example, $|det|$=2, which is coefficient of the linear transformation. Therefore, by minimizing the $\mathcal{L}_{density}$ when G is invertible, we are enforcing that mapping function G have similar volume changing for all samples.

Density Estimator We use the recommended hyper-parameters by BNAF: we set the hidden dim as 10, polyak averaging rate as 0.998. We use the AMSGrad optimizer with learning rate 0.001, β_1=0.9, β_2 =0.999. To reduce the computation cost, we set the number of blocks in BNAF to 1. Since there are 5 layers used to extract patch representations, we employ 5 BNAF for each domain in total. The number of parameters of the 5 BNAFs is 4.280 M. As a reference, the number of parameters of the generator is 11.378M.

2 PatchDist

We develop PatchDist as a strong baseline to demonstrate that the density is more useful than the pairwise distance quantity. DistanceGAN is trained to preserve the pairwise distance between images. Then we modify DistanceGAN such that it is trained to preserve the pairwise distance between all patch representations. PatchDist preserves the pairwise distance between patch representations while our method is trained to preserve the density information. The improvement of PatchDist over DistanceGAN suggests that using patch level representation can be more beneficial to unpaired image translation.

3 Ablation on patch size

We also perform ablation experiments on the patch size. During normal training, we are using 0,4,8,12,16 layer of the generator to extract patch representations and the representations correspond to patch size of 1,9,15,35 and 99. Now we run 5 experiments and we apply our density regularization in each experiment only one specific layer from 0,4,8,12,16. The results are shown in Table. 1. We can observe that our regularization works on different patch sizes and when the patch size is 9, it works pretry good. If we adopt all patch layers, we can obtain the best result.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Table 1: Abaltion results on the patch size

PatchSize	mAP	PixAcc	ClsAcc
Base	21.86	53.85	28.81
1	26.03	62.40	34.24
9	28.33	72.41	35.91
15	24.18	58.80	32.36
35	23.52	55.42	32.19
99	28.07	67.63	36.30
Full	30.97	72.93	39.33

4 Societal Impact

Image-to-image translation is a double-edged sword: On the one hand, it allows creative applications, such as the selfie→anime task and label→city. It also has great potential in related tasks, such as image super-resolution, medical image analysis and domain adaptation. On the other hand, it becomes easier to manipulate image data. In particular, DeepFakes have been used to create fake celebrity videos, fake news and malicious hoaxes. How to avoid such misuse remains an important research problem.

5 Resource Usage

We run our models on NVIDIA-A5000, A6000, V100 mostly. The entire project consumed approximately 1500 GPU hours.

6 Dataset License

The dataset used in our experiments are all existing datasets collected by different authors. All of them are freely available to academic and non-academic entities for non-commercial purposes such as academic research.

7 More Experiment Results

7.1 More Runs

We follow the protocols in CUT [2] and run each task once. Now we provide results of more runs to further justify the effectiveness of our method. If we disable our density regularization, i.e., set $\lambda_{density} = 0$, we denote our method as Base-GAN. As shown in Table. 2, our method performs consistently better than the Base-GAN model by a large margin. The stable improvements suggests that our proposed density regularization is effective in preserving semantic information in the input images.

Table 2: Repeated runs on label→city task.

Method	**mAP** ↑	**pAcc** ↑	**cAcc** ↑
CUT [2]	27.79	70.70	35.90
QS-Attn [1]	29.75	71.76	37.95
Base-GAN	21.86	53.85	28.81
Base-GAN	21.47	51.68	28.53
Base-GAN	25.68	59.45	34.14
DECENT	**30.97**	72.93	39.33
DECENT	30.48	**73.27**	39.28
DECENT	30.96	71.95	**39.81**

7.2 More Comparisons

Due to limited space, we only provide several samples in the main paper. Now we provide more samples for comparison.

label→city. We compare with the most recent methods in CVPR2022: QS-Attn [1] and MoNCE [4]. The results are present in Figure. 6 and 7. We can observe that the QS-Attn suffers the label flipping issues on the generated samples, i.e., the gray area is mapped to the trees sometimes (see, the first, second and last rows). Unlike QS-Attn, our method is able to find the correct mapping with our density changing regularization since it preserves the neighboring information. Although MoNCE [4] suffers less flipping issue in Figure. 7, we can observe that it generates unrealistic human photos. In contrast, our method generates more realistic human photos. We also present two samples generated by CycleGAN [5]. However, it suffers severe label flipping issue. It suggests that cycle consistency may not be effective when the city domain has more information than the label domain. In summary, our method achieved best performance with our proposed density changing regularizations.

cat→dog. We provide samples in Figure. 8. The Base-GAN model suffers the mode collapse issue. In contrast, our method generates realistic dog photos compared to the Base-GANN model. We can also observe that NEGCUT [3] and our method generates better samples than other baselines and dogs generated by our method are slightly better than NEGCUT, which is consistent with the quantivative results (NEGCUT obtained FID as 55.9 and our method is 55.2). It is also worth noting that our method only requires 56% training time of NEGCUT.

horse→zebra. We present more samples in Figure. 9. We can observe that the cycle consistency may be over restrictive. Our method achieved comparable results with existing SOTA methods.

selfie→anime. In addition to the three benchmark datasets used in CUT [2], we additionally conduct experiments on selfie→anime to verify the effectiveness of our method. As shown in Figure. 10, the Base-GAN model suffers the mode collapse issue. In contrast, our method is able to generate higher-quality anime faces. Our density regularization improves the BaseGAN model from 3.09 to 1.42 on KID metric.

8 Learned Densities

In this section, we present learned densities by our method in Figure. 12, 2, 3 and 4. Firstly, we can observe that our method can learn the densities from real images accurately on different tasks. For example, on the label→city task, our flow density estimators can tell that the purple segmentation (road) is of high density, which is consistent with human judgements. Secondly, we can observe that our method effectively avoids the label flipping issue in the label→city task, mode collapse in cat→dog and selfie→anime tasks. The encouraging results demonstrate the superiority of our method.

9 Limitation and Future Work

In the main paper, we have a brief discussion about the violation of assumption on horse→zebra task. We argue that it is because the unmatched dataset statistics since it is reported that horse takes 18% pixel of the image and zebra takes 37% pixels in the dataset [2]. Unlike other datasets, the objects we are interested (horse and zebra) are of low density over the domain. Therefore, our density changing regularization may not be so effective. We visualize two successful examples by our method in Figure.4. Our method discourages the mapping from the sky to the zebra because the sky is of high density and zebra is of low density. Therefore, it penalizes such background changes. However, we also provide failure cases in Figure. 5. On the left, the boy is mapped to a zebra. The reason is that the boy is uncommon in the horse dataset and it is of low density. However, the zebra is also of low density in the target domain. Our density changing regularization may not be able to correct such mapping because the patch of low density (boy) is mapped to the patch of low density (zebra). On the right, we can observe background changes. The reason is that beach is also of high density in the horse domain. However, it is not of high density in the zebra domain due to different habitats of horse and zebra. Therefore, our density changing regularization encourages the mapping from beach (high density in horse domain) to the grass (high density in zebra domain) This behaviour can be unwanted when we are interested in horse and zebra manipulation. As we proposed in the main

paper, we may consider to use attention module to address this issue. With attention module, we may focus on the horse and zebra objects directly. Then, we can apply our density regularization to further benefit the horse→zebra task.

References

[1] Xueqi Hu, Xinyue Zhou, Qiusheng Huang, Zhengyi Shi, Li Sun, and Qingli Li. Qs-attn: Query-selected attention for contrastive learning in i2i translation. *arXiv preprint arXiv:2203.08483*, 2022.

[2] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, pages 319–345. Springer, 2020.

[3] Weilun Wang, Wengang Zhou, Jianmin Bao, Dong Chen, and Houqiang Li. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14020–14029, 2021.

[4] Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Rongliang Wu, and Shijian Lu. Modulated contrast for versatile image synthesis. *arXiv preprint arXiv:2203.09333*, 2022.

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

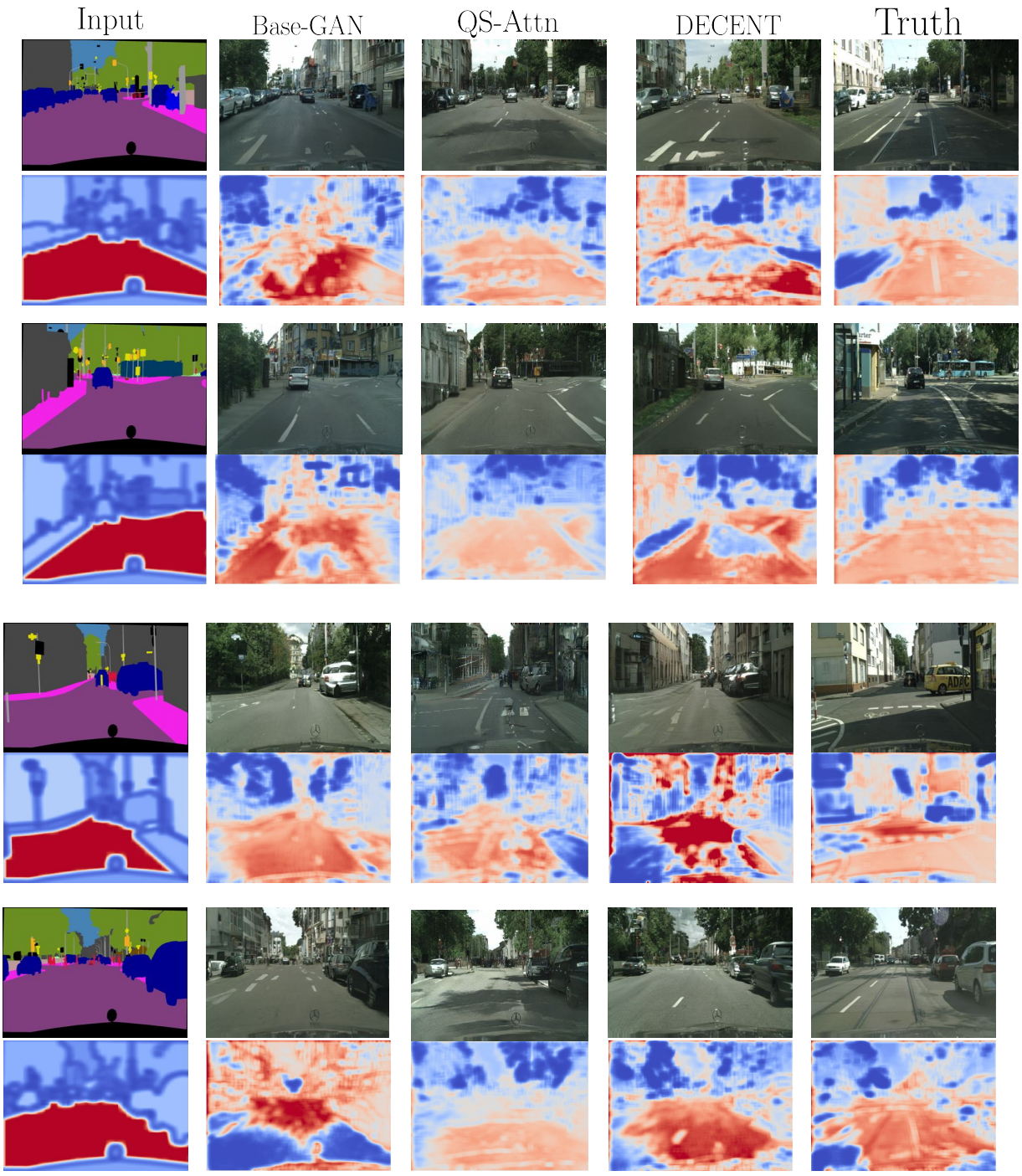

Figure 1: Learned densities on label→city task. Base-GAN suffers label flipping and always maps regions of high density to regions of low density. In contrast, our method effectively estimates the density as well as achieving our assumption.

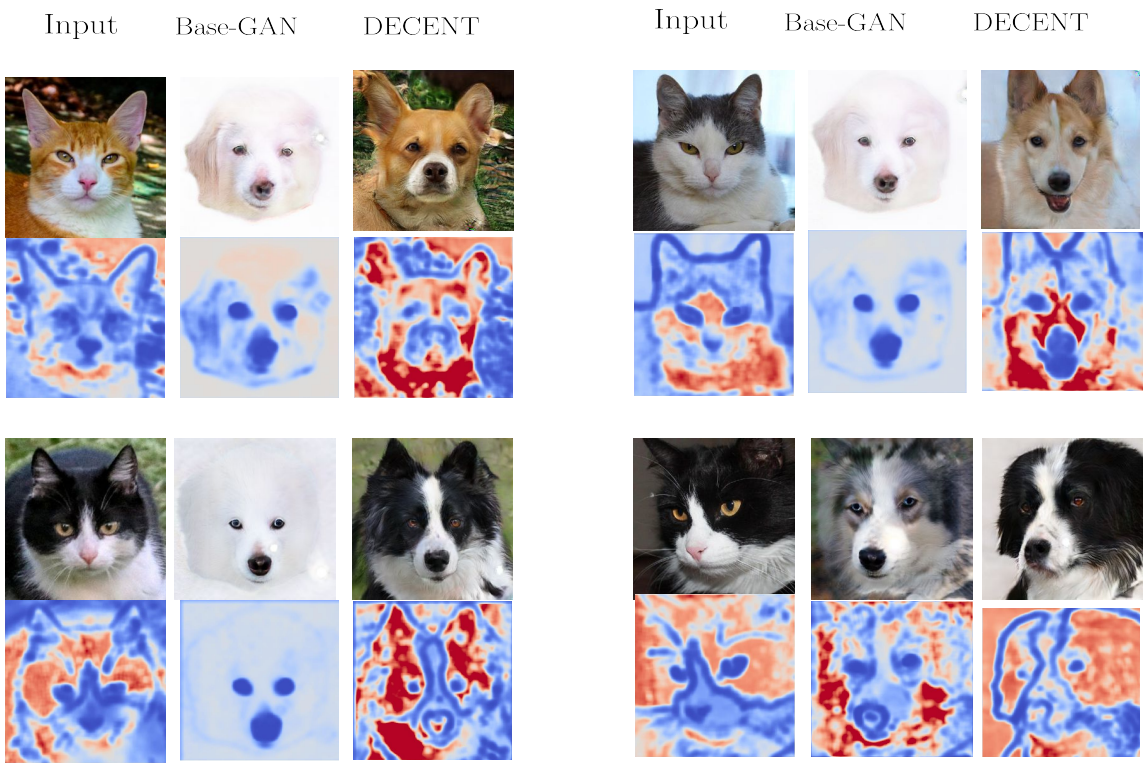

Figure 2: Learned densities on cat→dog task. With our density regularization, our method effectively avoids the mode collapse issue.

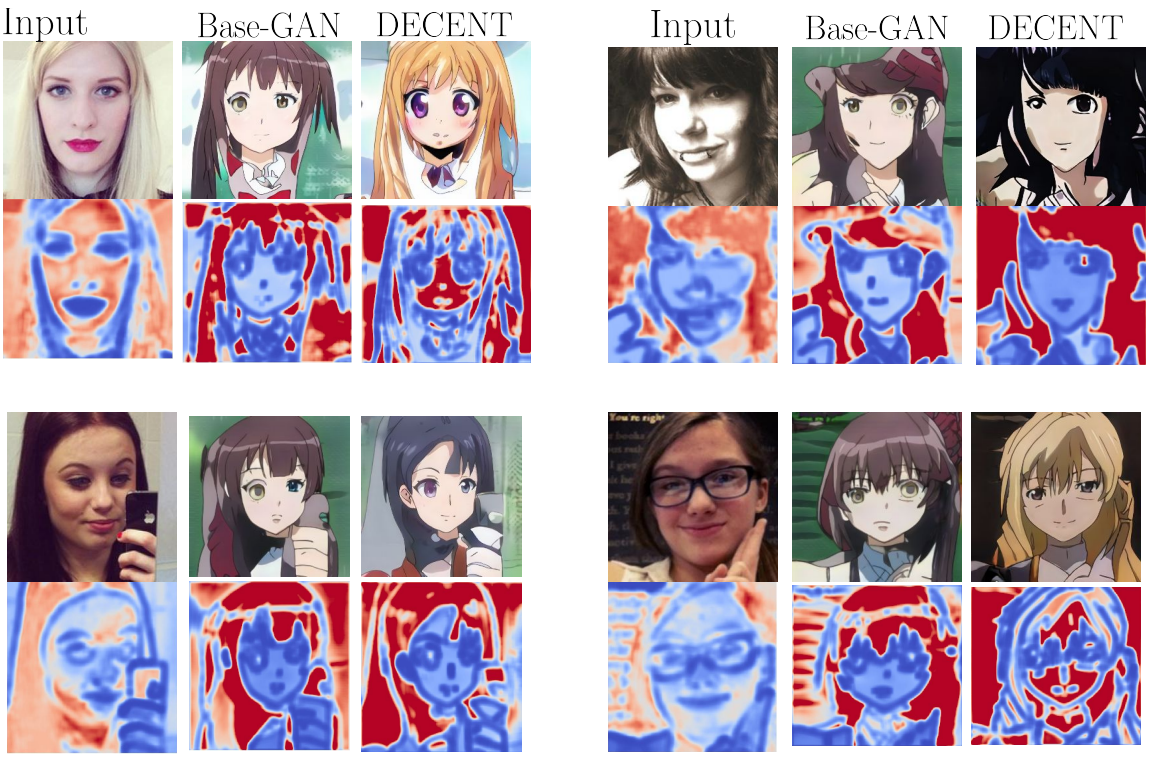

Figure 3: Learned densities on selfie→anime task. With our density regularization, our method can help preserve content and avoid mode collapse.

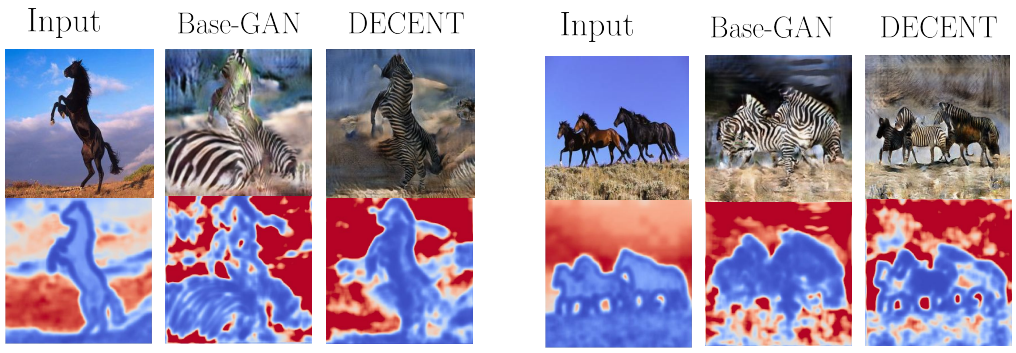

Figure 4: Learned densities on horse→zebra task. On some examples, our method can effectively regularize the output zebra images.

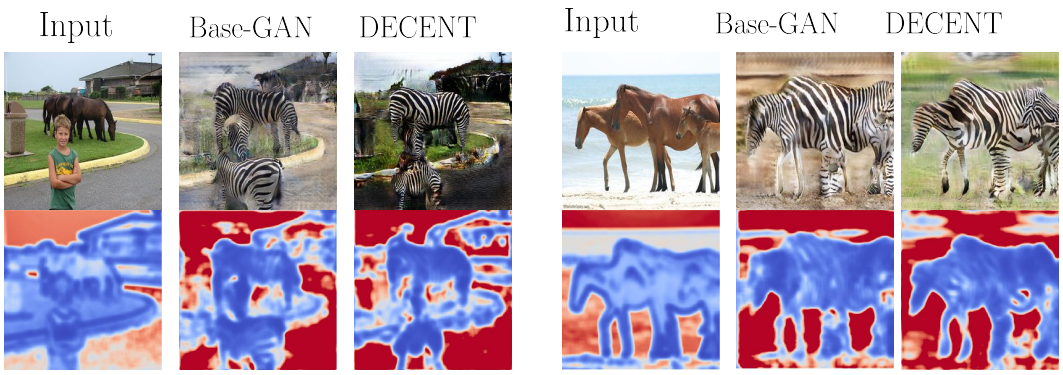

Figure 5: Some failure cases of our method on horse→zebra task. On the first example, the boy is mapped to a zebra. The reason is that the patch of the boy is of low density in the horse domain while zebra stripe is also of low density. Therefore, our density changing regularization cannot penalize such behaviour. On the right, we can see that the beach is mapped to the grass by our example. The reason is that patches of beach are common in the horse domain and are of high density. However, there is usually a few patches of beach in the zebra domain. Therefore, our method encourages patches of high density (beach) mapped to patches of high density(grass).

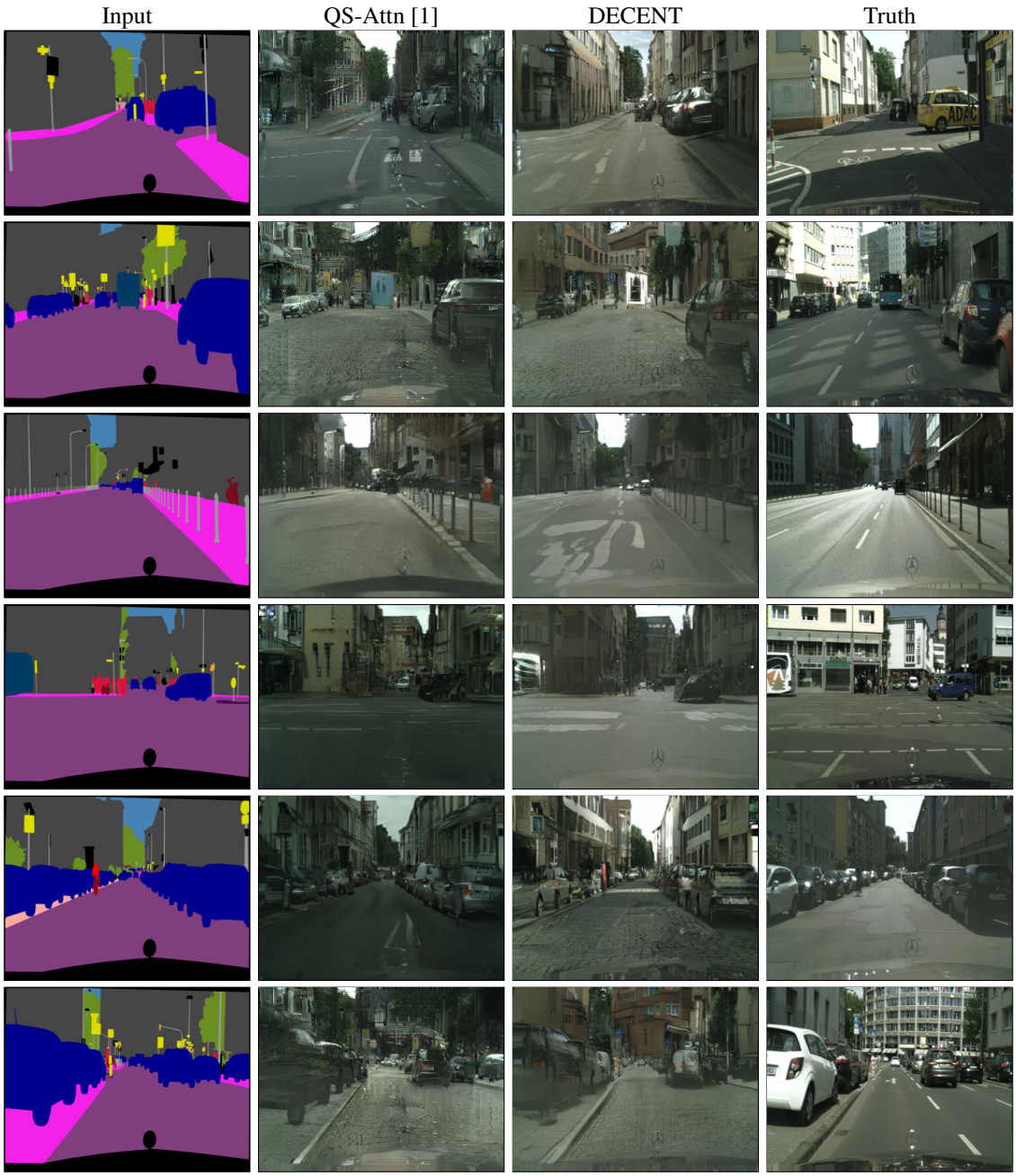

Figure 6: QS-Attn [1] suffers more label flipping issue while our method avoids it effectively.

Figure 7: MoNCE generates unrealistic human photos and CycleGAN suffers the label flipping issue.

Figure 8: Comparison with baseline methods. Without our density regularization, Base-GAN suffers the mode collapse issue. Our method generates higher-quality images when compared to other baselines.

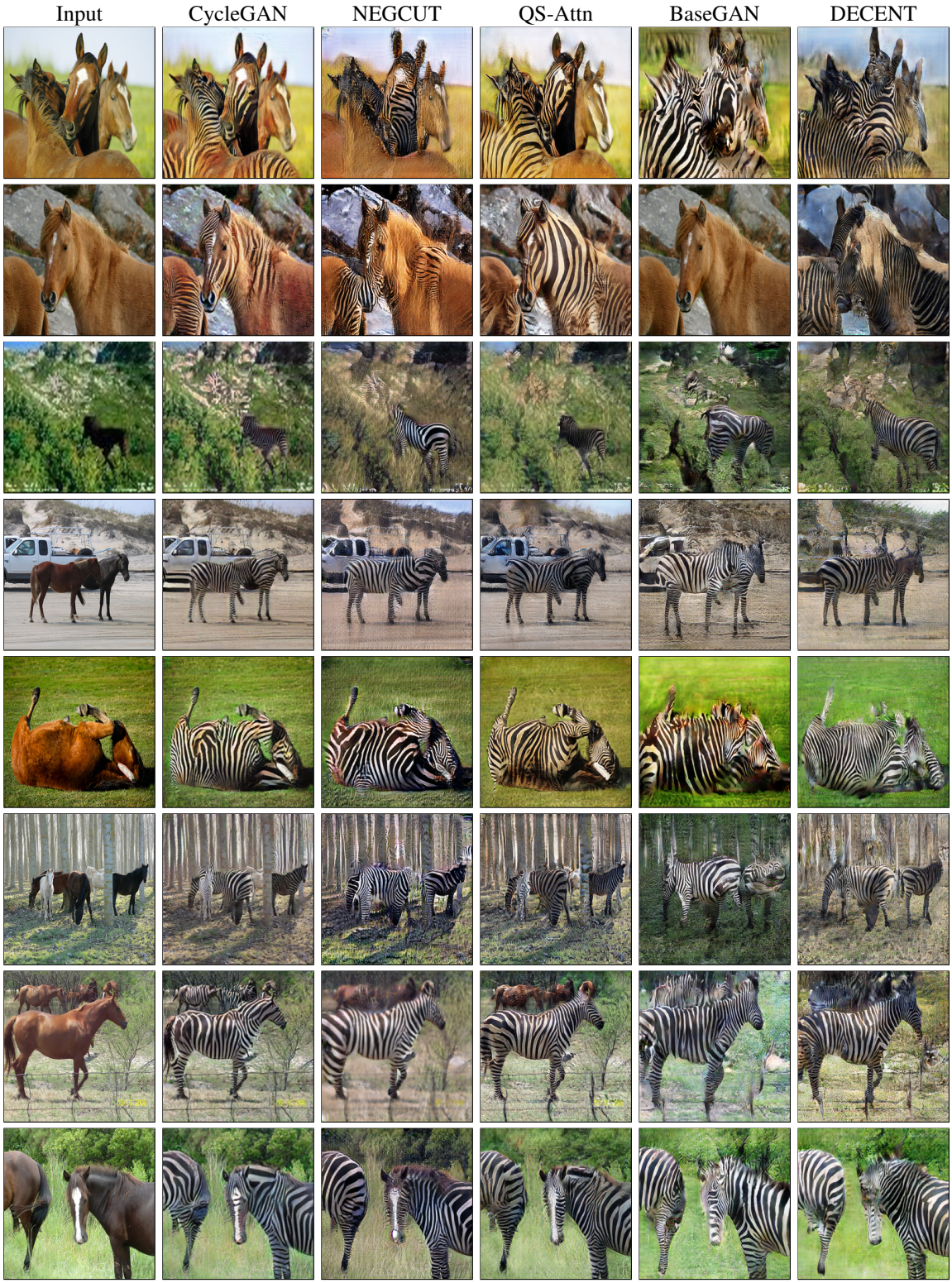

Figure 9: Our method achieved comparable results with existing state-of-the-art methods on horse→zebra.

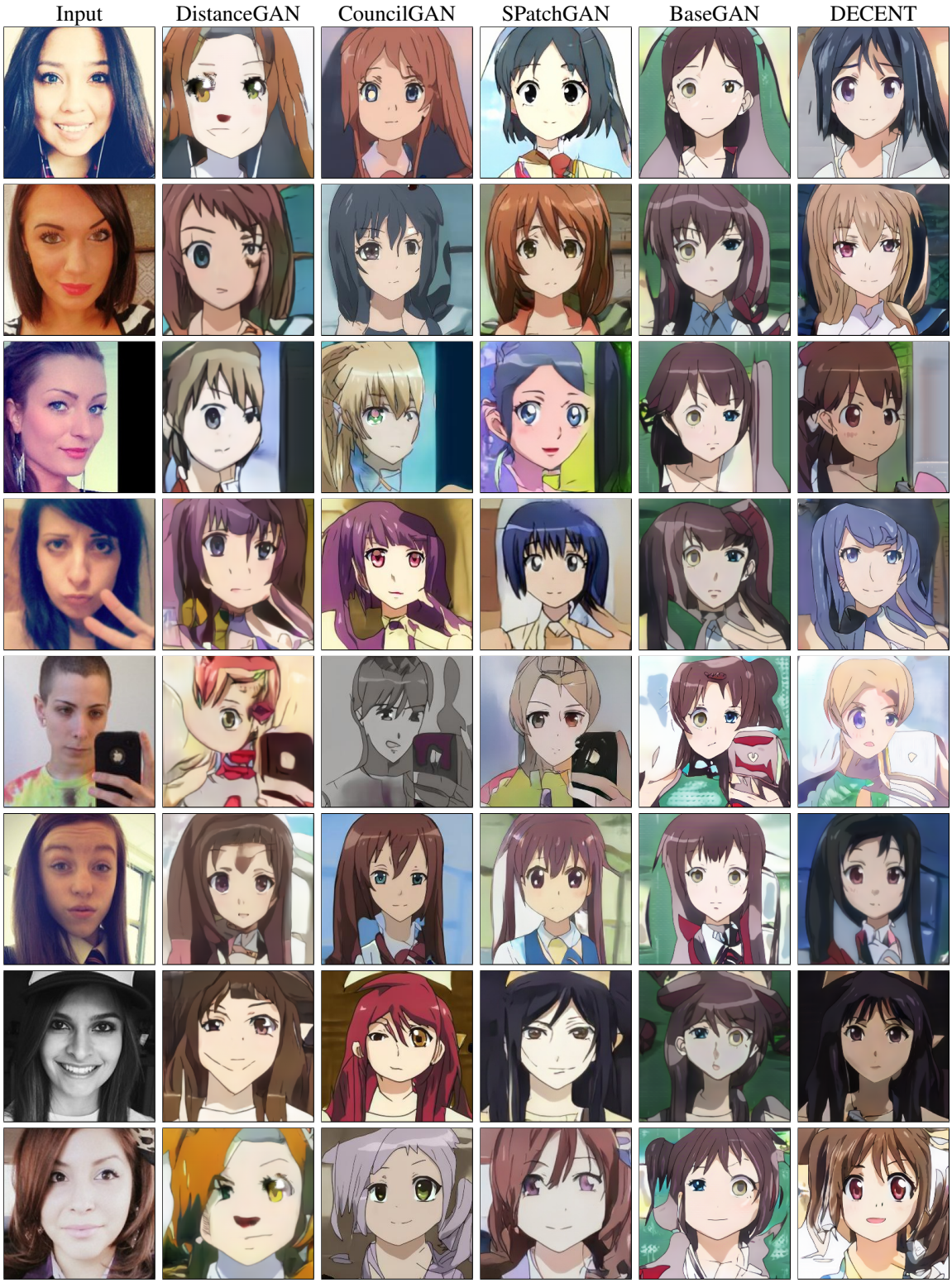

Figure 10: Comparison with baselines on selfie→anime task. Without our density regularization, BaseGAN model suffers the mode collapse issue. In contrast, our method generates realistic anime faces while important human identity information are preserved.

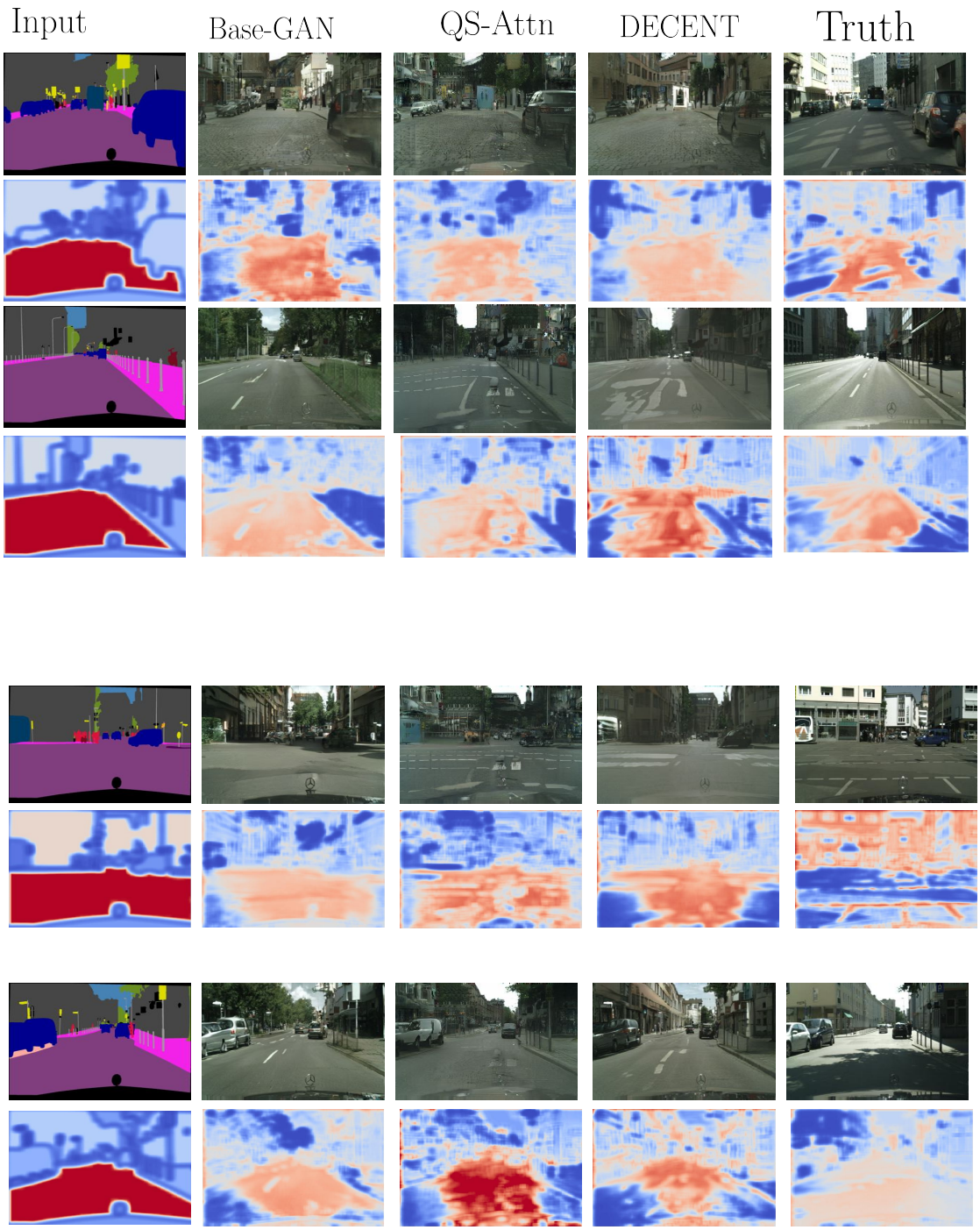

Figure 11: Learned densities on label→city task. Without our density changing regularization, the SOTA method QS-Attn still flips the building and trees. There is no explicit density changing regularization in QS-Attn method. As a consequence, QS-Attn generates low-density objects (tree) in the high density region (building).

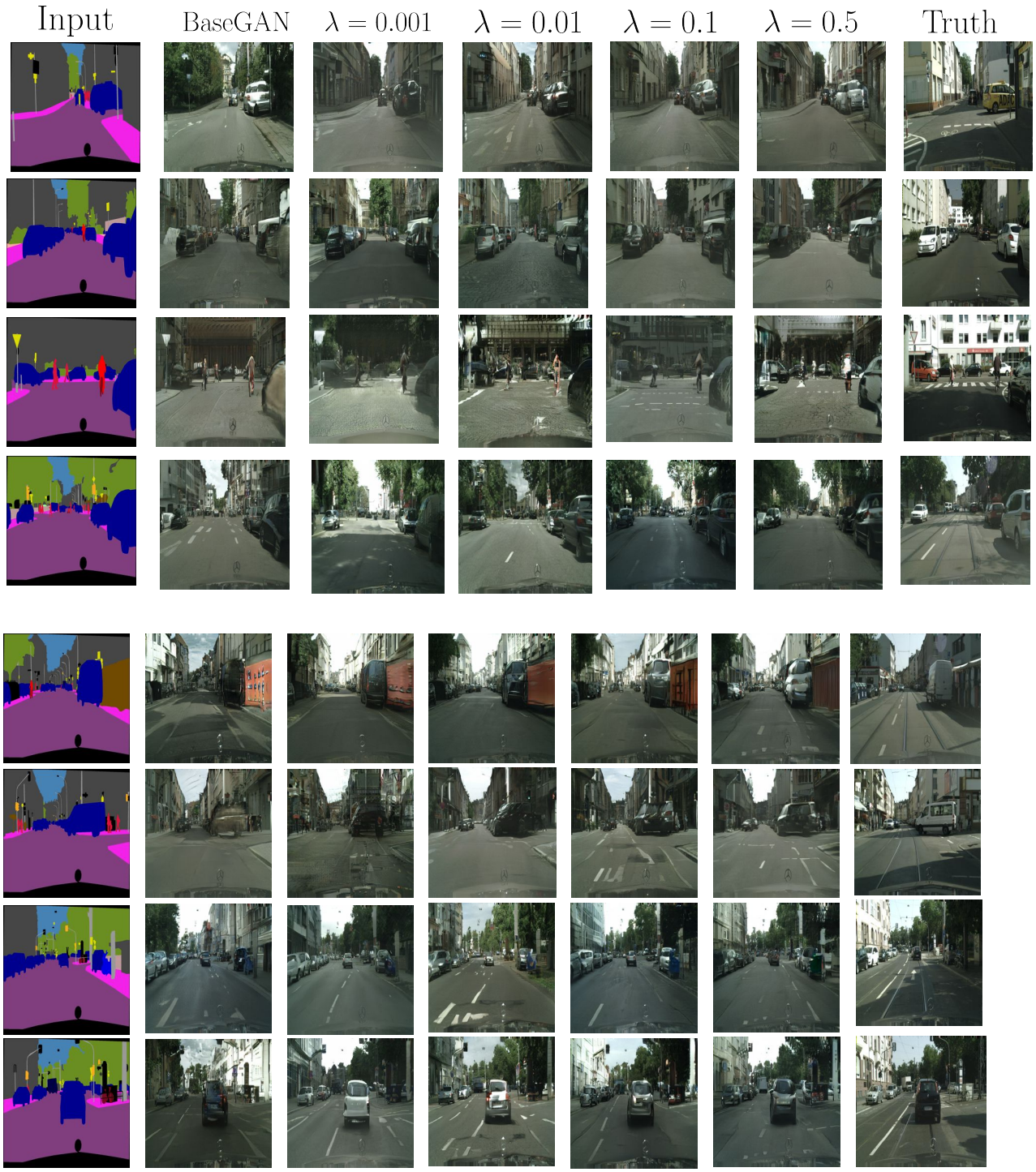

Figure 12: Qualitative ablation results on label→city task. We can clearly observe that the generations by Base-GAN often flips the building and tree. By contrast, our method is robust under different values of the hyper-parameter λ.