Panoptic-aware Image-to-Image Translation Supplementary Material

Liyun Zhang[1], Photchara Ratsamee[1,2], Bowen Wang[1], Zhaojie Luo[1], Yuki Uranishi[1], Manabu Higashida[1] and Haruo Takemura[1]
[1]Osaka University, Japan liyun.zhang@lab.ime.cmc.osaka-u.ac.jp
[2]Osaka Institute of Technology, Japan photchara@ime.cmc.osaka-u.ac.jp

1. Overview

In this supplementary, we first describe the cLSTM [12] component of our model in Section 2. The detailed setting of evaluation metrics is described in Section 3. We also show the description of our contributed dataset in Section 4. In Section 5, we provide the discussion of model efficiency. We show the additional experimental results in Section 6. Also, the limitation discussion is provided in Section 7.

2. cLSTM [12] component

After feature masking, the masked object feature maps need to be fused into a well-hidden representation for generating a realistic target image. Therefore, we need to integrate all objects in the desired locations and coordinate object feature maps based on other objects in the image. As shown in Fig. 1, the convolutional Long-Short-Term Memory (cLSTM) [12] is a multi-layer convolutional LSTM network, where the hidden states and cell states are both feature maps rather than vectors different from the traditional LSTM [2]. The computation of different gates is also done by convolutional layers. Therefore, cLSTM can better preserve spatial information compared with the traditional vector-based LSTM. It can integrate each object feature maps $\{F_{obj_i}\}_{i=1}^{m}$ one-by-one along the object sequence of $1 \sim m$ obtained by panoptic perception. The last output of cLSTM is used as the fused hidden representation H_{obj}. Different objects are sequentially fused together while keeping their spatial locations in the image.

3. Evaluation metrics

We chose the Human Preference (HP), Inception Score (IS) [8], Fréchet Inception Distance (FID) [7] and Diversity Score (DS) metrics instead of Peak Signal-to-Noise Ratio (PSNR) and Structure Similarity Index (SSIM) [11] metrics to evaluate the image quality. Because for images generated by GANs learning models, traditional PSNR and SSIM metrics deviate from human visual perception [13]. Also,

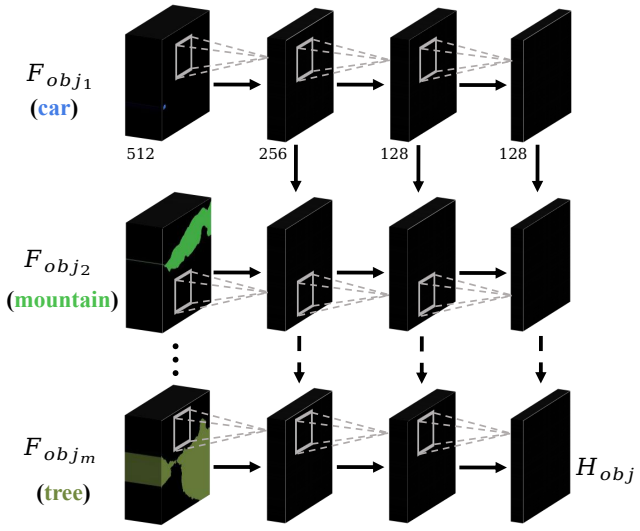

Figure 1. Illustration of convolutional Long-Short-Term Memory (cLSTM) component. We use three layers cLSTM for fusing all object feature maps together into the hidden feature maps H_{obj}. The number of channels in each layer of cLSTM is 256, 128, 128, respectively. The residual blocks are omitted for clarity. The first of each row is object feature maps of $F_{obj} = \{F_{obj_i}\}_{i=1}^{m}$.

we chose Panoptic Quality (PQ) [5] series metrics instead of instance segmentation and object detection to evaluate object recognition performance. Because PQ series metrics combine mean Intersection over Union (mIoU) in segmentation quality (SQ) and average precision (AP) in recognition quality (RQ) for more comprehensive scores. Also, since our framework is based on panoptic perception, using PQ series metrics will be more appropriate than the traditional object recognition metrics.

Human Preference (HP) compares the image quality through human cognition. We divided each evaluation set results into 5 groups to show to 5 persons in 20 participants (average age: 29.90, std: 23.49) in turn. Each group's images are selected with the best three results corresponding to

realism, object sharpness and scene similarity respectively with unbiased weights (1:1:1) to ensure adequate fairness of evaluation. The total number of selections is calculated as a final comprehensive percentage score.

Inception Score (IS) [8] uses an Inception V3 network pre-trained on the ImageNet-1000 classification benchmark and computes a statistics score of the network's outputs [10]. The higher the IS is, the better a generator model is.

Fréchet Inception Distance (FID) [7] also uses an Inception V3 network pre-trained on the ImageNet like IS to compute the Frechet distance [1] between two Gaussian distributions fitted to synthesized images and real images respectively [10]. The lower the FID is, the better a generator model is.

Diversity Score (DS) measures the differences between paired images generated from the same input by computing the perceptual similarity in deep feature space [14]. We used the LPIPS metric [13] for diversity scoring and pre-trained AlexNet [6] for feature extraction.

Panoptic Quality (PQ) is adopted to evaluate object recognition performance, PQ combines segmentation quality (SQ) and recognition quality (RQ) [5],

$$\mathrm{PQ} = \underbrace{\frac{\sum_{(p,g)\in TP} \mathrm{IoU}(p,g)}{|TP|}}_{\text{segmentation quality (SQ)}} \times \underbrace{\frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}}_{\text{recognition quality (RQ)}} \quad (1)$$

where SQ sums up all of the Intersection over Union (IoU) ratios for True Positives (TP) and evaluates how closely matched predicted segments are with their ground truths. RQ is a blend of precision and recall, where all True Positives, half False Positives (FP), and False Negatives (FN) are divided. It combines precision and recalls to identify how effective a trained model is at getting a prediction right.

4. Dataset contribution

The unaugmented source data from our contributed dataset contains 2,026 pairs of thermal and color images based on the partial KAIST-MS [3] dataset, it was annotated via the Segments.ai platform for the panoptic segmentation annotation by three professionally trained annotators. The annotated datasets can be augmented by various image manipulations for a variety of different tasks. We show the overview of annotated dataset via this link[1], please refer to the insights section on the overview tab to check the distribution of the categories and number of annotated objects ('thing' and 'stuff'). Also, for the annotation quality and detail of images, we show the samples of the dataset on paired color images via this link[2].

[1]**Overview:** https://segments.ai/panoptic/visible/

[2]**Samples:** https://segments.ai/panoptic/visible/samples

Model	Params (M)	FLOPs (G)	Avg PT (ms)		
			t2c	d2n	s2w
TSIT	116.1	**50.8**	19.1	19.7	**18.4**
INIT	130.3	62.9	22.3	21.0	21.6
Ours	**113.6**	51.4	**17.6**	**19.0**	19.9

Table 1. The floating-point operations (FLOPs) and parameters (Params) evaluate the computational cost and model complexity; The average processing time (Avg PT) per image evaluates processing speed. The lower the better.

5. Model efficiency

The model efficiency will influence practical applications, it is mainly measured from computational cost, model complexity and processing time. For the computational cost and model complexity, we used floating-point operations (FLOPs) and parameters (Params) as the evaluation indicators respectively. For the processing time, we calculate the average processing times (Avg PT) per image for different networks of competing baselines. As shown in Table 1, we present the model efficiency comparisons between our scores and the best baselines' scores, here we list the competing TSIT [4] (with +Seg, just not shown) and INIT [9] baselines. Table 1 shows that the Params of our model are lower than other models, and FLOPs are only slightly higher than the TSIT model. On different I2I translation tasks (summer-to-winter, day-to-night and thermal-to-color), our model spends on average less processing time than other models. Our method overall outperforms baselines since we use panoptic-level perception to avoid losing too much information in the translation, meanwhile, our model does not incur substantial computational cost and model complexity, also average processing time keeps competitive.

6. Additional experiment results

The experimental setting is the same as the main paper. From the additional results provided in Fig. 2, we further verify that most of the translation results generated by our method are better than other methods in image quality. As illustrated in Fig. 3, we also provide the comparison for the details of translated objects from different methods, the results demonstrate that translated objects from our method have sharper boundaries, adequate coloring, and also maintain a certain diversity, *e.g.*, the style of cars. Compared with competing methods on object recognition performance in Fig. 4, our model can also obtain significant improvement.

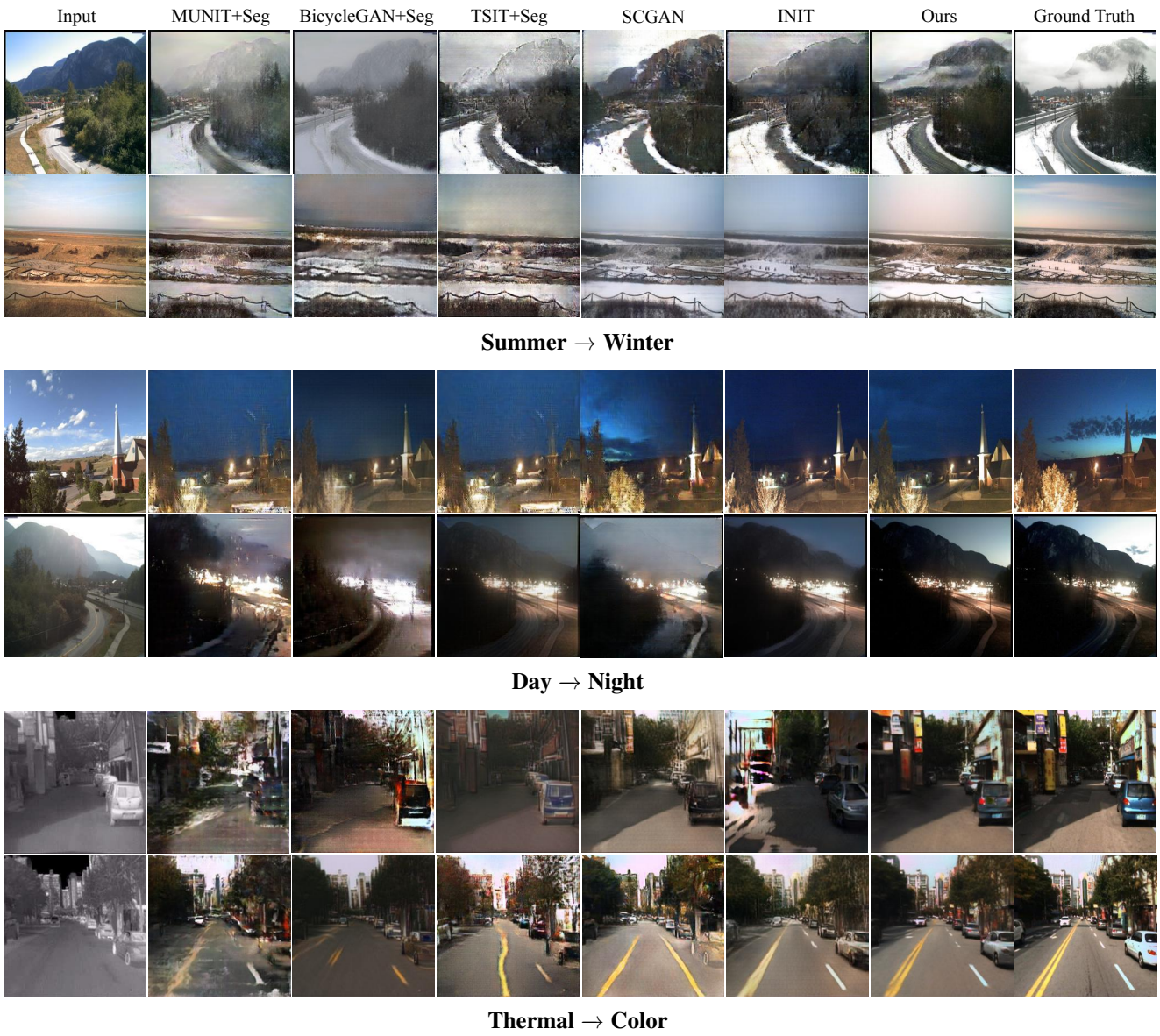

Figure 2. Comparison of the image quality of translated images. Top group are results of the summer-to-winter I2I translation task; Middle group are results of the day-to-night I2I translation task; Bottom group are results of the thermal-to-color I2I translation task.

7. Limitations

As we finely perceive the foreground object instances 'thing' and background semantic regions 'stuff' [5] to learn the translation model. For 'thing' (*e.g.*, car), it can generate different details for high diversity. However, for 'stuff' (*e.g.*, road), if generated 'stuff' texture is highly different from the ground truth (*e.g.*, the road has largely different lane markings and zebra crossings), it may decrease the whole image quality to some extent and thus affect the object recognition performance as well. This is because the 'stuff' normally has a larger region than 'thing', which should be considered with the majority of the whole image context.

References

[1] 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. *Advances in neural information processing systems*, 30, 2017.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

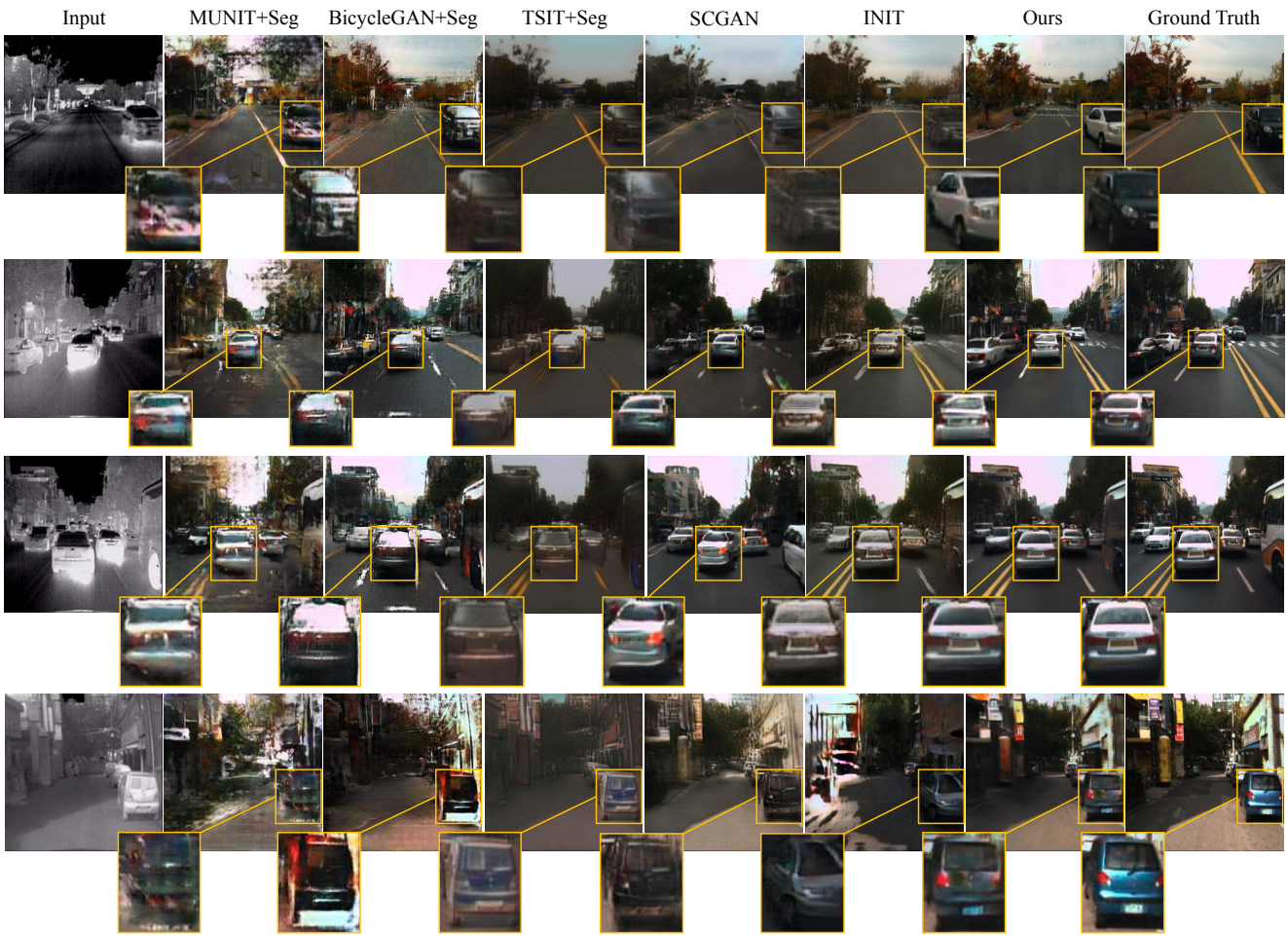

Figure 3. Comparison results on the details of translated objects from different approaches.

[3] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1037–1045, 2015.

[4] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. Tsit: A simple and versatile framework for image-to-image translation. In *European Conference on Computer Vision*, pages 206–222. Springer, 2020.

[5] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[7] Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. *Advances in neural information processing systems*, 32, 2019.

[8] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.

[9] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S Huang. Towards instance-level image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3683–3692, 2019.

[10] Wei Sun and Tianfu Wu. Learning layout and style reconfigurable gans for controllable image synthesis. *arXiv preprint arXiv:2003.11571*, 2020.

[11] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[12] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

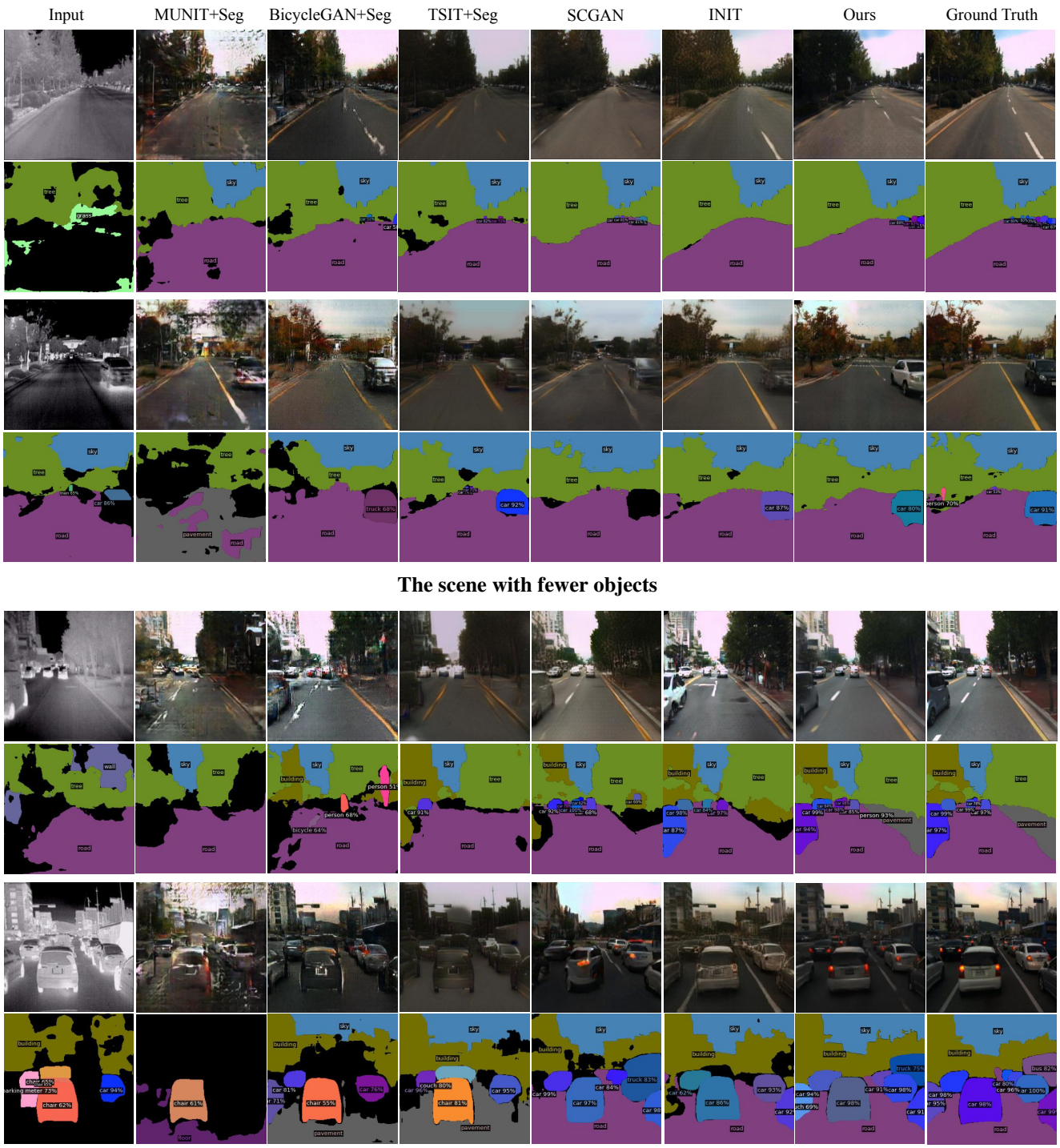

Figure 4. The object recognition performance of translated images for different approaches. Top group denotes the scene with fewer objects; Bottom group denotes the scene with multiple discrepant objects. In each group, upper row are translated images from different approaches, lower row are the results of the corresponding panoptic segmentation.

[13] 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 recogni-*

tion, pages 586–595, 2018.

[14] Bo Zhao, Lili Meng, Weidong Yin, and Leonid Sigal. Image generation from layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8584–8593, 2019.