Multi-Channel Attention Selection GAN with Cascaded Semantic Guidance for Cross-View Image Translation
- Supplementary Document -

Hao Tang[1,2*] Dan Xu[3*] Nicu Sebe[1,4] Yanzhi Wang[5] Jason J. Corso[6] Yan Yan[2]

[1]DISI, University of Trento, Trento, Italy [2]Texas State University, San Marcos, USA
[3]University of Oxford, Oxford, UK [4]Huawei Technologies Ireland, Dublin, Ireland
[5]Northeastern University, Boston, USA [6]University of Michigan, Ann Arbor, USA

This supplementary document provides additional results supporting the claims of the main paper. First, we provide detailed experimental results about the influence of the number of attention channels (Sec. 1). Additionally, we compare our two-stage model with one-stage model (Sec. 2). We also provide the visualization results of the generated uncertainty maps (Sec. 3) and the arbitrary cross-view image translation experiments on Ego2Top dataset [1] (Sec. 4). Finally, we compare our SelectionGAN with the state-of-the-arts methods, *i.e.* Pix2pix [2], X-Fork [3] and X-Seq [3]. Specifically, we compare the results of the generated segmentation maps (Sec. 5), and visualize the comparison results on Dayton [4], CVUSA [5] and Ego2Top [1] datasets (Sec. 6).

1. Influence of the Number of Attention Channels N

We investigate the influence of the number of attention channels N in Equation 3 in the main paper. Results are shown in Table 1. We observe that the performance tends to be stable after $N = 10$. Thus, taking both performance and training speed into consideration, we have set $N = 10$ in all our experiments.

Table 1: Influence of the number of attention channels N.

n	SSIM	PSNR	SD
0	0.5438	22.9773	19.4568
1	0.5522	23.0317	19.5127
5	0.5901	23.8068	**20.0033**
10	**0.5986**	23.7336	19.9993
32	0.5950	**23.8265**	19.9086

*Equal contribution.

Table 2: Results of coarse-to-fine generation. The best results are marked in blue color.

Baseline	Stage I	Stage II	SSIM	PSNR	SD
F	√		0.5551	23.1919	19.6311
F		√	**0.5989**	**23.7562**	**20.0000**
G	√		0.5680	23.2574	19.7371
G		√	**0.6047**	**23.7956**	**20.0830**
H	√		0.5567	23.1545	19.6034
H		√	**0.6167**	**23.9310**	**20.1214**

2. Coarse-to-Find Generation

We provide more comparison results of coarse-to-fine generation in Table 2 and Figures 1, 2 and 3. We observe that our two-stage method generate much visually better results than the one-stage model, which further confirms our motivations.

3. Visualization of Uncertainty Map

In Figures 1, 2, 3 and 4, we show some samples of the generated uncertainty maps. We can see that the generated uncertainty maps learn the layout and structure of the target images.

4. Arbitrary Cross-View Image Translation

We also conducted the arbitrary cross-view image translation experiments on Ego2Top dataset. As we can see from Figure 4, given an image and some novel semantic maps, SelectionGAN is able to generate the same scene but with different viewpoints in both outdoor and indoor environments.

5. Generated Segmentation Maps

Since the proposed SelectionGAN can generate segmentation maps, we also compare it with X-Fork [3] and X-Seq [3] on Dayton dataset. Following [3], we compute

Table 3: Per-class accuracy and mean IOU for the generated segmentation maps on Dayton dataset. For both metric, higher is better. (*) These results are reported in [3].

Method	a2g	
	Per-Class Acc.	mIOU
X-Fork [3]	0.6262*	0.4163*
X-Seq [3]	0.4783*	0.3187*
SelectionGAN (Ours)	**0.6415**	**0.5455**

per-class accuracies and mean IOU for the most common classes in this dataset: "vegetation", "road", "building" and "sky" in ground segmentation maps. Results are shown in Table 3. We can see that the proposed SelectionGAN achieves better results than X-Fork [3] and X-Seq [3] on both metrics.

6. State-of-the-art Comparisons

In Figures 5, 6, 7, 8 and 9, we show more image generation results on Dayton, CVUSA and Ego2Top datasets compared with the state-of-the-art methods *i.e.*, Pix2pix [2], X-Fork [3] and X-Seq [3]. For Figures 5, 6, 7, 8, we reproduced the results of Pix2pix [2], X-Fork [3] and X-Seq [3] using the pre-trained models provided by the authors[1]. As we can see from all these figures, the proposed SelectionGAN achieves significantly visually better results than the competing methods.

References

[1] Shervin Ardeshir and Ali Borji. Ego2top: Matching viewers in egocentric and top-view videos. In *ECCV*, 2016. 1

[2] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 2

[3] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *CVPR*, 2018. 1, 2

[4] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *ECCV*, 2016. 1

[5] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *ICCV*, 2015. 1

[1] https://github.com/kregmi/cross-view-image-synthesis

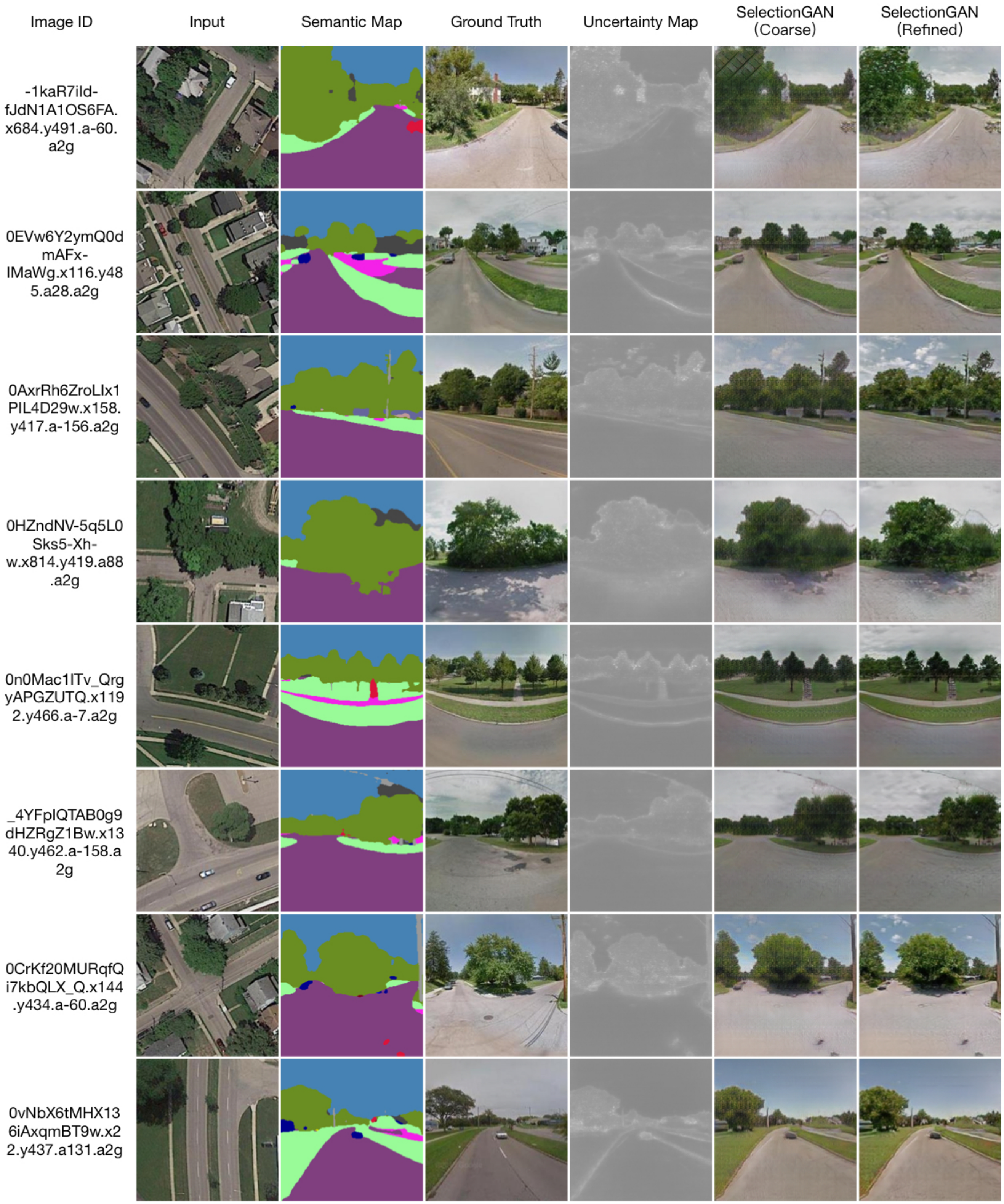

Figure 1: Results generated by our SelectionGAN in 256×256 resolution in a2g direction on Dayton dataset. These samples were randomly selected for visualization purposes.

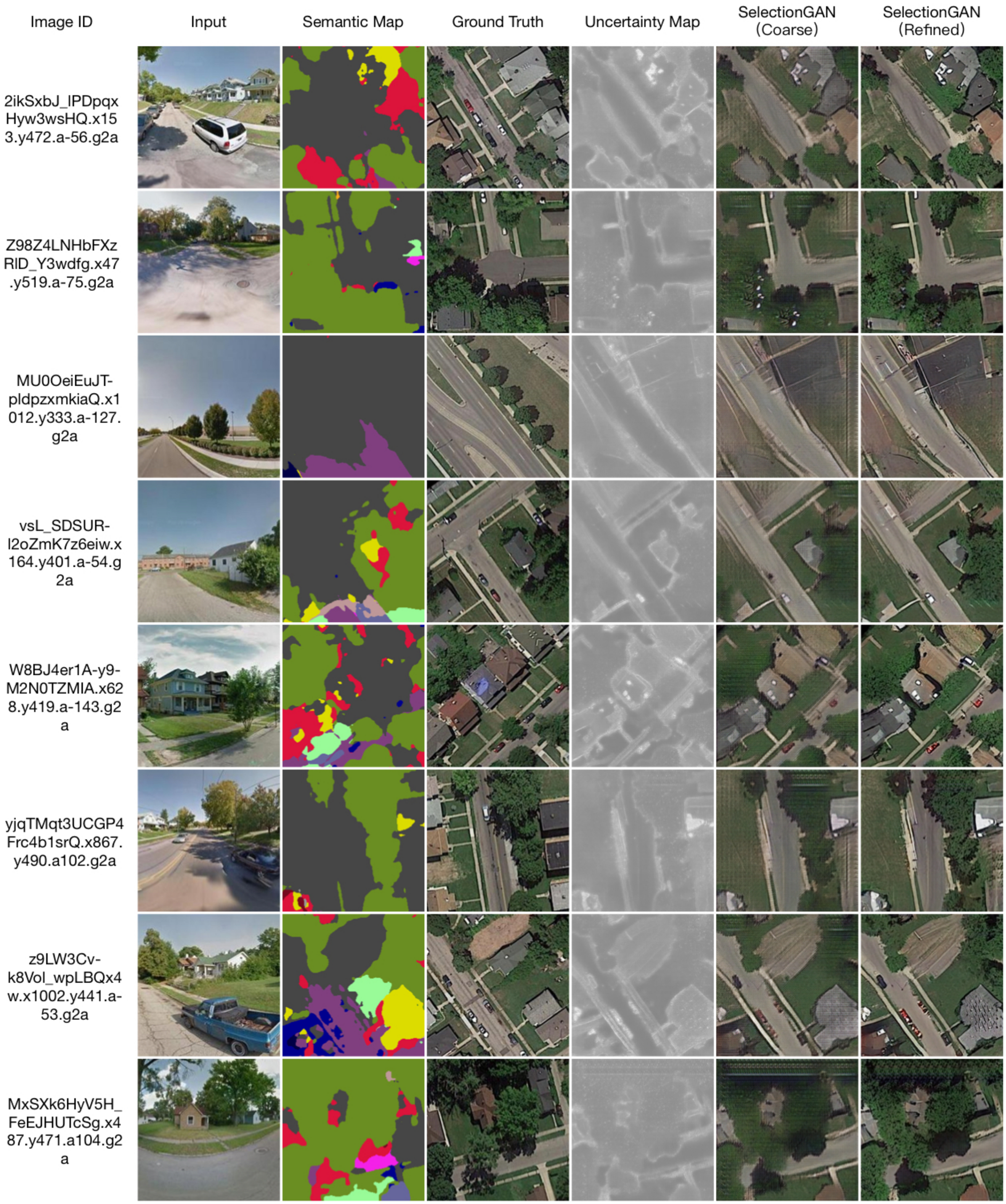

Figure 2: Results generated by our SelectionGAN in 256×256 resolution in g2a direction on Dayton dataset. These samples were randomly selected for visualization purposes.

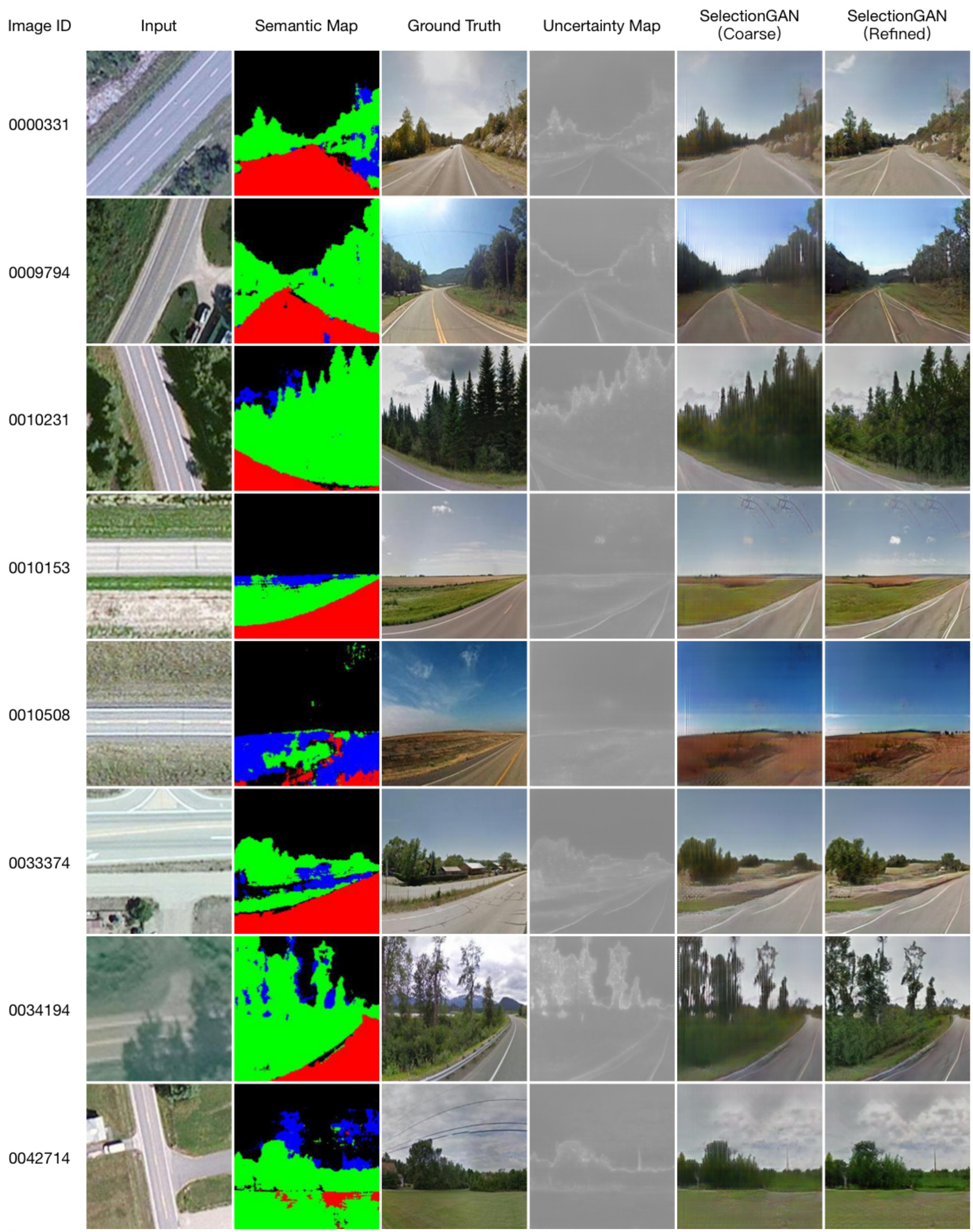

Figure 3: Results generated by our SelectionGAN in 256×256 resolution in a2g direction on CVUSA dataset. These samples were randomly selected for visualization purposes.

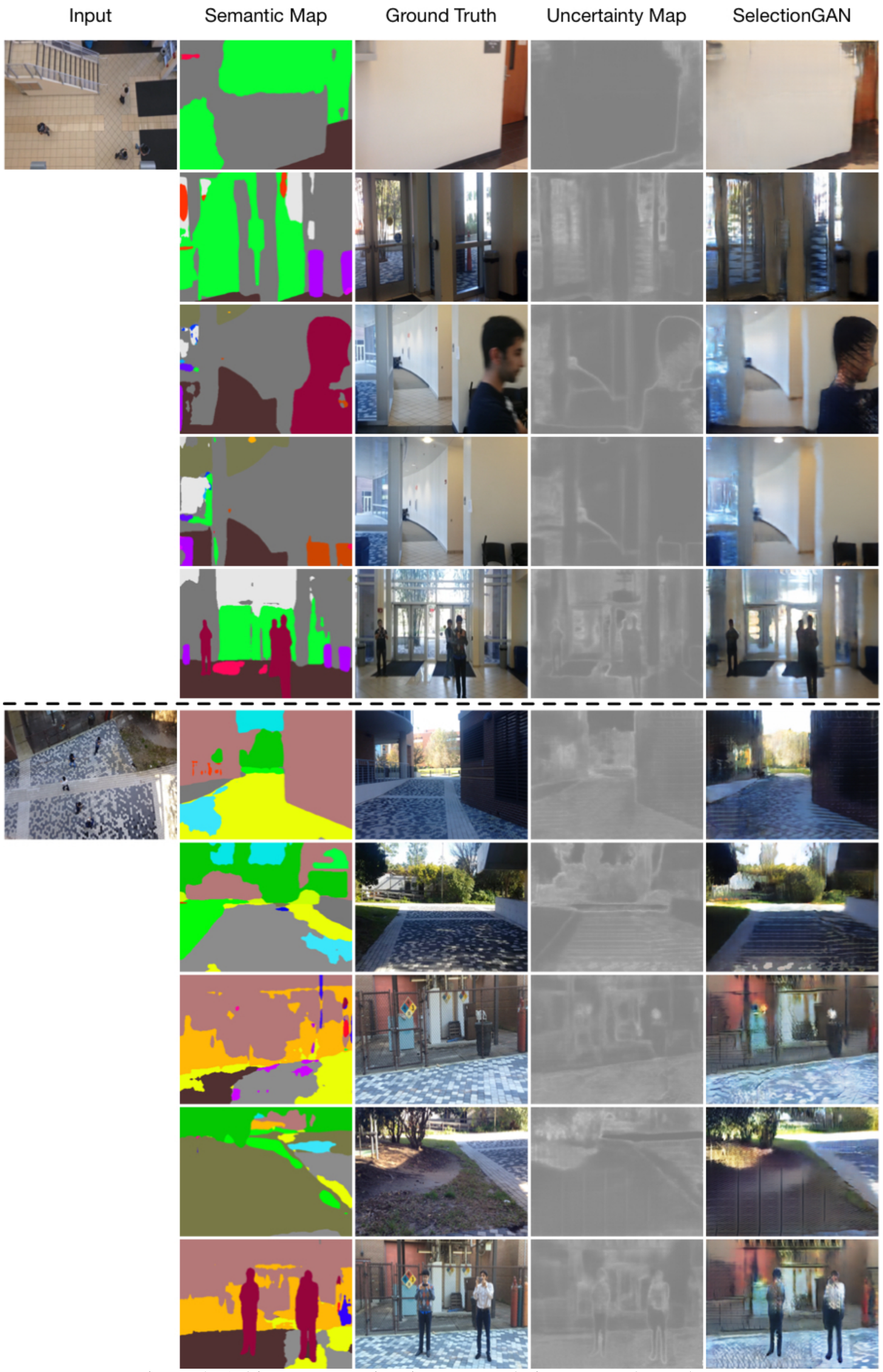

Figure 4: Arbitrary cross-view image translation on Ego2Top dataset.

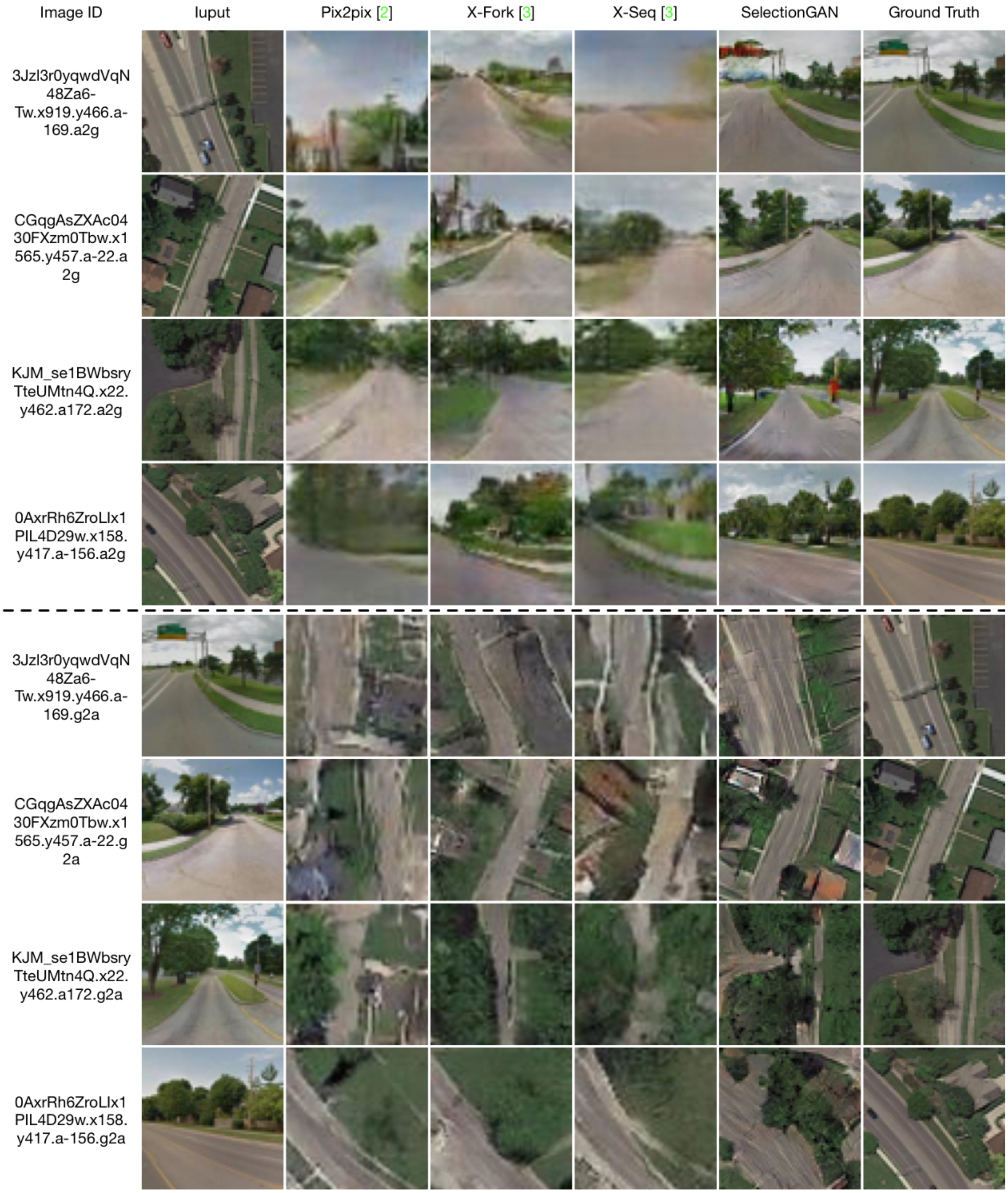

Figure 5: Results generated by different methods in 64×64 resolution in both a2g (Top) and g2a (Bottom) directions on Dayton dataset. These samples were randomly selected for visualization purposes.

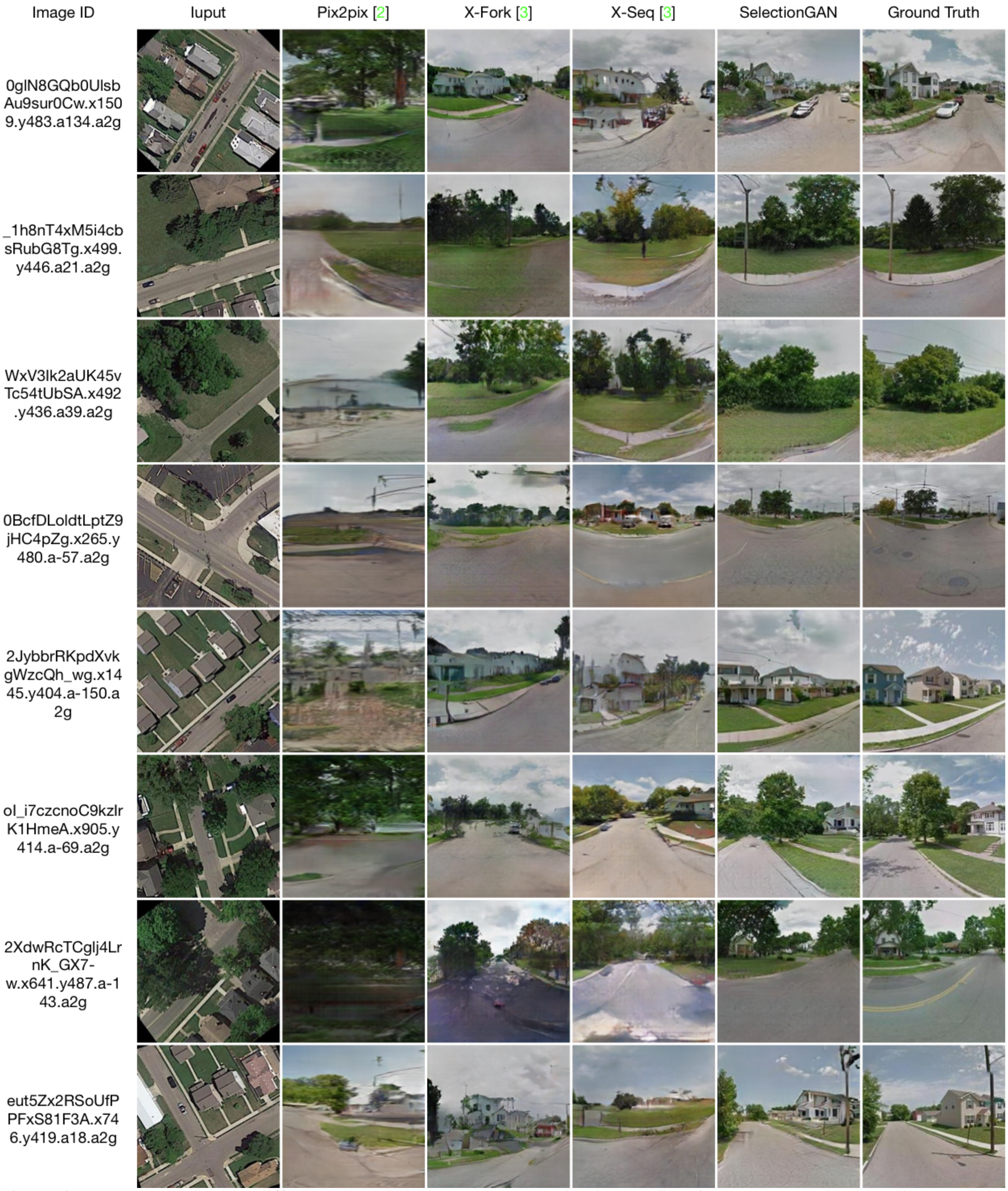

Figure 6: Results generated by different methods in 256×256 resolution in a2g direction on Dayton dataset. These samples were randomly selected for visualization purposes.

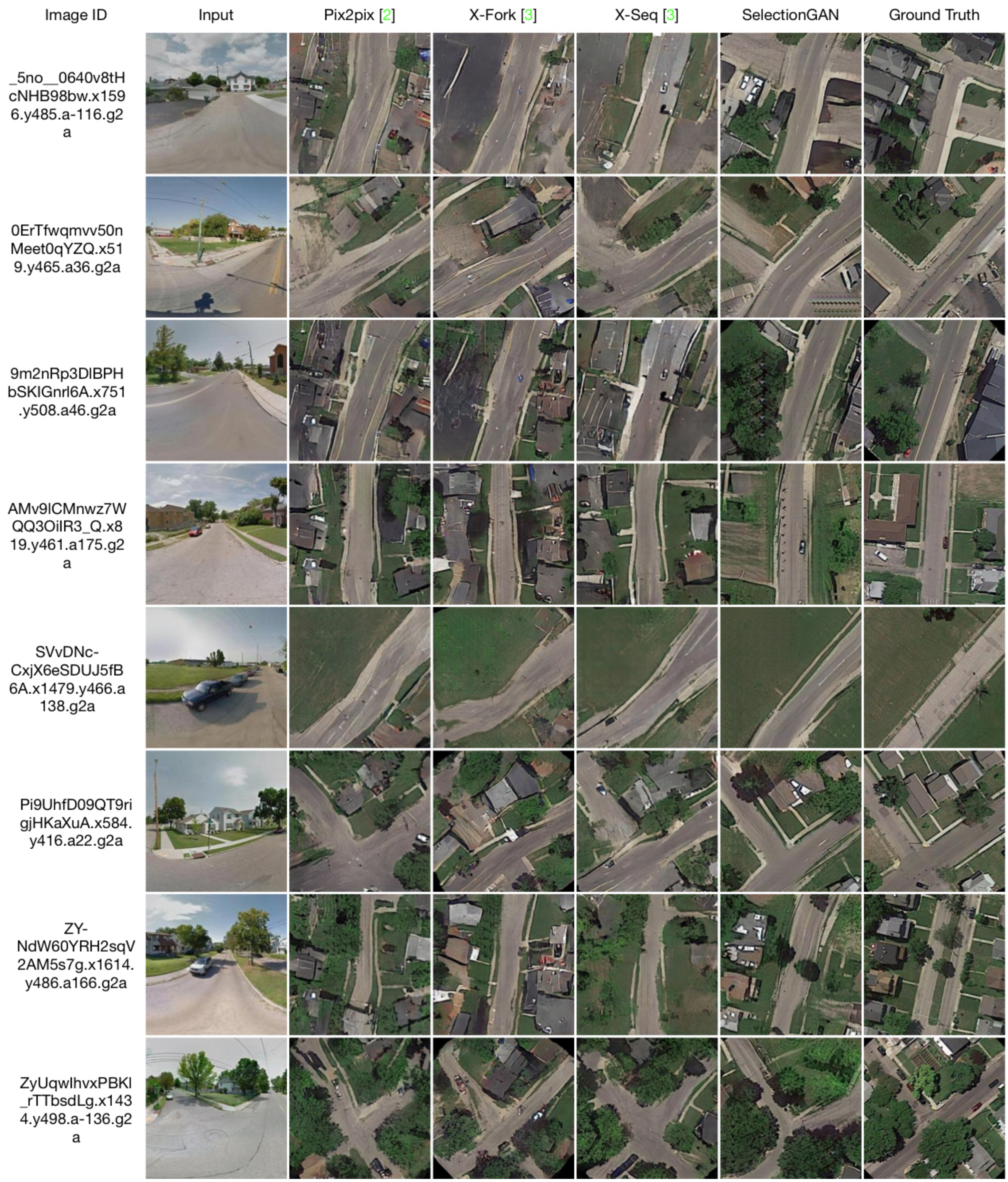

Figure 7: Results generated by different methods in 256×256 resolution in g2a direction on Dayton dataset. These samples were randomly selected for visualization purposes.

Figure 8: Results generated by different methods in 256×256 resolution in a2g direction on CVUSA dataset. These samples were randomly selected for visualization purposes.

Figure 9: Results generated by different methods in 256×256 resolution on Ego2Top dataset. These samples were randomly selected for visualization purposes.