

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 Wang⁵ Jason J. Corso⁶ Yan Yan²

¹DISI, University of Trento, Trento, Italy ²Texas State University, San Marcos, USA

³University of Oxford, Oxford, UK ⁴Huawei Technologies Ireland, Dublin, Ireland

⁵Northeastern University, Boston, USA ⁶University 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-Fine 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 authors¹. As we can see from all these figures, the proposed SelectionGAN achieves significantly visually better results than the competing methods.

References

- [1] Shervin Ardeshtir 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>

Image ID	Input	Semantic Map	Ground Truth	Uncertainty Map	SelectionGAN (Coarse)	SelectionGAN (Refined)
-1kaR7ild-fJdN1A1OS6FA.x684.y491.a-60.a2g						
0EVw6Y2ymQ0dmAFx-IMaWg.x116.y485.a28.a2g						
0AxrRh6ZroLlx1PIL4D29w.x158.y417.a-156.a2g						
0HZndNV-5q5L0Sks5-Xh-w.x814.y419.a88.a2g						
0n0Mac1ITv_QrgyAPGZUTQ.x1192.y466.a-7.a2g						
_4YFplQTAB0g9dHZRgZ1Bw.x1340.y462.a-158.a2g						
0CrKf20MURqfQi7kbQLX_Q.x144.y434.a-60.a2g						
0vNbX6tMHX136iAxqmBT9w.x22.y437.a131.a2g						

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.



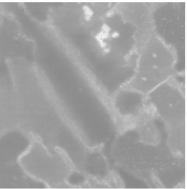




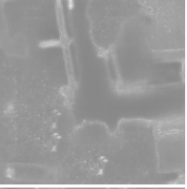
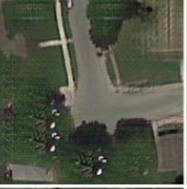




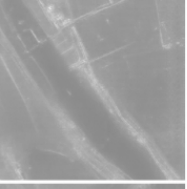

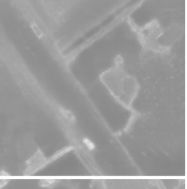



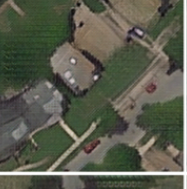




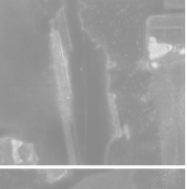





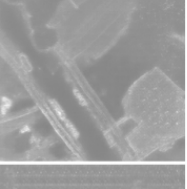




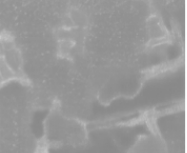


Image ID	Input	Semantic Map	Ground Truth	Uncertainty Map	SelectionGAN (Coarse)	SelectionGAN (Refined)
2ikSxbJ_IPDpqx Hyw3wsHQ.x15 3.y472.a-56.g2a						
Z98Z4LNHbFXz RID_Y3wdfg.x47 .y519.a-75.g2a						
MU0OeiEuJT- pdpzxmkiQ.x1 012.y333.a-127. g2a						
vsL_SD SUR- l2oZmK7z6eiw.x 164.y401.a-54.g 2a						
W8BJ4er1A-y9- M2N0TZMIA.x62 8.y419.a-143.g2 a						
yjqTMqt3UCGP4 Frc4b1srQ.x867. y490.a102.g2a						
z9LW3Cv- k8Vol_wpLBQx4 w.x1002.y441.a- 53.g2a						
MxSXk6HyV5H_ FeEJHUTcSg.x4 87.y471.a104.g2 a						

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.


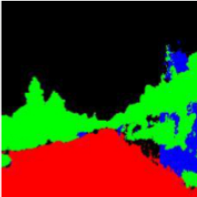

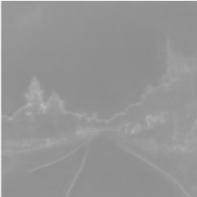



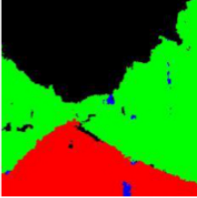

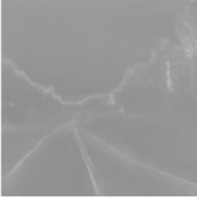



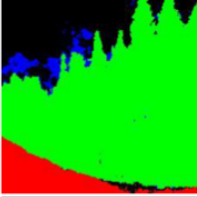
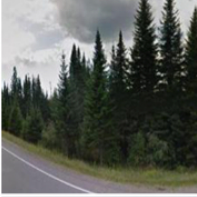
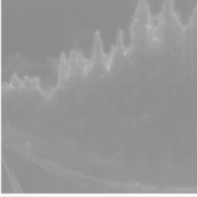



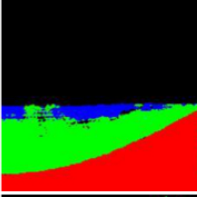




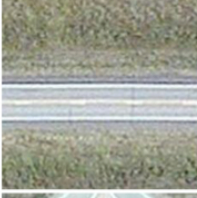
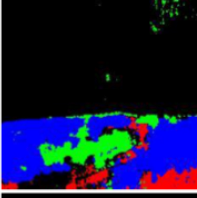
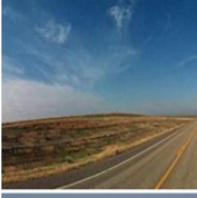
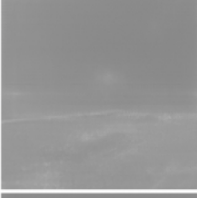



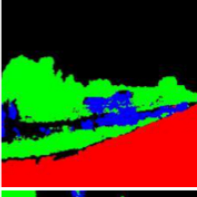




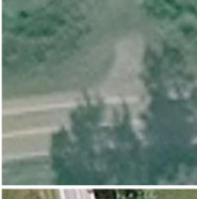
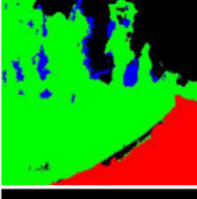

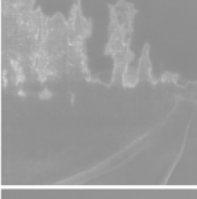



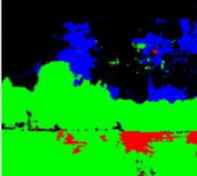




Image ID	Input	Semantic Map	Ground Truth	Uncertainty Map	SelectionGAN (Coarse)	SelectionGAN (Refined)
0000331						
0009794						
0010231						
0010153						
0010508						
0033374						
0034194						
0042714						

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.

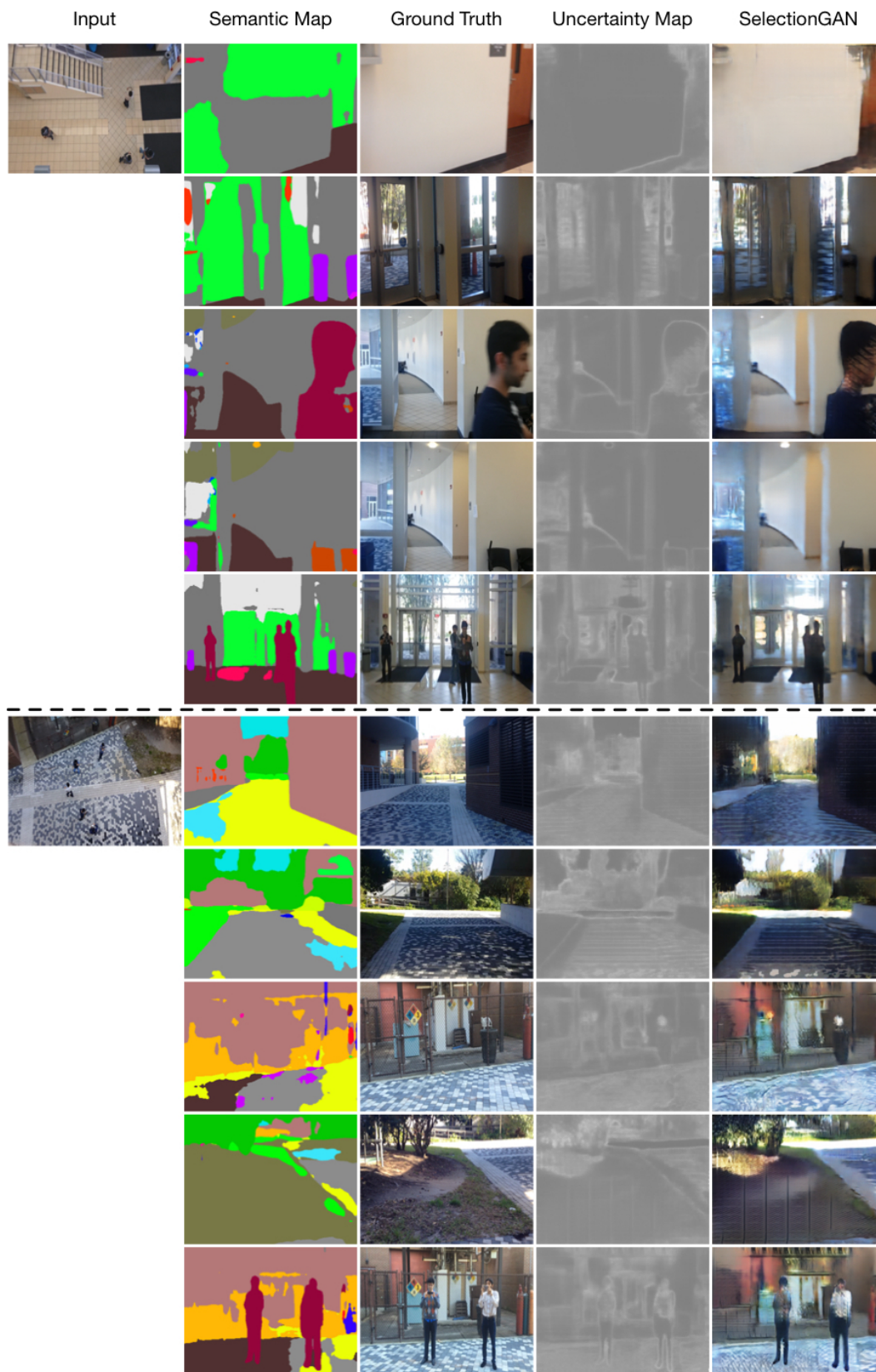


Figure 4: Arbitrary cross-view image translation on Ego2Top dataset.

Image ID	Input	Pix2pix [2]	X-Fork [3]	X-Seq [3]	SelectionGAN	Ground Truth
3Jzl3r0yqwdVqN 48Za6- Tw.x919.y466.a- 169.a2g						
CGqgAsZXAc04 30FXzm0Tbw.x1 565.y457.a-22.a 2g						
KJM_se1BWbsry TteUMtn4Q.x22. y462.a172.a2g						
0AxrRh6ZroLlx1 PIL4D29w.x158. y417.a-156.a2g						
3Jzl3r0yqwdVqN 48Za6- Tw.x919.y466.a- 169.g2a						
CGqgAsZXAc04 30FXzm0Tbw.x1 565.y457.a-22.g 2a						
KJM_se1BWbsry TteUMtn4Q.x22. y462.a172.g2a						
0AxrRh6ZroLlx1 PIL4D29w.x158. y417.a-156.g2a						

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.

Image ID	Input	Pix2pix [2]	X-Fork [3]	X-Seq [3]	SelectionGAN	Ground Truth
0gIN8GQb0Ulsb Au9sur0Cw.x150 9.y483.a134.a2g						
_1h8nT4xM5i4cb sRubG8Tg.x499. y446.a21.a2g						
WxV3Ik2aUK45v Tc54tUbSA.x492 .y436.a39.a2g						
0BcfDLoldtLptZ9 jHC4pZg.x265.y 480.a-57.a2g						
2JybbrRKpdXvk gWzcQh_wg.x14 45.y404.a-150.a 2g						
ol_i7czcnoC9kzlr K1HmeA.x905.y 414.a-69.a2g						
2XdwrCtCglj4Lr nK_GX7- w.x641.y487.a-1 43.a2g						
eut5Zx2RSoUfP PFxS81F3A.x74 6.y419.a18.a2g						

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.

Image ID	Input	Pix2pix [2]	X-Fork [3]	X-Seq [3]	SelectionGAN	Ground Truth
_5no__0640v8tH cNHB98bw.x151 6.y485.a-116.g2 a						
0ErTfwqmvv50n Meet0qYZQ.x51 9.y465.a36.g2a						
9m2nRp3DIBPH bSKIGnrl6A.x751 .y508.a46.g2a						
AMv9ICMnwz7W QQ3OIIIR3_Q.x8 19.y461.a175.g2 a						
SVvDNc- CxjX6eSDUJ5fB 6A.x1479.y466.a 138.g2a						
Pi9UhfD09QT9ri gjHKaXuA.x584. y416.a22.g2a						
ZY- NdW60YRH2sqV 2AM5s7g.x1614. y486.a166.g2a						
ZyUqwlhvxBKI _rTTbsdLg.x143 4.y498.a-136.g2 a						

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.

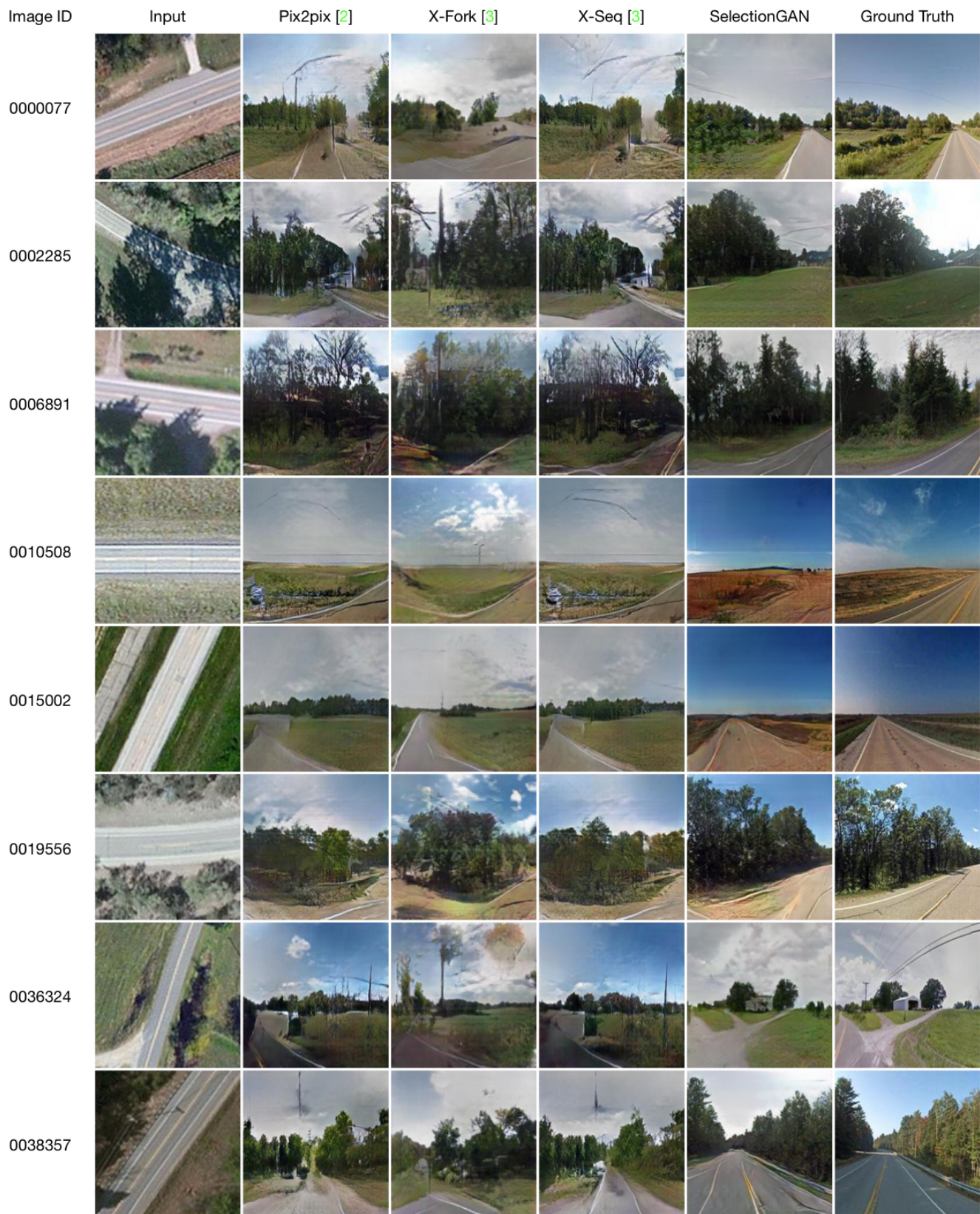


Figure 8: Results generated by different methods in 256×256 resolution in a2g direction on CVUSA dataset. These samples were randomly selected for visualization purposes.

Image ID	Input	Pix2pix [2]	X-Fork [3]	X-Seq [3]	SelectionGAN	Ground Truth
Case44_TopView_00451_Egocentric_1_00161						
Case46_Egocentric_5_00339_Egocentric_4_00008						
Case16_Egocentric_5_00298_Egocentric_5_00138						
Case24_Egocentric_4_00304_Egocentric_1_00001						
Case9_Egocentric_6_00202_Egocentric_4_01326						
Case19_TopView_00426_Egocentric_5_00264						
Case29_Egocentric_6_00150_Egocentric_5_00249						
Case39_Egocentric_5_00424_Egocentric_1_00481						

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