Geometry-Aware Satellite-to-Ground Image Synthesis for Urban Areas Supplementary Material

Xiaohu Lu[1*] Zuoyue Li[2*] Zhaopeng Cui[2†] Martin R. Oswald[2]
Marc Pollefeys[2,3] Rongjun Qin[1†]
[1]The Ohio State University [2]ETH Zürich [3]Microsoft

In this supplementary material, we provide further information about the generation of our dataset, the detailed network architecture, and additional experimental results.

1. Dataset

We select an approximate 4km×6km area centered in the City of London as the region of interest. Then the ground-truth depth and semantic images are generated from stereo matching [1, 5, 4] and supervised classification [7], respectively. We downloaded the corresponding street-view images via the Google Street View Static API[1]. Originally, we obtained 30k street-view panoramas in total with longitude, latitude, and orientation information. Fig. 1 shows the coverage of the dataset, where the red lines indicate the trajectories of the downloaded Google street-view panoramas. As introduced in Sec. 3.4.1 in the main paper, there are some misalignments between satellite images and street-view panoramas due to the error in the GPS information. To select the well aligned image pairs, we calculate the overlapping ratios of sky pixels on the geo-transformed street-view semantic image and the real street-view semantic image. The image pairs with overlapping ratio higher than 0.9 are considered as well aligned (*e.g.* the first seven rows in Fig. 5), while the rest are considered as not-aligned (the last four rows in Fig. 5). Therefore, we eventually obtain approximate 2k well-aligned satellite-street-view image pairs for the training phase.

2. Network Architecture

Tab. 1 provides a detailed description of the input and output tensor sizes of each sub-network in our pipeline. We use the same UNet structure for both UNet_{sat} and UNet_{str}, of which the network parameters are further detailed in Tab. 2. The BicycleGAN [9] we used consists of a UNet generator, two 3-layer discriminators and a ResNet encoder. We use the default network settings from BicycleGAN's official implementation[2] but changed the dimension of the latent vector to 32 and changed the number of filters in the first convolutional layer of all networks to 96. The external encoder has exactly the same structure as the encoder in BicycleGAN [9].

3. Additional Experimental Results

Additional qualitative comparison. Fig. 6 shows additional results of our method on the test dataset in comparison to Regmi *et al.* [6], and Pix2Pix [8, 2]. As can be seen, our proposed network performs significantly better on the semantic image generation than the other baseline. Especially regarding the correctness, it can be observed that the generated semantic image of [6] is only a rough guess of the street-view layout (*e.g.* row 2, 5, and 7) from the satellite image without any geometric guidance. While for our method, the street-view layout is almost the same as the ground truth, because the geometric layout of the scene is well preserved from the satellite to the street-view via our geo-transformation layer. Furthermore, the estimated position of the sidewalk in the result of [6] randomly appears in front of buildings. In contrast, our network can learn the relationship between sidewalk and building very well since the geometric information between the two classes can be better learned with the transformed street-view depth image. The improvements of our semantic images eventually also lead to significantly better RGB outputs of our method compared to the baselines. The results of [6] typically show less artifacts and more plausible street-view images than Pix2Pix. Nevertheless, the generated images from both baseline methods have many more blurred texture details and only some parts of the scene reflect the actual geometric scene configuration corresponding to the satellite image.

Pinhole-camera Test. To testify the flexibility of our network, we also present results for perspective images on the testing dataset which have been generated with the same

*These authors contributed equally to this work.
†Corresponding authors.
[1]https://developers.google.com/maps/documentation/streetview/intro
[2]https://github.com/junyanz/BicycleGAN

Figure 1. **Area coverage of our dataset.** Originally the satellite images cover an approximately 4km×6km area centered in the City of London, while the amount of street-view panoramas is around 30k in total.

network without any retraining by replacing the previous panoramic camera model. To this end, we generated images for a virtual pinhole camera using the same optical camera center as the panorama image with the height, width, focal length equal to 256, 256, 64 pixels, and a principle direction heading to the center of the panorama. Looking at the results presented in Fig. 2, it is apparent that our network generalizes well and produces pleasing street-view images also for the pinhole camera setting. Please note that the baselines Regmi *et al*. [6] and Pix2Pix [8, 2] do not generalize and require retraining in order to change the camera model.

Discussion on the failure cases. We also provide some failure examples in the last three rows of Fig. 6. We notice that these results do not have a similar skyline layout compared to the ground truth street-view panoramas. This

Table 1. **Sub-network overview.** We detail the input and output dimensions for all major parts in our pipeline.

Model/Layer	IO	Description	Tensor Dimension
UNet_{sat}	Input	Satellite RGB	$H_{sat} \times W_{sat} \times 3$
	Output	Satellite Depth + Semantics	$H_{sat} \times W_{sat} \times (1+3)$
UNet_{str}	Input	Transformed Street-view Depth + Semantics and Resized Satellite RGB	$H_{str} \times W_{str} \times (1+3+3)$
	Output	Street-view Semantics	$H_{str} \times W_{str} \times 3$
BicycleGAN	Input	Street-view Semantics + Depth and Encoded Latent Vector	$H_{str} \times W_{str} \times (3+N_{enc})$
	Output	Street-view RGB	$H_{str} \times W_{str} \times 3$
Geo-transformation	Input	Satellite Depth + Semantics	$H_{sat} \times W_{sat} \times (1+3)$
	Output	Transformed Street-view Depth + Semantics	$H_{str} \times W_{str} \times (1+3)$
Inv. Geo-transformation	Input	Street-view RGB + Depth	$H_{str} \times W_{str} \times (3+1)$
	Output	Inv. Transformed Satellite RGB	$H_{sat} \times W_{sat} \times 3$
External Encoder	Input	Satellite RGB	$H_{sat} \times W_{sat} \times 3$
	Output	Encoded Latent Vector	N_{enc}

Table 2. Detailed UNet network architecture used for the two networks: $\text{UNet}_{\{sat,str\}}$.

Part	Layer	Parameters	Output Dimension
Encoder	Conv1+BN+LeakyReLU	4×4, 64, 2	$H/2 \times W/2 \times 64$
	Conv2+BN+LeakyReLU	4×4, 128, 2	$H/4 \times W/4 \times 128$
	Conv3+BN+LeakyReLU	4×4, 256, 2	$H/8 \times W/8 \times 256$
	Conv4+BN+LeakyReLU	4×4, 512, 2	$H/16 \times W/16 \times 512$
	Conv5+BN+LeakyReLU	4×4, 512, 2	$H/32 \times W/32 \times 512$
	Conv6+BN+LeakyReLU	4×4, 512, 2	$H/64 \times W/64 \times 512$
	Conv7+BN+LeakyReLU	4×4, 512, 2	$H/128 \times W/128 \times 512$
	Conv8+BN+LeakyReLU	4×4, 512, 2	$H/256 \times W/256 \times 512$
Decoder	Decov1+BN+ReLU	4×4, 512, 2	$H/128 \times W/128 \times 512$
	Concat1	cat(Conv7,Decov1)	$H/128 \times W/128 \times 1024$
	Decov2+BN+ReLU	4×4, 512, 2	$H/64 \times W/64 \times 512$
	Concat2	cat(Conv6,Decov2)	$H/64 \times W/64 \times 1024$
	Decov3+BN+ReLU	4×4, 512, 2	$H/32 \times W/32 \times 512$
	Concat3	cat(Conv5,Decov3)	$H/32 \times W/32 \times 1024$
	Decov4+BN+ReLU	4×4, 512, 2	$H/16 \times W/16 \times 512$
	Concat4	cat(Conv4,Decov4)	$H/16 \times W/16 \times 1024$
	Decov5+BN+ReLU	4×4, 256, 2	$H/8 \times W/8 \times 256$
	Concat5	cat(Conv3,Decov5)	$H/8 \times W/8 \times 512$
	Decov6+BN+ReLU	4×4, 128, 2	$H/4 \times W/4 \times 128$
	Concat6	cat(Conv2,Decov6)	$H/4 \times W/4 \times 256$
	Decov7+BN+ReLU	4×4, 64, 2	$H/2 \times W/2 \times 64$
	Concat7	cat(Conv1,Decov7)	$H/2 \times W/2 \times 128$
	Decov8+BN+ReLU	4×4, {4,3}, 2	$H \times W \times 3$
	Tanh	-	$H \times W \times \{4, 3\}$

is mainly due to the error in the satellite depth estimation and also the misalignment of the satellite image and ground truth images. Moreover, we also noticed that there are very small artifacts in our generated panoramas. Actually, this kind of artifacts are a known problem of GANs which use transposed convolution (deconvolution) in the decoder. According to [3], this problem could be addressed by replacing the transposed convolution layer to a combination of bilinear upsamplling and a general convolution.

Qualitative comparison between with and without inv. geo-transformation layer. In Fig. 3, we provide two qualitative examples generated by the models with and without the loss respectively. As can be seen, the model with the inverted geo-transformation layer can yield more visible white lane lines while the roads generated by the model without the loss have a relatively uniform gray with sparse lane lines which is not very obvious. We believe that the layer can help to utilize the evidence provided by satellite images and can better preserve street patterns. Since we do not have ground truth lane marks, we did not include the quantitative results.

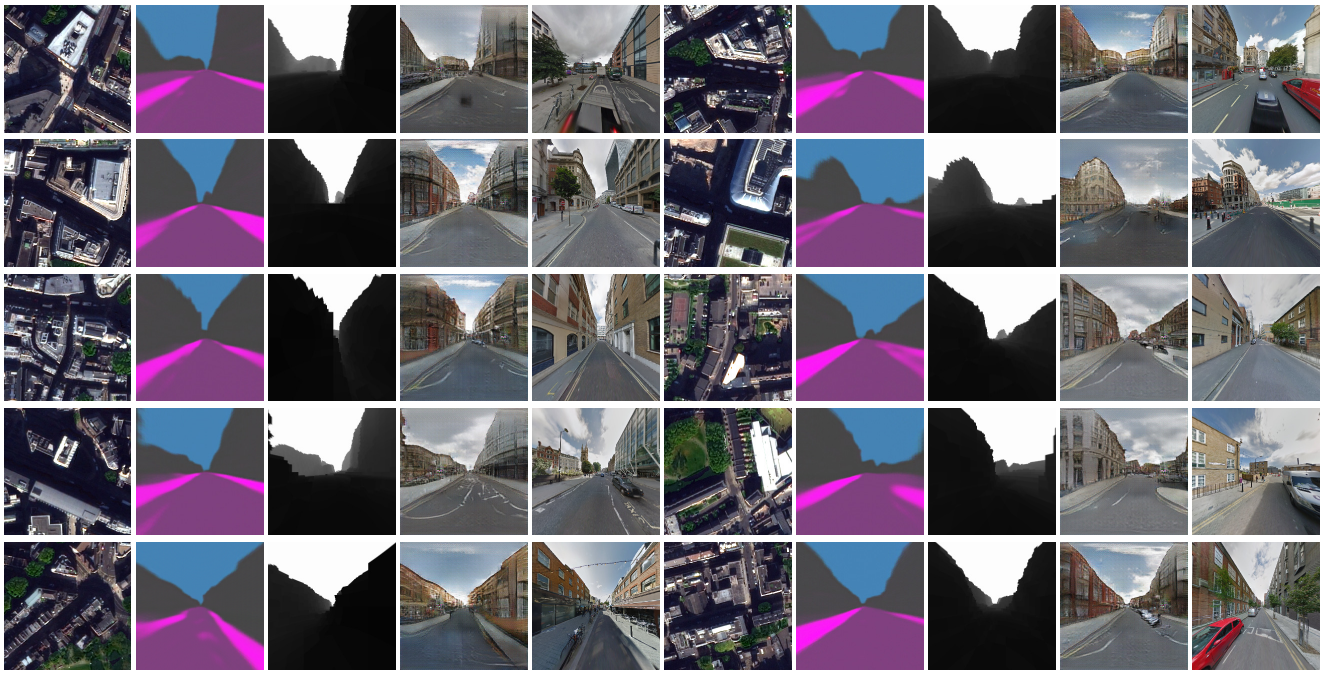

Figure 2. **Image generation results with a pinhole camera model.** Given the satellite RGB input, our method can also predict perspective semantic and depth images, and also generate geometrically correct perspective pinhole-camera images with good texture.

(a) w/o inverted geo-transf. (b) with inverted geo-transf.

Figure 3. Qualitative comparison between with and without inverted geo-transformation layer.

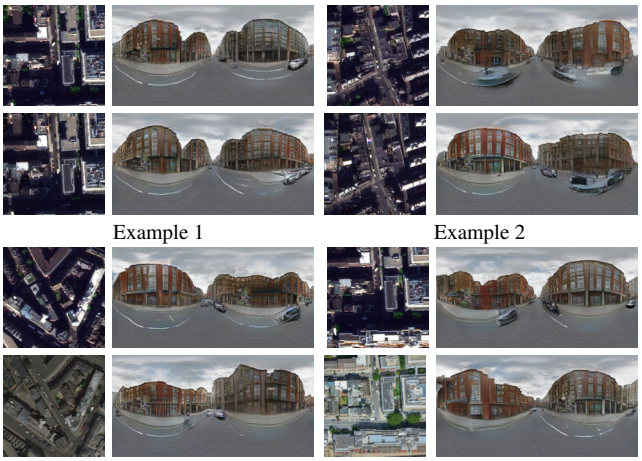

Figure 4. Qualitative comparison regarding to satellite content and photometric changes.

Robustness to content and photometric changes. Strong photometric changes can influence the performance of the estimated geometry, but typically still lead to plausible panoramas. Satellites with optical cameras are often in a sun-synchronous orbit and visit the same place in approximately the same time of the day. As a result, there are little photometric changes on satellite images captured within close dates (e.g. within a few weeks). The examples 1 and 2 in Fig. 4 show a test of our method on two satellite images with similar contents, leading to very little differences in both texture and layout of the resulting panorama. In contrast, seasonal changes significantly affect sun angles, illumination and structural changes (e.g. cast shadows, snow) which potentially impact the predicted depth. The examples 3 and 4 in Fig. 4 show satellite images which were taken 9 months apart. For image regions with significant appearance differences our trained model generates different depth models and the corresponding predicted street-view panorama will no longer preserve the correct geometry of the scene's layout. Fortunately, the texture of the predicted street-view panorama is still of good quality, which means that our satellite stage network domains the geometry while the street-view stage controls the texture of the final predicted street-view panorama.

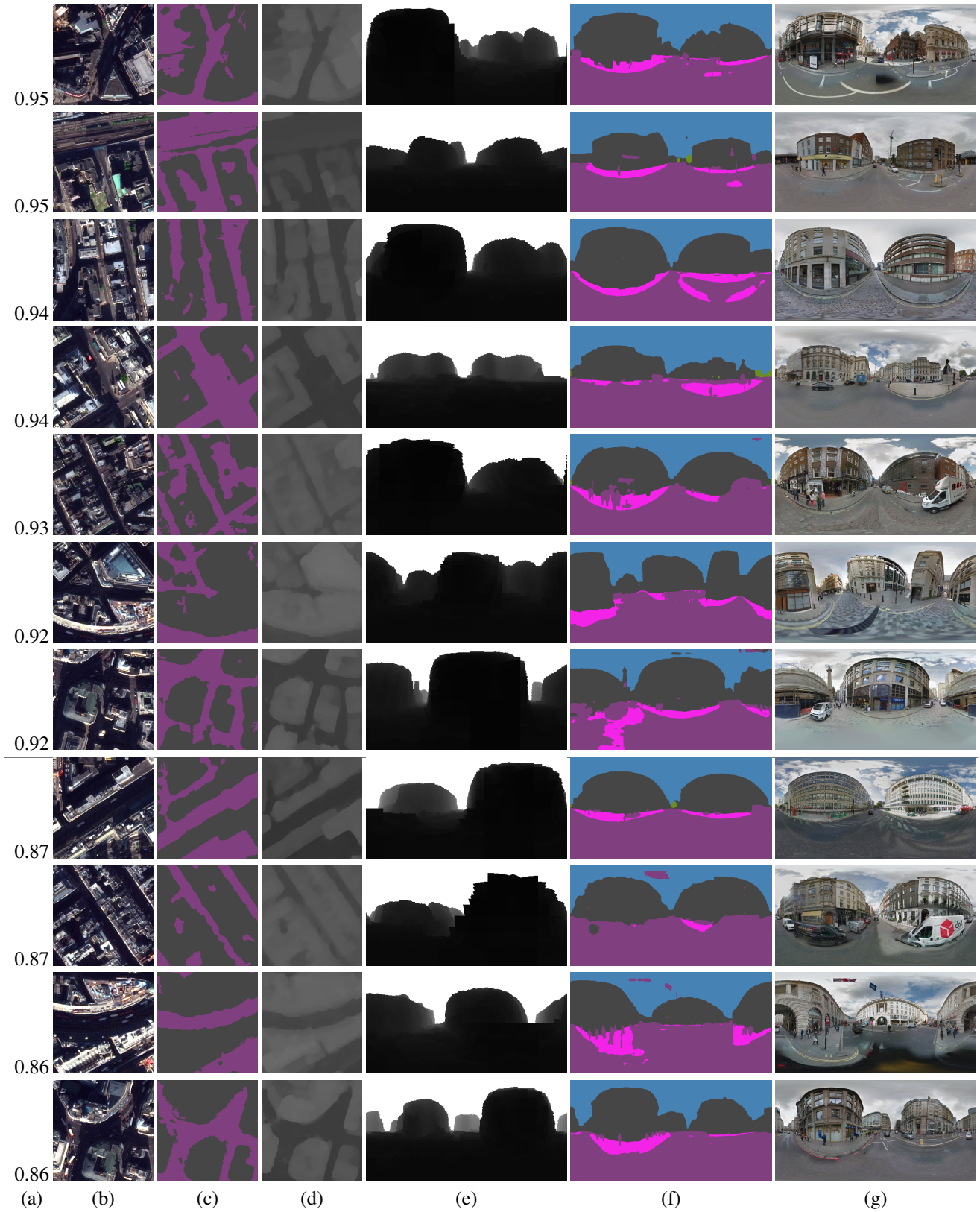

Figure 5. **Samples of our training dataset.** Left to right: **(a)** is the overlapping ratio (the higher the better aligned) of the image, **(b)** the satellite image, **(c)** the ground-truth satellite semantic segmentation, **(d)** the ground-truth satellite depth, **(e)** the transformed depth of (d), **(f)** the ground-truth street-view semantic, and **(g)** the ground-truth street-view image, respectively.

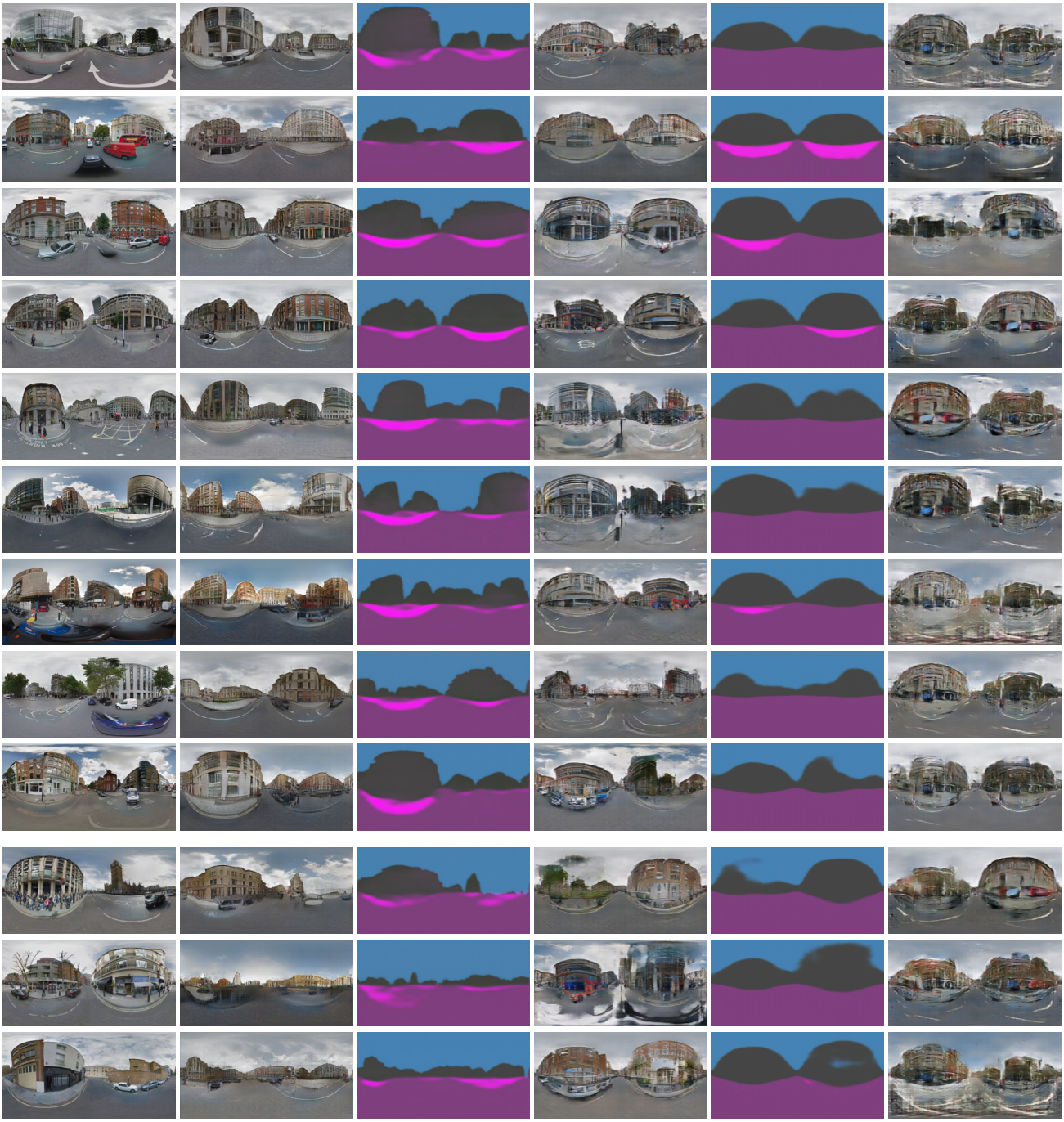

Figure 6. **Additional qualitative comparisons.** We present additional test results of our method, in comparison to Regmi *et al*. [6], and Pix2Pix [2]. The last three rows illustrate failure cases.

References

[1] Heiko Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(2):328–341, 2008. 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, 2, 6

[3] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. 3

[4] Rongjun Qin. Rpc stereo processor (rsp)–a software package for digital surface model and orthophoto generation from satellite stereo imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 3:77, 2016. 1

[5] Rongjun Qin. Automated 3d recovery from very high resolution multi-view satellite images. In *ASPRS (IGTF) annual Conference*, page 10, 2017. 1

[6] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3501–3510, 2018. 1, 2, 6

[7] Qian Zhang, Rongjun Qin, Xin Huang, Yong Fang, and Liang Liu. Classification of ultra-high resolution orthophotos combined with dsm using a dual morphological top hat profile. *Remote Sensing*, 7(12):16422–16440, 2015. 1

[8] 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. 1, 2

[9] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017. 1