

Sat2Vid: Street-view Panoramic Video Synthesis from a Single Satellite Image

Supplementary Material

Zuoyue Li¹ Zhenqiang Li² Zhaopeng Cui^{3*} Rongjun Qin⁴ Marc Pollefeys^{1,5} Martin R. Oswald^{1*}
¹ETH Zürich ²The University of Tokyo ³Zhejiang University ⁴The Ohio State University ⁵Microsoft

In this supplementary material, we provide further information about the detailed network architecture and additional quantitative / qualitative experiment results. Please find our project webpage at <https://github.com/lizuoyue/sat2vid>.

A. Additional Implementation Details

As a supplement to Fig. 2 in the main paper, the structure of the light-weight upsampling module is illustrated in Fig. A, where two outputs with different resolutions are used to compute the losses of the GAN. The main purpose of the upsampling module is better scalability for high-resolution video output. Since the geometrically and temporally consistent 3D point-based scene generation is more expensive due to the 3D setting, we can still maintain this consistency while performing efficient upsampling in the 2D setting to increase the output resolution. The upsampling module is designed rather light-weight in order to only perform small modifications on the output data to avoid breaking the scene consistency given by the point cloud.

Tab. A further provides a detailed description of the input and output tensor sizes of each sub-network in our pipeline. We use the same U-Net [5] structure as used in S2G [3] for the satellite depth and semantics prediction. For SparseConvNet [1] which is used as the coarse generator, we follow its official U-Net-backbone implementation⁶ with (64, 128, 192, 256, 320) channels and residual blocks enabled (2 repeats). The fine generator, RandLA-Net [2], adopts an unofficial PyTorch implementation⁷, where the default hyper-parameters are used but we set the number of nearest neighbors to 8, and the decimation ratio in its local feature aggregation module to 4 as mentioned in the main paper (Sec. 3.2).

The training uses the default network settings from the official implementation⁸ of BicycleGAN [8] with two 3-

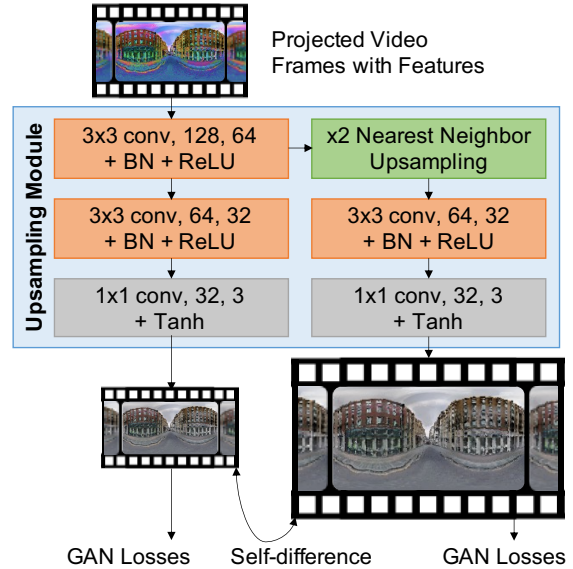


Figure A: **Details of the upsampling module.** The module uses few convolutional layers and outputs both the low- and high-resolution RGB videos, which are further used to compute GAN losses.

layer discriminators and an internal encoder. The loss of the generator consists of 3 parts: the L_1 differences between the low- and high-resolution outputs and their corresponding ground truth, the L_1 difference between themselves (downsampling the higher-resolution one), as well as the perceptual losses [7] performed on both outputs.

Because the accuracy of the satellite depth is barely high enough, it is hard to ensure that the point cloud estimated from the satellite depth is well aligned with the generated ground-truth panorama video. For instance, the distant points (geometrically *sky* points) might be assigned with colors or semantics of the building if the estimated height of that building does not agree with the panorama. Especially, for the objects like street lamps and cars which are nearly invisible in the satellite image, the misalignment is inevitable. We simply use the semantic information to determine the misalignment for which pixels will be given zero weights when computing these losses.

*Corresponding author.

Please use **Adobe Reader** / **KDE Okular** to view *animations*.

⁶<https://github.com/facebookresearch/SparseConvNet>

⁷<https://github.com/aRI0U/RandLA-Net-pytorch>

⁸<https://github.com/junyanz/BicycleGAN>

Model/Layer	IO	Description	Tensor Dimension
UNet [5]	Input Output	Sat. RGB Sat. Dep. + Sem.	$H_s \times W_s \times 3$ $H_s \times W_s \times (1 + 1)$
Transformation	Input Output	Sat. Dep. + Sem. Trajectory Coord. Point Cloud with Sem. + Coord. Point-pixel Correspondence	$H_s \times W_s \times (1 + 1)$ $T \times 2$ $N_p \times (1 + 3)$ $T \times H_v/2 \times W_v/2$
SparseConvNet [1]	Input Output	Point Cloud with Sem. + Coord. + Latent Vec. Point Cloud with Coarse Ft.	$N_p \times (1 + 3 + D_e)$ $N_p \times D_c$
RandLA-Net [2]	Input Output	Point Cloud with Sem. + Coord. + Latent Vec. + Coarse Ft. Point Cloud with Fine Ft.	$N_p \times (1 + 3 + D_e + D_c)$ $N_p \times D_f$
Concatenation & Projection	Input Output	Point Cloud Coarse & Fine Ft. Point-pixel Correspondence Low-resolution Frames with Ft.	$N_p \times (D_c + D_f)$ $T \times H_v/2 \times W_v/2$ $T \times H_v/2 \times W_v/2 \times (D_c + D_f)$
Upsampling Module	Input Output	Low-resolution Frames with Ft. High-resolution RGB Video	$T \times H_v/2 \times W_v/2 \times (D_c + D_f)$ $T \times H_v \times W_v \times 3$
Multi-class Encoder [8]	Input Output	Sat. RGB Per-class Encoded Latent Vec.	$H_s \times W_s \times 3$ $N_c \times D_e$

Table A: **Sub-networks overview.** We detail the input and output dimensions for all major parts in our pipeline. H_s, W_s : height and width of the satellite images; T, H_v, W_v : number of frames, height and width of the output video; N_p : number of points in the point cloud; D_c, D_f, D_e : dimension of the coarse, fine features and encoded latent vectors. In our experiment, $H_s = W_s = 256, T = 15$ or $60, H_v = 256, W_v = 512, N_p$ varies on different scenes, $D_c = D_f = 64$ and $D_e = 16$.

Sat. Ground Truth S2G-F [3] S2G-I [3] Vid2Vid [6] WC-Vid2Vid [4] Sat2Vid (Ours)

Figure B: **Qualitative baseline comparison as supplements (animations).**

Sat. Ground Truth Vid2Vid [6] WC-Vid2Vid [4] Sat2Vid(Ours)

Figure C: **Qualitative lateral motion trajectories (animations).**

Method	PSNR \uparrow	SSIM \uparrow	Sharp Diff. \uparrow	P _{Alex} \downarrow	P _{Squeeze} \downarrow	P _{VGG} \downarrow
Vid2Vid [6]	22.099	0.794	31.328	0.108	0.072	0.198
WC-Vid2Vid [4]	18.244	0.458	26.486	0.274	0.179	0.371
Sat2Vid (Ours)	22.969	0.818	34.090	0.111	0.070	0.174

Table B: **Quantitative cross-sample consistency.** The evaluation is based on the videos generated from two opposite directions.

Method	PSNR \uparrow	SSIM \uparrow	Sharp Diff. \uparrow	P _{Alex} \downarrow	P _{Squeeze} \downarrow	P _{VGG} \downarrow
Vid2Vid [6]	13.547	0.395	25.596	0.491	0.365	0.549
WC-Vid2Vid [4]	13.869	0.348	25.585	0.527	0.380	0.566
Sat2Vid (Ours)	14.993	0.414	25.857	0.482	0.342	0.537

Table C: **Quantitative lateral motion trajectories.** The evaluation is based on videos generated from lateral motion trajectories.

Sat. Azimuth turning left (↶) Azimuth turning right (↷)
Figure D: **Videos with complex trajectory (animations).**

B. Additional Experiment Results

B.1. More Qualitative Results

As a supplement to Fig. 3 in the main paper, we present more qualitative results in Fig. B.

In addition to the simple forward path we used in the main paper, our method also allows for more complex camera motion paths. The videos in Fig. D show a forward moving camera which also **goes down and up** in combination with different azimuth angle changes.

B.2. Lateral motion Trajectories

We further conducted an experiment with lateral-motion trajectories, *i.e.*, the moving direction is in the left of the viewing direction, and they are mutually perpendicular. The quantitative and qualitative results are illustrated in Tab. C and Fig. C, respectively. It can be seen that the results and conclusion are consistent with the one with forwarding paths shown in the main paper, *i.e.*, our method surpasses the baseline methods.

B.3. Inverse Trajectories

Besides the temporal self-consistency evaluation via a u-turn-shape trajectory from a single run (see main paper Sec. 4.4), we also conduct another experiment for the global consistency across different trajectories in the same scene. The basic reasoning behind this is that the videos generated from the same satellite image but different trajectories should be globally consistent.

Specifically, we reverse the original forward-motion trajectory and generate the video again, and then compare the frames along with the same locations with forward and backward motions (*e.g.*, the starting point of the forward one and the ending point of the backward one).

The quantitative results are illustrated in Tab. B, where we outperform other state-of-the-art methods in almost all the metrics. We also show three examples in Fig. E. Please note that for better visual comparison, the frames of the

Sat. Vid2Vid [6] WC-Vid2Vid [4] Sat2Vid(Ours)

Figure E: **Qualitative evaluation of the cross-sample consistency (animations).** Results are generated based on two opposite trajectories. **For each example - top row:** forward motion; **mid row:** backwards-played and shifted backward motion; and **bottom row:** the pixel-wise RGB differences.

backward trajectory (the second line of each example) are played backward and shifted according to the locations and viewing directions of the forward trajectory.

We notice that Vid2Vid [6] has better cross-sample consistency than WC-Vid2Vid [4], which might mainly be due to its generated fixed patterns (*e.g.*, the windows in the buildings). From the third line of each example, we can also see that our results have significantly lower cross-sample differences.

References

- [1] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2
- [2] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2
- [3] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R. Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2
- [4] Arun Mallya, Ting-Chun Wang, Karan Sapra, and Ming-Yu Liu. World-consistent video-to-video synthesis. In *Eur. Conf. Comput. Vis.*, 2020. 2, 3
- [5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015. 1, 2
- [6] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *Proceedings of Conference on Neural Information Processing Systems (NeurIPS)*, 2018. 2, 3
- [7] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 1
- [8] 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, 2