Image-to-Voxel Model Translation for 3D Scene Reconstruction and Segmentation Supplementary material

Vladimir V. Kniaz[1,2] [0000−0003−2912−9986],
Vladimir A. Knyaz[1,2] [0000−0002−4466−244X],
Fabio Remondino[3] [0000−0001−6097−5342],
Artem Bordodymov[1] [0000−0001−8159−2375], and
Petr Moshkantsev[1] [0000−0001−9624−4322]

[1] State Res. Institute of Aviation Systems (GosNIIAS), Moscow, Russia
[2] Moscow Institute of Physics and Technology (MIPT), Russia {knyaz, vl.kniaz, bordodymov, moshkantsev}@gosniias.ru
[3] Bruno Kessler Foundation (FBK), Trento, Italy
remondino@fbk.eu

1 Pose6DoF Discriminator

Our Pose6DoF discriminator works by processing an input concatenated from a fruxel model and an input image (see Figure 1). Different from modern volumetric discriminators [1], that qualify the input voxel model as being either 'real' or 'fake,' our Pose6DoF discriminator estimates 6DoF poses of objects in the scene and their perceptual realism. Hence, the architecture of our Pose6DoF discriminator fuses a pose estimation model and a discriminator. The architecture of our Pose6DoF discriminator is based on the inverted volumetric residual blocks.

Firstly, we concatenate the fruxel model (either real or predicted by generator) with the input image. We use 'copy-inflate' skip connections to match the 2D-3D dimensions. After that, our Pose6DoF processes the input using blocks of 3D inverted residual blocks. Finally, it produces a tensor with $16 \times 16 \times 16$ cells. Each cell contains five groups of parameters representing annotations of poses of five objects the could be located in the cell. A single group of parameters includes object's pose $t = \{x_c, y_c, z_c\}$ in normalized fruxel space coordinates, object's rotation quaternion $q = \{q_1, q_2, q_3, q_4\}$, object's dimensions $d = \{w, h, d\}$, and the probability r of object being either 'real' or 'fake.' We cluster annotations hypnotizes similar to [2].

2 SemanticVoxels Dataset

Our SemanticVoxels Dataset includes 116k samples of 3D and 2D data. Each data sample represents a single camera pose. It consists of a color image, a semantic frustum voxel model, a depth map, a camera pose, and an object pose annotations for all classes. We made our dataset consistent with the NuScenes

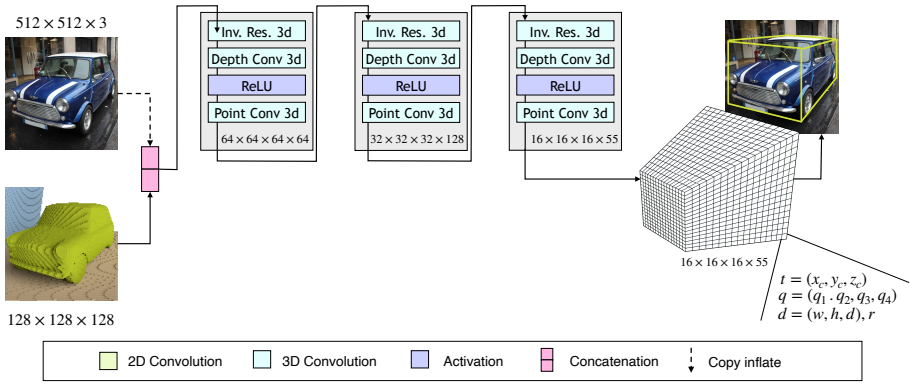

Fig. 1. `Pose6DoF` discriminator.

dataset format [3]. Our dataset is divided into two splits: real and synthetic. The real split was generated using a Structure-from-Motion (SfM) technique similar to [4]. It contains 16k images. Example scenes from the dataset are shown in Figure 2.

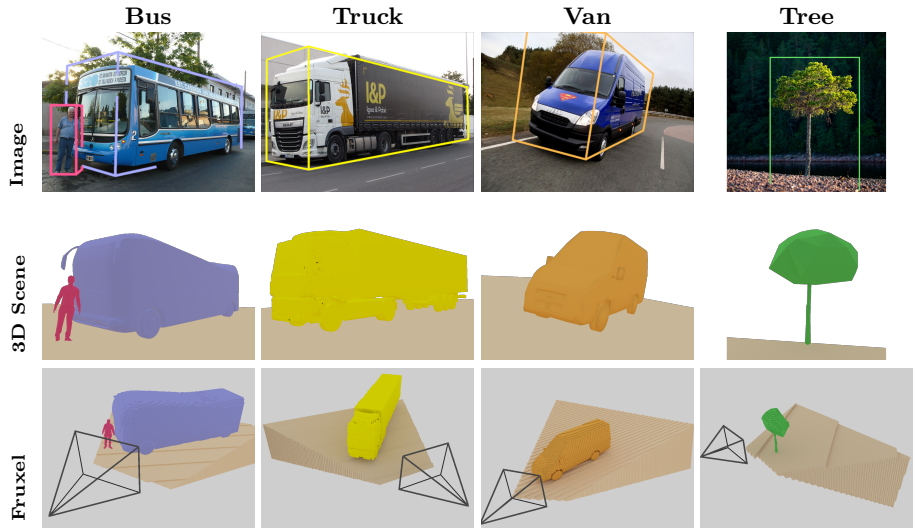

Fig. 2. Examples of color images with 6D pose annotations and ground truth semantic voxel models from our SemanticVoxels dataset.

3 Qualitative Evaluation

We present additional qualitative results of 3D reconstruction using `DISN` [5], `Pix2Vox` [6], `3D-R2N2` [7], and our `SSZ` on our SemanticVoxels dataset in Figure 3.

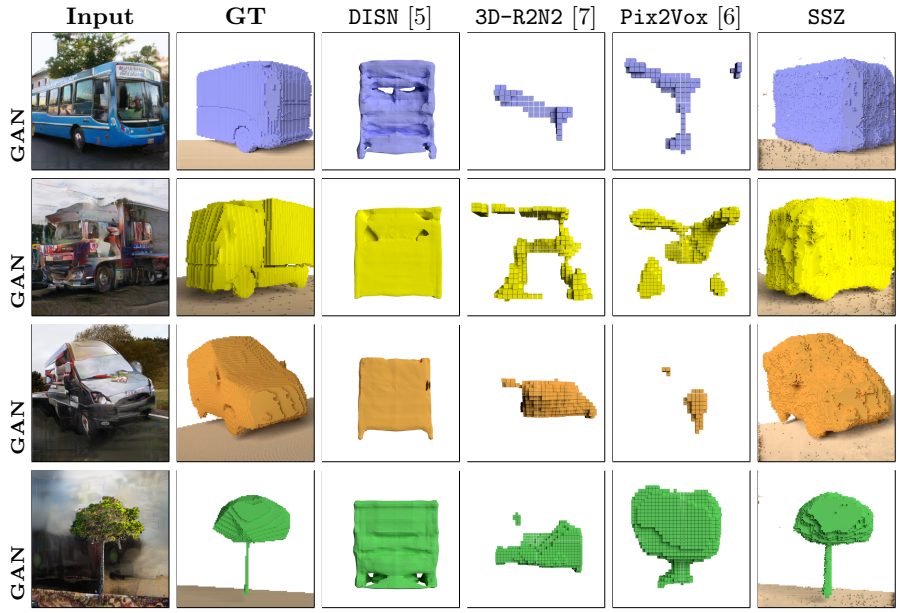

Fig. 3. Example of 3D reconstruction using `DISN` [5], `Pix2Vox` [6], `3D-R2N2` [7], and our `SSZ` on our SemanticVoxels dataset.

References

1. Wu, J., Zhang, C., Xue, T., Freeman, W.T., Tenenbaum, J.B.: Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: Advances in Neural Information Processing Systems. (2016) 82–90
2. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. (2017) 6517–6525
3. Hodaň, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. IEEE Winter Conference on Applications of Computer Vision (WACV) (2017)
4. Locher, A., Havlena, M., Gool, L.V.: Progressive structure from motion. In: Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IV. (2018) 22–38
5. Xu, Q., Wang, W., Ceylan, D., Mech, R., Neumann, U.: Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In Wallach, H., Larochelle,

H., Beygelzimer, A., dÁlché-Buc, F., Fox, E., Garnett, R., eds.: Advances in Neural Information Processing Systems 32. Curran Associates, Inc. (2019) 492–502
6. Xie, H., Yao, H., Sun, X., Zhou, S., Zhang, S.: Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In: The IEEE International Conference on Computer Vision (ICCV). (October 2019)
7. Choy, C.B., Xu, D., Gwak, J., Chen, K., Savarese, S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In: Proceedings of the European Conference on Computer Vision (ECCV). (2016)