LEGO: Learning Edge with Geometry all at Once by Watching Videos Suppelmentary Material

Zhenheng Yang[1] Peng Wang[2] Yang Wang[2] Wei Xu[3] Ram Nevatia[1]

[1]University of Southern California [2]Baidu Research
[3]National Engineering Laboratory for Deep Learning Technology and Applications

1. Edge ground truth generation

The edge ground truth of Cityscapes dataset is generated from semantic segmentation ground truth. Some semantic categories share the same 3D surface and are connected in geometrical sense. These geometrically-consistent categories are combined and the geometrical edges are extracted from combined segmentation results. The edges between different instances are preserved in this process. There are four groups of such combining categories as shown in Tab. 1. Examples of the generation of geometrical ground truth are presented in Fig. 1.

Table 1: Four groups of semantic categories are combined.

Combined Category	Combining Categories
'ground'	'ground', 'road', 'sidewalk', 'parking'
'pole'	'pole', 'polegroup', 'traffic light', 'traffic sign'
'rider'	'rider', 'motorcycle', 'bicycle'
'wall'	'wall', 'fence', 'guard rail'

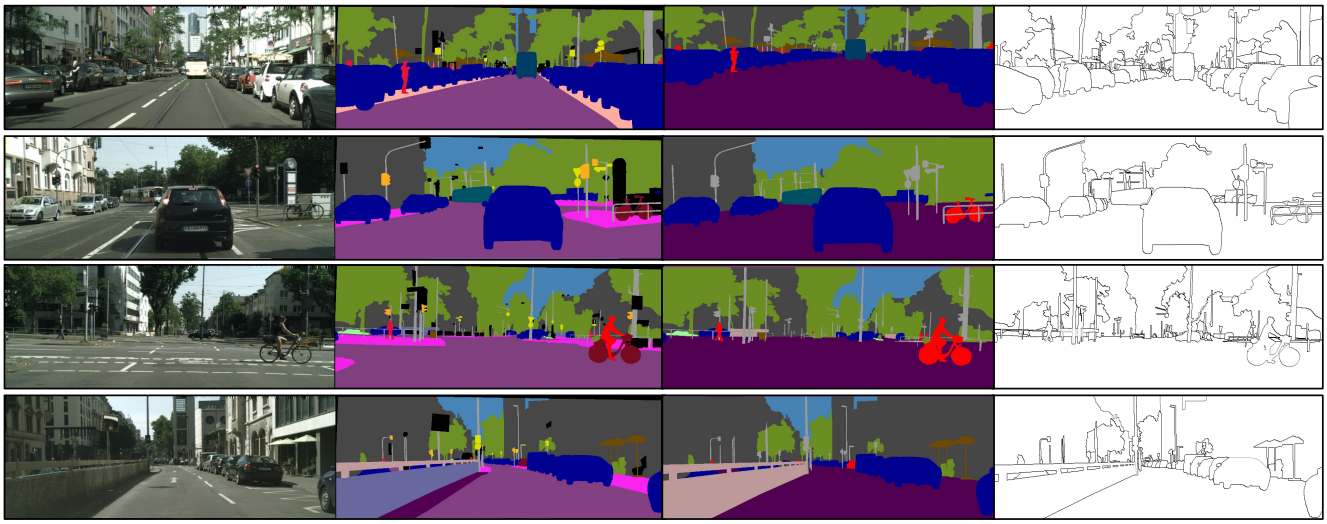

Figure 1: The process of geometrical edge ground truth generation. From left to right: RGB images, semantic segmentation ground truth, combined-category segmentation results, geometrical edge ground truth.

2. Inference between Eqn. 3 and Eqn. 4 / Eqn. 5

From Eqn. 3 to Eqn. 4. For any two points x_i x_j that lie on the same 3D surface S, the surface normal direction should be the same for the two points, which is constrained in Eqn. 4.

From Eqn. 3 to Eqn. 5. For three points p_i, p_j, p_k that lie on the 3D line, the gradient between any two points should be the same.

From Eqn. 5 to Eqn. 3. For any three points p_i, p_j, p_k, the gradients between p_i, p_j and p_j, p_k are the same. Assume the 3D line linking p_i, p_j and p_j, p_k are represented as:

$$a_1 x + b_1 y + c_1 z = 1$$
$$a_2 x + b_2 y + c_2 z = 1$$

The gradients are the same for the two lines, thus:

$$\frac{a_1}{c_1} = \frac{a_2}{c_2}, \frac{b_1}{c_1} = \frac{b_2}{c_2}$$

Considering that p_j lies on both lines, thus these two lines are identical. Thus Eqn. 3 and Eqn. 5 are mutually necessary and sufficient conditions.

3. Example outputs

LEGO jointly estimates depth, surface normal and geometrical edge. Some example results are shown in Fig. 2.

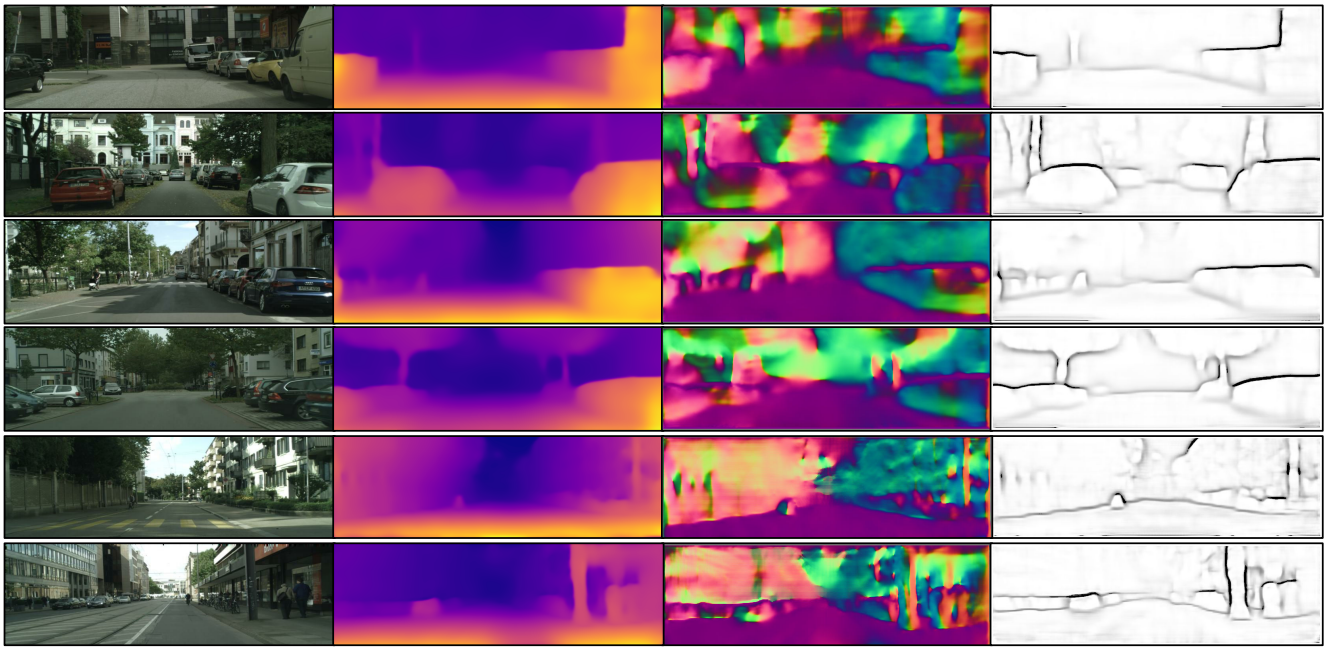

Figure 2: Example outputs of LEGO. From left to right: input image, predicted depth, predicted normal, predicted edge.

4. Comparison with previous methods

We provide visual comparison with both [2] and [1] in Fig. 3 (see next page). LEGO generates depth and normal results of better structure and preserves aligned object boundaries.

5. Qualitative results

Some qualitative results on Cityscapes dataset are shown in the attached video. Cityscapes dataset provides a 30-frame snippet around the key frames. We show 10 snippets in validation set from diverse scenes. The video of higher resolution is available at this link (https://youtu.be/40-GAgdUwI0)

References

[1] Z. Yang, P. Wang, W. Xu, L. Zhao, and N. Ram. Unsupervised learning of geometry from videos with edge-aware depth-normal consistency. In *arXiv preprint arXiv:1711.03665*, 2017. 2, 3

[2] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017. 2, 3

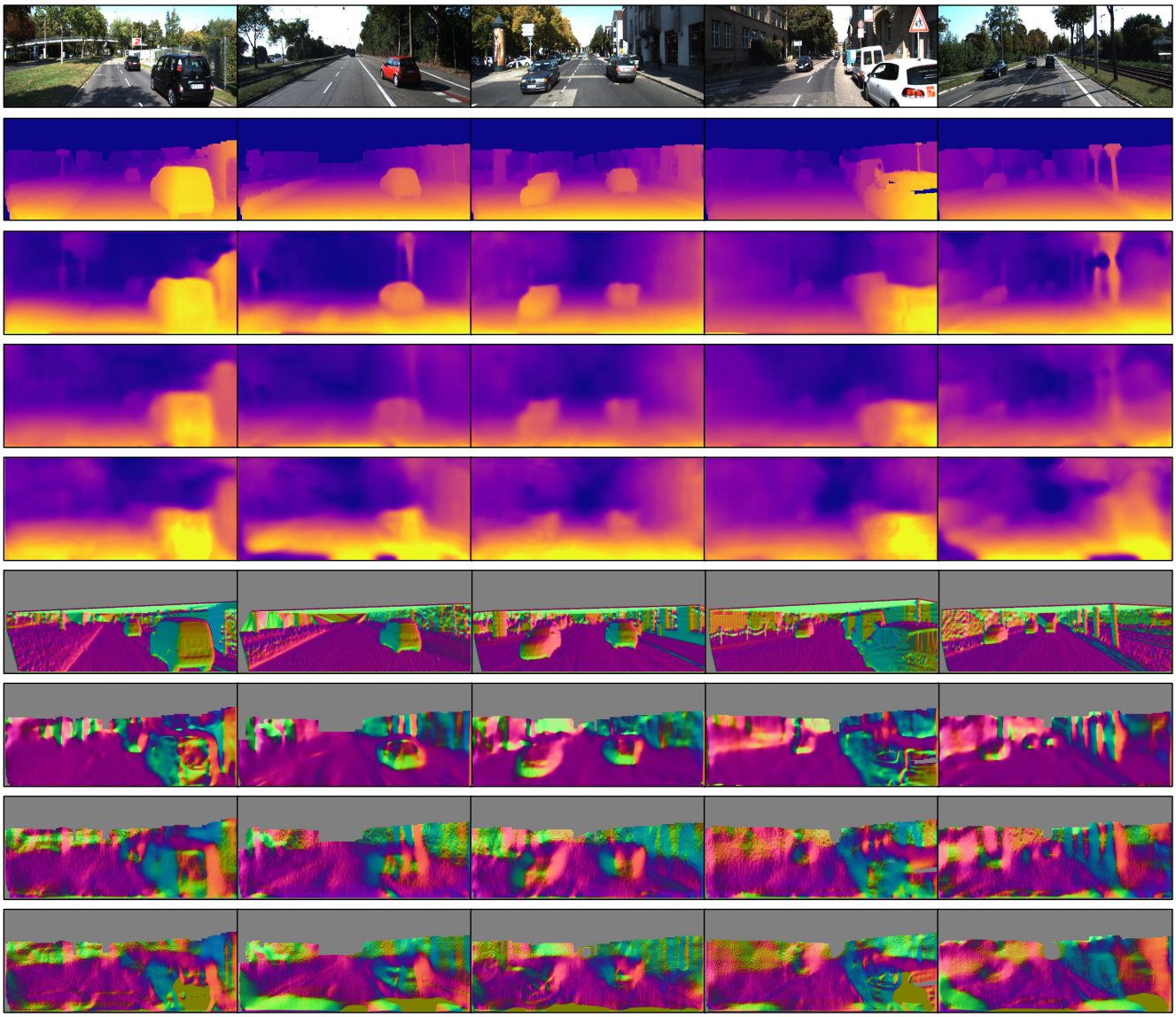

Figure 3: Visual results of depth and surface normal by different methods. From top to bottom: input image, depth ground truth, LEGO depth, depth by [1], depth by [2], normal ground truth, LEGO normals, normals by [1], normals by [2]