Pix2Point: Learning Outdoor 3D Using Sparse Point Clouds and Optimal Transport
Supplementary Material

This document aims to provide further details regarding the implementation and additional results to element discussed in the main paper. We also provide an ablation study of our proposed method.

1 Implementation Details

This section provides additional details regarding the implementation of loss functions and evaluation metrics.

1.1 Loss functions

Chamfer distance The chamfer distance is the average of squared euclidean distances to the nearest neighbour from one set to the other. It is defined between two point sets S_1 and S_2 as follows:

$$d_C(S_1, S_2) = \frac{1}{|S_1|} \sum_{x \in S_1} \Delta(x, S_2) + \frac{1}{|S_2|} \sum_{y \in S_2} \Delta(y, S_1) \tag{1}$$

with $\Delta(\,\cdot\,, S) = \min_{y \in S} \| \,\cdot\, - y\|_2^2$.

In our implementation, we performed the computation of the chamfer distance using the `pytorch-3d` python library [11].

Optimal transport or OT distance This distance allows to compare point sets distributions, therefore two 3D point clouds showing a rather small OT distance value reveal 3D distributions that are close to each other. It is defined straightforwardly for point sets with equal cardinality as:

$$d_{OT}(S_1, S_2) = \min_{\phi : S_1 \to S_2} \sum_{x \in S_1} \|x - \phi(x)\|_2^2 \tag{2}$$

where ϕ is a one-to-one mapping and can be generalised to handle cardinality-unbalanced point sets.
Solving (2) comes with a heavy computational cost, hence we consider the Sinkhorn divergence, an efficient regularised approximation [3, 8]. The computation were performed using the `geomloss` library [5]

1.2 Evaluation Metrics

Here we provide a formal definition of the performance criteria we are considering for the 3D reconstruction task.

Completeness is the coverage, in per cent, of the target point cloud by the predicted points. A target point is covered if a predicted point lies in its surrounding (*i.e.* fixed radius ball). We evaluate completeness values for radius of 50 cm, 25 cm and 10 cm.

Accuracy is the distance d, in meter, from the r-th percentile of the distances to the nearest neighbour, from the predicted point cloud to the ground-truth point cloud. It measures the greatest distance to the nearest neighbour among the predicted points closest to the ground truth. We choose $r < 90\%$ to include most of the points and discard eventual outliers.

Relative accuracy is similar to the accuracy, where every distance to their nearest neighbour is divided by the norm of the corresponding target point. It provides a higher penalty to short-range predictions.

Formally, we define P_j and T_j the predicted and target point clouds of the j-th scene respectively. Every point from one point cloud is provided with a nearest neighbour distance to the other point cloud. $\delta_i^{(j)} = \min_{y \in T_j} \|x_i - y\|$ where $x_i \in P_j$, is the distance from a predicted point to the closest target point. Reciprocally we note $\gamma_i^{(j)} = \min_{x \in P_j} \|x - y_i\|$ where $y_i \in T_j$. Also we provide relative distance $\delta_i^{rel(j)} = \frac{\min_{y \in T_j} \|x_i - y\|}{\|y\|}$.
Let $\Delta = \{\delta_i^{(j)}, \forall j, i\}$ and $\Gamma = \{\gamma_i^{(j)}, \forall j, i\}$, we define the proposed criteria as follows:

$$\begin{aligned} \textbf{Completeness} \quad & d \mapsto \tfrac{|\Gamma < d|}{|\Gamma|} \\ \textbf{Accuracy} \quad & r \mapsto d \text{ s.t. } \tfrac{|\Delta < d|}{|\Delta|} = r \\ \textbf{Rel. accuracy} \quad & r \mapsto d \text{ s.t. } \tfrac{|\Delta^{rel} < d|}{|\Delta^{rel}|} = r \end{aligned}$$

where d is the nearest neighbour distance threshold, $r \in]0, 1]$ and $|\,\cdot\,|$ denotes the cardinality.

2 Ablation Study

This section is dedicated to an ablation study of Pix2Point.

We present in Table 1 every increment and their respective performance from bottom to top. Starting with the prediction of N=2500 points using the PSGN approach [4] with the network parameters optimised to minimise either the chamfer distance or the OT distance. The last two rows show the performances of these approaches in terms of accuracy and completeness for several thresholds. PSGN minimising the chamfer distance provides better precision than PSGN-OT while showing equivalent or slightly better ground-

Table 1. Incremental performance evaluation from bottom to top. P2P refers to the proposed Pix2Point method, OT stands for Optimal Transport loss, and C for chamfer distance loss. N is the number of predicted 3D points

N	Increment	Complete. ↑(in %)			Accuracy ↓	
		50cm	25cm	10cm	in m	rel.
10k	P2P-ResNet-OT	**71.3**	**48.8**	**15.1**	1.92	0.18
	P2P-VGG-OT	67.4	47.7	14.7	1.79	0.19
	P2P-VGG-C	64.4	36.0	8.0	**0.85**	**0.05**
	DensePCR [7]	59.9	23.5	3.5	*1.77*	0.18
2.5k	PSGN-OT[4]	61.53	20.79	1.72	2.47	0.36
	PSGN-C[4]	60.19	32.98	6.09	0.97	0.06

truth completeness for the chosen threshold. However current completeness thresholds do not highlight the OT benefits as PSGN-OT prevail over PSGN-C for greater completeness threshold values. To overcome the poor number of points predicted by the PSGN unit, a PointNet-like [9, 10] module already trained to map from 2500 element point cloud to a corresponding 10000 element point cloud is added. This densification module is trained by minimisation of the chamfer distance in a supervised fashion to augment the number of points by a given factor. We report in Table 1 the DensePCR-like approach [7] that includes PSGN-OT followed by the densification module. A moderate gain is noticeable for small completeness threshold values and the accuracy has further improved as the number of predicted points per scene increases. Instead of training each unit individually, we propose to learn the parameters of both units in an end-to-end fashion. First, by using a VGG backbone and minimising the chamfer distance, referred to as P2P-VGG-C in Table 1, we obtain significantly better completeness, as well as good absolute and relative accuracy, which is expected as the minimisation of the chamfer distance implies a low value for accuracy. Alternately, the same architecture is optimised to minimise the OT distance. This model referred to as P2P-VGG-OT, reveals better completeness while preserving the same accuracy as DensePCR. Our approach achieves moderately better completeness performances by replacing the VGG backbone with a ResNet Backbone, P2P-ResNet-OT.

3 Data Processing

3.1 Data Augmentation

Geometric: KITTI scenes exhibit chirality, hence, usual image flipping augmentation during training would highly deteriorate predictions at test time.

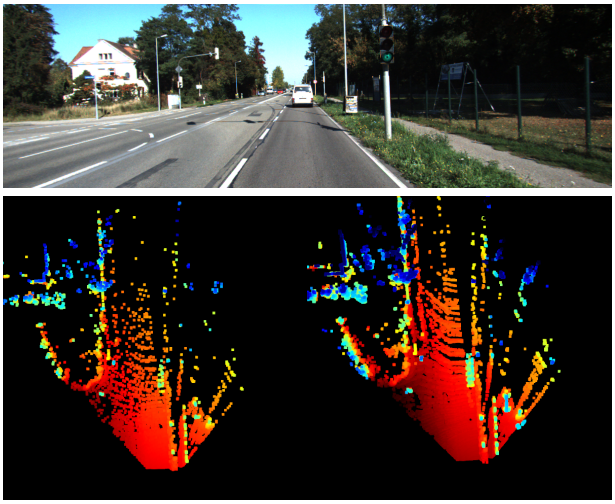

Figure 1. Clockwise: RGB image of a scene (top), bird's eye view of the corresponding full reference point cloud (right) and its 10K points subsampling used as training target (left). Blue denotes highest points of the scene while red denotes lowest points near the ground

Therefore geometric augmentation was not used.

Colour: Pix2Point architecture comprise instance normalisation layers that would demean usual colour transformations, therefore we did not consider colour augmentation.

3.2 Point cloud sub-sampling

We constructed the 10k points training dataset by randomly choosing 10k non-zeros pixels in corresponding KITTI depth maps. By doing so, training point clouds are built to match the full reference point cloud spatial distribution. Fig. 1 shows that resulting target point clouds and reference point clouds share the same spatial distribution of points.

4 Additional Results

In this section, we display predictions on more scenes for AdaBins [1], BTS [6], Pix2Point-VGG-OT, Pix2Point-VGG-C and Pix2Point-ResNet-OT in Figure 2. As stated in the main paper, these figures show the Pix2Point architectures achieve a better coverage by predicting points further than AdaBins and BTS, especially for OT variants. They also display lower error globally.

References

[1] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[2] Lucas Caccia, Herke van Hoof, Aaron Courville, and Joelle Pineau. Deep generative modeling of LiDAR data. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[3] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NIPS*, 2013.

[4] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017.

[5] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

[6] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From Big to Small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

[7] Priyanka Mandikal and Venkatesh Babu Radhakrishnan. Dense 3D point cloud reconstruction using a deep pyramid network. In *2019 Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019.

[8] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 2019.

[9] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2017.

[10] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems 30*, 2017.

[11] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

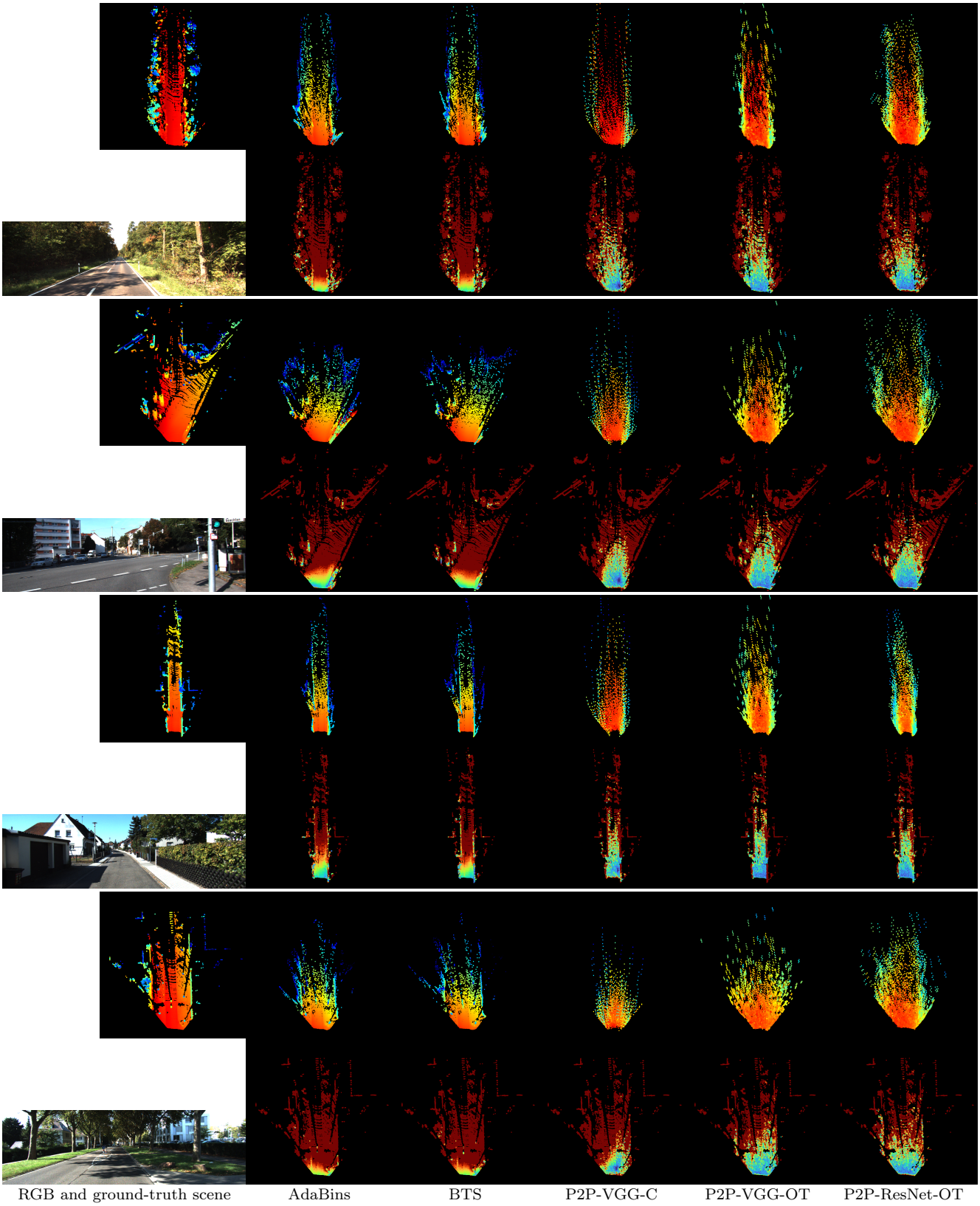

Figure 2. For each scene, first row: 3D ground-truth and predictions for the RGB image according to AdaBins [1], BTS [6], our Pix2Point VGG-chamfer, VGG-OT and ResNet-OT, all trained on 10k points. 3D representation following [2]. Bird's eye view where the colour encodes the altitude. Second row: the input RGB image and the ground-truth-to-prediction error map for each method. errors are from 0 (blue) to 50cm (red).