2021 The 21st International Conference on Control, Automation and Systems (ICCAS 2021)
Ramada Plaza Hotel, Jeju, Korea, Oct. 12∼15, 2021

GBNet: Gradient Boosting Network for Monocular Depth Estimation

Daechan Han and Yukyung Choi†

The Robotics and Computer Vision Lab., Sejong University, South Korea
dchan@rcv.sejong.ac.kr, ykchoi@sejong.ac.kr

Abstract: Recently, neural networks have shown promising results in estimating depth from a single image. A large amount of per-pixel ground truth depth data is required to train the neural network in supervised learning. However, the dense depth data of ground truth is challenging to collect in realistic dynamic environments. To solve this problem, many researchers propose self- and semi-supervised learning as a credible alternative. This paper proposes a novel self- and semi-supervised monocular depth estimation method, inspired by the gradient boosting method. The existing gradient boosting method provides training to several sequential, additive, and gradual models for minimizing the error. Similarly, we design our proposed network to refine the predicted depth map sequentially and gradually generate a high-quality depth map via multi-stack CNN structures. Our method shows the state-of-the-art results for monocular depth estimation on a DDAD (Dense Depth for Autonomous Driving) dataset.

Keywords: Monocular Depth Estimation, Self-Supervised Learning, Semi-Supervised Learning

1. INTRODUCTION

Depth information is a crucial component in computer vision and robotics society with potential applications such as autonomous driving and drone technology. However, it is not easy to obtain dense depth information in the outdoor environment. In the case of LiDAR, a representative 3D sensor for autonomous vehicles, it can provide accurate depth information, but sensing data cannot cover the entire area and can only provide sparse information. Recently, to solve this limitation, various deep learning-based methods have been proposed, and attempts to provide high-quality and dense depth information are increasing. As an example of that, various approaches such as monocular-, binocular-, and multi-view stereo-based methods have been proposed. In this paper, we only focus on monocular depth estimation, which could 1) produce dense depth maps and 2) replaces the expensive sensor such as LiDAR.

Supervised learning method in monocular depth estimation has shown successful results on the various benchmark [1][3][12]. This learning method learns distance to object by direct comparing ground truth (dense depth map) to an inferred depth map. Therefore, supervised learning requires a significant amount of high-quality depth maps to achieve good performance. For overcoming these problems, *self- and semi-supervised methods* have been introduced that do not require dense depth during training. The self-supervised approach [5][6] uses geometric relationships between sequence or stereo images to train the network that predicts the dense map. The semi-supervised method [7][8] uses both the self-supervised method and the supervised method, and for supervised learning, a sparse LiDAR is used instead of a dense depth map. This semi-supervised method has both the advantages of 1) supervised learning that can use physically very accurate and sparse depth information during training and 2) self-supervised learning that accurate dense depth map can be used during training, resulting in significant performance improvement.

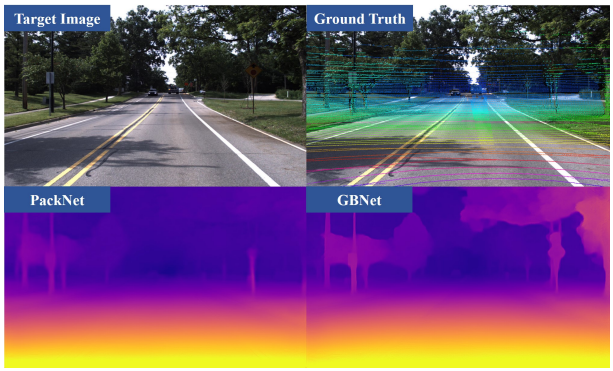

Fig. 1.: **Result of monocular depth estimation.** Our proposed network, GBNet, produced sharp and high quality depth map.

In this paper, we propose a novel self- and semi-supervised depth estimation method, called Gradient Boosting Network (GBNet), that leads to performance improvement. The GBNet is a coarse to fine network for estimating high-quality depth maps, inspired by the gradient boosting method. This method sums each inferred depth map to learn how to assemble the correct depth pixels. Finally, the proposed method has the advantage of being compatible with the existing monocular depth estimation methods.

The remainder of this paper is organized as follows. In Section 2., we present a novel architecture and loss function to which self- or semi-supervised learning methods can be applied. Comparisons and analysis results with the state-of-the-art methods are described in Section 3.. Finally, in Section 4., we conclude our paper with a brief discussion.

† Corresponding author
This work won 1_{st} and 3_{rd} place at CVPR2021-Dense Depth for Autonomous Driving (DDAD) challenge.
https://github.com/sejong-rcv/GBNet

978-89-93215-21-2/21/$31.00 ©ICROS

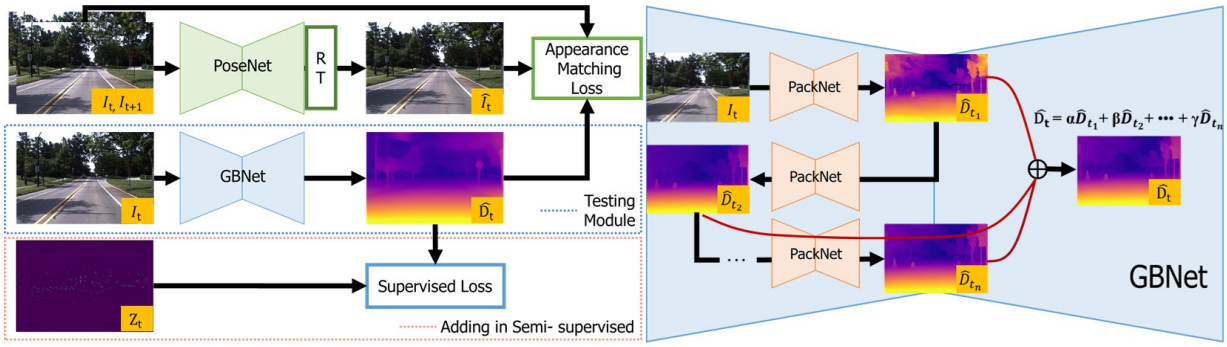

Fig. 2.: **GBNet-SfM & GBNet** (a) Our proposed self- and semi-supervised monocular SFM architecture. (b) Our proposed monocular depth network (GBNet).

2. GRADIENT BOOSTING NETWORK

This section describes our self- and semi-supervised monocular GBNet-SfM system (Fig. 2-(a)) that takes a target images I_t and infers a depth $\hat{D}_t$. We aim to learn models as follow: (*i*) a monocular depth estimation model $G(I) = D$ that estimates the well refined depth; (ii) a monocular ego-motion estimator $E(I_t, I_{t+1}) = e_{t\to t+1}$, that estimates the relative pose $e_{t\to t'}$ of each source image $I_{t+1} \in [I_t, I_{t+1}]$, with respect to the I_t's pose. We train (i) and (ii) models with self- and semi- supervised learning methods which are not used dense depth maps as the role of supervision. We first explain our proposed depth estimation network and then training loss for self- and semi-supervised learning separately.

2.1 Proposed Architecture

Existing monocular depth estimation methods are focusing a single generator on enhancing performance. However, using only one generator model can not refine the output. To solve this problem, we propose a novel architecture, called GBNet (Fig. 2-(b)), that is inspired by the gradient boosting method. This method combines weak models into a single strong learner in an iterative fashion. It builds the model in a stage-wise fashion, and it reduces prediction errors when blended with previous ones as stages progress. Similar to the gradient boosting method, we design a hierarchical and residual network to refine $\hat{D}_t$. Our proposed GBNet $G(I)$ gradually minimizes the depth error every time it passes each single network P. The single network is based on PackNet [4], *i.e.* encoder-decoder architecture, with skip connections, preserving and recovering important spatial information. We construct $G(I)$ into N single network. The first single network P_1 use I_t as input and the other single networks P_i, $i \in (2, N)$ are worked with $\hat{D}_{t(i-1)}$. To leverage the abundant information of all images and make them complementary relationships, we finally estimate $\hat{D}_t$ as follow:

$$\hat{D}_t = \sum_{i=1}^{N} \lambda_i \hat{D}_{ti}, \tag{1}$$

where, λ is a static value that increases the effect of more refined depth maps by adjusting the influence of each single network. Z is the ground truth of LiDAR to estimate the supervised loss.

2.2 Self-Supervised Objective

Following the work of Monodepth2 [5], we simultaneously train the depth and pose models. The overall self-supervised constraints consist of an appearance matching loss term L^p_{sl} to make the synthesized target image $\hat{I}_{t+1}$ and the target image I_t and a depth smoothness loss term L^s_{sl} to encourage estimated depth $\hat{D}_t$ to be locally smooth.

Appearance Matching Loss We use a combination of an L1 distance and the Structural Similarity (SSIM) [10] term to aim to increase the pixel-level similarity between the target image I_t and the synthesized image $\hat{I}_{t+1}$, by Eq. 2.

$$L^p_{sl}(I_t, \hat{I}_{t+1}) = \alpha \frac{1 - SSIM(I_t, \hat{I}_{t+1})}{2} + (1-\alpha)\|I_t - \hat{I}_{t+1}\| \tag{2}$$

Eq. 2 is a robust learning method for self-supervision typically. However, the error of parallax in the scene makes the out-of-view and occluded pixels. It causes an undesirable effect incurred to the learning. We use the per-pixel minimum re-projection loss to solve the out-of-view pixels and occluded pixels problems, as shown in Monodepth2 [5] It alleviates the undesirable problems by calculating the minimum loss per pixel for each source image I_{t+1}. It means that the same pixel is not out-of-view and occluded in the synthesized target images.

$$L^p_{sl}(I_t, I_{t+1}) = \min L^p_{sl}(I_t, \hat{I}_{t+1}) \tag{3}$$

We also apply the auto-masking static pixels methods suggested in Monodepth2 [5]. Due to the static pixels have a minor appearance matching loss and can make an infinite depth hole when assuming no ego-motion between frames, we use auto-masking to ignore the static pixels. We find the pixels that satisfy having $L_{sf}(I_t, I_{t+1}$ higher than $L_{sf}(I_t, \hat{I}_{t+1})$ in order to produce a mask.

$$M = \min L^p_{sl}(I_t, I_{t+1}) > \min L^p_{sl}(I_t, \hat{I}_{t+1}) \tag{4}$$

Table 1.: **Quantitative performance comparison on the DDAD dataset**

Type	Method	Abs Rel ↓	Sqr Rel ↓	RMSE ↓	RMSE log ↓	$\delta_{1.25}$↑	$\delta_{1.25^2}$↑	$\delta_{1.25^3}$↑
self-supervised	PackNet	0.175	5.002	15.326	0.263	0.793	0.921	0.962
	GBNet$_{total}$	0.148	3.329	14.471	0.244	0.818	0.930	0.967
semi-supervised	PackNet	0.122	2.477	13.200	0.220	0.849	0.941	0.973
	GBNet$_{total}$	0.124	2.476	13.276	0.220	0.846	0.940	0.973

Table 2.: **Ablation study on the GBNet architecture**

Type	Method	Abs Rel ↓	Sqr Rel ↓	RMSE ↓	RMSE log ↓	$\delta_{1.25}$↑	$\delta_{1.25^2}$↑	$\delta_{1.25^3}$↑
self-supervised	GBNet$_1$	0.254	6.055	17.582	0.331	0.606	0.871	0.950
	GBNet$_2$	0.150	3.394	14.661	0.246	0.812	0.928	0.967
	GBNet$_3$	0.164	3.467	15.148	0.256	0.786	0.923	0.966
	GBNet$_{total}$	0.148	3.329	14.471	0.244	0.818	0.930	0.967
semi-supervised	GBNet$_1$	0.191	4.411	16.757	0.275	0.740	0.913	0.962
	GBNet$_2$	0.130	2.609	13.335	0.223	0.841	0.939	0.973
	GBNet$_3$	0.143	2.855	14.460	0.238	0.817	0.929	0.969
	GBNet$_{total}$	0.124	2.476	13.276	0.220	0.846	0.940	0.973

Depth Smoothness Loss As suggested in Monodepth [5], the depth smoothness loss penalizes depth discontinuity in texture-less low-image gradient regions. We apply depth smoothness to our constraints.

$$L_{sl}^{s} = |\delta_x \hat{D}_t| e^{-|\delta_x I_t|} + |\delta_y \hat{D}_t| e^{-|\delta_y I_t|} \quad (5)$$

The self-supervised loss throughout the process is as follows:

$$L_{sl} = L_{sl}^{p}(I_t, I_{t+1}) \odot M + 0.001 * L_{sl}^{s} \quad (6)$$

where, $\odot$ denotes element-wise multiplication.

2.3 Semi-Supervised Objective

Similar to Kuznietsov *et al.* [8], we use supervised learning in semi-supervised objective to provide accurate depth information to the networks. To lead estimating more detailed than self-supervised objective, the LiDAR data Z which include sparse depth information is the ground truth of supervised loss. The constraints of supervised learning measure the deviation of the inferred depth map from the available ground truth at the pixels.

$$L_{sp}(I_t, Z_t) = \|G(I_t) - z_t\|, z_t = Z_t > 0 \quad (7)$$

We formulate a total semi-supervised loss (L_{sm}) function that incorporates supervised (L_{sp}) and self-supervised (L_{sl}) objective as follow:

$$L_{sm}(I_t, I_{t+1}, Z_t, Z_{t+1}) = \gamma L_{sp}(I_t, Z_t) + \beta L_{sl}(I_t, I_{t+1}) \quad (8)$$

where, $\gamma = 0.9$ and $\beta = 0.1$.

3. EXPERIMENTS

3.1 Datasets

We experiment with our proposed method on DDAD (Dense Depth for Autonomous Driving) datasets [4], which is the more realistic and challenging benchmark. This dataset composes a diverse scene of urban, highway, and residential, and it contains 12,350 images for training, 3,950 for validation, and 3,085 for evaluation color frames with the LiDAR of ground-truth depth maps. We evaluate our method with validation set at the CVPR 2021 Dense Depth for Autonomous Driving challenge.

3.2 Implementation Details

We implement our proposed model in PyTorch with all models trained across 8 Titan V100 GPUs. We use the Adam optimizer [11] with β_1=0.9 and β_2=0.999. We initialize our single depth generators P and monocular ego-motion estimator $E(I_t, I_{t+1})$ with PackNet weights pre-trained for KITTI [2] depth estimation. In addition, our proposed model is trained for 10 epochs, with a batch size of 4 in the DDAD dataset. We set the number of single network to $N = 3$ and, the SSIM weight to $\alpha = 0.85$.

3.3 Evaluation Metrics

We measure the accuracy of our proposed method in depth prediction using the 3D LiDAR ground truth on the test images. We follow depth evaluation metrics used by Eigen *et al.* [9].

Abs Rel: $\frac{1}{|T|}\sum_{y \in T} |y - y^*|/y^*$

Sqr Rel: $\frac{1}{|T|}\sum_{y \in T} \|y - y^*\|^2/y^*$

RMSE: $\sqrt{\frac{1}{T}\sum_{i=1}^{T} \|\rho(x_i)^{-1} - Z(x_i)\|}$

RMSE log: $\sqrt{\frac{1}{T}\sum_{i=1}^{T} \|log(\rho(x_i)^{-1}) - log(Z(x_i))\|}$

Threhold: $\%\ of\ y_i\ s.t.\ max(\frac{y_i}{y_i^*}, \frac{y_i^*}{y_i}) = \delta < thr$

where, T is the number of pixels with ground-truth in the test set.

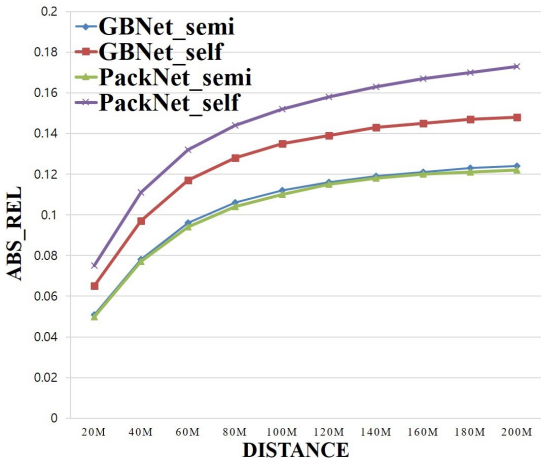

Fig. 3.: **Absolute relative error (ABS_REL) comparison for each depth scale.**

3.4 Ablation Study

We analyze the effect of every single network in our proposed recursive network in Table 2. To demonstrate the progressive minimization of the depth error of our methodology, we evaluate the inferred depth maps from every single network preceding the ensemble method in GBNet. The GBNet$_i$, i $\in$ $(1, 3)$ is the single network that the estimated output is $\hat{D}_{ti}$. In results of self-supervised, GBNet$_2$ achieves higher depth map prediction (2.921 m in RMSE) than GBNet$_1$, and also is more accruate depth (3.422 m in RMSE) in semi-supervised. It means that adding the single network performance can be boosted in overall learning methods. The ensemble of every single network improve entire performance (average 1.574m in RMSE).

3.5 Depth Estimation Performance

In Table 1, we present the performance of our proposed method for the self- and semi-supervised monocular depth estimation on the DDAD datasets. We separately evaluate our proposed method by self- and semi-supervised learning and compare it with our based network PackNet. The results show that our monocular method outperforms the PackNet significantly in self-supervised for all metrics. When evaluating in semi-supervised, our methods are 0.001 more accurate in Sqr Rel than the results reported by PackNet. The benefits of GBNet are larger than semi-supervised in self-supervised which LiDAR as a source of supervision is not used. We illustrate the qualitative results of comparing methods in Fig. 4. In overall scene, our monocular method appear more detailed and sharper than PackNet. It can be seen that more accurate depth is predicted even compared the ground-truth to methods. In addition, We analyze the performance of each method according to different depth intervals in Fig. 3. It shows that the performance of our self-supervised GBNet is higher than self-supervised PackNet in all of the depth intervals, the performance gap becoming more apparent as the distance increases.

4. CONCLUSION

We present a novel architecture GBNet for self- and semi-supervised monocular depth estimation, achieving state-of-the-art performance. Our work leverages multi-stack CNN networks to reduce error in the predicted depth map sequentially and gradually. As a result of the analysis, our proposed network provides detailed and sharp results of distant objects than the baseline method PackNet. We show that our results are superior in both quantitative and qualitative evaluations of the DDAD dataset, based on this result, we won 1st (semi-supervised track) and 3rd (self-supervised track) place at CVPR 2021-Dense Depth for Autonomous Driving challenge.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2020M3F6A1109603) and Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT). (2021-0-02067).

REFERENCES

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (CVPR), 2012

[2] Andreas Geiger, Philip Lenz, Christoph Stiller and Raquel Urtasun. "Vision meets robotics: The kitti dataset," in *The international journal of robotics research* (IJRR), vol. 32, no. 11, pp. 1231–1237, 2013.

[3] Daniel Scharstein, et al. "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proceedings of the german conference on pattern recognition* (GCPR), 2014.

[4] Vitor Guizilini et al. "3D Packing for Self-Supervised Monocular Depth Estimation," in *Proceeding of international conference on computer vision and pattern recognition* (CVPR), 2020.

[5] Godard, et al. "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE international conference on computer vision* (ICCV), 2019.

[6] Tinghui Zhou, et al. "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (CVPR), 2017.

[7] Vitor Guizilini, et al. "Robust semi-supervised monocular depth estimation with reprojected distances," in *Proceedings of the conference on robot learning,* (CoRL), 2020.

[8] Yevhen Kuznietsov, Jorg Stuckler, and Bastian Leibe. "Semi-Supervised Deep Learning for Monocular Depth Map Prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (CVPR), 2017.

[9] David Eigen, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network," in *arXiv preprint arXiv:1406.2283*, 2014.

[10] Zhou Wang, et al. "Image quality assessment: from error visibility to structural similarity," in *IEEE transactions on image processing* (TIP), vol. 13, no. 4, pp 600-612, 2004.

[11] Kingma, Diederik P, and Jimmy Ba. "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.

[12] Menze Moritz and Andreas Geiger. "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (CVPR), 2015.

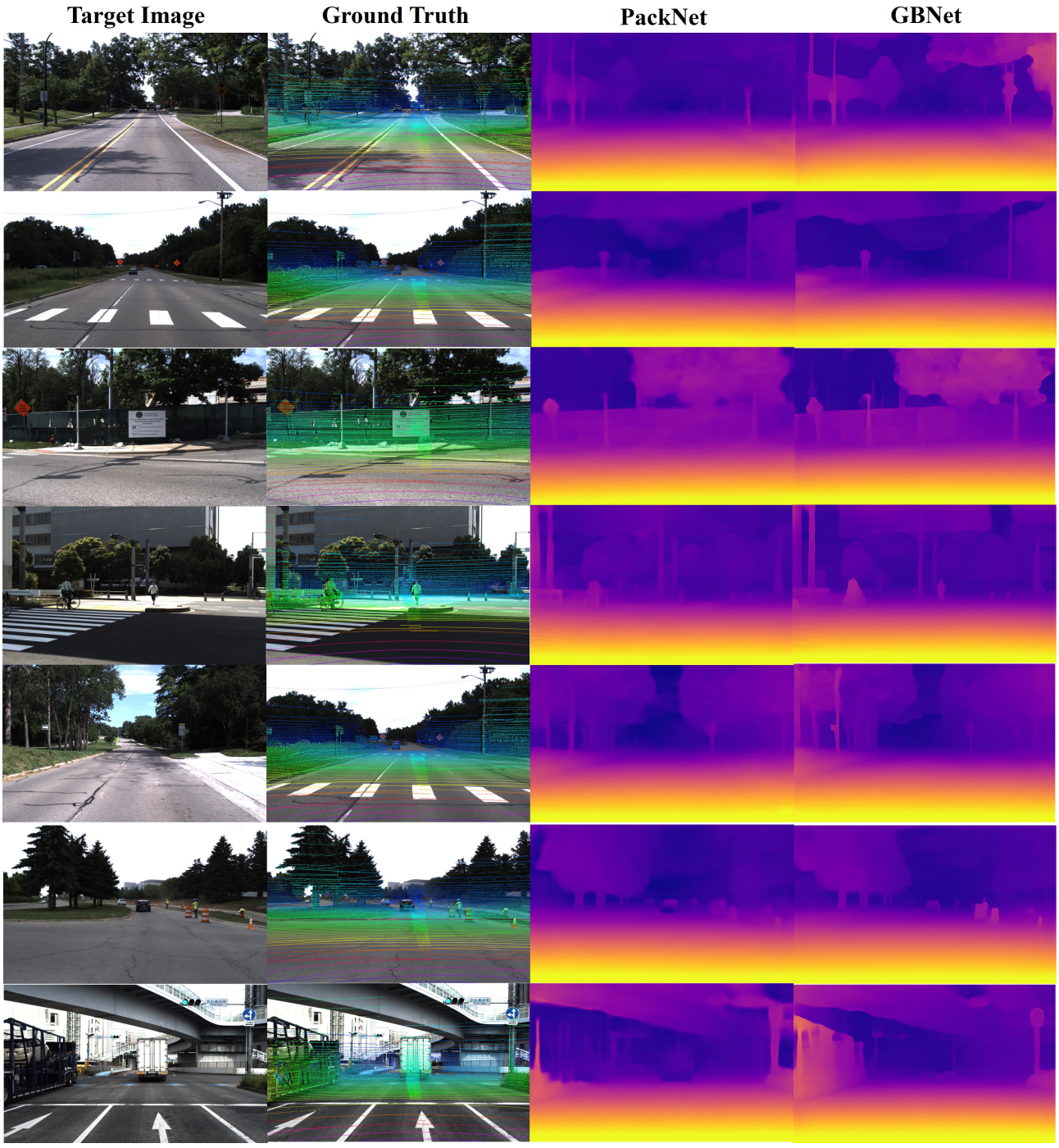

Fig. 4.: **Qualitative Results of the Monocular Depth Estimation.** Each column indicates (a) RGB image (b) 3D LiDAR (c) Comparison method called PackNet (d) Proposed method called GBNet.