ROBUST LIGHTWEIGHT DEPTH ESTIMATION MODEL VIA DATA-FREE DISTILLATION

Zihan Gao[1] *Peng Gao*[2] *Wei Yin*[3] *Yifan Liu*[4] *Zengchang Qin*[1]

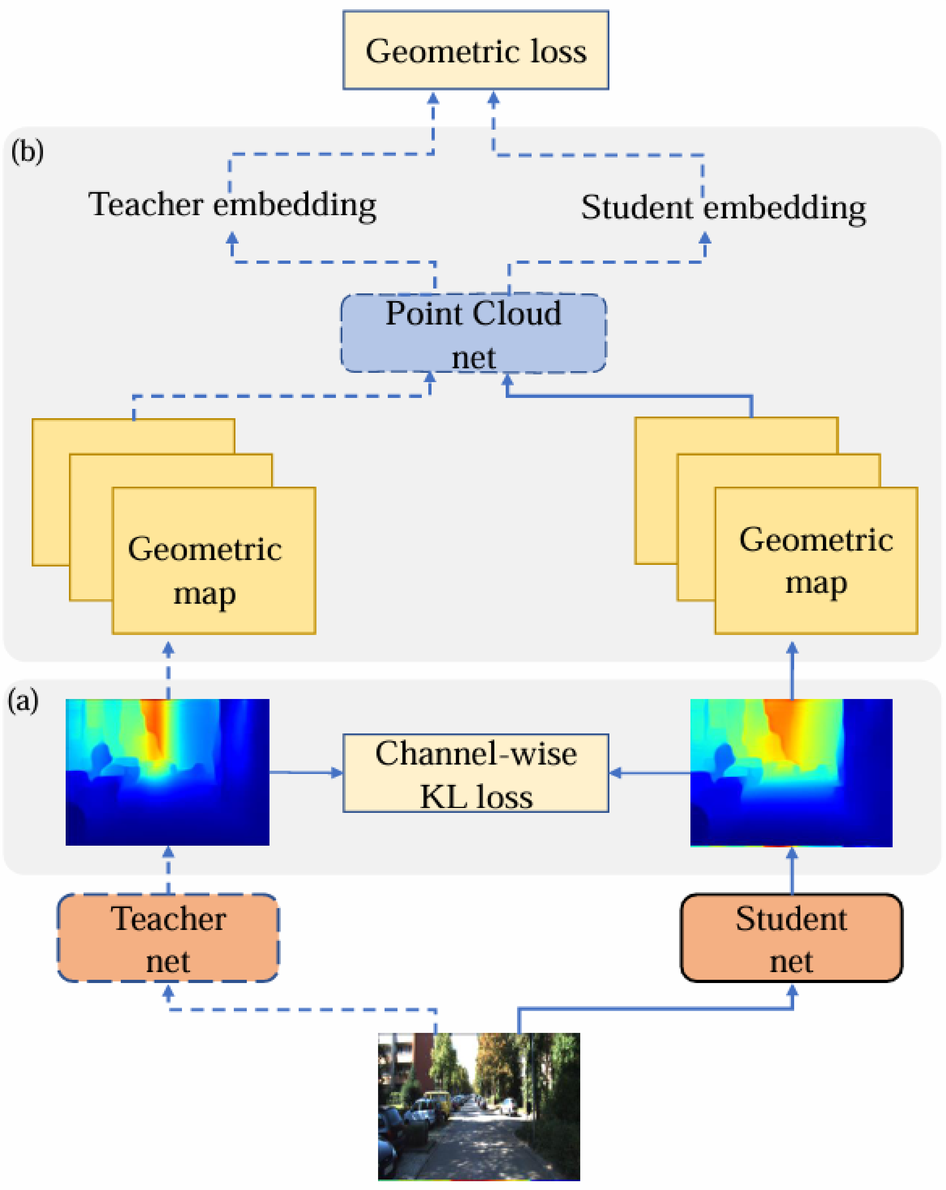
[1]Beihang University [2]Hangzhou Dianzi University

[3]Shenzhen DJI Sciences and Technologies Ltd [4]The University of Adelaide

Background

- Existing Monocular depth estimation methods often use large and complex neural networks. Despite the advanced performance of these methods, we consider the efficiency and generalization for practical applications with limited resources.
- Previous distillation methods tend to improve the accuracy in specific domains. We explore the domain-agnostic generalization ability transfer by using the knowledge distillation on depth estimation task.
- The labeled depth is hard and costly to obtain compared to RGB data. For some downstream tasks, only RGB images are accessible. So we consider label-free distillation method.

Label-free Distillation Method



(a) **Channel-wise Distillation.** Denote the teacher and student networks as T and S, the predicted depth map from T and S are d^T and d^S, first normalize the depth and then compute KL divergency loss.

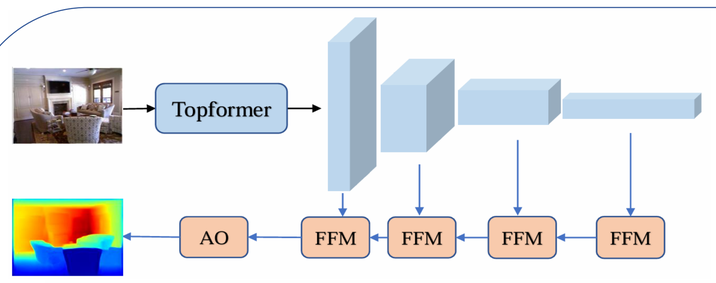
$$\phi(d_i) = \frac{\exp(\frac{d_i}{\mathcal{T}})}{\sum_{i=1}^{W\cdot H}\exp(\frac{d_i}{\mathcal{T}})} \qquad \ell_{cwd}(d^T, d^S) = \mathcal{T}^2 \sum_{i=1}^{W\cdot H} \phi(d_i^T)\cdot\log\left[\frac{\phi(d_i^T)}{\phi(d_i^S)}\right]$$

(b) **Geometric Distillation.** denote the teacher and student networks as T and S, the predicted depth map from T and S are d^T and d^s, first convert the 2D depth map into 3D position map, then use PointNet to extract features (input and feature), then compute the MSE loss.

$$Z = D, \quad X = \frac{(u-u_0)Z}{f_x}, \quad Y = \frac{(v-v_0)Z}{f_y} \qquad \ell_{geod} = \ell_{input} + \ell_{feature}$$

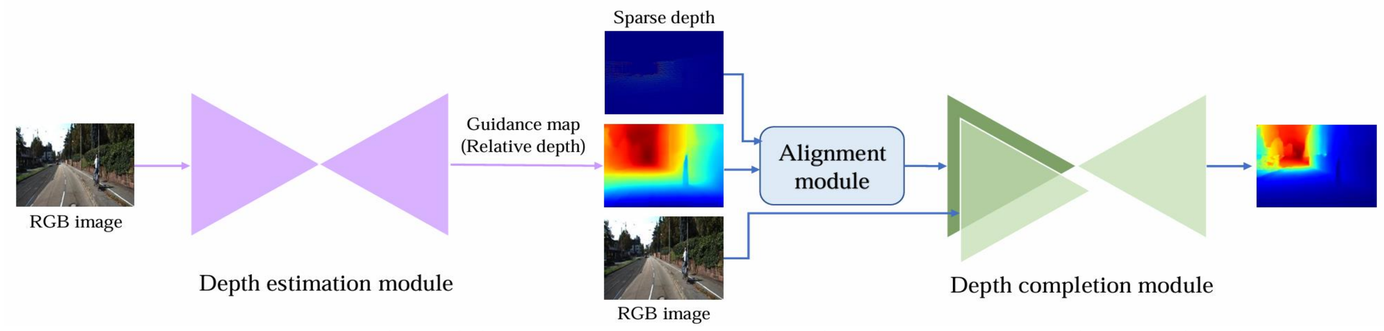
Conclusion

- We propose a robust lightweight monocular depth estimation network, which is available to apply in practical applications. We show that our proposed lightweight depth estimation model can generalize well to unseen zero-shot datasets.
- Considering the limitation of labeled data, we proposed a geometric label-free distillation method to further improve the generalization performance of our lightweight model in unlabeled domains.
- We apply our lightweight network to a two-stage depth completion task. Our method shows superior cross-domain generalization ability against state-of-the-art depth completion methods.



Model Architecture

Lightweight Depth estimation Model takes a RGB image as input and it outputs a relative depth**.** Encoder is Token Pyramid Vision Transformer (Topformer) to extract multi-scale features. Decoder is Multi-layer Feature Fusion Module (FFM) to fuse features from encoder progressively.



Depth Completion Model consists of three components. Depth estimation module predicts relative depth from a single RGB, alignment module aligns the relative depth with the sparse depth up to the metric scale and shift, depth completion module outputs a dense depth with the guidance of RGB image and aligned depth.

Results and Analysis

Method		RMSE↓	MAE↓	AbsRel↓	δ_1 ↑	δ_2 ↑
Teacher		0.0440	0.0199	0.0834	92.31	98.20
Student		0.0561	0.0297	0.1504	79.30	92.36
Logits	Feature	–	–	–	–	–
MSE	–	0.0532	0.0276	0.1370	81.64	94.16
CN+MSE	–	0.0562	0.0298	0.1516	78.96	92.16
CN+KL	–	0.0521	0.0269	0.1333	82.31	94.41
MSE	CN+KL	0.0530	0.0276	0.1372	81.66	94.27
MSE	CN+MSE	0.0534	0.0278	0.1399	80.95	93.70
MSE	SN+KL	0.0530	0.0274	0.1353	82.12	94.36
MSE	PWD	0.0529	0.0273	0.1353	82.00	94.42
CN+KL	PWD	**0.0520**	**0.0266**	**0.1306**	**82.95**	**94.83**
Ours		**0.0465**	**0.0210**	**0.0879**	**91.70**	**97.91**

Model	FLOPs(G)	Params(M)
Teacher	96.14	52.13
Student	20.92	11.49

FLOPs and Params of Teacher and Student model is shown in the **above** Table. The FLOPs are calculated with a fixed input resolution of 448×448.

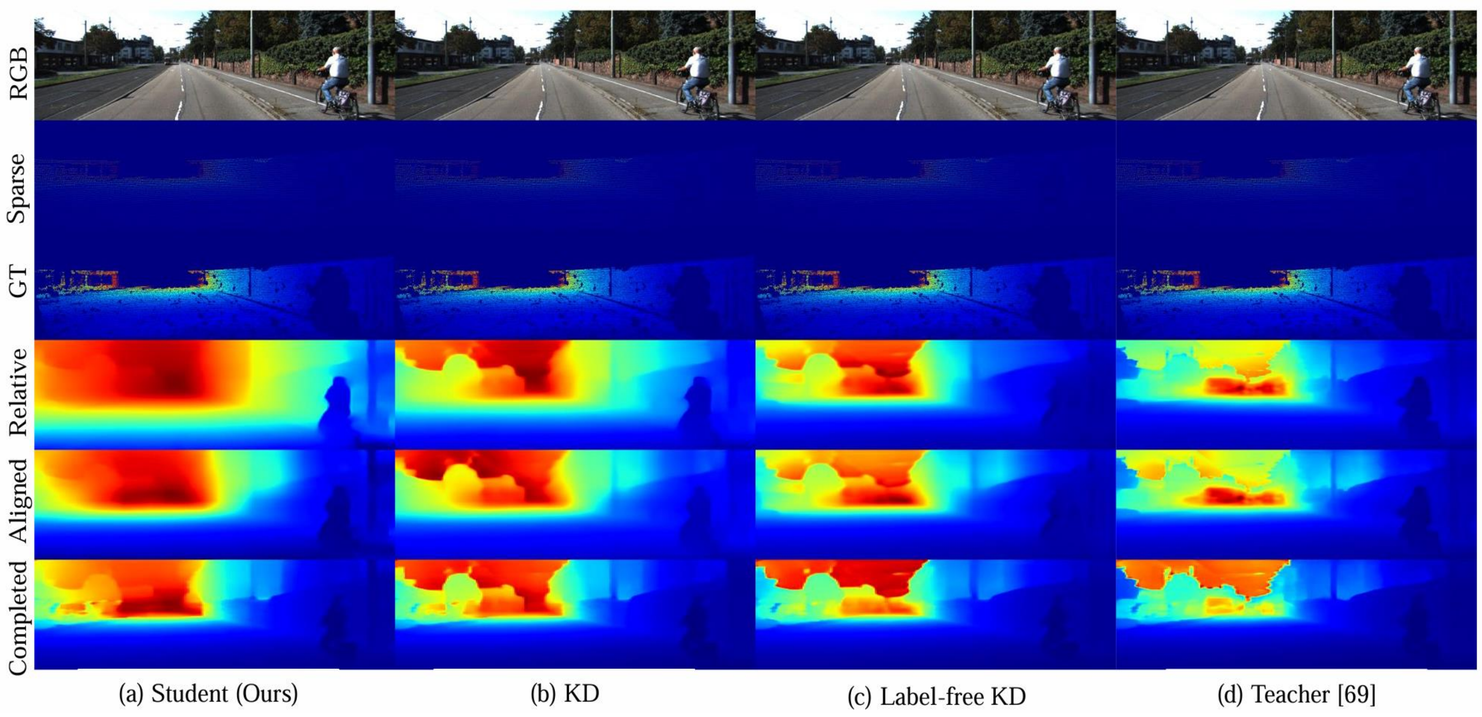
Ablation study of multiple knowledge distillation methods on KITTI test dataset is shown in the **left** Table . The labeled data is not seen in our method, while other distillation methods exploit labeled data for additional supervision. Except for CN+MSE logits distillation method, all other distillation methods can improve the generalization performance of our lightweight model by learning large model's knowledge and our proposed label-free distillation method achieves best performance.

Methods	NYUv2				DIODE			KITTI				ScanNet		
	RMSE↓	AbsRel↓	δ_1 ↑	δ_2 ↑	δ_1 ↑	δ_2 ↑	δ_3 ↑	RMSE↓	MAE↓	AbsRel↓	δ_2 ↑	δ_1 ↑	δ_2 ↑	δ_3 ↑
Yin *et al.* [7]	0.0454	0.0562	95.45	99.08	78.43	89.68	94.61	0.0440	0.0199	0.0834	98.20	97.26	99.61	99.92
Ours	0.0436	0.0564	95.68	99.28	77.23	89.35	94.54	0.0561	0.0297	0.1504	92.36	96.76	99.55	99.92

Monocular Depth Estimation results of lightweight and cumbersome models on zero-shot test datasets is shown in the **above** Table. On NYUv2, our lightweight model slightly outperforms the cumbersome one. On DIODE and ScanNet, the lightweight and cumbersome models are comparable with minimal differences. Our lightweight model show comparable performance compared to the cumbersome model on zero-shot test while maintaining high efficiency.

Method	NYUv2			DIODE			KITTI			ScanNet		
	δ_1 ↑	δ_2 ↑	δ_3 ↑	δ_1 ↑	δ_2 ↑	WHDR↓	δ_1 ↑	δ_2 ↑	AbsRel↓	δ_1 ↑	δ_2 ↑	AbsRel↓
ENet [16]	50.75	60.83	67.79	72.58	81.57	0.2131	99.79	99.94	0.0116	93.20	96.70	0.0835
PENet [16]	45.09	59.58	69.90	71.39	81.38	0.2173	**99.79**	**99.94**	**0.0114**	92.06	96.42	0.0941
NLSPN [17]	99.36	**99.93**	**99.98**	**87.33**	**93.74**	0.1213	67.60	83.45	0.3254	98.72	99.55	0.0556
Yin *et al.* [14]	99.34	99.89	99.98	85.71	92.66	0.0977	96.59	98.75	0.0568	**99.61**	**99.90**	**0.0101**
Ours	**99.39**	99.90	**99.98**	86.35	93.13	**0.0950**	97.04	98.83	0.0501	99.60	**99.90**	0.0102

Depth completion results on zero-shot test datasets is shown in the **above** Table. Our method achieves on par or even better performance with state-of-the-art methods. Enet and PENet are trained on KITTI. NLSPN is trained on NYUv2. Note that Yin et al. and our method are trained on large-scale mixed datasets and have never seen the 4 zero-shot datasets.



(a) Student (Ours) (b) KD (c) Label-free KD (d) Teacher [69]

The visualization of two-stage depth completion on KITTI is shown in **above** figure. From left to right, we change the stage one module: (a) is our lightweight model, (b) is our lightweight model after simple knowledge distillation, (c) is student after geometric label-free distillation, (d) is the teacher model. From top to bottom and from left to right, the depth map becomes progressively clearer(the distant scenery and the outlines of the utility poles), which shows the predicted results improve, aligning with the numerical experimental findings.