Towards Lightweight Underwater Depth Estimation

Keyu Zhou, Jin Chen, Shuangchun Gui
School of System Design and Intelligent Manufacturing, Southern University of Science and Technology,
Shenzhen, China
{12112943,12112241,12132667}@mail.sustech.edu.cn

Zhenkun Wang
School of System Design and Intelligent Manufacturing & Department of Computer Science and Engineering, Southern University of Science and Technology,
Shenzhen, China
wangzhenkun90@gmail.com

***Abstract*—Underwater depth estimation is crucial in the applications of marine robotics. It can provide environment information for target tracking, robot navigation, and 3D reconstruction of underwater terrain. Existing works transform underwater images into in-air conditions to adapt methods that are designed for natural images. However, this may result in expensive computational resources. To overcome this limitation, we propose a lightweight knowledge distillation framework for underwater depth estimation. We utilize a powerful model designed for underwater images as the teacher model and a lightweight CNN model as the student model. We distill global features to enable the student to acquire both local and global information, thereby improving estimation performance. Our framework includes a global transformation module for efficient global feature distillation and a global-local fusion module to combine local and global information for final estimation. Experimental results on the FLSea dataset demonstrate that our student model is lighter than the teacher model while outperforming lightweight in-air models. Our network is 60% lighter than the teacher model and achieves a 3.1% improvement in the δ_1 metric compared to the lightweight in-air model.**

***Index Terms*—Underwater depth estimation, knowledge distillation, lightweight network.**

I. Introduction

Underwater depth estimation aims to automatically determine the distance from the camera to different objects present in a submerged image. It is crucial for various applications of marine robots, including 3D reconstruction, object tracking, and robot navigation [1]–[5]. There are two primary categories of learning-based approaches. The first focuses on valuable priors [6]–[8], enhancing the visual quality by leveraging object blurriness [6], colorized depth maps [7], and sparse depth priors [8]. The second category designs frameworks to transfer underwater images to their in-air counterparts, a process known as underwater image enhancement [9]. For instance, Gupta *et al*. [10] proposed two connected dense-block-based autoencoders, and Ye *et al*. [11] trained a framework in an adversarial manner to jointly estimate scene depth. Despite their success in advancing estimation performance, these methods are computationally expensive, limiting their execution on devices with constrained resources. To deploy the models on embedded hardware, many works propose lightweight approaches for in-air images [12]–[15]. These methods resize the inputs to increase throughput and employ the pre-trained MobileNetV2 to simplify the network architecture [14], [15]. To strike a balance between low resolution and fine-grained details, Rudolph *et al*. [16] integrate the guided upsampling block (GUB) into the decoder, while relying on the output of a lightweight encoder. Nevertheless, these methods are specifically designed for in-air images, overlooking the inherent complexity of the underwater optical environment.

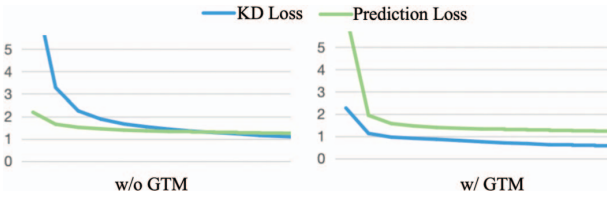

Fig. 1. Loss convergence of knowledge distillation on GuideDepth.

To solve these problems, we propose a novel knowledge distillation (KD) framework for lightweight underwater depth estimation. Specifically, we employ UwDepth [8], a powerful model designed for submerged images, as the teacher, while the student model is implemented by a lightweight CNN-based network, denoted as GuideDepth [16]. As a CNN-Transformer hybrid architecture, UwDepth can capture valuable environment-specific features, such as global underwater environment information. By distilling this knowledge, our lightweight model can obtain both global and local information, thereby enhancing the estimation performance in the underwater environment.

To achieve this, we first train a teacher model for environment-specific feature learning. Subsequently, we conduct feature-level distillation on the student model via a global transformation module (GTM). In contrast to distilling directly, Fig. 1 shows that GTM can mitigate the bias of distillation loss, thus facilitating task loss convergence. Once the local and global information are obtained, we employ a global-local fusion module (GLFM) to fuse these features for final estimation. Experimental evaluations are conducted on the FLSea dataset following [8]. The results show that

This work is supported by the National Natural Science Foundation of China (Grant No. 62106096), Shenzhen Technology Plan (Grant No. JCYJ20220530113013031), Characteristic Innovation Project of Colleges and Universities in Guangdong Province (Grant No. 2022KTSCX110), and Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation ("Climbing Program" Special Funds) (Grant No. pdjh2023c21602). (*Keyu Zhou and Jin Chen contributed equally to this work.*) (*Corresponding author: Shuangchun Gui and Zhenkun Wang.*)

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00258

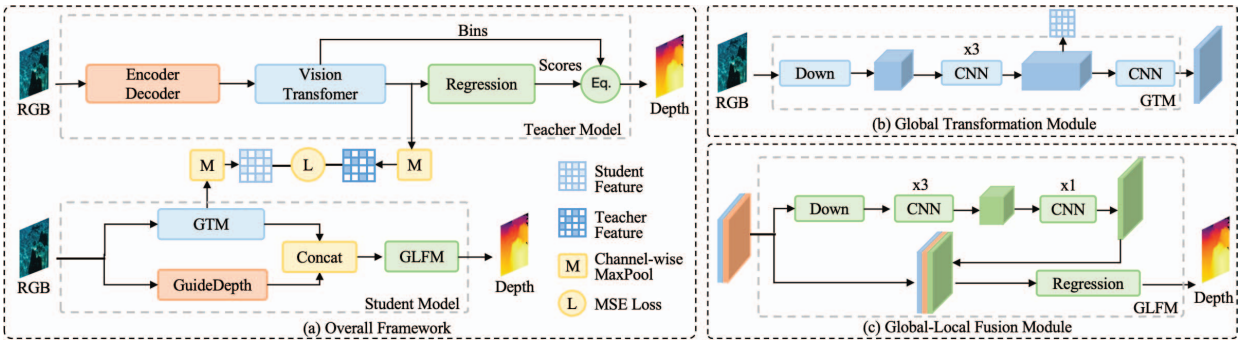

Fig. 2. Overall architecture.

our student model is substantially lighter than the SOTA underwater model and the estimation performance is better than the lightweight in-air model. Our key contributions are as follows:

- We propose a novel knowledge distillation framework for lightweight underwater depth estimation.
- We devise GTM and GLFM to distill the teacher's features efficiently.
- Experimental results show that our model is lighter than the SOTA underwater model and outperforms the lightweight in-air model.

II. Methods

In this paper, we denote the training set as $\mathcal{T}_r = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^{H\times W\times 3}$ is an RGB image and $\mathbf{y}_i \in \mathbb{R}^{H\times W}$ is the ground-truth depth map. Our objective is to train a depth estimation model with minimized prediction error on the test set $\mathcal{T}_s = \{\overline{\mathbf{x}}_i, \overline{\mathbf{y}}_i\}_{i=1}^{\overline{N}}$.

A. Overall Framework

Fig. 2 (a) illustrates the workflow of our proposed KD framework. In the training process, we first feed the input images to the teacher model for depth estimation. Subsequently, the environment-specific knowledge from the teacher model is transferred to the student model during feature-level distillation. Specifically, this is implemented in three steps: employing 1) the GuideDepth model for local depth mapping, 2) GTM for global depth mapping, and 3) GLFM to fuse these maps for final depth estimation. During the inference process, the student model can estimate the depth independently without relying on the teacher model, enabling a fast-response estimation.

B. Teacher Model Training

The Fig. 2 (a) shows the architecture of the teacher model. Following UwDepth [8], we feed an RGB image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ into the CNN-Transformer architecture $\mathcal{G}(\cdot)$, obtaining the teacher features. These features are subsequently fed to a regression head $\mathcal{R}(\cdot)$ for depth prediction. By minimizing the task loss between the predictions and ground truth labels, we obtain a powerful teacher model $\mathcal{G}^*(\cdot)$. The corresponding teacher features can be acquired as follows:

$$F^T = \mathcal{G}^*(\mathbf{x}), \tag{1}$$

where $F^T \in \mathbb{R}^{D^T\times H/2\times W/2}$ represents the teacher feature map, and D^T is the feature dimension.

C. Student Model Training

1) Local Context Modeling: To enable a lightweight student model, we employ GuideDepth [16] as the local student model $\mathcal{F}_l(\cdot)$. It is a convolutional encoder-decoder network, which is commonly used for in-air depth estimation. By feeding an RGB image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ to the local estimation model $\mathcal{F}_l(\cdot)$, the student depth map can be acquired as

$$M_l^S = \mathcal{F}_l(\mathbf{x}). \tag{2}$$

2) Global Context Modeling: During feature-level distillation, we propose GTM to learn student global features. As depicted in Fig. 2 (b), this module consists of three components: 1) a downsample layer and three 3×3 convolution layers for feature encoding, 2) a 3×3 convolution layer to align the feature dimension between the teacher and student, and 3) a regression head to generate global depth maps. Formally, the RGB image is denoted as $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$. By sequentially feeding $\mathbf{x}$ to the downsample layer and three convolutional layers, we obtain features $F_1^S \in \mathbb{R}^{D^S\times H/4\times W/4}$ with the dimension of D^S. To align with the teacher feature dimension D^T, F_1^S are fed to a convolution layer to obtain the student global features $F_g^S \in \mathbb{R}^{D^T\times H/4\times W/4}$. These features are guided through the mean absolute error (MAE) loss:

$$\mathcal{L}_{kd} = \frac{1}{N}\sum_{i=1}^{N}\left|\mathcal{P}(F^T) - \mathcal{P}(F_g^S)\right|, \tag{3}$$

where $\mathcal{P}(\cdot)$ is a max pooling layer. We utilize a regression head to acquire the global depth map $M_g^S = \mathcal{R}(F_g^S)$.

3) Global-Local Map Fusion: To effectively incorporate the global and local context information for final depth estimation, we propose a lightweight GLFM. As shown in Fig. 2 (c), this module consists of three components: 1) a downsample

TABLE I
ABLATION STUDY OF DIFFERENT MODULES COMBINATIONS

Method	Resolution	RMSE ↓	rel ↓	log_{10} ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑	# Params (M) ↓
Teacher (UwDepth)	240 × 320	44.9	4.0	1.8	97.1	98.9	99.5	15.59
Baseline (GuideDepth)	480 × 640	66.1	15.4	6.3	80.1	97.8	99.2	5.83
+ KD	480 × 640	**66.0**	15.3	6.4	80.3	97.8	99.2	5.83
+ KD-GTM	480 × 640	66.7	14.6	6.1	81.6	97.9	99.2	6.03
+ KD-GTM + GLFM ***(Ours)***	480 × 640	66.3	**13.6**	**5.8**	**83.2**	**98.0**	**99.2**	6.16

layer and three 3×3 convolution layers for fused global-local feature encoding, 2) a 3×3 convolution regression head to generate global-local fused depth maps, 3) a regression head to selectively incorporate global and local depth maps for final depth estimation. Specifically, the local depth map M_l^S and global depth map M_g^S are first concatenated as M_{gl}^S. By sequentially feeding M_{gl}^S to the downsample layer and three convolutional layers, hidden features $F_2^S \in \mathbb{R}^{D^S \times H/4 \times W/4}$ with the dimension of D^S can be obtained. After that, F_2^S is fed to the convolutional regression head, obtaining a global-local fused depth map $M_f^S \in \mathbb{R}^{1 \times H \times W}$. These maps are concatenated with M_{gl}^S at depth level, formulating a composite global-local map $M_{glf}^S \in \mathbb{R}^{3 \times H \times W}$. M_{glf}^S is subsequently fed to the regression head to calculate the depth map $\hat{y} = \mathcal{R}(M_{glf}^S)$. To optimize the models, we calculate prediction loss using the ground truth. This loss is a weighted sum of root mean squared error (RMSE) and scale invariant logarithmic loss (SILog loss) [17]:

$$\mathcal{L}_{pred} = \lambda_1 \mathcal{L}_{RMSE} + \lambda_2 \mathcal{L}_{SILog}, \tag{4}$$

where λ_1 and λ_2 are set to 0.3 and 0.7, respectively. RMSE is a standard loss function for minimizing errors in actual depth values, while SILog loss emphasizes errors at close range. We employ a parameterized adaptation of SILog loss [8], [18] to balance accurate metric scale estimation and effective relative depth prediction. These losses are formulated as follows:

$$\mathcal{L}_{RMSE} = \sqrt{\frac{1}{N}\sum_{i}^{N}(\hat{y}_i - y_i)^2}, \tag{5}$$

$$\mathcal{L}_{SILog} = \beta \sqrt{\frac{1}{N}\sum_{i}^{N} g_i{}^2 - \frac{\lambda}{N^2}(\sum_{i}^{N} g_i)^2}, \tag{6}$$

where $g_i = \log \hat{y}_i - \log y_i$. The overall training loss is

$$\mathcal{L} = \mathcal{L}_{pred} + \alpha \mathcal{L}_{kd}, \tag{7}$$

where α is empirically set to 0.2.

III. EXPERIMENTS

A. Dataset and Evaluation Metrics

In this study, we utilize the publicly available FLSea dataset [19] for model training and evaluation, consisting of raw RGB images, corrected color images, and ground-truth depth labels. Collected from 12 different locations, the dataset contains 22,451 image frames. Adhering to UwDepth's [8] methodology, we allocate 10 videos for training and 2 for evaluation. Although color-corrected images are provided, we solely use raw images during training and testing. To evaluate the performance, we employ 6 commonly-used metrics in current literature [16], *i.e.*, root mean square error (RMSE), mean absolute relative error (rel), scale-invariant error (log_{10}), and 3 threshold accuracy (δ_1,δ_2, and δ_3).

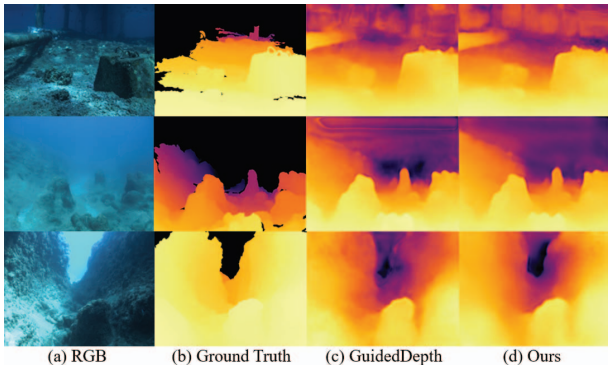

Fig. 3. Depth prediction examples on the FLSea dataset.

B. Implementation Details

The framework is developed using PyTorch [20], and all experiments are conducted on a single NVIDIA Tesla V100 32GB GPU. Following [8], the input images for both teacher and student training undergo light data augmentations, including resizing images to 480×640, horizontal flips, color, brightness, and depth scaling. Throughout training, models are optimized using AdamW [21] with a base learning rate of 1e-4 and a weight decay of 0.9. For teacher training, we employ the pre-trained MobileNetV2 as the image encoder and randomly initialize the other components in UwDepth. These components are trained for 22 epochs with a batch size of 6. In student training, we utilize the pre-trained DDRNet [22] to encode the images, training the model for 30 epochs with a batch size of 16. Our model converges at around 20 epochs, for a fair comparison, we choose the results from the last epoch for testing across all experiments.

TABLE II
SENSITIVITY STUDY

α	RMSE ↓	rel ↓	log_{10} ↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
0.2	**66.3**	**13.6**	**5.8**	**83.2**	**98.0**	**99.2**
0.4	**66.3**	15.3	6.3	80.6	97.9	99.1
0.6	66.4	15.9	6.4	80.9	97.5	99.1
0.8	67.5	14.9	6.2	82.2	97.9	**99.2**

C. Results

We serve GuideDepth as the baseline and conduct ablation experiments as follows: "+ KD" refers to distilling the teacher features directly, "+ KD-GTM" refers to distilling the teacher features by using the proposed GTM, and "+ KD-GTM + GLFM ***(Ours)***" utilizes both the GTM and GLFM. Fig. 3 presents visualization results and Table I illustrates the metric scores. The results indicate that individually learning teacher features with a single module is beneficial. The last metric "# Params" shows that the student network is nearly 60% smaller than the teacher network. Furthermore, our designed lightweight modules only increase by 0.3M compared to the baseline but result in a 3.1% improvement in the δ_1 metric.

D. Sensitivity Study

In Eq.7, α is used to balance the contributions of the features and task losses. To study the impact of different loss combinations on distillation performance, we test the performance using different α settings. We sample α uniformly in the range of 0.2 to 0.8 with intervals of 0.2. Table II shows all the performance metrics achieved under different α. We can find when $\alpha \in [0.4, 0.8]$, the performance drops. The reason can be it causes overemphasis on global features learning, making the student model neglect the acquisition of local features. Based on these experimental results, this work adopts $\alpha = 0.2$.

IV. CONCLUSION

This paper introduces a lightweight knowledge distillation framework for underwater depth estimation. In this framework, UwDepth [8], a powerful underwater depth estimation model, serves as the teacher model for learning environment-specific features. The student model, GuideDepth [16], a lightweight in-air CNN model excelling at capturing local features. To enhance the learning process, GTM is introduced as a distinctive knowledge distillation method to learn global context from the teacher model. Additionally, we propose GLFM for the profound integration of global and local knowledge for final depth prediction. Our model achieves outstanding improvement in both lightweight aspects and overall performance. The ablation study is also conducted to validate the effectiveness of each key component of our model.

REFERENCES

[1] S. T. Digumarti, G. Chaurasia, A. Taneja, R. Siegwart, A. Thomas, and P. Beardsley, "Underwater 3d capture using a low-cost commercial depth camera," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[2] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3227–3234, 2020.

[3] M. Roznere and A. Q. Li, "Underwater monocular image depth estimation using single-beam echosounder," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 1785–1790.

[4] S. Rahman, A. Q. Li, and I. Rekleitis, "Svin2: An underwater slam system using sonar, visual, inertial, and depth sensor," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1861–1868.

[5] B.-Y. Raanan, J. Bellingham, Y. Zhang, M. Kemp, B. Kieft, H. Singh, and Y. Girdhar, "Detection of unanticipated faults for autonomous underwater vehicles using online topic models," *Journal of Field Robotics*, vol. 35, no. 5, pp. 705–716, 2018.

[6] Y.-T. Peng, X. Zhao, and P. C. Cosman, "Single underwater image enhancement using depth estimation based on blurriness," in *2015 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 4952–4956.

[7] P. L. Drews, E. R. Nascimento, S. S. Botelho, and M. F. M. Campos, "Underwater depth estimation and image restoration based on single images," *IEEE Computer Graphics and Applications*, vol. 36, no. 2, pp. 24–35, 2016.

[8] L. Ebner, G. Billings, and S. Williams, "Metrically scaled monocular depth estimation through sparse priors for underwater robots," *arXiv preprint arXiv:2310.16750*, 2023.

[9] J. Zhou, T. Yang, and W. Zhang, "Underwater vision enhancement technologies: a comprehensive review, challenges, and recent trends," *Applied Intelligence*, vol. 53, no. 3, pp. 3594–3621, 2023.

[10] H. Gupta and K. Mitra, "Unsupervised single image underwater depth estimation," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 624–628.

[11] X. Ye, Z. Li, B. Sun, Z. Wang, R. Xu, H. Li, and X. Fan, "Deep joint depth estimation and color correction from monocular underwater images based on unsupervised adaptation networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3995–4008, 2019.

[12] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "Towards real-time unsupervised monocular depth estimation on cpu," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 5848–5854.

[13] F. Aleotti, G. Zaccaroni, L. Bartolomei, M. Poggi, F. Tosi, and S. Mattoccia, "Real-time single image depth perception in the wild with handheld devices," *Sensors*, vol. 21, no. 1, p. 15, 2020.

[14] D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze, "Fastdepth: Fast monocular depth estimation on embedded systems," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6101–6108.

[15] X. Tu, C. Xu, S. Liu, R. Li, G. Xie, J. Huang, and L. T. Yang, "Efficient monocular depth estimation for edge devices in internet of things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2821–2832, 2020.

[16] M. Rudolph, Y. Dawoud, R. Güldenring, L. Nalpantidis, and V. Belagiannis, "Lightweight monocular depth estimation through guided decoding," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2344–2350.

[17] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems*, vol. 27, pp. 2366–2374, 2014.

[18] S. F. Bhat, I. Alhashim, and P. Wonka, "Adabins: Depth estimation using adaptive bins," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4009–4018.

[19] Y. Randall, "Flsea: Underwater visual-inertial and stereo-vision forward-looking datasets," Ph.D. dissertation, University of Haifa (Israel), 2023.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, pp. 8026–8037, 2019.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[22] Y. Hong, H. Pan, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes," *arXiv preprint arXiv:2101.06085*, 2021.