EndoDepthL: Lightweight Endoscopic Monocular Depth Estimation with CNN-Transformer

Yangke Li
Imperial College London, London, United Kingdom
yangke.li23@imperial.ac.uk

***Abstract*—In this study, we address the key challenges concerning the accuracy and effectiveness of depth estimation for endoscopic imaging, with a particular emphasis on real-time inference and the impact of light reflections. We propose a novel lightweight solution named EndoDepthL that integrates Convolutional Neural Networks (CNN) and Transformers to predict multi-scale depth maps. Our approach includes optimizing the network architecture, incorporating multi-scale dilated convolution, and a multi-channel attention mechanism. We also introduce a statistical confidence boundary mask to minimize the impact of reflective areas. To better evaluate the performance of monocular depth estimation in endoscopic imaging, we propose a novel complexity evaluation metric that considers network parameter size, floating-point operations, and inference frames per second. We comprehensively evaluate our proposed method and compare it with existing baseline solutions. The results demonstrate that EndoDepthL ensures depth estimation accuracy with a lightweight structure.**

***Index Terms*—Endoscopic Image Processing, Monocular Depth Estimation, Biomedical Image Analysis**

I. Introduction

In contemporary surgical practices, monocular depth estimation is critical for endoscopic procedures, demanding both accuracy and efficiency [1]. Monocular cameras such as endoscopes are widely utilized but encounter the challenge of losing depth information from a 3D scene to a 2D image. This loss compels surgeons to rely heavily on their experience to discern the depth within the field of view, intensifying the complexity and decision-making pressure of the procedure [1]. While sensors like lidar could provide precise spatial positions, their integration into endoscopes is fraught with difficulties [2]. The transition of monocular depth estimation methods such as Monodepth2 [3] and LiteMono [4] from the autonomous driving field is not straightforward, as endoscopic scenarios include unique challenges like inconsistent lighting. With the advancement in machine learning techniques, supervised learning has been explored to learn the 3D structure from 2D images [5]. However, obtaining labelled endoscopic training data is prohibitively costly and challenging. Therefore, self-supervised learning, which extracts signals from image data without the need for additional labels, has gained increasing attention [3], [6]–[8]. Besides, reflections due to smooth organ surfaces and the real-time acquisition and computation of image data remain to be addressed, and maintaining a consistent frame rate is also critical. The unique combination of these challenges highlights the need for specialized depth estimation solutions in endoscopic applications.

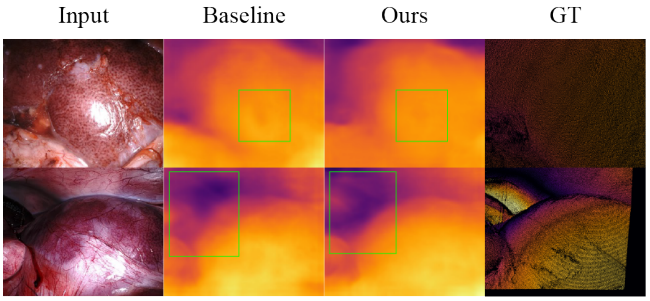

Fig. 1. **Comparison of EndoDepthL with the baseline.** Our method effectively deals with the challenges in endoscopy, such as uneven lighting.

The precision of depth estimation is crucial for endoscopic image analysis and surgery operations. Existing deep learning methods exhibit limitations in providing accurate and fast depth estimation results. Earlier solutions primarily hinged on Convolutional Neural Network(CNN), but the restrictions of the convolutional kernel make it hard to extract global features from images. Recently, numerous Transformer-based approaches have been proposed [4], [9]–[17]. While these methods effectively enhance the acquisition of global features, they also largely increase the network's parameter size, bringing down the inference speed significantly. This will affect the medical diagnosis results. Since algorithms must operate in real-time on edge devices like endoscopes, we cannot directly apply algorithms developed for specialized graphical computing devices to such equipment [18], [19]. Hence, there is an urgent to find a method to process endoscopic images in real-time while improving the accuracy and robustness of depth estimation, especially under strong light reflections.

To this end, we propose a novel depth estimation method that concentrates on two main areas: firstly, we optimize the network architecture design to reduce the parameters of the network by designing a more lightweight network architecture; specifically, we incorporate multi-scale dilation convolution, and a multi-channel attention mechanism in the encoder to extract image features more efficiently. Moreover, we establish a confidence mask to minimize the influence of light-reflective regions on the training process and guide the network to focus on non-reflective regions.

Compared with existing endoscopic depth estimation methods, the contributions of our work can be summarized as follows:

a) Lightweight CNN-Transformer Encoder: We design a method combining a Convolutional Neural Network (CNN) and a Transformer for predicting multi-scale depth maps from input images. This method combines dilated convolution with a cross-covariance attention mechanism to broaden the sensory field and capture global features without augmenting the network parameters. EndoDepthL ensures comparable performance to existing methods while facilitating faster inference speed.

b) Reflective Mask: We propose a masking mechanism to address the common issue of reflections in endoscope environments. This mechanism effectively reduces the impact of reflective regions on depth estimation. Specifically, when calculating the loss function, the contribution of reflective regions is minimized to nearly zero. This enables EndoDepthL to prioritize depth estimation in non-reflective regions.

c) Complexity Evaluation: We introduce the metric of parameter size, floating point operations and inference frames per second in our evaluation metrics to provide a more comprehensive evaluation of depth estimation over endoscopic images. To the best knowledge of the author, this is the first study in the field of endoscopic depth estimation that evaluates from the complexity perspective. We compare EndoDepthL with existing methods from both accuracy and efficiency perspectives, providing a benchmark for the practical application of depth estimation over endoscopic images.

In summary, our study provides a lightweight solution for endoscopic depth estimation that can mitigate the effect of reflections and is expected to further enhance the efficiency and safety of laparoscopic surgery.

II. Related Work

A. Self-supervised Mono-Depth Estimation

Self-supervised learning shows potential in depth estimation, with significant advancements due to monocular methods. Godard et al. [6] introduced a self-supervised monocular network, using disparity for supervision. Zhou et al. [20] integrated multi-frame video sequences, estimating depth while learning camera motion.

However, challenges still exist in scenarios with dynamic scenes and changing lighting conditions. Zou et al. [21] address these issues by introducing an unsupervised joint learning method of depth and flow called Df-net, which leverages cross-task consistency to model the relative motion within the scene. On the other hand, Li et al. [22] tackle the same challenges by proposing Megadepth, a method that learns single-view depth prediction from a large and diverse set of Internet photos. There are extra issues with estimating the depth of small objects, as mentioned by Sattler et al. [23] and Wang et al. [19].

Recent research addresses these issues from the loss function and neural network structures. Wang et al. [24] proposed an occlusion-aware loss function, while Guizilini et al. [25]leveraged fixed pre-trained semantic segmentation networks to guide self-supervised representation learning via pixel-adaptive convolutions. There have been some efforts to enhance network structures for more efficient feature extraction. Zhao et al. [26] focused on geometric consistency to aid depth perception, integrating geometry-based constraints within their network structure. Yin et al. [27] solved the challenge by enforcing strong supervisory signals from the underlying 3D geometry, creating an alignment between monocular depth estimation and surface normals. Fu et al. [28] proposed a deep ordinal regression network, employing ordinal depth ranking among pixels to enable a more robust and discriminative representation of depth information. Additionally, Guizilini et al. [29] introduced PackNet, a novel network structure that employs spatial packing and unpacking within convolutional layers. Similarly, Yang et al. [30] developed the LEGO (Learning Edge with Geometry all at Once) framework, which incorporates geometric constraints such as edges, planes, and vanishing points to improve depth estimation accuracy.

Transformer models could be beneficial in self-supervised monocular depth estimation. Vaswani et al. [31]has built based on the attention mechanism by Dosovitskiy et al. [32] for image recognition. Carion et al. [33] proposed DETR, a Transformer-based object detection model. Transformers have shown potential in semantic segmentation [34] and deep estimation [35]. A recent work [16] combined plain convolutions with Transformer blocks to enhance local feature extraction and global information understanding in visual tasks.

Despite the potential advantages, Transformers still face challenges with high-resolution images [36] and insufficient training data [37]. These issues are particularly relevant for specific applications like endoscopic image depth estimation.

B. Endoscopic Image Analysis

Endoscopic image depth estimation is challenging due to distinct lighting conditions, complex backgrounds, and precision requirements. These difficulties have led to the emergence of various research strategies.

The unique lighting conditions in endoscopy affect image brightness and color, impacting depth estimation. Kohler et al. [38] proposed a new color constancy method by separating spectral information of endoscopic images. Ma et al. [39] used Generative Adversarial Network (GAN) to enhance endoscopic images under inconsistent illumination.

Endoscopic images often contain complex backgrounds like blood and tissue debris, posing challenges to depth estimation since image noise affects the performance of depth estimation. To mitigate this, researchers have employed semantic segmentation techniques. Seo et al. [40]and Zhu et al. [41] used deep learning to effectively separate regions of interest from complex medical images, hence improving the estimation accuracy in the following.

High-resolution images are necessary for precision in endoscopic surgery, demanding real-time, efficient depth estimation methods. Tang et al. [42] proposed MobileNets, using depthwise separable convolution to improve model efficiency. Zhang et al. [43]introduced ShuffleNet, aiming to enhance model efficiency significantly.

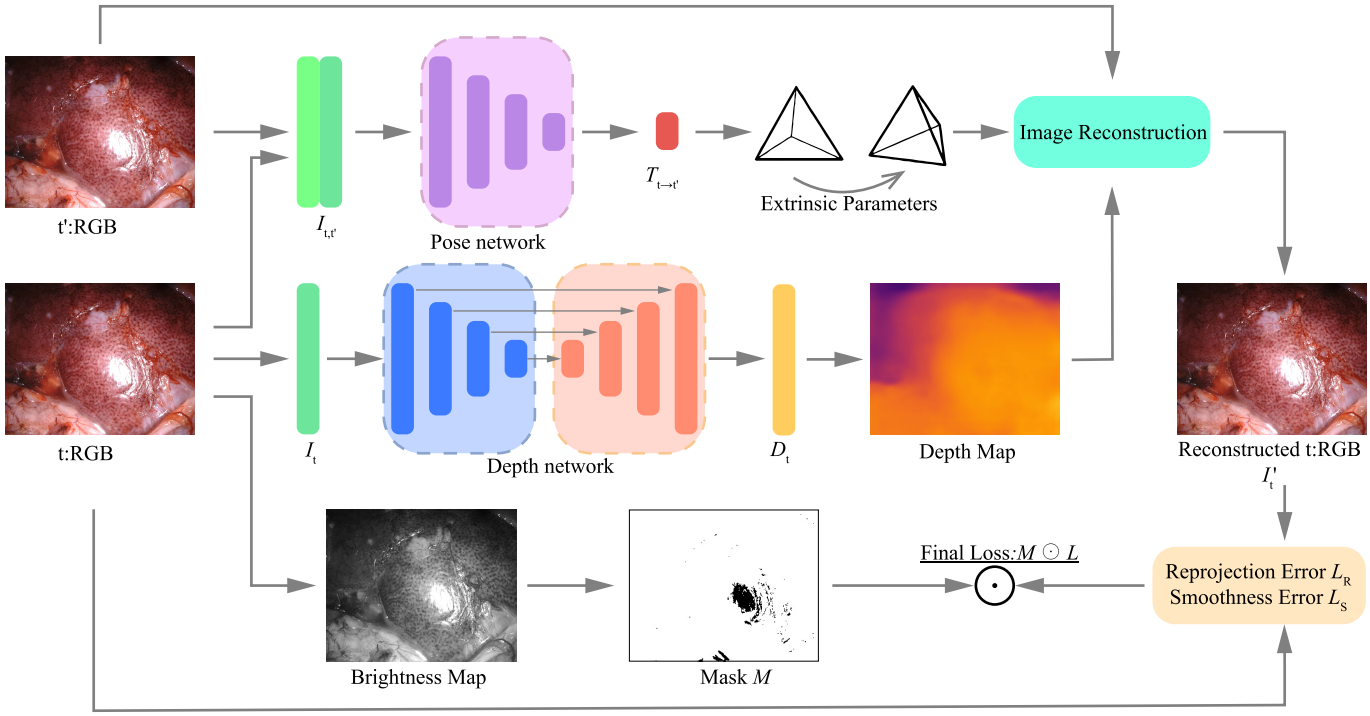

Fig. 2. **Overview of the proposed method.** Put the source and target frames into the pose network and the target frame into the depth network. Each network extracts respective features: the pose network determines the transition from the source to the target, and the depth network produces initial depth predictions. Then reduces the reconstruction error by leveraging the camera's inherent parameters. To wrap up, a Statistical Confidence Boundary Mask is used to counteract the effects of light reflection, ensuring a more precise and stable result.

Despite existing solutions providing several feasible ways to get depth estimation results, endoscopic image depth estimation remains an open question, requiring further work to improve accuracy and reliability.

III. Methodology

A. Self Supervised Loss Function

EndoDepthL is based on a self-supervised principle, as illustrated in Fig. 2. Define two sequential images I_t, I_{t+1}, where I_t is the reference frame and I_{t+1} is the target frame, as described by [6] and [20]. The depth network D predicts the per-pixel depth value d_t of the reference frame and converts it into a point cloud X_t:

$$X_t = d_t K^{-1} p_t, \tag{1}$$

where K is the known camera intrinsic matrix, and p_t is the pixel coordinates. The pose network T predicts the relative pose from the reference frame to the target frame. Using this pose, we can transform the point cloud X_t into the coordinate system of the target frame:

$$X_{t+1} = RX_t + t. \tag{2}$$

The transformed point cloud X_{t+1} is then projected back onto the image plane to obtain the pixel coordinates p_{t+1} of the target frame:

$$p_{t+1} = KX_{t+1}. \tag{3}$$

The reprojection pixel value $I_t^{'}$ is calculated using bilinear interpolation at the p_{t+1} location of the target frame image I_{t+1}. The difference between this value and the original pixel value I_t of the reference frame is used to calculate the reprojection error:

$$L_R = I_t - I_t^{'}. \tag{4}$$

Minimizing this error allows the network to learn more accurate depth and relative pose during training.

Reconstruction uses both frames I_{t-1} and I_{t+1}, and the smallest reconstruction error is selected for minimization:

$$L_R = min(L_{R,t-1}, L_{R,t+1}), \tag{5}$$

where $L_{R,t-1}$ and $L_{R,t+1}$ are the reconstruction errors with I_{t-1} and I_{t+1}, respectively.

Following [3], a binary mask μ is defined to handle pixels that cannot be correctly projected:

$$\mu = min(L_{R,t-1}, L_{R,t+1}) < min(I_t - I_{t-1}, I_t - I_{t+1}), \tag{6}$$

The associated loss function is:

$$L_p = \mu * L_R, \tag{7}$$

To encourage smoother depth map prediction, a smoothness loss based on the first-order derivatives of the image and depth map is added:

$$L_s = |\partial_x d_t| e^{-|\partial_x I_t|} + |\partial_y d_t| e^{-|\partial_y I_t|}, \tag{8}$$

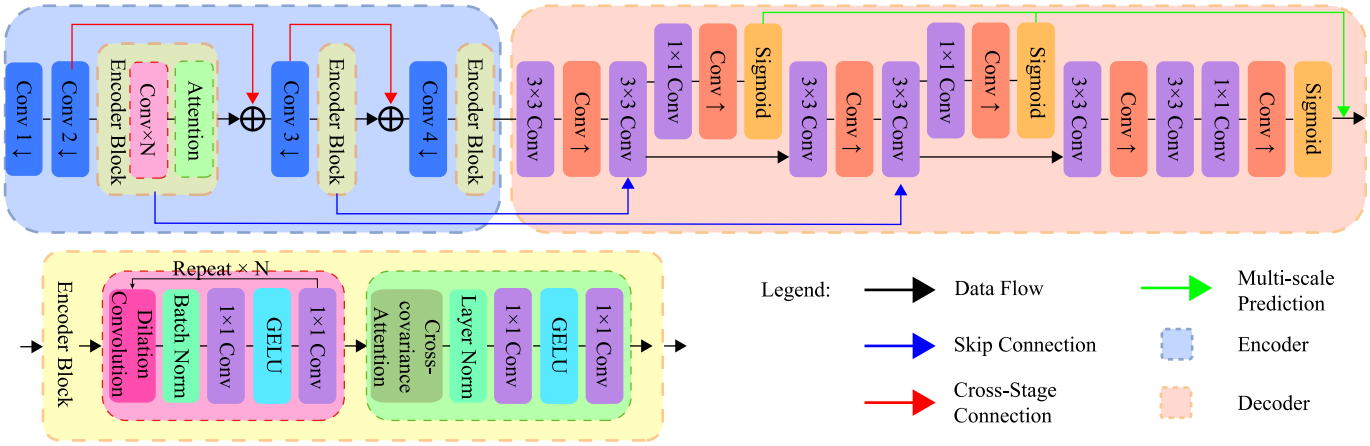

Fig. 3. **Depth network.** We've enhanced feature extraction in the Encoder by incorporating an Encoder Block, consisting of convolution and attention components. We propose two Encoder network sizes(efficiency and performance) to meet varied requirements, as detailed in Table I.

TABLE I
CONVOLUTIONAL PARAMETERS

	Efficiency	Performance
Input	320×256×3	
cov1	160×128×32	160×128×64
cov2	80×64×32	80×64×64
dilation conv	rate =1,2,3	
cov3	40×32×64	40×32×128
dilation conv	rate =1,2,3	
cov4	20×16×128	20×16×256
dilation conv	rate =1,2,3,2,4,6	rate =1,2,3,2,4,6,3,6,9

The total loss function can be represented as:

$$L = L_p + \lambda L_s, \tag{9}$$

where λ is an adjustable weight parameter. In summary, the described method transforms unsupervised depth estimation into a process that jointly estimates the camera's depth and relative pose. This is based on the optimization of reprojection error, incorporation of a binary mask, and the application of a smoothness loss.

B. CNN-Transformer Lightweight Depth Network

The proposed architecture comprises a DepthNet with Encoder-Decoder and a PoseNet with only Encoder. DepthNet, as illustrated in Fig.3, involves predicting multi-scale depth maps from an input image, while PoseNet is dedicated solely to predicting camera motion between sequential frames. Once these predictions are accomplished, a reconstructed target image is generated, and loss for model optimization is computed.

Traditional convolution operations are limited by their receptive field. To solve this, dilated convolution [44] is introduced into the model. This method expands the receptive field without extra parameters by interspersing gaps within the kernel elements, which can be formally represented as follows:

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k] w[k], \tag{10}$$

where $w[k]$ refers to a filter of length K, and r is the dilation rate. Dilated convolution's applied [4] allows the model to grasp a broader contextual understanding and enhance the feature representation.

As shown in Table I, we set two sets of convolution parameters, corresponding to efficiency mode (Eff.) and performance mode (Perf.). Our model applies varying channel counts for downsampling convolutions, followed by three dilation convolutions with increasing rates. After the fourth downsampling, we use additional dilation convolutions with larger rates to obtain features in larger scales, totaling six iterations. In performance mode, we add three further dilation iterations.

Inspired by the Transformer [31], a global feature extraction method is employed that not only provides local features but also encompasses global information. Drawing from approaches like [4], this method utilizes a cross-covariance attention mechanism [45]. It processes the attention between feature channels by linearly projecting the input feature map to derive the Query (Q), Key (K), and Value (V) components.The process can be depicted as:

$$\hat{X} = \text{Attention}(Q, K, V) + X, \tag{11}$$

where

$$\text{Attention}(Q, K, V) = V \cdot \text{softmax}(Q^T \cdot K). \tag{12}$$

To further enhance the non-linearity of features, a GELU [47] activation function is applied to the feature map. Then merged with the original input feature map to produce the final output feature map. Like the strategy proposed by [48], the enhanced feature map is combined with the original input features, leading to a richer feature map.

The following sections discuss a strategy to reduce the impact of reflections in depth estimation. We use a mask to help the model concentrate on key image areas. This approach ignores regions that reflections might distort.

TABLE II
COMPARATIVE EXPERIMENTAL RESULTS

Method	Data	Accuracy							Complexity		
		Abs Rel	Sq Rel	RMSE	RMSE log	δ <1.25	δ <1.25^2	δ <1.25^3	Param.	FLOPs	FPS
Monodepth2-Res18 [3]	M	0.159	1.796	21.905	0.216	0.742	0.938	0.999	14.329M	8.038G	56.6
Monodepth2-Res50 [3]	M	0.143	1.466	21.589	0.191	0.757	0.981	**>0.999**	32.522M	16.663G	32.82
AF-SfMLearner [46]	M	0.101	0.678	5.416	0.133	0.881	**>0.999**	**>0.999**	14.329M	5.359G	53.65
LiteMono [4]	M	0.133	1.225	7.332	0.167	0.791	0.993	**>0.999**	3.069M	3.355G	60.9
LiteMono-8m [4]	M	0.112	0.96	6.556	0.135	0.888	0.998	**>0.999**	8.766M	7.475G	45.67
EndoDepthL-Eff.	M	0.104	0.727	5.38	0.135	0.883	0.998	**>0.999**	**2.143M**	**1.894G**	**62.8**
EndoDepthL-Perf.	M	**0.094**	**0.635**	**5.229**	**0.113**	**0.953**	0.998	**>0.999**	10.882M	8.211G	45.01

TABLE III
METRICS FOR ACCURANCY

Metric	Formula
Abs Rel	$\frac{1}{N}\sum_i \frac{\lvert d_i-\hat{d}_i\rvert}{d_i}$
Sq Rel	$\frac{1}{N}\sum_i \frac{(d_i-\hat{d}_i)^2}{d_i}$
RMSE	$\sqrt{\frac{1}{N}\sum_i \left(d_i-\hat{d}_i\right)^2}$
RMSE log	$\sqrt{\frac{1}{N}\sum_i \left(\log d_i-\log \hat{d}_i\right)^2}$
δ	$\frac{1}{N}\sum_i \left[\max\left(\frac{d_i}{\hat{d}_i},\frac{\hat{d}_i}{d_i}\right)<\theta\right]$

C. Statistical Confidence Boundary Mask

Reflection and shadows may cause inconsistency in pixel intensities, violating the photometric consistency assumption. The masking mechanism can effectively mitigate this by excluding reflection areas from loss computation. The intensity map L for an input image I is computed:

$$L = 0.299 \cdot I_r + 0.587 \cdot I_g + 0.114 \cdot I_b, \tag{13}$$

where I_r, I_g, I_b represent the red, green, blue channels of image I. A threshold τ distinguishes reflection from non-reflection areas. The intensity map L is normalized to the [0, 1] interval, producing L_n:

$$L_n = \frac{L - \min(L)}{\max(L) - \min(L)}, \tag{14}$$

A mask M is generated via a logistic function:

$$M = \frac{1}{1 + e^{-k\cdot(\tau - L_n)}}, \tag{15}$$

where k controls the transition smoothness. This mask transitions smoothly from 1 to 0 in reflection areas, whereas in non-reflection areas, values approach 1.

The loss function L' incorporates the mask:

$$L' = M \odot L, \tag{16}$$

where $\odot$ represents the Hadamard product. As reflection areas in the mask have values close to 0, their contribution to the loss function is negligible, effectively allowing the model to focus on non-reflective areas. The subsequent section presents experiments on a public dataset to validate these methods.

IV. EVALUATION AND VALIDATION

A. Experiment Setup

a) Dataset: SCARED [49] (Surgeries with CAmeras and Rigid Endoscopes Dataset) consist of endoscopic surgical videos collected using a da Vinci Xi endoscope and projector on fresh porcine cadaver abdominal anatomy to obtain high-quality depth maps. This process is performed at 5-10 different camera positions, following specific coded structured light imaging methods [50]. The values in the depth maps are in millimeters, and invalid pixels are masked out. Since the camera must remain stationary during each structured light projection, the dataset is expanded with camera motion and warped depth maps using known camera poses from the da Vinci Xi kinematics. These poses are released as a 4x4 matrix, along with the stereo camera calibration for the sequence. The dataset is partitioned into training (15,351 frames), validation (1,705 frames), and testing sets (1,243 frames). The known intrinsic parameters of the endoscope guide the self-supervised training process, and during the evaluation phase, depth prediction remains within a 150mm constraint, simulating the physical limitations of endoscopic devices. Following [3], we introduce data augmentation procedures, including horizontal flip, brightness, saturation, contrast adjustment, and hue jitter—with each occurring with a 50% probability.

b) Hyperparameters: The experiments were conducted on a system equipped with an 8-core CPU, 30GB of memory, and an NVIDIA T4 GPU, a mid-range unit resonating with the computational capabilities of edge devices. The system was hosted on Google Cloud and employed PyTorch version 1.12 for data processing and training. The training parameters include a batch size of 8 and the AdamW optimizer [51], with a weight decay of 1×10^{-2}. The initial learning rate was set at 5×10^{-4}, adhering to a cosine learning rate schedule. A monocular training session spanning 30 epochs took approximately 70 hours. The chosen configuration and methodological approach facilitated accurate modeling within the constraints of a medium-performance computational environment.

c) Evaluation Metrics: The assessment employs standard monocular depth estimation metrics: Abs Rel, Sq Rel, RMSE, RMSE log, $\delta < 1.25$, $\delta < 1.25^2$and$\delta < 1.25^3$. As shown in Table III, d_i represents the true depth values, $\hat{d}_i$ denotes

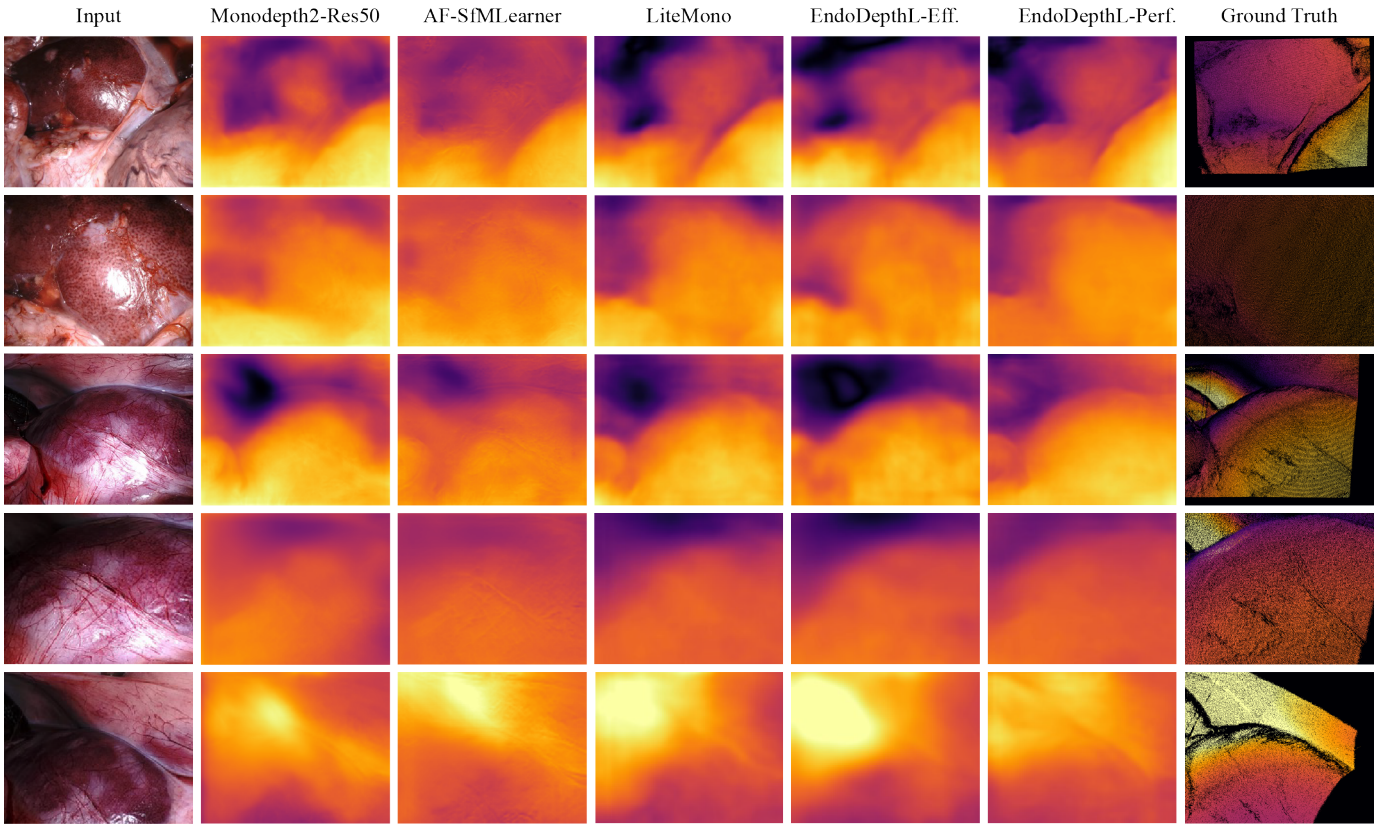

Fig. 4. **Experimental result for our analysis.** We extracted representative frames from two distinct video segments. These carefully chosen frames encompass various perspectives, including frontal and lateral viewpoints, and capture different degrees of organ exposure. In some instances, the organs are fully visible, while in others, they are partially obscured or covered. From this figure, we can see that the EndoDepthL performance model is better with smoother and more accurate depth estimation.

the predicted depth values, N is the total number of pixels, and θ is a threshold. Additionally, we add an efficiency evaluation, considering the algorithm's overhead, including parameter size, floating point operations, and inference frames per second.

B. Baseline Methods

To highlight EndoDepthL's performance, we compared it with some popular baselines: Monodepth2, LiteMono, and AF-SfMLearner.

Monodepth2 [3], a typical classic method that includes versions ResNet18 and ResNet50, for benchmarking our study. Known for handling moving objects and occlusions, we followed its original parameter settings in experiments.

Lite-Mono [4], designed for autonomous driving challenges, blends CNN's local processing with Transformers' global capabilities. We adapted this method for endoscopic dataset, including fine-tuning the input size.

AF-SfMLearner [46], selected from open-source works, tackles endoscopic challenges such as inconsistent illumination. Its technique "Appearance Flow" aligns with our study's unique challenges.

TABLE IV
ABLATION STUDY RESULT

Method	AbsRel	SqRel	RMSE	RMSElog	δ<1.25
EndoDepthL-Eff.	0.104	0.727	5.380	0.135	0.883
w/o mask	0.119	1.011	6.449	0.172	0.833
EndoDepthL-Perf.	0.094	0.635	5.229	0.113	0.953
w/o mask	0.102	0.694	5.312	0.132	0.908

C. Experimental Results

We compare EndoDepthL with baseline methods, and the results are listed in Table II and Fig. 4, covering performance and efficiency. "M" denotes SCARED monocular videos. Our experiments are trained from scratch, and the best results are marked in bold. Our model demonstrates enhanced performance compared to the baseline methods. Specifically, it attains results akin to those of AF-SfMLearner but operates with reduced complexity. The efficiency of EndoDepthL allows for a 30% reduction in network size while improving performance, or alternatively, it can decrease both the size and complexity by 3-5 times while maintaining similar levels of effectiveness.

We also compared with algorithms from autonomous driving, such as Litemono. EndoDepthL shows greater stability, which is essential for handling reflections on smooth organ

surfaces in endoscopy. This stability leads to superior performance, with a four-fold increase in efficiency.

D. Ablation Study

We conducted an ablation study to validate the proposed mask module's effectiveness. The results are presented in Table IV and underscore the crucial function of this component. The experimental conditions were consistent with those of the previous comparative study.

The confidence mask is a vital part of our architecture, designed to diminish the impact of reflections during training. Without it, the reflections substantially affect the performance of EndoDepthL, increasing extreme values and having a noticeable influence on the Root Mean Square Error (RMSE). In our tests, removing the confidence mask led to an approximately 10% increase in RMSE, accentuating the importance of managing reflections. Moreover, we observed a more pronounced decline in the processing ability of smaller, efficiency-oriented models when the mask was absent. This ablation study further substantiates the essential role of neutralizing reflection effects and establishes that our mask module is integral to improving the performance of lightweight networks.

V. Conclusion

This paper proposes a novel lightweight monocular depth estimation approach tailored for endoscopic applications. Utilizing a hybrid CNN and Transformer architecture, EndoDepthL adeptly extracts multi-scale local and global features from the endoscopic images. In addition, by integrating the confidence mask, EndoDepthL efficiently mitigates the detrimental effects of reflections, which is a standout challenge in endoscopic depth estimation. Experimental validation on the SCARED dataset demonstrates underscores our method's capability to balance low computational complexity with high estimation accuracy, paving the way for real-world deployment of depth estimation techniques in endoscopy.

Acknowledgment

The author would like to thank Jiaping Xiao for the help and valuable feedback on this work.

References

[1] A. P. Stegemann, K. Ahmed, J. R. Syed, S. Rehman, K. Ghani, R. Autorino, M. Sharif, A. Rao, Y. Shi, G. E. Wilding *et al.*, "Fundamental skills of robotic surgery: a multi-institutional randomized controlled trial for validation of a simulation-based curriculum," *Urology*, vol. 81, no. 4, pp. 767–774, 2013.

[2] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[3] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[4] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 537–18 546.

[5] A. Saxena, S. Chung, and A. Ng, "Learning depth from single monocular images," *Advances in neural information processing systems*, vol. 18, 2005.

[6] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 270–279.

[7] A. Gordon, H. Li, R. Jonschkowski, and A. Angelova, "Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8977–8986.

[8] C. Wang, J. M. Buenaposada, R. Zhu, and S. Lucey, "Learning depth from monocular videos using direct methods," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2022–2030.

[9] J. Bae, S. Moon, and S. Im, "Deep digging into the generalization of self-supervised monocular depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 187–196.

[10] V. Guizilini, R. Ambruș, D. Chen, S. Zakharov, and A. Gaidon, "Multi-frame self-supervised depth with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 160–170.

[11] A. Karpov and I. Makarov, "Exploring efficiency of vision transformers for self-supervised monocular depth estimation," in *2022 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2022, pp. 711–719.

[12] Z. Liu, R. Li, S. Shao, X. Wu, and W. Chen, "Self-supervised monocular depth estimation with self-reference distillation and disparity offset refinement," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[13] Y. Mao, R. Zhao, T. Zhang, and H. Zhao, "Bevscope: Enhancing self-supervised depth estimation leveraging bird's-eye-view in dynamic scenarios," *arXiv preprint arXiv:2306.11598*, 2023.

[14] Y. Shi, H. Cai, A. Ansari, and F. Porikli, "Ega-depth: Efficient guided attention for self-supervised multi-camera depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 119–129.

[15] A. Varma, H. Chawla, B. Zonooz, and E. Arani, "Transformers in self-supervised monocular depth estimation with unknown camera intrinsics," *arXiv preprint arXiv:2202.03131*, 2022.

[16] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "Monovit: Self-supervised monocular depth estimation with a vision transformer," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 668–678.

[17] C. Zhuang, Z. Lu, Y. Wang, J. Xiao, and Y. Wang, "Spdet: Edge-aware self-supervised panoramic depth estimation transformer with spherical geometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[18] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1655–1674, 2019.

[19] F. Wang, M. Zhang, X. Wang, X. Ma, and J. Liu, "Deep learning for edge computing applications: A state-of-the-art survey," *IEEE Access*, vol. 8, pp. 58 322–58 336, 2020.

[20] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[21] Y. Zou, Z. Luo, and J.-B. Huang, "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 36–53.

[22] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2041–2050.

[23] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe, "Understanding the limitations of cnn-based absolute camera pose regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3302–3312.

[24] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4884–4893.

[25] V. Guizilini, R. Hou, J. Li, R. Ambrus, and A. Gaidon, "Semantically-guided representation learning for self-supervised monocular depth," *arXiv preprint arXiv:2002.12319*, 2020.

[26] S. Zhao, H. Fu, M. Gong, and D. Tao, "Geometry-aware symmetric domain adaptation for monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9788–9798.

[27] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5684–5693.

[28] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2002–2011.

[29] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2485–2494.

[30] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia, "Lego: Learning edge with geometry all at once by watching videos," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 225–234.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[33] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[34] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6881–6890.

[35] Z. Cheng, Y. Zhang, and C. Tang, "Swin-depth: Using transformers and multi-scale fusion for monocular-based depth estimation," *IEEE Sensors Journal*, vol. 21, no. 23, pp. 26 912–26 920, 2021.

[36] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are rnns: Fast autoregressive transformers with linear attention," in *International conference on machine learning*. PMLR, 2020, pp. 5156–5165.

[37] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[38] H. Köhler, A. Kulcke, M. Maktabi, Y. Moulla, B. Jansen-Winkeln, M. Barberio, M. Diana, I. Gockel, T. Neumuth, and C. Chalopin, "Laparoscopic system for simultaneous high-resolution video and rapid hyperspectral imaging in the visible and near-infrared spectral range," *Journal of Biomedical Optics*, vol. 25, no. 8, pp. 086 004–086 004, 2020.

[39] Y. Ma, J. Liu, Y. Liu, H. Fu, Y. Hu, J. Cheng, H. Qi, Y. Wu, J. Zhang, and Y. Zhao, "Structure and illumination constrained gan for medical image enhancement," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3955–3967, 2021.

[40] K. Seo, J.-H. Lim, J. Seo, L. S. Nguon, H. Yoon, J.-S. Park, and S. Park, "Semantic segmentation of pancreatic cancer in endoscopic ultrasound images using deep learning approach," *Cancers*, vol. 14, no. 20, p. 5111, 2022.

[41] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, "Anatomynet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Medical physics*, vol. 46, no. 2, pp. 576–589, 2019.

[42] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[43] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6848–6856.

[44] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[45] A. Ali, H. Touvron, M. Caron, P. Bojanowski, M. Douze, A. Joulin, I. Laptev, N. Neverova, G. Synnaeve, J. Verbeek *et al.*, "Xcit: Cross-covariance image transformers," *Advances in neural information processing systems*, vol. 34, pp. 20 014–20 027, 2021.

[46] S. Shao, Z. Pei, W. Chen, W. Zhu, X. Wu, D. Sun, and B. Zhang, "Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue," *Medical image analysis*, vol. 77, p. 102338, 2022.

[47] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[49] M. Allan, J. Mcleod, C. Wang, J. C. Rosenthal, Z. Hu, N. Gard, P. Eisert, K. X. Fu, T. Zeffiro, W. Xia *et al.*, "Stereo correspondence and reconstruction of endoscopic data challenge," *arXiv preprint arXiv:2101.01133*, 2021.

[50] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*. Springer, 2014, pp. 31–42.

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