Simultaneous Monocular Endoscopic Dense Depth and Odometry Estimation Using Local-Global Integration Networks

Wenkang Fan[1][0000−0002−8364−5159], Wenjing Jiang[1], Hao Fang[1,2], Hong Shi[3], Jianhua Chen[3], and Xiongbiao Luo[1,2,*][0000−0001−7906−8857]

[1] Department of Computer Science and Technology, Xiamen University, Xiamen 361102, China. xbluo@xmu.edu.cn, * *indicates the corresponding author*
[2] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361102, China
[3] Fujian Medical University Cancer Hospital, Fuzhou 350014, China. endoshihong@hotmail.com

Abstract. Accurate dense depth prediction of monocular endoscopic images is essential in expanding the surgical field and augmenting the perception of depth for surgeons. However, it remains challenging since endoscopic videos generally suffer from limited field of view, illumination variations, and weak texture. This work proposes LGIN, a new architecture with unsupervised learning for accurate dense depth recovery of monocular endoscopic images. Specifically, LGIN creates a hybrid encoder using dense convolution and pyramid vision transformer to extract local textural features and global spatial-temporal features in parallel, while building a decoder to effectively integrate the local and global features and use two-heads to estimate dense depth and odometry simultaneously, respectively. Additionally, we extract structure-valid regions to assist odometry prediction and unsupervised training to improve the accuracy of depth prediction. We evaluated our model on both clinical and synthetic unannotated colonoscopic video images, with the experimental results demonstrating that our model can achieve more accurate depth distribution and more sufficient textures. Both the qualitative and quantitative assessment results of our method are better than current monocular dense depth estimation models.

Keywords: Monocular depth estimation · Transformer · Endoscopy · Unsupervised learning · Colonoscopy.

1 Introduction

Endoscopy is an essential diagnostic and therapeutic tool in minimally invasive surgery. However, monocular endoscopy with a limited field of view and a lack of depth perception of the surgical scene increases operative time and surgical risks. To this end, dense depth recovery for endoscopic field 3-D reconstruction is widely discussed to expand the endoscopic viewing of the surgeon [7].

Deep learning approaches are commonly used for dense depth estimation. Supervised learning usually requires large annotated data which is particularly unrealistic for monocular endoscopic videos. Therefore, recent researches [14, 5] introduce self-supervised learning methods in endoscopy images by using sparse depth supervision (e.g.SfM), which greatly depends on the quality of sparse reconstruction. Fortunately, unsupervised learning approaches employing the photometric loss for training to simultaneously estimate dense depth and camera poses are more convenient [15, 20, 9, 26, 11].

Convolutional neural networks (CNNs) are widely used for dense depth estimation [12, 15, 28, 13]. Although CNNs extract abundant local spatial texture features, they are inadequate to extract global features because convolution only focuses on small patches. Recently, vision transformers (ViT) are increasingly discussed in dense depth prediction task [18, 22, 1, 25]. However, texture details obtained by these models were insufficient compared to CNNs because the transformer focuses too much on long-distance relationships while neglecting fine textures. Therefore, aggregating global features with local features is a promising way for accurate dense depth estimation. Currently, many locally enhanced ViT models have been proposed for depth prediction [8, 17, 16, 27, 3, 24]

This work explores the performance of CNN and the transformer for monocular dense depth prediction and proposes a new model LGIN with unsupervised learning for depth and odometry prediction. The contributions of our work are as follows. Firstly, we create a hybrid encoder combining dense convolution and pyramid vision transformer to extract features in parallel which can obtain more sufficient and accurate local texture features and global depth distribution features. Secondly, we build a powerful decoder to effectively integrate local and global features from coarse to fine and use dual heads to estimate dense depth and odometry simultaneously. The feature-shared way can improve the relevance of depth and pose prediction. Finally, we extract structure-valid regions through automatic motion-driven photo-difference to assist odometry prediction and unsupervised training to improve the accuracy of depth prediction.

2 Methods

This section details our proposed local textural and global spatial-temporal features integration networks (LGIN) with unsupervised learning for monocular endoscopic dense depth and odometry estimation, as shown in Fig 1.

2.1 Hybrid Encoder of LGIN

Given a set of images $\mathcal{V} = \{V_1, \cdot V_t \cdots, V_T\} \in \mathbb{R}^{H\times W\times 3\times T}$ (where each frame V_t is $\in \mathbb{R}^{H\times W\times 3}$ and T represents the number of frames), the hybrid encoder combines dense convolution and pyramid vision transformer to extract local and global features from the image set $\mathcal{V}$.

Dense Convolution. Integrating local textural features into global features is a promising way to predict dense depth with accurate texture and structure.

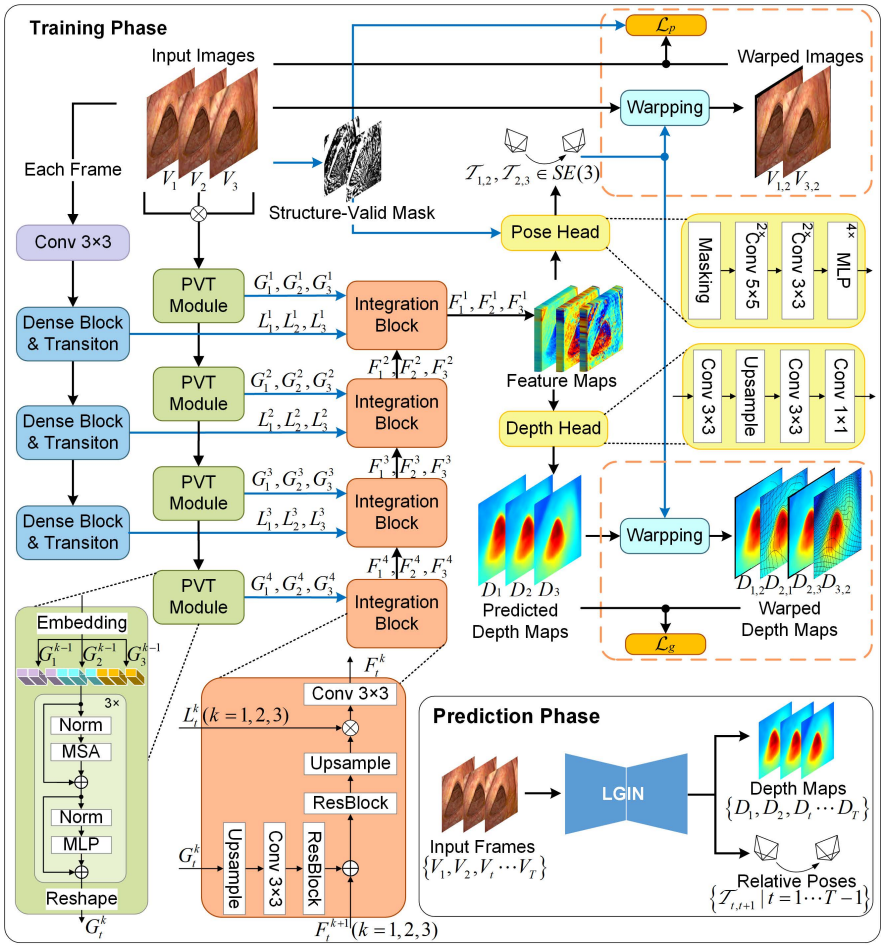

Fig. 1. The overall framework of the proposed LGIN.

To this end, we introduce DenseNet [10] as an encoder due to its excellent local feature reuse and reservation mechanism.

For each frame $V_t \in \mathcal{V}$, the dense encoder first performs a 3×3 convolution to obtain the initial feature map which has C channels, which is then sent to 3 dense blocks with a transition-down module [10] for multi-scale local features extraction. The dense block comprises 4 convolutions with skip connections and a transition-down module (convolution and pooling) for purpose of increasing the receptive field of local features while reducing parameters. After the local feature extraction, we further obtain 3 local feature maps $L_t^1 \in \mathbb{R}^{H/2\times W/2\times(C+4r)}$, $L_t^2 \in \mathbb{R}^{H/4\times W/4\times(C+8r)}$, and $L_t^3 \in \mathbb{R}^{H/8\times W/8\times(C+12r)}$, where r represents the growrate of the dense block. These local textural feature maps are then sent to fuse with global spatial-temporal features.

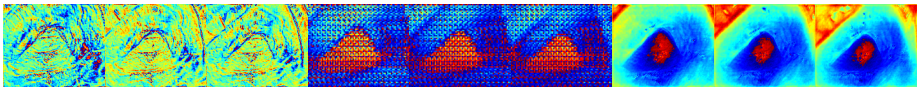

Fig. 2. A sample of local features, global features, and integrated features

Pyramid Vision Transformer. ViT [4] is generally adapted to capture long-dependence relationships and extract global features. It is proven that pyramid vision transformer is more effective than ViT in dense prediction [21]. Hence, We employ PVT as another encoder to extract both global spatial features and temporal information between consecutive frames [23] to perceive global depth range and illumination variations for better training with unsupervised learning.

The PVT encoder is of 4 stages and each PVT module contains patch and position embeddings, transformer blocks, and reshaping, as depicted on the bottom left of Fig 1. The first PVT module embeds the frame set $\mathcal{V}$ into tokens $\in \mathbb{R}^{N\times C_1}$, where $N = H \times W \times T/P^2$ represents the number of patches, P denotes the initial patch size which is set to 8, and each token has C_1 dimensions. After that, it uses three transformer blocks with multi-head self-attention (MSA) and multi-layer perceptron (MLP) for global spatial-temporal feature extraction[4]. Finally, the tokens of frame V_t is reshaped into image-like global feature map $G_t^1 \in \mathbb{R}^{H/8\times W/8\times C_1}$. The following PVT modules are similar but embed the feature map from the previous stage with a patch size of 2 to reduce the resolution of feature maps while increasing the dimension of tokens. In this way, we can obtain the other 3 feature maps $G_t^2 \in \mathbb{R}^{H/16\times W/16\times C_2}$, $G_t^3 \in \mathbb{R}^{H/32\times W/32\times C_3}$, and $G_t^4 \in \mathbb{R}^{H/64\times W/64\times C_4}$.

2.2 Feature Integration Decoder of LGIN

As described in Fig 1, the decoder first uses 4 integration blocks to aggregate 3 local textural features and 4 global spatial-temporal features, and thus obtain integration feature F_t^1 from coarse to fine for each frame

$$F_t^4 = \mathrm{Conv}(\mathcal{U}(\Theta(\Theta(\mathrm{Conv}(\mathcal{U}(G_t^4)))))), \tag{1}$$

$$F_t^i = \mathrm{Conv}(\mathcal{U}(\Theta(F_t^{i+1} \oplus \Theta(\mathrm{Conv}(\mathcal{U}(G_t^i))))) \otimes L_t^i), i = 1, 2, 3 \tag{2}$$

where $\mathcal{U}$, Θ, $\oplus$, $\otimes$, and Conv represent upsample (bilinear interpolation), residual block, addition, concatenation, and 3×3 convolution, respectively. All global feature maps are first upsampled, and the channel numbers are fixed to $\hat{C}$ through Conv. Note that the other $\mathcal{U}$ is placed before concatenation and 3×3 convolution so that local texture features can compensate upsampled global features for coarse granularity. Moreover, we directly add integrated features from the previous layer into the current layer, which enables the decoder to estimate the depth information in a coarse-to-fine mode. A sample of the hybrid feature maps is displayed in Fig. 2, the local features contain more texture information while global features represent the overall depth range and distribution information.

Then the decoder uses dual-heads (depth and pose head) to simultaneously estimate dense depth $\{D_1, D_2, \cdots, D_T\}$ and odometry $\{\mathcal{T}_{1,2}, \mathcal{T}_{2,3}, \cdots, \mathcal{T}_{T-1,T}\}$

from integrated features $\{F_1^1, F_2^1, \cdots, F_T^1\}$. The *yellow* blocks in Fig. 1 illustrate the pose head and depth head in detail. Note that we will calculate a structure-valid mask to eliminate interference from relatively static or untextured irrelevant regions with pose predictions

$$M_{i,j}^{sv}(x,y) = \begin{cases} False & |V_i(x,y) - V_j(x,y)| < \tau \\ True & |V_i(x,y) - V_j(x,y)| \geq \tau \end{cases}, \tag{3}$$

where x, y represents the pixel position in the frame and τ indicates the threshold to distinguish whether each pixel position is static between two frames according to [6]. The obtained mask $M_{i,j}^{sv}$ will be used to extract the structure-valid region in the feature maps F_i^1 and F_j^1 for pose prediction.

2.3 Unsupervised Learning

This work employs unsupervised learning to train our LGIN. Photometric loss is commonly used for unsupervised training which aims to measure the photometric inconsistency between two endoscopic images and their warped images [15]. However, the traditional photometric loss cannot effectively supervise the network due to illumination variations on endoscopic images. Hence, we introduce the minimum photometric supervision [6] to solve this problem. Additionally, to eliminate the influence of relatively static regions and untextured regions on the calculation of photometric error especially in endoscopic images of tubular organs, we only use structure-valid regions of images for calculation. Specifically, we use three consecutive frames to calculate the loss

$$\mathcal{L}_p(V_t, V_i, V_j) = 1 - \sum (M_{t,i}^{sv} \cup M_{i,j}^{sv}) \text{Max}(\Omega(V_{t,i}, V_i), \Omega(V_{j,i}, V_i)), \tag{4}$$

where $\Omega(\cdot)$ is the SSIM function to compute the similarity for accurate photometric supervision, $V_{t,i}$ means the warped image from V_t to V_i. We also introduce geometric consistency loss [2] to ensure the scale consistency and smooth the depth structure

$$\mathcal{L}_g(D_i, D_j) = \sum \frac{(D_{i,j} - D_j)^2}{D_{i,j}^2 + D_j^2} + \sum \frac{(D_{j,i} - D_i)^2}{D_{j,i}^2 + D_i^2}, \tag{5}$$

where $D_{i,j}$ means the warped depth map from two adjacent frames D_i to D_j.

3 Experimental settings

We used two kinds of data in the experiment: Synthetic and clinical colonoscopic video data. The public synthetic data [19] simulate weak texture and illumination variations of colonoscopic images with ground truth depth and odometry. The size of the virtual image is 475×475 and there are 33 image sequences. Each sequence contains 600 frames. The clinical data were recorded during colonoscopies from 80 patients. We manually selected frames with a relatively large

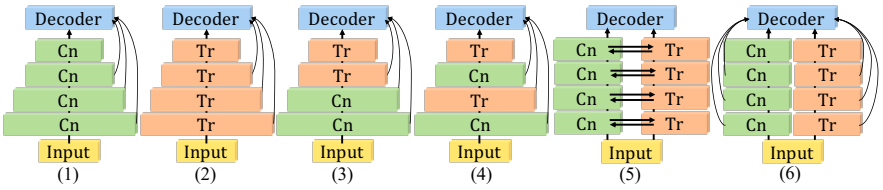

Fig. 3. Six feature extraction manners using the CNN (Cn) and the transformer (Tr).

camera motion for better unsupervised learning. Finally, 160 colonoscopic image sequences were used and each sequence contains about 15 to 20 frames for model training. All images were downsampled to a size of 320 × 256.

For LGIN, we set the first convolution channel C to 48, dense-block growing rate r to 12, and the dimension C_1, C_2, C_3, and C_4 of tokens are 384, 768, 768, and 768, respectively. The channel $\hat{C}$ of global feature maps is fixed to 256 in integration blocks. We used the stochastic gradient descent algorithm as an optimizer with a momentum of 0.9 during training. We divided the dataset into training and testing by a ratio of 7:3 and used cross-validation to obtain diverse experimental results. The balance coefficients of two losses $\mathcal{L}_p$ and $\mathcal{L}_g$ were 0.8 and 0.2. For a fair comparison, we inputted three frames ($T = 3$) at a time to train all models. The learning rate from 10^{-4} to 10^{-3} and the batch size, epoch, and iterations were set to 2, 200, and 500, respectively.

We first compare our LGIN (6) to five convolution or transformer-based depth prediction models: (1) EndoSLAM [15], (2) MonoFormer [1], (3) DPT-H [18], (4) LiteMono [27], and (5) DSCT [3]. Note that models (2)-(5) use PoseNet of EndoSLAM. We can categorize the six models according to different manners of feature extraction as shown in Fig. 3. (1) is convolution-based feature extraction while (2) is transformer-based, (3)-(6) integrate the convolution and transformer for feature extraction. Specifically, (3) uses the transformer to extract global features based on the local feature maps; (4) inserts convolution modules before multi-head self-attention for feature extraction; (5) uses CNN and the transformer for feature extraction and performing feature interaction; and our LGIN (6) uses CNN and the transformer for feature extraction respectively. Then, an ablation study is conducted to verify the effectiveness of two-heads and structure-valid regions extraction in LGIN: (7) LGIN w/o PH: only using depth prediction head and introducing PoseNet of EndoSLAM; and (8) LGIN w/o M^{sv}: removing the structure-valid masking operation of the pose head.

Since there is no ground truth for clinical colonoscopic images, we employ two metrics SSIM and PSNR for quantitative assessment. Specifically, we warp one frame into another by model-estimated camera poses and dense depth maps and calculate SSIM and PSNR between original images and warped images. For the synthesis data, we scale the predicted depth map through a median ratio with the ground truth and then use four classical metrics of absolute relative error (AbsRel), square relative error (SqRel), root mean square error (RMSE), and

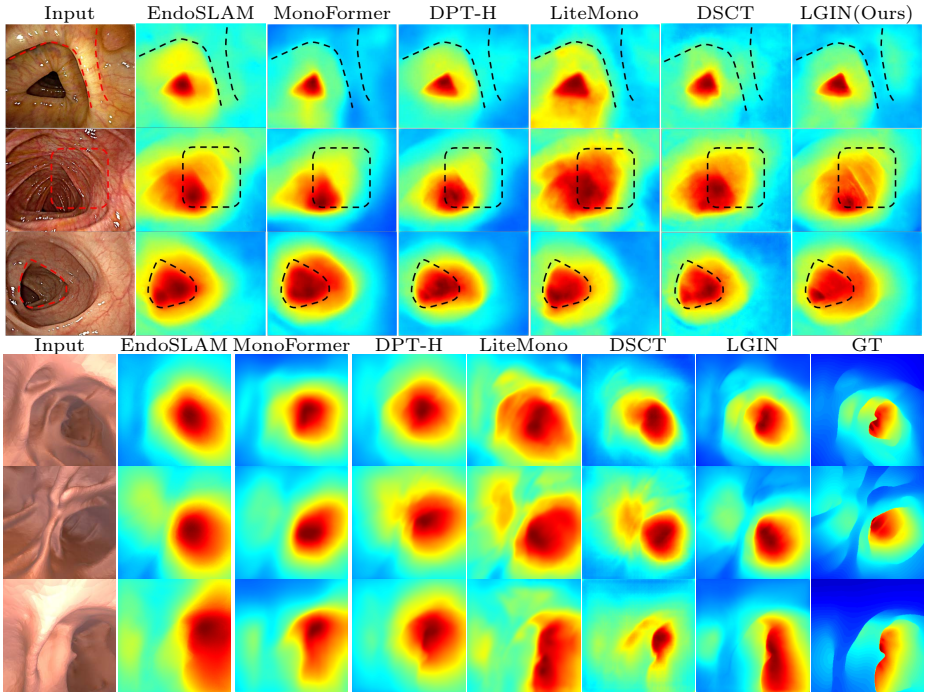

Fig. 4. Qualitative comparison of monocular dense depth estimated by the six models introduced in Experimental settings with the same unsupervised learning method. *Rows* 1∼4 correspond to clinical data and *Rows* 5∼8 correspond to synthetic data.

proportion of distribution consistency δ_t to evaluate the depth accuracy and the popular metrics absolute trajectory error (ATE) and relative pose error (T_{RPE} and R_{RPE}) to evaluate the odometry prediction.

4 Results and Discussion

Fig. 4 displays estimated colonoscopic dense depth maps. We can see the fully convolutional method EndoSLAM [15] limit itself with inaccurate depth distribution and global structure. Transformer-based networks MonoFormer [1] can predict more accurate depth distribution but with insufficient details. DPT-H [18], Litemono [27], and DSCT [3] can also not extract better local texture details and global depth coherence. Our LGIN can generally estimate global depth distribution and depth details in local structures. Table. 1 demonstrates the quantitative assessment results of different methods. EndoSLAM performs not better than other methods. MonoFormer performs better than DPT-H, Lite-Mono, and DSCT for clinical data but not for synthetic data. LGIN w/o M^{sv} is a little better than LGIN w/o PH and Our LGIN outperforms other methods. Fig. 5 demonstrates the local and global feature maps extracted by the convolution and the transformer. All local feature maps extracted by four models show

Table 1. Comparison of quantitative assessment results of using the eight methods.

Datasets	Clinical data		Synthetic colonoscopic data								
Types	Depth		Depth						Pose		
Metrics	SSIM↑	PSNR↑	AbsRel↓	SqRel↓	RMSE↓	δ_1↑	δ_2↑	δ_3↑	$ATE^{(cm)}$↓	$T^{(mm)}_{RPE}$↓	$R^{(\circ)}_{RPE}$↓
EndoSLAM [15]	0.572	19.439	0.281	4.474	12.90	0.561	0.854	0.938	14.268	1.281	2.352
MonoFormer [1]	0.590	19.797	0.250	3.861	12.54	0.624	0.876	0.957	8.755	0.692	1.868
DPT-H [18]	0.588	19.743	0.247	3.278	12.73	0.613	0.861	0.958	11.212	0.759	1.913
LiteMono [27]	0.582	19.776	0.256	3.568	13.18	0.621	0.869	0.955	12.439	0.816	2.236
DSCT [3]	0.583	19.654	0.245	3.212	12.56	0.625	0.875	0.956	12.981	0.701	1.871
LGIN w/o PH	0.591	19.875	0.207	2.869	11.38	0.706	0.893	0.968	10.251	0.682	1.922
LGIN w/o M^{sv}	0.594	19.882	0.198	2.380	10.55	**0.710**	0.914	0.982	6.922	0.452	1.547
LGIN	**0.605**	**20.012**	**0.188**	**2.279**	**10.03**	0.709	**0.926**	**0.987**	**6.375**	**0.385**	**1.424**

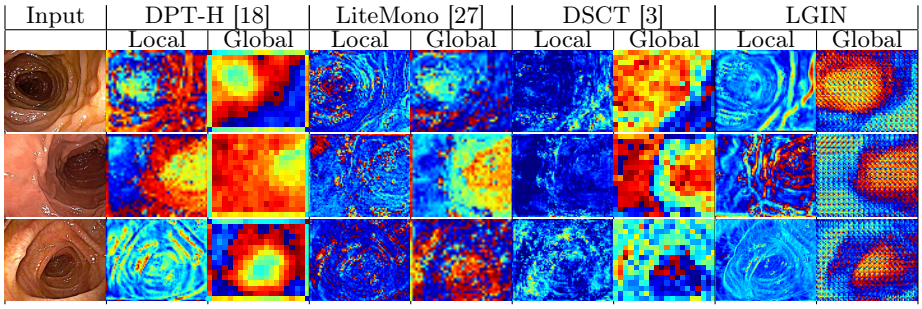

Fig. 5. The local and global feature maps extracted by the convolution and transformer of DPT-H [18], Litemono [27], DSCT [3], and our LGIN. We selected and colored some feature maps to show. More can be seen in the supplementary material.

sufficient local texture information. Our extracted global features can represent the depth distribution information better than the other three models.

This work aims to integrate the advantages of CNN and the transformer to propose an effective depth and odometry prediction model for monocular endoscopic images. The effectiveness of our methods is discussed as follows. First, LGIN creates a hybrid encoder to extract abundant local textural features and global spatial-temporal features in parallel and employs a powerful decoder to integrate these local and global features. Such a way of feature extraction and fusion is more effective than DPT-H [18], Litemono [27], and DSCT [3]. Second, the feature-shared way of the dual-head mechanism improves the relevance and accuracy of depth and pose prediction in unsupervised learning. Finally, structure-valid region extraction improves the accuracy of odometry prediction by eliminating interference from irrelevant areas. Additionally, we use transformers to extract temporal features and introduce minimum photometric loss can reduce the influences of illumination variations.

Our method inevitably suffers from some limitations. Firstly, although our model can generally predict accurate dense depth for monocular endoscopic images, accurate odometry estimation is still a challenge for deep learning models. Secondly, the computational efficiency of our model should be improved to meet

real-time applications. Our future work will explore a lightweight network encoder for convolutional extraction and acceleration of multi-head self-attention. In summary, this work proposes a new deep learning model LGIN with unsupervised learning for accurate endoscopic dense depth prediction.

Acknowledgement This work was supported partly by the National Natural Science Foundation of China under Grants 82272133 and the Fujian Provincial Technology Innovation Joint Funds under Grant 2019Y9091.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bae, J., Moon, S., Im, S.: Deep digging into the generalization of self-supervised monocular depth estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 187–196 (2023)
2. Bian, J.W., Zhan, H., Wang, N., Li, Z., Zhang, L., Shen, C., Cheng, M.M., Reid, I.: Unsupervised scale-consistent depth learning from video. International Journal of Computer Vision **129**(9), 2548–2564 (2021)
3. Chen, M., Zhang, L., Feng, R., Xue, X., Feng, J.: Rethinking local and global feature representation for dense prediction. Pattern Recognition **135**, 109168 (2023)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR). pp. 1–21 (2021)
5. Fan, W., Zhang, K., Shi, H., Chen, J., Chen, Y., Luo, X.: Deep triple-supervision learning unannotated surgical endoscopic video data for monocular dense depth estimation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
6. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3828–3838 (2019)
7. Gottlieb, K., Daperno, M., Usiskin, K., Sands, B.E., Ahmad, H., Howden, C.W., Karnes, W., Oh, Y.S., Modesto, I., Marano, C., et al.: Endoscopy and central reading in inflammatory bowel disease clinical trials: achievements, challenges and future developments. Gut **70**(2), 418–426 (2021)
8. Han, W., Yin, J., Jin, X., Dai, X., Shen, J.: Brnet: Exploring comprehensive features for monocular depth estimation. In: European Conference on Computer Vision. pp. 586–602. Springer (2022)
9. Huang, B., Zheng, J.Q., Nguyen, A., Xu, C., Gkouzionis, I., Vyas, K., Tuch, D., Giannarou, S., Elson, D.S.: Self-supervised depth estimation in laparoscopic image using 3d geometric consistency. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 13–22. Springer (2022)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4700–4708 (2017)

11. Li, W., Hayashi, Y., Oda, M., Kitasaka, T., Misawa, K., Mori, K.: Multi-view guidance for self-supervised monocular depth estimation on laparoscopic images via spatio-temporal correspondence. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 429–439. Springer (2023)
12. Liu, X., Sinha, A., Ishii, M., Hager, G.D., Reiter, A., Taylor, R.H., Unberath, M.: Dense depth estimation in monocular endoscopy with self-supervised learning methods. IEEE Transactions on Medical Imaging **PP**(99), 1–1 (2019)
13. Liu, Y., Zuo, S.: Self-supervised monocular depth estimation for gastrointestinal endoscopy. Computer Methods and Programs in Biomedicine p. 107619 (2023)
14. Ma, R., Wang, R., Zhang, Y., Pizer, S., McGill, S.K., Rosenman, J., Frahm, J.M.: Rnnslam: Reconstructing the 3d colon to visualize missing regions during a colonoscopy. Medical Image Analysis **72**, 102100 (2021)
15. Ozyoruk, K.B., Gokceler, G.I., Bobrow, T.L., Coskun, G., Incetan, K., Almalioglu, Y., Mahmood, F., Curto, E., Perdigoto, L., Oliveira, M., et al.: Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Medical Image Analysis **71**, 102058 (2021)
16. Papa, L., Russo, P., Amerini, I.: Meter: a mobile vision transformer architecture for monocular depth estimation. IEEE Transactions on Circuits and Systems for Video Technology (2023)
17. Piccinelli, L., Sakaridis, C., Yu, F.: idisc: Internal discretization for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21477–21487 (2023)
18. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12179–12188 (2021)
19. Rau, A., Bhattarai, B., Agapito, L., Stoyanov, D.: Bimodal camera pose prediction for endoscopy. IEEE Transactions on Medical Robotics and Bionics (2023)
20. Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., Zhang, B.: Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. Medical Image Analysis **77**, 102338 (2022)
21. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 568–578 (2021)
22. Wang, Y., Shi, M., Li, J., Huang, Z., Cao, Z., Zhang, J., Xian, K., Lin, G.: Neural video depth stabilizer. arXiv preprint arXiv:2307.08695 (2023)
23. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8741–8750 (2021)
24. Yang, Z., Pan, J., Dai, J., Sun, Z., Xiao, Y.: Self-supervised lightweight depth estimation in endoscopy combining cnn and transformer. IEEE Transactions on Medical Imaging (2024)
25. Yuan, W., Gu, X., Li, H., Dong, Z., Zhu, S.: Monocular scene reconstruction with 3d sdf transformers. arXiv preprint arXiv:2301.13510 (2023)
26. Yue, H., Gu, Y.: Tcl: Triplet consistent learning for odometry estimation of monocular endoscope. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 144–153. Springer (2023)
27. Zhang, N., Nex, F., Vosselman, G., Kerle, N.: 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. pp. 18537–18546 (2023)

28. Zheng, Q., Yu, T., Wang, F.: Dcu-net: Self-supervised monocular depth estimation based on densely connected u-shaped convolutional neural networks. Computers & Graphics **111**, 145–154 (2023)