Unsupervised Structure-Geometric Consistency for Monocular Endoscopic Depth Overestimation

Wenkang Fan[1], Enqi Qiu[2], Hongzhi Xu[2], and Xiongbiao Luo[1,3,4,*]

[1] Department of Computer Science and Engineering, Xiamen University, Xiamen 361102, China
[2] Zhongshan Hospital, Xiamen University, Xiamen 361004, China
[3] National Institute for Data Science in Health and Medicine, School of Medicine, Xiamen University, Xiamen 361102, China
[4] Discipline of Intelligent Instrument and Equipment, Xiamen University, Xiamen 361102, China
xiongbiao.luo@gmail.com, Asterisk indicates the corresponding author

Abstract. Monocular endoscopic depth estimation is a key to expand the surgical field and visually navigate the endoscope, augmenting the perception of surgeons and reducing inadvertent damages during robotic surgery. Unfortunately, current deep learning methods still suffer from a limited field of view, moving and limited artificial optic-fiber light sources (illumination variations), and weak textures or structures in monocular endoscopic video images collected from complex surgical scenarios, as well as they also get trapped in depth overestimation. This work first explores a small deep learning model of densely convolved pyramid transformer to simultaneously predict monocular depth and pose of the endoscope without using any annotation data. Specifically, this small model employs dense convolution and hierarchical transformer to encode multiscale local and global features, while it uses residual attention to effectively fuse or decode these features. Then, a photometric structure-aware consistency mechanism is introduced to deal with the problems of weak texture and depth overestimation, refining endoscopic depth and pose estimation. We evaluated our methods on both synthetic and clinical colonoscopic video images, with the experimental results showing that our unsupervised learning methods can attain higher accurate depth distribution and more sufficient textures, and better qualitative and quantitative results than state-of-the-art monocular depth estimation models.

Keywords: Monocular depth estimation · Unsupervised learning · Vision transformer · 3D reconstruction · Endoscopic vision.

1 Introduction

Robotic-assisted minimally invasive surgery routinely uses endoscopes to visually diagnose and treat various diseases in the body. Augmented endoscopic vision is essential to enhance the perception of surgeons, improve surgical outcomes and reduce surgical risks and operating times during robotic surgery. Monocular endoscopic depth estimation for 3D reconstruction of surgical fields is a promising way to augment surgical vision for precise surgical navigation.

Deep learning approaches are increasingly used for monocular depth estimation. Supervised learning requires a large amount of annotated data, which is particularly unrealistic or impractical for monocular endoscopic videos acquired from the operating room. Many researchers work on self-supervised learning methods [7, 2] that use sparse depth supervision (e.g., structure from motion) for depth prediction, but they depend critically on the quality of sparse reconstruction. Recently, unsupervised learning approaches are attractively discussed to simultaneously estimate dense depth and camera poses [8, 10, 4, 12, 6, 11, 5]. Unfortunately, these unsupervised methods still suffer from illumination variations (caused by moving and limited artificial optic-fiber light sources), weak textures or structures, specific endoscope movements, artifacts (e.g., bleeding) in endoscopic images, leading to depth estimation uncertainty and overestimation.

The motivation of this work is to explore unsupervised learning methods to precisely predict monocular endoscopic depth and address the problem of depth overestimation. Several technical contributions are clarified as follows. First, we construct a simple, small deep learning model of densely convolved hierarchical transformer. Specifically, this model combines dense convolution and hierarchical transformer in a parallel way that can obtain more sufficient and accurate local texture features and global depth distribution features, while it can effectively integrate local and global features in a coarse-to-fine mode to simultaneously estimate monocular endoscopic dense depth and pose in a feature-shared way that can improve the relevance of depth and pose prediction. More interestingly, we discover, demonstrate and address a problem of unsupervised monocular endoscopic depth overestimation. Monocular endoscopic depth prediction typically suffers from insufficient or weak texture, illumination variations, characteristics of tubular organs and camera movements in the body, leading to depth overestimation. Then, we formulate a new multiframe photometric structure-aware consistency function to address depth overestimation in static regions caused by camera movements, and a 3D geometric consistency function to supervise the depth prediction in structureless or untextured regions.

2 Methods

This section details our unsupervised learning model for precise monocular endoscopic depth and pose estimation, while a photometric structure-geometric consistency constraint is introduced to address the problem of overestimation.

2.1 Monocular Depth and Pose Estimation

Densely Convolved Hierarchical Transformer We explore a simple, small deep learning architecture of densely convolved hierarchical transformer (DCHT) for unsupervised monocular endoscopic depth and pose estimation. DCHT mainly consists of hybrid encoders and fusion decoders for feature representation.

The encoder combines a dense convolution block (DCB) and a hierarchical transformer block (HTB) to extract multiscale local and global features. Dense convolution is the capacity of excellent local feature reuse and reservation. DCB contains 4 convolutions with skip connections and a transition-down module (convolution and pooling) that can increase the receptive field of local features and reduce model parameters. HTB can capture long-dependence relationships and extract global features [1]. It mainly includes patch and position embeddings, a transformer block, and reshaping, extracting both global spatial features and temporal information between consecutive frames and perceiving global depth range and illumination variations for robust training with unsupervised learning.

The decoder uses four residual attention fusion (RAF) blocks to aggregate three local features and four global feature maps. RAF fuses local features $\mathbf{X}_k^i$ and global features $\mathbf{H}_k^i$ to generate feature map $\mathbf{F}_k^i$ of frame k at stage i

$$\mathbf{F}_k^4 = \mathrm{Conv}(\mathrm{US}(\mathrm{RB}(\mathrm{RB}(\mathrm{Conv}(\mathrm{US}(\mathbf{H}_k^4)))))), \tag{1}$$

$$\mathbf{F}_k^i = \mathrm{Conv}(\mathrm{US}(\mathrm{RB}(\mathbf{F}_k^{i+1} \oplus \hat{\mathbf{F}}_k^i)) \otimes \mathbf{X}_k^i), i = 1, 2, 3, \tag{2}$$

where $\hat{\mathbf{F}}_k^i = \mathrm{RB}(\mathrm{Conv}(\mathrm{US}(\mathbf{H}_k^i)))$; US, RB, Conv, $\oplus$ and $\otimes$ are upsample (bilinear interpolation), residual block, 3×3 convolution, addition and concatenation, respectively. Note that US is placed before concatenation and 3×3 convolution at stage i (i=1,2,3) so that local texture features can compensate upsampled global features for coarse granularity. Moreover, we directly add fused features from the previous layer into the current layer, which enables the decoder to estimate the depth information in a coarse-to-fine mode.

Unsupervised Training We train DCHT in an unsupervised mode. Photometric consistency constraint (PCC) loss is commonly used for unsupervised training [8]. We introduce minimum PCC loss $\mathcal{L}_P(\mathbf{I}_{k-1}, \mathbf{I}_k, \mathbf{I}_{k+1})$ to deal with illumination variations between three consecutive frames $\mathbf{I}_{k-1}$, $\mathbf{I}_k$, $\mathbf{I}_{k+1}$:

$$\mathcal{L}_P = \sum_p \min(|\mathbf{I}_{k-1}^k - \mathbf{I}_k|, |\mathbf{I}_{k+1}^k - \mathbf{I}_k|), \tag{3}$$

where p is a pixel in $\mathbf{I}_k$, warped images $\mathbf{I}_{k-1}^k$ and $\mathbf{I}_{k+1}^k$. Moreover, we introduce geometric consistency constrain (GCC) loss $\mathcal{L}_G^{2D}(\mathbf{D}_{k-1}, \mathbf{D}_k, \mathbf{D}_{k+1})$ between three consecutive 2D depth maps $\mathbf{D}_{k-1}$, $\mathbf{D}_k$, $\mathbf{D}_{k+1}$ to smooth the depth structure:

$$\begin{aligned}\mathcal{L}_G^{2D} = &\sum_p \frac{(\mathbf{D}_k^{k-1} - \mathbf{D}_{k-1})^2}{(\mathbf{D}_k^{k-1})^2 + \mathbf{D}_{k-1}^2} + \sum_p \frac{(\mathbf{D}_{k-1}^k - \mathbf{D}_k)^2}{(\mathbf{D}_{k-1}^k)^2 + \mathbf{D}_k^2} \\ &+ \sum_p \frac{(\mathbf{D}_{k+1}^k - \mathbf{D}_k)^2}{(\mathbf{D}_{k+1}^k)^2 + \mathbf{D}_k^2} + \sum_p \frac{(\mathbf{D}_k^{k+1} - \mathbf{D}_{k+1})^2}{(\mathbf{D}_k^{k+1})^2 + \mathbf{D}_{k+1}^2},\end{aligned} \tag{4}$$

where $\mathbf{D}_\alpha^\beta$ means the warped depth map from $\mathbf{D}_\alpha$ to $\mathbf{D}_\beta$.

2.2 Overestimation and Solution

This section proposes an overestimation problem and shows a solution to address it. Fig. 1 illustrates our framework to solve the problem of overestimation.

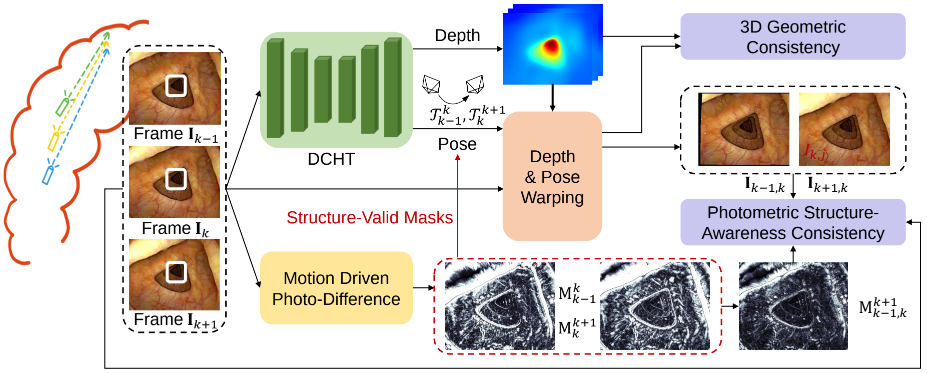


Fig. 1. Our unsupervised structure-geometric learning to address overestimation.

Overestimation Problem To calculate PCC, we warp two frames: $\mathbf{I}_{k-1} \rightarrow \mathbf{I}_k$ and $\mathbf{I}_{k+1} \rightarrow \mathbf{I}_k$). Let $p = [u_k, v_k, 1]^{\mathrm{T}}$ (T is transpose), d_j, $\mathbf{Q}$ and $[\mathbf{R}_{k-1}^{k}\ \mathbf{P}_{k-1}^{k}]$ be a pixel in frame $\mathbf{I}_k$, its depth, camera intrinsic matrix and relative pose (rotation matrix $\mathbf{R}_{k-1}^{k}$ and position $\mathbf{P}_{k-1}^{k}$) between two consecutive frames, we compute 3D point $[x_k^{k-1}, y_k^{k-1}, z_k^{k-1}]^{\mathrm{T}}$ of pixel p in the camera coordinate system:

$$\begin{bmatrix} x_k^{k-1} \\ y_k^{k-1} \\ z_k^{k-1} \end{bmatrix} = (\mathbf{R}_k^{k-1})^{\mathrm{T}} \left(\mathbf{Q}^{-1} d_k \begin{bmatrix} u_k \\ v_k \\ 1 \end{bmatrix} - \mathbf{P}_k^{k-1} \right), \tag{5}$$

which is reprojected on $\mathbf{I}_{k-1}$ to obtain its pixel $[u_k^{k-1}, v_k^{k-1}, 1]^{\mathrm{T}}$ and depth d_k^{k-1}

$$d_k^{k-1} \begin{bmatrix} u_k^{k-1} \\ v_k^{k-1} \\ 1 \end{bmatrix} = (\mathbf{R}_{k-1}^{k})^{\mathrm{T}} \left(d_k \begin{bmatrix} u_k \\ v_k \\ 1 \end{bmatrix} + \mathbf{Q}\mathbf{P}_{k-1}^{k} \right). \tag{6}$$

Eqs. (5) and (6) generates warped images $I_{i,j}$ to compute PCC loss (Eq. (3)).

In endoscopic surgery, surgeons usually move the endoscopic camera along the centerlines of the tubular organs in the body and barely change the direction of the endoscope. In this way, we can observe some regions orthogonal to the direction of the endoscope trajectory (white-square regions in Fig. 1) are relatively static in consecutive endoscopic frames. These static regions will cause a problem of overestimation in the PCC calculation, i.e., unsupervised deep models usually estimate large depth for these regions. In warping, pixel p in these static regions should be reprojected at the same position from one to another frame, ensuring to accurately compute PCC, i.e., $[u_k^{k-1}, v_k^{k-1}, 1]$ should be equal to $[u_k, v_k, 1]$ in these regions. Actually, this equality is violated in the PCC calculation.

Since surgeons barely change the direction of the endoscope in moving, i.e., $\mathbf{R}_{k-1,k}$ can be considered as the identity matrix, we rewrite Eq. (6) as

$$d_k^{k-1}\begin{bmatrix}u_k^{k-1}\\ v_k^{k-1}\\ 1\end{bmatrix} \approx d_k \begin{bmatrix}u_k\\ v_k\\ 1\end{bmatrix} + \mathbf{Q}\mathbf{P}_{k-1}^k. \tag{7}$$

To let $[u_k^{k-1}, v_k^{k-1}, 1]$ approximate $[u_k, v_k, 1]$ in Eq. (7) for precise PCC calculation, either $\mathbf{P}_{k-1}^k$ approximates 0 or d_k approximates ∞. Obviously, $\mathbf{P}_{k-1}^k \rightarrow 0$ is certainly violated since the endoscope is moving. So, PCC-trained models can only estimate infinite depth ($d_k \rightarrow \infty$) for these relatively static regions in consecutive images to satisfy Eq. (7), ensuring that PCC loss converges to zero. This is the problem of overestimation, i.e., PCC-trained models try to estimate pixel depth d_k in these static regions to ∞: $d_k \rightarrow \infty$. Additionally, the deepest regions can affect the other regions since unsupervised models usually integrate the smoothness loss with PCC in training, which also leads to depth overestimation.

To address this problem, while introducing a 3D geometric consistency constraint to retain DCHT, we typically employ structural masks to reformulate Eq. (3) to define a new loss of photometric structure-aware consistency (PSC).

Photometric Structure-Aware Consistency To build PSC, we first use a convenient motion driven photo-difference method to identify these static regions in frames $\mathbf{I}_{k-1}$ and $\mathbf{I}_k$ to generate structure mask $\mathbf{M}_{k-1}^k(p)$

$$\mathbf{M}_{k-1}^k(p) = \begin{cases} \frac{|\mathbf{I}_{k-1}(p)-\mathbf{I}_k(p)|}{\tau} & |\mathbf{I}_{k-1}(p) - \mathbf{I}_k(p)| < \tau \\ 1 & |\mathbf{I}_{k-1}(p) - \mathbf{I}_k(p)| \geq \tau \end{cases}, \tag{8}$$

where $\mathbf{M}_{k-1}^k(p)$ ranges in $[0,1]$ and threshold τ restricts the maximum of photo-difference $|\mathbf{I}_{k-1}(p) - \mathbf{I}_k(p)|$ since there are many reflective highlight points on endoscopic images where the photo-difference is too large. $\mathbf{M}_{k-1}^k(p) < 1$ means that the photo-difference at pixel p is small, and p is considered as static and excluded from the PCC. Similarly, we compute $\mathbf{M}_k^{k+1}(p)$ from frames k and $k+1$.

Based on structure-valid masks $\mathbf{M}_{k-1}^k(p)$ and $\mathbf{M}_k^{k+1}(p)$, we reformulate Eq. (3) and define the PSC loss $\mathcal{L}_{PS}(\mathbf{I}_{k-1}, \mathbf{I}_k, \mathbf{I}_{k+1})$ to compute the photometric consistency between three consecutive frames $\mathbf{I}_{k-1}$, $\mathbf{I}_k$, $\mathbf{I}_{k+1}$:

$$\mathcal{L}_{PS} = \sum_p \mathbf{M}_{k-1,k}^{k+1}(p)\mathit{\Theta}(p), \mathbf{M}_{k-1,k}^{k+1}(p) = \min(\mathbf{M}_{k-1}^k(p), \mathbf{M}_k^{k+1}(p)), \tag{9}$$

$$\mathit{\Theta}(p) = \min(1 - \mathit{\Omega}(\mathbf{I}_{k-1}^k, \mathbf{I}_k), 1 - \mathit{\Omega}(\mathbf{I}_{k+1}^k, \mathbf{I}_k)), \tag{10}$$

where we further employ the structural similarity index measure $\mathit{\Omega}(\cdot)$ to compute the similarity for photometric supervision. Note that structure-valid masks $\mathbf{M}_{k-1}^k(p)$ and $\mathbf{M}_k^{k+1}(p)$ also exclude weak-texture or structureless regions since these structureless regions are insensitive to camera movement, deteriorating the PCC calculation during DCHT training.

3D Geometric Consistency While $\mathcal{L}_{PS}$ can accurately supervise the depth prediction of structure-valid regions while the depth of structureless regions is unsupervised. Hence, we introduce a 3D geometric consistency loss $\mathcal{L}_G^{3D}$ to predict depth more smoothly. By the camera intrinsic matrix $\mathbf{Q}$, estimated depth maps $\mathbf{D}_k^{k-1}$, $\mathbf{D}_{k-1}$, $\mathbf{D}_{k-1}^k$, $\mathbf{D}_k$ $\mathbf{D}_{k+1}^k$, $\mathbf{D}_k^{k+1}$, $\mathbf{D}_{k+1}$, and poses $\mathcal{T}_k^{k-1}$, $\mathcal{T}_{k-1}$, $\mathcal{T}_{k-1}^k$, $\mathcal{T}_k$, $\mathcal{T}_k^{k+1}$, $\mathcal{T}_{k+1}^k$, $\mathcal{T}_{k+1}$, we generate seven 3D point sets $\{V_k^{k-1}\}$, $\{V_{k-1}\}$, $\{V_{k-1}^k\}$, $\{V_k\}$, $\{V_{k+1}^k\}$, $\{V_k^{k+1}\}$, $\{V_{k+1}\}$. Similar to GCC, we define $\mathcal{L}_G^{3D}$ as computing Euclidean distance between these 3D point sets:

$$\begin{aligned}\mathcal{L}_G^{3D} = &\sum_p \|V_k^{k-1} - V_{k-1}\| + \sum_p \|V_{k-1}^k - V_k\| \\ &+ \sum_p \|V_{k+1}^k - V_k\| + \sum_p \|V_k^{k+1} - V_{k+1}\|.\end{aligned} \tag{11}$$

Our experiments will demonstrate that $\mathcal{L}_G^{3D}$ generates more smooth (less noises) depth maps than $\mathcal{L}_G^{2D}$. Eventually, the total loss $\mathcal{L}_{total}$ combines $\mathcal{L}_{PS}$ and $\mathcal{L}_G^{3D}$

$$\mathcal{L}_{total} = \mu\mathcal{L}_{PS} + (1-\mu)\mathcal{L}_G^{3D}, \tag{12}$$

where μ is a weight to balance two training-loss functions.

3 Validation

We used a public synthetic colonoscopic database [9] and our in-house clinical colonoscopic database to evaluate our methods. The public synthetic database was annotated as ground truth depth for quantitative evaluation. It simulated different lighting conditions and tissue textures of colonoscopic scenarios, and includes 33 colonoscopic video sequences (600 frames per sequence). Our in-house colonoscopic database was built by collecting clinical monocular colonoscopic videos from the operating room in different medical centers. While parameters $\tau = 30$ and $\lambda = 0.8$, we set the learning rate from 10^{-6} to 10^{-5} and used the stochastic gradient descent algorithm as an optimizer with a momentum of 0.9 during training. The batch size, epoch, and iterations were set to 2, 300, and 500, respectively. We divided all data into 7:3 for training and testing.

We train our DCHT model in two ways of (1) DCHT1 using $\mathcal{L}_P$ and $\mathcal{L}_G^{2D}$ and (2) DCHT2 using $\mathcal{L}_{total}$ (Eq. 12) to demonstrate the problem of overestimation and the effectiveness of $\mathcal{L}_{PS}$ (Eq. 9) to address it. Moreover, we compare our method with the following unsupervised learning approaches: (1) Modepth2 [3], (2) EndoSL [8], (3) M3depth [4], (4) AppeFlow [10] introduces an auxiliary module to predict appearance flow to compensate for illumination variations, and (5) TCL [12]. We qualitatively visualize and compare predicted depth maps for our in-house colonoscopic data, while use five classical metrics of absolute relative error (Abs Rel), square relative error (Sq Rel), root mean square error (RMSE), RMSE log, and proportion of distribution consistency δ_* to quantitatively evaluate the estimated depth results from the synthesis database.

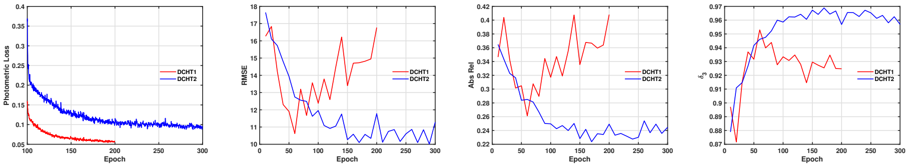


Fig. 2. Training loss, RMSE, Abs Rel and δ_3 curves generated by DCHT1 and DCHT2.

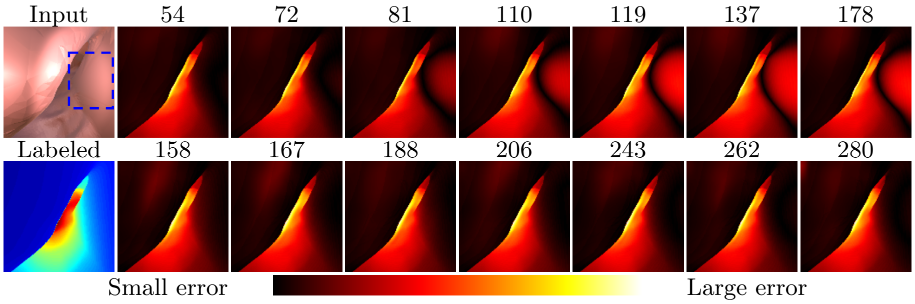


Fig. 3. Visually compared depth-error maps of DCHT1 (Row 1) and DCHT2 (Row 2) tested on the synthetic database: The numbers above the maps denote training epochs.

4 Results and Discussion

Results. Fig. 2 shows the changes of the training loss, RMSE, AbsRel, and δ_3 when increasing the epochs in DCHT1 and DCHT2 training. Both DCHT1 and DCHT2 were trained to gradually converge by PCC or PSC. While RMSE and Abs Rel of DCHT1 were initially descending and then ascending, the errors of DCHT2 were always descending. Moreover, DCHT2 attains much higher accuracy δ_3 than DCHT1. Fig. 3 compares the estimated depth-error maps of DCHT1 and DCHT2 when increasing the training epochs. Obviously, the predicted depth error of DCHT1 becomes larger and larger when the epochs increase. This is because DCHT1 overestimates the depth in the static regions (blue rectangle). DCHT2 remains almost the same depth error when increasing the epochs since it does not overestimate the depth in the static regions. Fig. 4 further compares the estimated depth maps and depth errors of using DCHT1 and DCHT2 tested on the public database. It can be observed that the movement of the colonoscope introduces static regions (blue rectangle). While DCHT1 overestimated the depth in the blue rectangle with large depth errors, DCHT2 precisely estimated the depth with small depth errors.

Fig. 5 compares the predicted depth maps of using the six unsupervised learning methods. EndoSL [8], M3depth [4] and AppeFlow [10] generate overestimated and incorrect depth maps in some regions in both public and in-house colonoscopic images. Modepth2 [3] and TCL [12] work better than EndoSL [8],

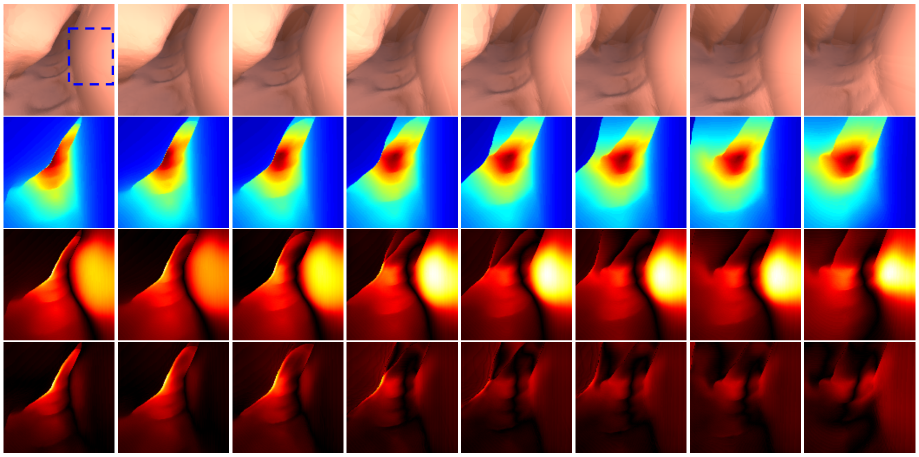


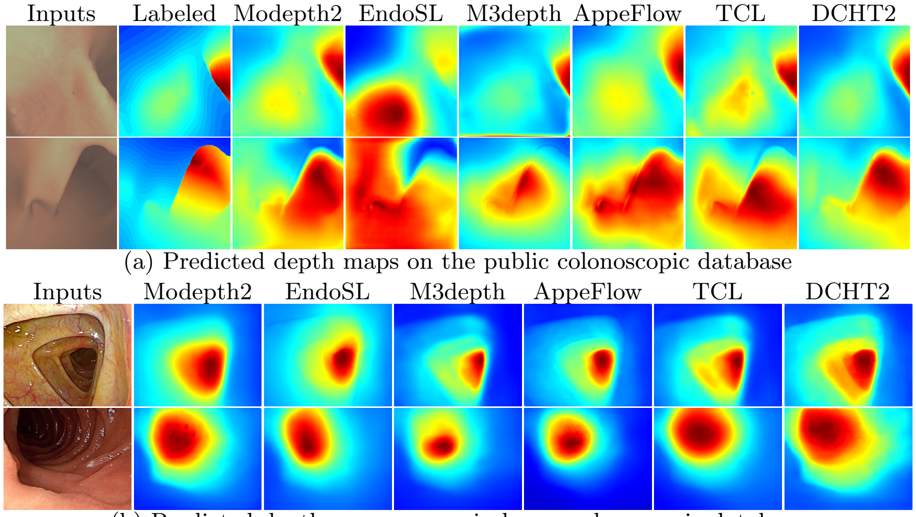
Fig. 4. DCHT1 and DCHT2 predicted depth-error maps: Rows 1∼4 correspond to input synthesis or virtual images, ground truth, DCHT1's and DCHT2's error maps.

Table 1. Comparison of the quantitative results tested on the public synthetic colonoscopic database: *Magenta* and *blue* indicate the better and best results, respectively.

Methods	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ_1↑	δ_2↑	δ_3↑
Modepth2 [3]	0.257	3.081	10.789	0.309	0.609	0.858	0.946
EndoSL [8]	0.331	4.302	12.973	0.493	0.432	0.760	0.875
M3depth [4]	0.323	4.197	12.620	0.484	0.453	0.762	0.887
AppeFlow [10]	0.295	3.793	11.298	0.376	0.519	0.784	0.903
TCL [12]	0.275	3.651	11.811	0.337	0.537	0.812	0.935
DCHT1	0.291	3.704	11.620	0.384	0.525	0.779	0.897
DCHT2	0.228	2.747	10.267	0.255	0.689	0.907	0.967

M3depth [4] and AppeFlow [10]. Our DCHT2 model significantly outperforms the other methods. It introduces very few depth errors in the public synthetic data and predicts more reasonable depth in our clinical data. Particularly, Moreover, DCHT2 can estimate more coherent depth distribution and more texture details (e.g., intestinal folds) than other methods. Table 1 lists the quantitative assessment results. They are generally consistent with the qualitative results.

Discussion The effectiveness of our method lies in: (1) DCHT can encode both local textural and global spatial-temporal features and effectively fuse them for monocular endoscopic depth and pose estimation and (2) the photometric consistency integrates with structure-valid masks to formulate a new photometric structure-aware consistency loss function, which can successfully deal with the problems of depth overestimation, illumination variations and weak texture. Our method still suffers from certain limitations. First, our method gets trapped in incorrect monocular endoscopic pose estimation. Next, our structure-valid mask



(a) Predicted depth maps on the public colonoscopic database

(b) Predicted depth maps on our in-house colonoscopic database

Fig. 5. Visual comparison of the monocular endoscopic depth maps predicted by Modepth2 [3], EndoSL [8], M3depth [4], AppeFlow [10], TCL [12] and DCHT2 tested on the public synthetic and in-house clinical colonoscopic sequences

extraction method is problematic. Moreover, the mask extraction method also introduces non-static regions, which possibly deteriorate unsupervised training.

Conclusion This work formulates a problem of depth overestimation and demonstrates it leads to largely incorrect and inaccurate depth prediction. We proposed unsupervised structure-geometric learning to successfully address the problems of depth overestimation, weak textures, and illumination variations.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 82272133, in part by the High-Quality Development Science and Technology Major Project of Xiamen Health Commission under Grant 2024GZL-ZD03, and in part by Ningbo 2035 Key Research and Development Program under Grant 2024Z127.

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

References

1. Dong, B., Wang, W., Fan, D.P., et al.: Polyp-PVT: Polyp segmentation with pyramid vision transformers. CAAI Artificial Intelligence Research **2**, 9150015 (2023)
2. Fan, W., Jiang, W., Shi, H., Zeng, H.Q., Chen, Y., Luo, X.: Triple-supervised convolutional transformer aggregation for robust monocular endoscopic dense depth estimation. IEEE Transactions on Medical Robotics and Bionics (2024)

3. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3828–3838 (2019)
4. 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)
5. Ke, B., Obukhov, A., Huang, S., et al.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9492–9502 (2024)
6. 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)
7. 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)
8. 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)
9. Rau, A., Bhattarai, B., Agapito, L., Stoyanov, D.: Bimodal camera pose prediction for endoscopy. IEEE Transactions on Medical Robotics and Bionics (2023)
10. 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)
11. Yang, X., Ma, Z., Ji, Z., et al.: Gedepth: Ground embedding for monocular depth estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12719–12727 (2023)
12. 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)