Image Intrinsic-Based Unsupervised Monocular Depth Estimation in Endoscopy

Bojian Li, Bo Liu, Miao Zhu, Xiaoyan Luo and Fugen Zhou

***Abstract*— Unsupervised monocular depth estimation plays a vital role for endoscopy-based minimally invasive surgery (MIS). However, it remains challenging due to the distinctive imaging characteristics of endoscopy which disrupt the assumption of photometric consistency, a foundation relied upon by conventional methods. Distinct from recent approaches taking image pre-processing strategy, this paper introduces a pioneering solution through intrinsic image decomposition (IID) theory. Specifically, we propose a novel end-to-end intrinsic-based unsupervised monocular depth learning framework that is comprised of an image intrinsic decomposition module and a synthesis reconstruction module. This framework seamlessly integrates IID with unsupervised monocular depth estimation, and dedicated losses are meticulously designed to offer robust supervision for network training based on this novel integration. Noteworthy, we rely on the favorable property of the resulting albedo map of IID to circumvent the challenging images characteristics instead of pre-processing the input frames. The proposed method is extensively validated on SCARED and Hamlyn datasets, and better results are obtained than state-of-the-art techniques. Beside, its generalization ability and the effectiveness of the proposed components are also validated. This innovative method has the potential to elevate the quality of 3D reconstruction in monocular endoscopy, thereby enhancing the accuracy and robustness of augmented reality navigation technology in MIS. Our code will be available at: https://github.com/bobo909/IID-SfmLearner.**

***Index Terms*— unsupervised learning, monocular depth estimation, intrinsic image decomposition, Endoscopy.**

I. Introduction

MINIMALLY invasive surgery (MIS) offers significant advantages in terms of minimal tissue trauma and speedy wound healing, rendering it a widely employed technique in clinical practice. Endoscopy, including laparoscopy and gastroscopy, is commonly utilized for navigation in MIS procedures [1] which can be divided into monocular endoscopy with one imaging sensor [2], stereo endoscopy with at least two imaging sensors [3] and structured light endoscopy with precise lighting device [4]. Considering the size and cost of the endoscopy equipment, the monocular endoscopy is more widely used in clinical practice [5].

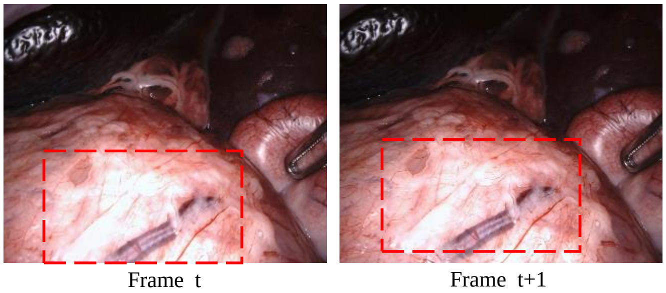


Fig. 1. Illustration of the appearance changes between two adjacent frames in SCARED dataset.

Nevertheless, there is a challenge for monocular endoscopy-based MIS due to its restricted spatial perception ability. To address this, augmented reality navigation systems have been studied extensively to provide surgeons with supplementary anatomical and positional information [6]. These systems typically rely on the reconstruction of intraoperative anatomical surfaces from endoscopic video to obtain spatial information and register with preoperative images [7]. Therefore, the 3D reconstruction is a vital step for the navigation system in which the estimation of depth information plays a pivotal role.

Endoscopic images present inherent challenges, including specular reflection, variations in lighting, low-texture, etc., which make endoscopic depth estimation very challenging. For example, for traditional monocular depth estimation methods, low-texture appearance causes many mismatches and results in degraded performance for the shape-from-motion (SfM) [8] and simultaneous localization and mapping (SLAM) [9] methods; the specular reflections compromise the accuracy greatly for the shape-from-shading (SfS) [10] method.

Recently, deep learning-based monocular depth estimation methods have found remarkable applications in the fields of autonomous driving, robot navigation, etc. Compared to their supervised counterparts, unsupervised methods is more attractive as it circumvents the need for large-scale labeled data [11]. Mainstream unsupervised monocular depth estimation methods utilize the similarity between the reconstructed and the original frames to guide the network training under the assumption of photometric consistency, which posits that the appearance of the same anatomical structure between adjacent frames remains constant [11]–[13]. However, the inherent challenging characteristics of endoscopic images make this assumption no longer valid [14]. As illustrated in Fig. 1,

This work was supported in part by the National Key R&D Program of China under Grant 2018YFA0704100 and Grant 2018YFA0704101; in part by the Beijing Natural Science Foundation under Grant L222034, Grant L232037, Grant L222104, and Grant 7232340; in part by the National Natural Science Foundation of China under Grant 61971443 and Grant 12375359. (Corresponding author: Bo Liu; Xiaoyan Luo)

Bojian Li, Bo Liu, Miao Zhu, Xiaoyan Luo and Fugen Zhou are with the Image Processing Center, School of Astronautics, Beihang University, Beijing, 100191, China(e-mail: bojian@buaa.edu.cn; bo.liu@buaa.edu.cn; ZY2215229@buaa.edu.cn; luoxy@buaa.edu.cn; zhfugen@buaa.edu.cn).

the appearance of the same anatomical structure between adjacent frames varies greatly due to the movement of the light source along with the endoscopic camera. Therefore, existing monocular depth estimation methods based on the photometric consistency assumption cannot be directly applied to endoscopic images to obtain adequate results.

Driven by the necessity of monocular depth estimation in endoscopy-based MIS, efforts have been dedicated to addressing this challenge in the past few years [15]–[17]. Typically, current methods employ a pre-processing strategy trying to align them more closely with the photometric consistency assumption by alterring the distribution of original data. For instance, Ozyoruk *et al.* [16] employ a linear image brightness adjustment strategy, and Shao *et al.* [17] utilize appearance flow to perform nonlinear adjustment. Though promising results have been demonstrated, this strategy relies on certain assumptions about the appearance variation model. However, endoscopic imaging inside the body presents complex characteristics as stated above, which are difficult to fully model. The performance of these methods is constrained by the quality of pre-processing.

In this work, we aim to explore an innovative solution to address the challenging characteristics of the endoscopic images. Specifically, we propose an unsupervised monocular depth estimation method grounded in the intrinsic image decomposition (IID) theory. The idea of intrinsic image decomposition (IID) was proposed by Barrow *et al.* [18] based on the observation that humans are able to derive and understand intrinsic characteristics from images even if they are not familiarized with the scene or the objects. Meanwhile, the experiments conducted by Land *et al.* [19], [20] shown that the color of an object is determined by the object's ability to reflect light, not by the intensity of the reflected light. The color of the object is not affected by the non-uniformity of illumination and has consistency, that is, color constancy. This strategy was widely recognized and applied by later scholars. They referred to intrinsic image decomposition as the problem of separating the reflectance (Albedo) and shading components, represented by the equation $I = A \otimes S$, where the operator $\otimes$ denotes the element-wise product [21]–[23]. While shading component is intricately tied to illumination conditions and influenced by external imaging factors, the albedo component is correlated with the material of the object and remains unaffected by changes in camera viewpoint and illumination conditions. In particular, the illumination-invariance property of the albedo component has been widely utilized in many fields. For example, Baslamisli *et al.* [24] utilized the illumination invariance of the albedo component to address the problem of imaging condition variation for outdoor semantic segmentation. Luo *et al.* [25] relied on the albedo to extract robust features for object detection under uneven and poor lighting conditions.

The illumination-invariance property of the albedo serves as a key motivation for our approach. As it is straightforward to utilize the reconstruction loss of the albedo instead of the original image to guide the network training, the challenges associated with illumination variation can be effectively circumvented. Specifically, we introduce an image intrinsic decomposition module designed to disentangle adjacent frames into their fundamental components. Based on the intrinsic invariance of albedo, we impose a robust constraint on the reconstruction error of the albedo map. Besides, within this framework, we develop additional dedicated losses to enhance the supervision of network training. In essence, our contributions can be summarized as follows:

- We pioneer to leverage the intrinsic image decomposition theory to address the depth estimation problem in monocular endoscopy.
- We present an end-to-end intrinsic-based unsupervised monocular depth learning framework which encompasses an image intrinsic decomposition module and a synthesis reconstruction module.
- Dedicated losses are carefully designed for network training based on intrinsic-based decomposition and the reconstruction of adjacent intrinsic components and frames.
- Experimental results demonstrate that our proposed method can achieve satisfactory depth estimation even in the presence of appearance variations. Additionally, it outperforms state-of-the-art methods on the SCARED dataset and Hamlyn dataset.

II. Related Works

A. Supervised Depth Estimation

Supervised learning methods are characterized by the utilization of neural networks to establish a correspondence between pixel intensity and depth information. In the context of depth estimation for natural images, Eigen *et al.* [26] are pioneers in introducing the use of a convolutional neural network for monocular depth estimation. Their approach employed a coarse-scale network to roughly extract image depth information at a global level, followed by the refinement of results in local regions using a fine-scale network. Cao *et al.* [27] treated the depth estimation problem as a pixel-level classification task and introduced a classification-oriented deep fully convolutional residual network framework. Zhang *et al.* [28] integrated depth estimation with temporal information by introducing a network based on convolutional long short-term memory. Li *et al.* [29] exploited long-range correlation and local information for accurate monocular depth estimation.

Obtaining accurate depth information for endoscopic images is inherently more challenging compared to natural scenes in which the straightforward use of radar and depth cameras is feasible. Presently, supervised depth estimation methods in endoscopy predominantly rely on training with synthetic dense depth maps generated from CT data or simulated data. Mahmood *et al.* [30] presented a novel architecture for monocular endoscopic depth estimation and topographical reconstruction, leveraging a joint CNN and CRF-based framework. Rau *et al.* [31] generated data from a simulation environment and proposed a modification to the well-known cGAN pix2pix to addresse the issue of domain shift between the real and synthetic. However, the learned shapes do not encompass all scenarios encountered during colonoscopy. In a different approach, Yang *et al.* [32] proposed a geometric consistency loss to distribute spatial information across sample grids, aiming to

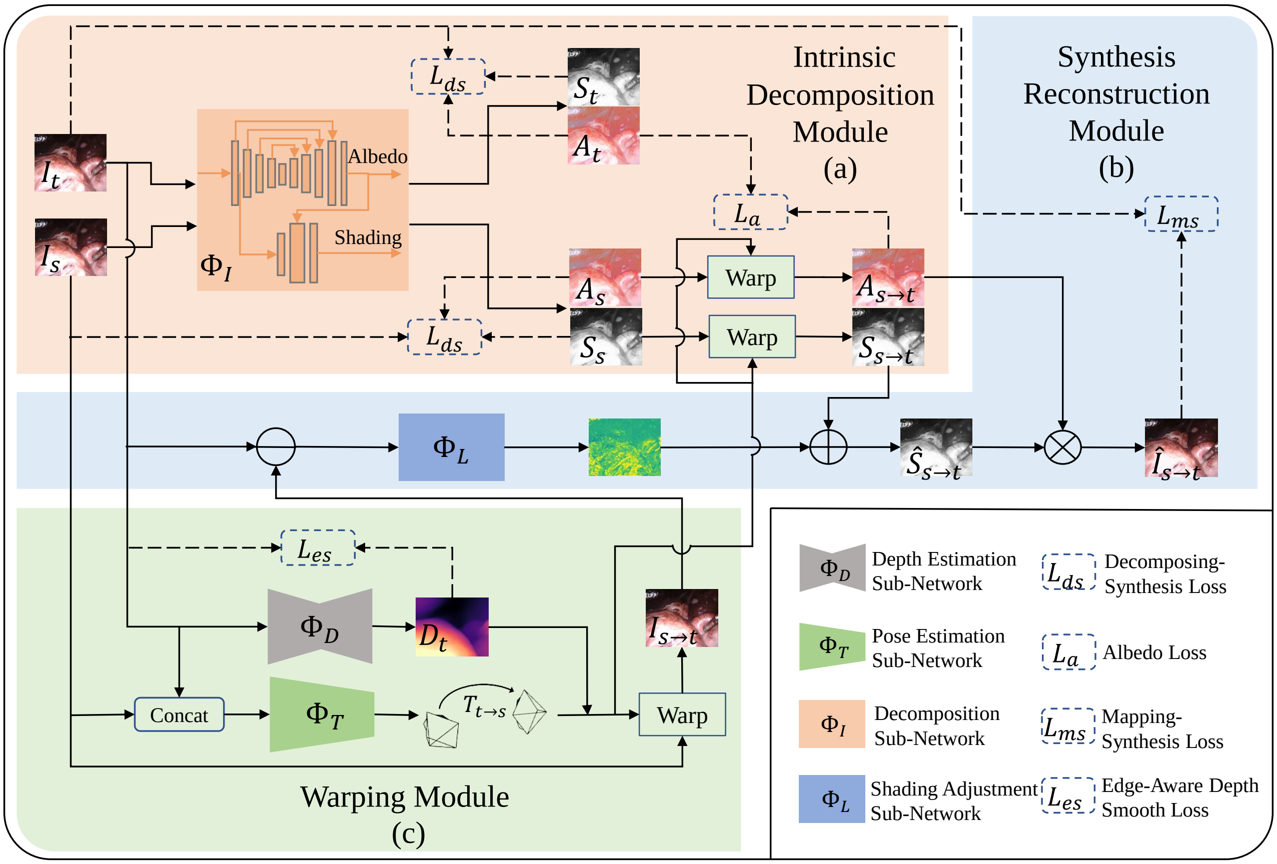


Fig. 2. The framework of the proposed method. a) The intrinsic decomposition module used to decompose the frames into albedo and shading components. b) The synthesis reconstruction module used to adjust the shading map and provide the mapping-synthesis loss. c) The warping module used to estimate the depth map and camera motion matrix, and to warp the source frame into the viewpoint of target frame.

constrain global geometric anatomy structures. Nevertheless, their depth predictions are afflicted by scale ambiguity. In summary, obtaining real-depth data for endoscopic images is a complex and costly process, thereby impeding the practical application of supervised methods.

B. Unsupervised Depth Estimation

To alleviate the demand for large-scale labeled training data, unsupervised depth learning methods have attracted a lot of attentions. In the field of depth estimation for outside scene, Godard *et al.* [33] pioneered the utilization of left-right consistency in stereo images to train depth estimation network. Afterwards, Zhou *et al.* [11] proposed to jointly estimate both depth and the ego-motion of the camera and Godard *et al.* [13] proposed a minimum reprojection loss and an auto-masking loss to ignore training pixels violating camera motion assumptions. To address the low image quality for depth estimation in night driving scenarios, Zheng *et al.* [34] proposed to improve the generalization performance by jointly learning a nighttime image enhancer and a depth estimator.

MovingIndoor [35] is the first work to investigate unsupervised depth estimation in indoor scenes, addressing non-textured regions with an optical flow estimation network and using dense flows to supervise depth learning. Afterward, Yu *et al.* [36] leveraged super-pixels as a plane prior to regularize the texture-less planar regions, but the fronto-parallel assumption limited its applicability.

While these methods have demonstrated promising results, their effectiveness is contingent upon the assumption of photometric consistency, rendering them unsuitable for endoscopic images. To tackle this, Li *et al.* [37] utilized a highlight mask to mask out regions that do not satisfy this assumption and, as a result, this method cannot estimate the depth for the masked-out area. Ozyoruk *et al.* [16] used a linear adjustment strategy to adjust the image, which has a restricted adjustment range. Shao *et al.* [17] addressed the issue by introducing the concept of appearance flow, yet precise extraction of appearance flow becomes challenging in cases of poor image quality. In this paper, we adopt a distinct approach and propose to address the issue via intrinsic image decomposition, which is shown to be more effective than previous strategies.

III. Methodology

As shown in Fig. 2, our method consists of two main modules: the intrinsic decomposition module and the synthesis reconstruction module, built upon the foundational unsupervised monocular depth estimation framework (the warping module). In this section, we begin by reviewing the fundamental concepts that underlie current unsupervised monocular depth estimation methods. Subsequently, we elaborate on the proposed image intrinsic-based depth estimation strategy.

A. The Basic Method for Unsupervised Monocular Depth Estimation (UMDE)

The basis of our method is the fundamental approach for UMDE, which involves estimating depth maps and camera pose information from two adjacent frames (a target frame and an adjacent source frame). Based on the estimated depth and pose information, a synthesized target frame is reconstructed from the source frame, and the reconstruction error is utilized for network training.

As illuminated in Fig. 2 (c), the basic UMDE consists of two subnetworks [11], [13]. One is the depth estimation subnetwork Φ_D which is responsible for estimating the depth map for the target frame I_t. The other is the pose estimation subnetwork Φ_T for estimating the camera motion matrix $T_{t\rightarrow s}$ between the target frame I_t and the source frame I_s. With the estimated $\Phi_D(I_t)$, the depth map D_t of I_t can be obtained by

$$D_t = \frac{1}{a\Phi_D\left(I_t\right) + b} \tag{1}$$

where a and b are factors used to scale the final depth value to a pre-defined range.

Then, based on the stereoscopic vision principle, the 3D spatial coordinate q_t of a pixel p_t in I_t can be determined by leveraging the pre-known camera intrinsic matrix K.

$$q_t = D_t\left(p_t\right) K^{-1} p_t \tag{2}$$

Its projection p_s in I_s can be computed through K after transforming q_t using the estimated camera motion matrix $T_{t\rightarrow s}$:

$$p_s \sim K T_{t\rightarrow s} q_t = K T_{t\rightarrow s} D_t\left(p_t\right) K^{-1} p_t \tag{3}$$

In this way, the position correspondence between the I_t and I_s is determined, and a synthesized frame $I_{s\rightarrow t}$ can be reconstructed via warping transformation:

$$I_{s\rightarrow t}\left(p_t\right) = I_s\left(p_s\right) \tag{4}$$

The training for UMDE primarily relies on the photometric loss, derived from the assumption of photometric consistency. The photometric loss measures the reconstruction error between the target frame I_t and the synthesized frame $I_{s\rightarrow t}$. It is commonly constructed as a weighted combination of L1 loss and SSIM loss as follows:

$$L_p\left(I_t, I_{s\rightarrow t}\right) = \alpha \frac{1 - \mathrm{SSIM}\left(I_t, I_{s\rightarrow t}\right)}{2} + (1-\alpha)\left\|I_t - I_{s\rightarrow t}\right\|_1 \tag{5}$$

where α is the weighting factor.

As we discuss above, in endoscopic imaging, due to the interference of moving light source, specular reflection and other factors, photometric consistency is no longer valid. The basic UMDE method fails to provide adequate results under such conditions. Therefore, it is necessary to develop a more robust depth estimation method that does not rely on the photometric consistency assumption.

B. Image Intrinsic-Based Unsupervised Monocular Depth Estimation

1) Method Overview: In this section, we will explore our innovative approach to integrating intrinsic image decomposition theory into endoscopic depth estimation. While the theory has found success in various applications [24], [25], [38], decomposing an image into intrinsic components is a complex ill-posed problem due to the inherent ambiguity and lack of unique solutions. The challenge is further complicated by the characteristics of endoscopic images, such as specular reflections. Moreover, the absence of ground truth data necessitates an unsupervised approach. However, despite those challenges, the sequential imaging nature of endoscopy makes the intrinsic decomposing of endoscopy image viable. As the adjacent frames exhibit significant scene overlap and can be aligned via the warping transform, we can acquire multiple images of the same scene with differing illumination. Therefore, the problem of intrinsic decomposition of endoscopy images aligns well with the methodology framework of multiple-input-based methods [39]–[41] and can be addressed by leveraging the constancy of scene albedo.

Considering the deep learning methods offer the advantage of reducing reliance on human priors [42], in this work, we propose a deep learning-based unsupervised method for endoscopic intrinsic decomposition and novelly integrate it with the UMDE framework. As shown in Fig. 2, our method mainly comprises an intrinsic decomposition module and a synthesis reconstruction module which are developed based on the warping module. The intrinsic decomposition module is responsible for decomposing adjacent frames into their corresponding albedo and shading components. A CNN subnetwork is fabricated to decompose images into two components, and two dedicated unsupervised losses are designed to ensure the decomposed components follow the property of albedo and shading, based on intrinsic reconstruction and the constancy of the albedo. Within this module, the warping module for basic UMDE procedure is fused to derive one of the training losses. Meanwhile, the synthesis reconstruction module is designed to compute the reconstruction loss for the synthesized target frame, a standard metric in UMDE methods. Unlike conventional approaches that reconstruct the synthesized target frame directly from the source frame, our module utilized the resulting albedo and shading components from the intrinsic decomposition module for reconstruction. This not only enhances the interaction between IID and UMDE procedures but also augments the supervision signals for the decomposition network. Overall, through the synergy of these two modules and the collaborative effect of four dedicated losses, both the tasks of IID and UMDE are accomplished in an end-to-end unsupervised way.

2) The Intrinsic Decomposition Module: As shown in the intrinsic decomposition module of Fig. 2(a), we employ a decomposition subnetwork Φ_I to separate the target frame I_t and the source frame I_s into their constituent components of albedo (A) and shading (S):

$$\{A_t, S_t\} = \Phi_I(I_t); \{A_s, S_s\} = \Phi_I(I_s) \tag{6}$$

Then, the synthesized albedo $A_{s\to t}$ and shading $S_{s\to t}$ can be obtained through the warping transformation of A_s and S_s based on the depth map and camera motion matrix estimated by the warping module.

Under this framework, dedicated losses are designed to supervise the network training. Firstly, according to the IID theory, the composed image from the albedo and shading map ($\hat{I} = A \otimes S$) should be the same as the original image. Therefore, we define a **Decomposing-Synthesis Loss** to provide a handy regularization for network training:

$$L_{ds}(\hat{I}, I) = \alpha \frac{1 - \mathrm{SSIM}(\hat{I}, I)}{2} + (1-\alpha)\|\hat{I} - I\|_1 \tag{7}$$

This loss is applied to both the target and source frames.

In addition, as we discuss above, the reflectance characteristics of the same scene should remain the same across different frames after warping. Therefore, the reconstructed $A_{s\to t}$ should closely resemble A_t. Consequently, we adopt an **Albedo Loss** to enforce material reflectivity consistency, providing another regularization for network training:

$$L_a\left(A_t, A_{s\to t}\right) = \|A_t - A_{s\to t}\|_1 \tag{8}$$

Totally, the training loss for the intrinsic decomposition module is the sum of the decomposing-synthesis loss L_{ds} and the albedo loss L_a:

$$L_d = L_{ds}(\hat{I}_t, I_t) + L_{ds}(\hat{I}_s, I_s) + L_a\left(A_t, A_{s\to t}\right) \tag{9}$$

Specifically, the decomposing-synthesis loss L_{ds} ensures the faithful reconstruction of target and source frames using their respective decomposed components. The albedo loss L_a compels the albedo components of the same object to remain as consistent as possible between adjacent frames corresponding to the illumination-invariance property of the albedo components. Therefore, among the decomposed components, the one used to compute L_a loss is considered to be the albedo components. Besides, it should be noted that, while the metrics employed in this loss are widely used, the proposed L_d is innovative in that as it supervises the reconstruction error of the albedo component instead of the original image.

3) The Synthesis Reconstruction Module: For unsupervised depth estimation methods, the depth information is embedded in the intensity disparity between adjacent frames. Therefore, most previous works impose supervision via mining latent information of the original target and source frames, through the reconstruction loss between the target frame and the synthesized target frame. In contrast, our proposed intrinsic decomposition module only incorporates the reconstruction loss of the albedo component and ignores the shading component. It should be beneficial to include additional supervisory signal associated with the reconstruction of the shading component.

Therefore, we propose a synthesis reconstruction module to meet this end, in which a synthesized target frame is reconstructed by composing the synthesized $A_{s\to t}$ and $S_{s\to t}$, and the reconstruct loss between the reconstructed and original target frames is used as an additional supervisory signal. Considering the shading component between frames may change according to the imaging factors such as lighting condition, a shading adjustment subnetwork Φ_L is proposed to adjust the $S_{s\to t}$ before reconstruction. As shown in Fig. 2, Φ_L takes the synthsized target frames and original target frames as input and extract the illumination changes between them:

$$L_{Adjust} = \Phi_L(|I_t - I_{s\to t}|) \tag{10}$$

The adjusted shading map $\hat{S}_{s\to t}$ is obtained by superimposing the estimated shading adjustment map on $S_{s\to t}$:

$$\hat{S}_{s\to t} = S_{s\to t} \oplus L_{Adjust} \tag{11}$$

where the operator $\oplus$ represents element-by-element addition.

Based on this, the target frame is reconstructed through

$$\hat{I}_{s\to t} = A_{s\to t} \otimes \hat{S}_{s\to t} \tag{12}$$

and a **Mapping-Synthesis Loss** is proposed to supervise the similarity between the reconstructed and the original target frame:

$$L_{ms}\left(\hat{I}_{s\to t}, I_t\right) = \alpha \frac{1 - \mathrm{SSIM}\left(\hat{I}_{s\to t}, I_t\right)}{2} + (1-\alpha)||\hat{I}_{s\to t} - I_t||_1 \tag{13}$$

As we can see in the Fig.2, the L_{ms} involves all components of our network and optimizing it will update the parameters of all subnetworks. It is similar to the standard photometric or reconstruction loss in other UMDE methods. However, instead of reconstructing the synthesized target frame directly from the source frame, our method utilizes the resulting albedo and shading components from the intrinsic decomposition module for reconstruction. This not only enhances the interaction between IID and UMDE procedures but also augments the supervision signals for the intrinsic decomposition network. As for the role of shading adjustment subnetwork, it adjusts the shading part of the source frame to provide a more reliable reconstruction loss, thus enhancing the efficacy of our method, as evidenced by the ablation experiment.

4) Total Loss: Besides the L_d and L_{ms} constructed in the above two modules, considering the characteristic of endoscopic images, additional loss and process are applied to better supervise the network training.

Edge-Aware Depth Smooth Loss. Considering that the depth map typically exhibits smoothness in non-edge regions, we adopt an edge-aware depth smoothness loss function that incorporates image gradient weighting [33]. This loss is designed to enhance smoothness in non-edge regions while preserving sharp edges and details:

$$L_{es}(D_t, I_t) = |\partial_x D_t| e^{-|\partial_x I_t|} + |\partial_y D_t|\, e^{-|\partial_y I_t|} \tag{14}$$

where $\partial_x D_t$ and $\partial_y D_t$ are the first derivative of the depth map along x and y directions, $\partial_x I_t$ and $\partial_y I_t$ are the first derivative of the target frame along x and y directions.

Automatic mask. The synthesized images via the warping transformation may contain missing regions because of the camera movement between frames, and these regions should not be considered in loss computation. To address this, we employ a masking technique by filling zero to the missing regions during warping. The complement of these zero-filled regions serves as mask M_a to prevent the network from learning erroneous information:

$$M_a = [I_{s\to t} > 0] \tag{15}$$

where $[\cdot]$ is the Iverson bracket. The mask is applied to all losses associated with warping transformation.

In summary, the total network loss is composed of three parts:

$$\begin{aligned} loss = \lambda_{ds}(L_{ds}(\hat{I}_t, I_t) + L_{ds}(\hat{I}_s, I_s)) + \lambda_a L_a \otimes M_a \\ + \lambda_{ms} L_{ms} \otimes M_a + \lambda_{es} L_{es} \end{aligned} \tag{16}$$

in which the decomposing-synthesis loss L_{ds} only influences the decomposition subnetwork Φ_I; the edge-aware depth smooth loss L_{es} only influences the depth estimation network Φ_D; the albedo loss L_a influences the decomposition subnetwork Φ_I, the depth estimation subnetwork Φ_D, and the pose estimation subnetwork Φ_T; the mapping-synthesis loss L_{ms} influences all subnetworks. The $\lambda_{ds}, \lambda_a, \lambda_{ms}$, and λ_{es} are the weighting factors and the symbol $\otimes$ denotes the element-wise product.

IV. Experiment

A. Implementation Details and Experimental Settings

1) Network Architecture: The decomposition subnetwork Φ_I adopts a U-shaped structure with skip connections, utilizing ResNet18 [43] without fully-connected layer as the encoder. The shading branch predicts the shading map based on low-dimensional features and the predicted albedo map. For the depth estimation subnetwork Φ_D and the pose estimation subnetwork Φ_T, we employ the same designs as monodepth2 [13]. The shading adjustment subnetwork Φ_L is implemented as a simple convolutional network consisting of four layers of convolution, and its last layer employs the tanh activation function to obtain an illumination adjustment map in the range of $[-1, 1]$.

2) Training Details: We implement our method using the PyTorch framework [44]. The model is trained end-to-end using the Adam optimizer [45] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We train for 30 epochs on a GeForce RTX 3090 GPU with mini-batches of size 8. The initial learning rate is set to $1e^{-4}$ and multiplied by a scale factor of 0.1 after 10 epochs. In our experiments, the weighting factor α is set to 0.85, the loss weights λ_{ds}, λ_a, λ_{ms}, and λ_{es} are set to 0.2, 0.2, 1, and 0.01, respectively. We also preform random data augmentation by adjusting the brightness, contrast, saturation, and hue of the images to prevent overfitting and ensure that the training data distribution remains unbiased across specific patients or cameras. The decomposition subnetwork, the depth estimation subnetwork, and the pose estimation subnetwork all employ a ResNet18 encoder pre-trained on ImageNet, which has been shown to be effective in reducing training time and improving depth estimation accuracy. In comparison with other competitive methods which adopts multi-stage training strategy [17], the training of our method is end-to-end and straightforward, which reduces the complexity of training.

TABLE I
THE CALCULATION OF APPLIED EVALUATION METRICS, IN WHICH $\boldsymbol{d}$ REPRESENTS THE PREDICTION DEPTH AND $\boldsymbol{d^*}$ REPRESENTS THE GROUND TRUTH.

Metric	Definition
Abs diff	$\frac{1}{\lvert N\rvert}\sum_{d\in N}\lvert d-d^*\rvert$
Abs Rel	$\frac{1}{\lvert N\rvert}\sum_{d\in N}\lvert d-d^*\rvert/d^*$
Sq Rel	$\frac{1}{\lvert N\rvert}\sum_{d\in N}\lvert d-d^*\rvert^2/d^*$
RMSE	$\sqrt{\frac{1}{\lvert N\rvert}\sum_{d\in N}(d-d^*)^2}$
RMSE log	$\sqrt{\frac{1}{\lvert N\rvert}\sum_{d\in N}(\ln d-\ln d^*)^2}$
$\delta1, \delta2$, and $\delta3$	$\max\left(\frac{d}{d^*}, \frac{d^*}{d}\right) < T, T\epsilon\{1.25, 1.25^2, 1.25^3\}$

3) Compared Methods and Evaluation Metrics: We compare our method with five unsupervised monocular depth estimation methods: SfmLearner [11], Monodepth2 [13], DeFeat-Net [46], EndoSfmLearner [16] and AF-SfmLearner [17] with AF-SfmLearner as the SOTA algorithm on the SCARED dataset. To ensure a fair comparison, we utilized the code supplied by the respective authors and adhered to the suggested training strategy to obtain the reported results. For example, all methods except for the SfmLearner used the encoder pretrained on ImageNet. Besides, as for the AF-SfmLearner, we also reported the results obtained using the model provided by the author along with the results reproduced by us.

We employ commonly used evaluation metrics, i.e., Abs diff, Abs Rel, Sq Rel, RMSE, RMSE log, and Threshold $\delta1$, $\delta2$, $\delta3$, to benchmark with the compared methods. The definition of these metrics is outlined in Table I. Similar to the method [17], during the evaluation, we scale the predicted depth map via the median scaling, which can be expressed as:

$$D_{scaled} = d * (median(d^*)/median(d)) \tag{17}$$

where d represents the prediction depth, and d^* represents the ground truth. The scaled depth map is capped at 150 mm for SCARED and Hamlyn dataset, which can cover all pixels.

B. Experimental Results on the SCARED Dataset

The SCARED dataset is obtained from fresh porcine cadaver abdominal anatomy using a da Vinci Xi endoscope [47]. During acquisition, a projector is employed to obtain high-quality depth maps of the scene as ground truth. The values in the depth map are in millimeters, and invalid pixels are masked out. In total, there are 9 scenes, each containing 4 or 5 sequences of frames. Following the division strategy of Shao *et al.* [17], we split SCARED dataset into 15351, 1705, 551 frames for training, validation, and testing. During training, we resize the image to a resolution of 256 × 320.

The quantitative results are presented in Table II. SfmLearner heavily relies on the assumption of photometric consistency, resulting in suboptimal performance when applied to endoscopic images. Monodepth2 employs masking to suppress regions with significant impact on the results, enhancing performance to some extent. DeFeat-Net constrains image feature consistency between frames, yet it still performs poorly on endoscopic images. EndoSfmLearner employs a linear method for image adjustment, resulting in some improvement but

TABLE II
QUANTITATIVE RESULTS ON THE SCARED DATASET. THE BEST RESULTS ARE PRESENTED IN BOLD.

Method	Abs diff↓	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta 1 < 1.25 \uparrow$	$\delta 2 < 1.25^2 \uparrow$	$\delta 3 < 1.25^3 \uparrow$
SfmLearner [11]	5.198	0.086	1.021	7.553	0.121	0.925	0.987	0.996
Monodepth2 [13]	3.833	0.066	0.577	5.781	0.093	0.961	0.995	0.999
DeFeat-Net [46]	5.224	0.090	0.984	7.672	0.124	0.921	0.990	0.997
EndoSfmLearner [16]	4.080	0.068	0.679	6.227	0.098	0.955	0.992	0.998
AF-SfmLearner [17]	3.502	0.062	0.493	5.213	0.086	0.964	**0.998**	**1.000**
	3.328†	0.060†	0.443†	4.964†	0.082†	**0.973**†	**0.998**†	**1.000**†
Ours	**3.223**	**0.058**	**0.435**	**4.820**	**0.080**	0.969	**0.998**	**1.000**

† The test results obtained using the model provided by the author.

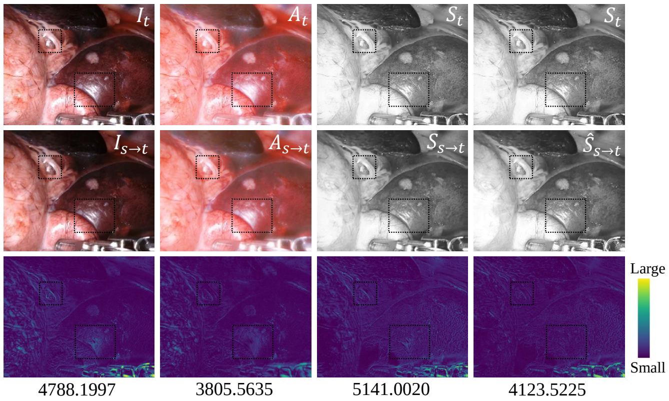
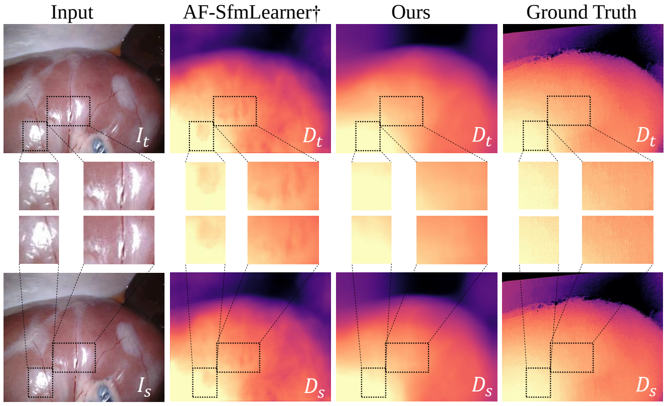


Fig. 3. The intermediate results of our method. The first, second, third, and fourth column compares the reconstructed target frame, albedo map, shading map, and adjusted shading map with their corresponding ground truth. The third row displays the difference between the reconstructed and ground truth pairs, with the L1 norm of the difference shown below.



† The test results obtained using the model provided by the author.

Fig. 4. Comparing the results of our method with those of AF-SfmLearner in regions with illumination changes demonstrates a clear superiority of the proposed approach.

with limited effect. AF-SfmLearner effectively adjusts images using appearance flow and obtains relatively better results. In contrast, our method archives the best results on all metrics except for $\delta 1$, for which the result of our method is lower than AF-SfmLearner by a mere 0.04.

The superior performance of our method is attributed to the image intrinsic-based strategy we adopted. As illustrated in Fig. 3, a comparison between the first and second columns reveals a noticeable reduction in discrepancy, aligning with our expectation that the reflectance characteristics of the same object should be consistent across adjacent frames. Furthermore, a comparison between the third and fourth columns demonstrates a reduction in differences in shading maps after adjustment. The effectiveness of the image decomposition and the proposed adjustment module leads to the superior performance compared to all comparative methods.

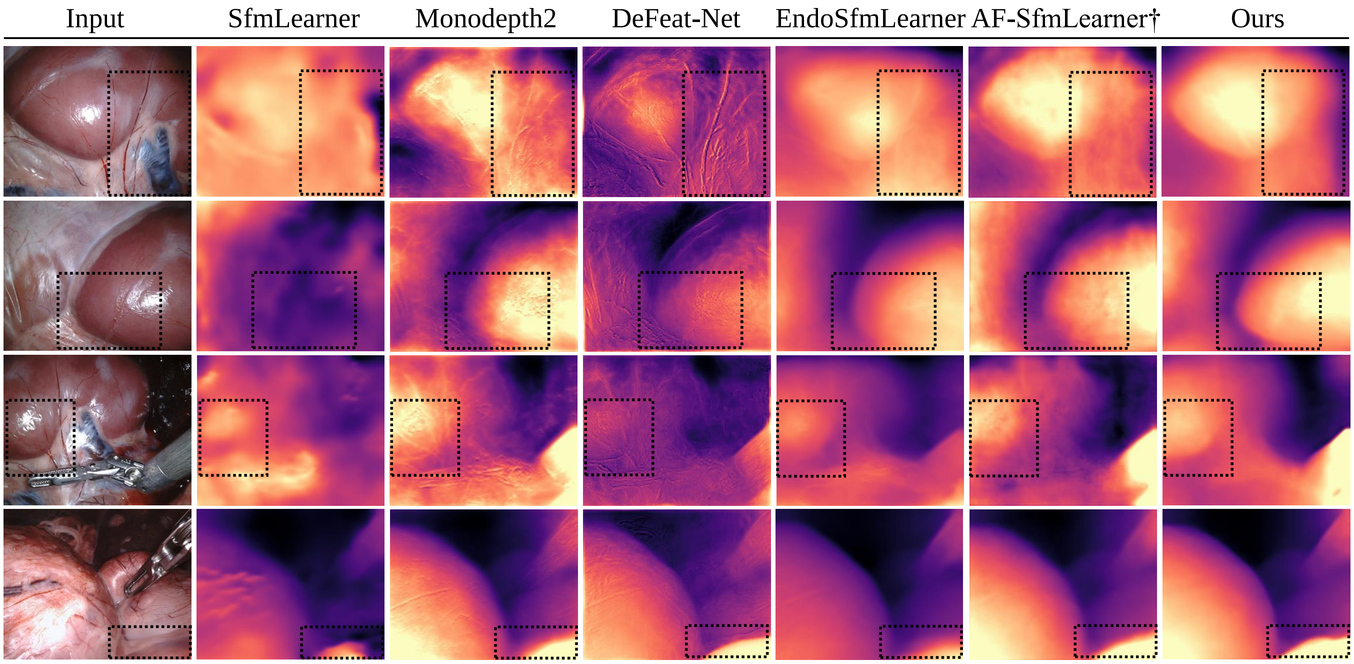
Fig. 5 shows some representative results for the compared methods. It is evident that our proposed approach yields smoother depth estimation results with enhanced clarity in boundary. To further elucidate the effectiveness of our proposed method in situations where the assumption of photometric consistency does not hold, Fig. 4 presents the depth estimation results of our approach and the most competitive method (AF-SfmLearner) for two adjacent frames with significant illumination variation. It can be observed that our method produces stable results in regions with illumination changes, whereas the AF-SfmLearner yields noticeably erroneous outcomes.

C. Experimental Results on the Hamlyn Dataset

The Hamlyn dataset is a public video dataset provided by the Hamlyn Center Laparoscopic at Imperial College London. It includes endoscopic videos of porcine abdomen and heart phantom. We utilize the dataset processed by Recasens *et al.* [48] in which the ground truth was generated using the stereo matching software Libelas. We use 9108, 1567, 1057 frames for training, validation, and testing. During training, we resize the image to a resolution of 256 × 288.

The Hamlyn dataset, characterized by lower data quality in contrast to the SCARED dataset, presents challenges such as reduced image brightness, lower image resolution, and more pronounced inter-frame camera movement. This imposes heightened demands on depth estimation methods. The experimental findings, detailed in Table III, reveal performance deterioration for all methods. Nevertheless, in contrast, our method consistently achieves superior performance compared to all other methods. Significantly, owing to the robustness of the intrinsic-based strategy employed, our method surpasses the performance of the compared method by a more substantial margin than on the SCARED dataset.

In addition, to assess the generalization ability, we compare the methods' performance by directly applying the model trained on the SCARED dataset to the Hamlyn dataset without fine-tuning. The quantitative results are presented in Table IV, and the result demonstrates that our method also achieves the best performance. As shown in Fig. 6, our method not



† The test results obtained using the model provided by the author.

Fig. 5. Qualitative comparison to other methods on the SCARED dataset. Our method obtains smoother depth with clearer boundaries as highlighted by the black rectangle.

TABLE III
QUANTITATIVE RESULTS ON THE HAMLYN DATASET. THE BEST RESULTS ARE PRESENTED IN BOLD.

Method	Abs diff↓	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta 1 < 1.25$ ↑	$\delta 2 < 1.25^2$ ↑	$\delta 3 < 1.25^3$ ↑
SfmLearner [11]	9.942	0.177	2.902	12.687	0.220	0.690	0.939	0.994
Monodepth2 [13]	7.089	0.122	1.546	9.482	0.158	0.851	0.982	0.999
DeFeat-Net [46]	9.412	0.162	2.636	12.158	0.204	0.746	0.947	0.994
EndoSfmLearner [16]	6.789	0.116	1.528	9.073	0.149	0.857	0.977	**1.000**
AF-SfmLearner [17]	7.667	0.133	1.940	10.342	0.172	0.828	0.966	0.997
Ours	**6.423**	**0.115**	**1.217**	**8.220**	**0.140**	**0.881**	**0.993**	**1.000**

TABLE IV
QUANTITATIVE RESULTS OF APPLYING THE MODEL TRAINED ON THE SCARED DATASET TO THE HAMLYN DATASET.THE BEST RESULTS ARE PRESENTED IN BOLD.

Method	Abs diff↓	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta 1 < 1.25$ ↑	$\delta 2 < 1.25^2$ ↑	$\delta 3 < 1.25^3$ ↑
SfmLearner [11]	7.441	0.125	1.754	9.930	0.163	0.821	0.974	0.999
Monodepth2 [13]	6.132	0.100	1.255	8.377	0.134	0.880	0.987	**1.000**
DeFeat-Net [46]	7.400	0.123	1.735	9.984	0.165	0.834	0.973	0.998
EndoSfmLearner [16]	6.423	0.110	1.237	8.353	0.137	0.881	**0.994**	**1.000**
AF-SfmLearner† [17]	6.257	0.104	1.390	8.299	0.131	0.890	0.987	0.999
Ours	**5.903**	**0.099**	**1.143**	**7.753**	**0.124**	**0.904**	0.992	**1.000**

† The test results obtained using the model provided by the author.

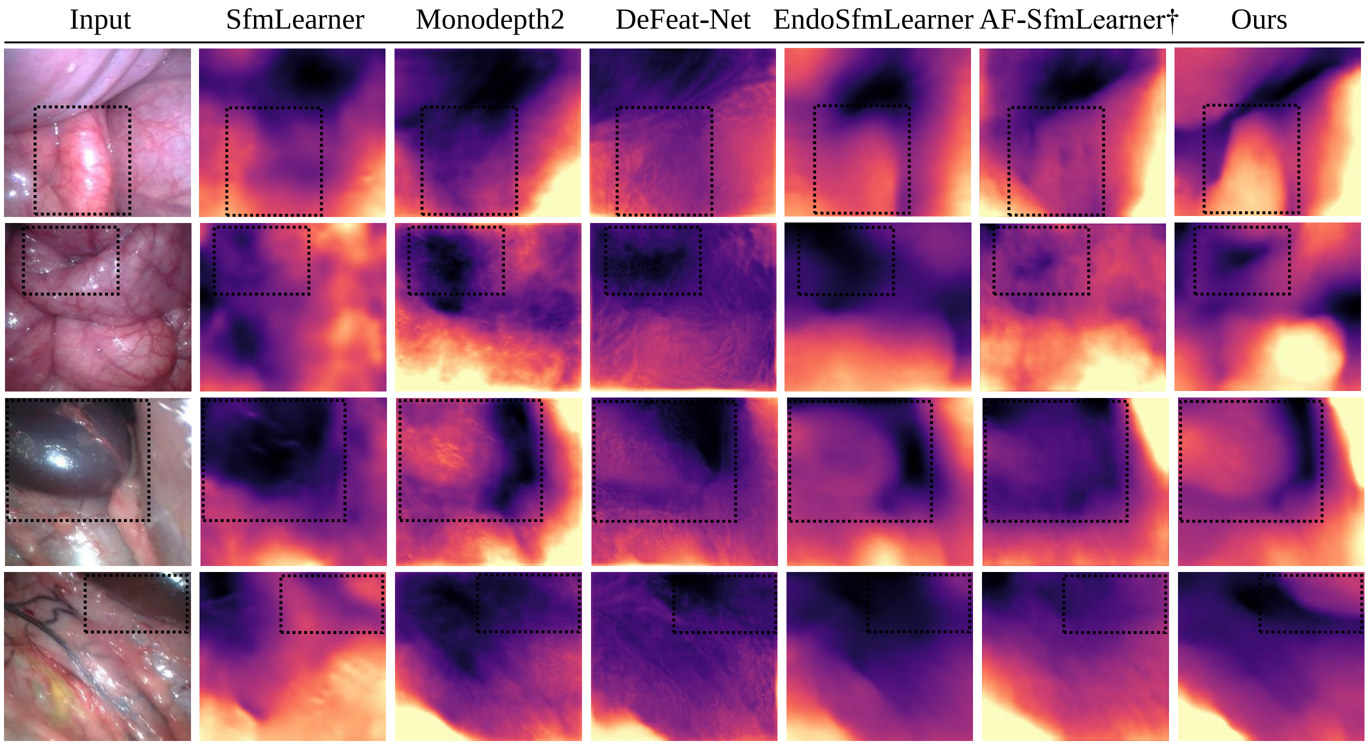
only avoids catastrophic errors but also presents more detailed boundary information.

It is intriguing to observe that the models trained on the SCARED dataset outperform those trained directly on the Hamlyn dataset for all methods. This can be attributed to several factors. Firstly, the SCARED dataset is notably larger in size compared to the Hamlyn dataset, allowing models to achieve better convergence during training. Additionally, the Hamlyn dataset generally exhibits lower image quality and significant variations across different video sequences, resulting in a marked disparity in data distribution between the training and testing sets. Consequently, the advantage gained from training on the Hamlyn dataset is significantly diminished. It makes sense that the model trained on SCARED dataset exhibits better performance on the Hamlyn dataset.

D. Ablation Study

In order to further investigate the effectiveness of the proposed components, ablation experiments were conducted on the SCARED dataset, and the results are presented in Table V.

The first set of configurations investigate the effectiveness of the proposed losses. A comparison between the first and fourth rows reveals a substantial improvement in performance with the inclusion of L_{ms} with a remarkable 16.7% reduction in RMSE, which shows that the synthesis reconstruction module



† The test results obtained using the model provided by the author.

Fig. 6. Qualitative comparison when applying the models trained on SCARED dataset to the Hamlyn dataset. Our approach consistently outperforms the others, demonstrating the ability to prevent critical errors and provide finer boundary details, as emphasized by the black rectangle.

TABLE V
THE RESULTS OF ABLATION STUDY.

Configurations			Abs diff↓	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta 1 < 1.25$ ↑
Loss function	1	$L_{ds}\&L_a\&L_{es}$	3.843	0.066	0.580	5.788	0.094	0.956
	2	$L_{ds}\&L_{es}\&L_{ms}$	3.414	0.062	0.500	5.148	0.086	0.961
	3	$L_a\&L_{es}\&L_{ms}$	3.318	0.059	0.462	5.056	0.083	0.969
	4	$L_{ds}\&L_a\&L_{es}\&L_{ms}$	**3.223**	**0.058**	**0.435**	**4.820**	**0.080**	**0.969**
Intrinsic Decomposition Module	5	With Star [22]	3.669	0.064	0.543	5.505	0.089	0.960
	6	With JieP [21]	3.505	0.062	0.501	5.245	0.086	0.963
	7	With Φ_I	**3.223**	**0.058**	**0.435**	**4.820**	**0.080**	**0.969**
Synthesis Reconstruction Module	8	Without Φ_L	3.390	0.061	0.457	4.986	0.084	0.967
	9	With Φ_L	**3.223**	**0.058**	**0.435**	**4.820**	**0.080**	**0.969**

plays an important role in this framework. Contrasting the second and fourth rows demonstrates that incorporating the albedo loss, L_a, leads to a notable 6.4% reduction in RMSE. Furthermore, comparing the third and fourth rows illustrates a 4.7% decrease in RMSE by introducing the Decomposing-Synthesis loss, L_{ds}. These findings collectively confirm the effectiveness of proposed losses.

The second set of experiments is uesd to verify the effectiveness of the decomposition subnetwork within the intrinsic decomposition module. In the comparative variant, we employ the method of structure and texture-aware (Star) model proposed by Xu *et al.* [22] and the method of joint intrinsic-extrinsic prior (JieP) model proposed by Cai *et al.* [21] to replace our decomposition subnetwork Φ_I. Their decomposition results are utilized for subsequent data flow. From the experimental results it can be seen that our decomposition method can achieve better results. At the same time, it can also be seen that using traditional methods for intrinsic decomposition can achieve relatively satisfactory results, further consolidating the validity and efficiency of the IID based strategy.

The third set of experiments is conducted to verify the effectiveness of the shading adjustment subnetwork within the synthesis reconstruction module. The results clearly indicate that the inclusion of the adjustment subnetwork is indeed effective, resulting in improvement for all metrics and a noteworthy 3.3% reduction in RMSE.

E. Camera Pose Estimation

The warping transformation hinges on the outcomes of both depth estimation and pose estimation, which are interconnected tasks. Consequently, we conducted additional assessments to evaluate the effectiveness of the proposed method specifically in the context of pose estimation. The Absolute Trajectory Error (ATE) served as the evaluation metric, calculated on 5-frame snippets and averaged over the entire sequence. We utilized three image sequences from the SCARED dataset for testing, including 448 frames from

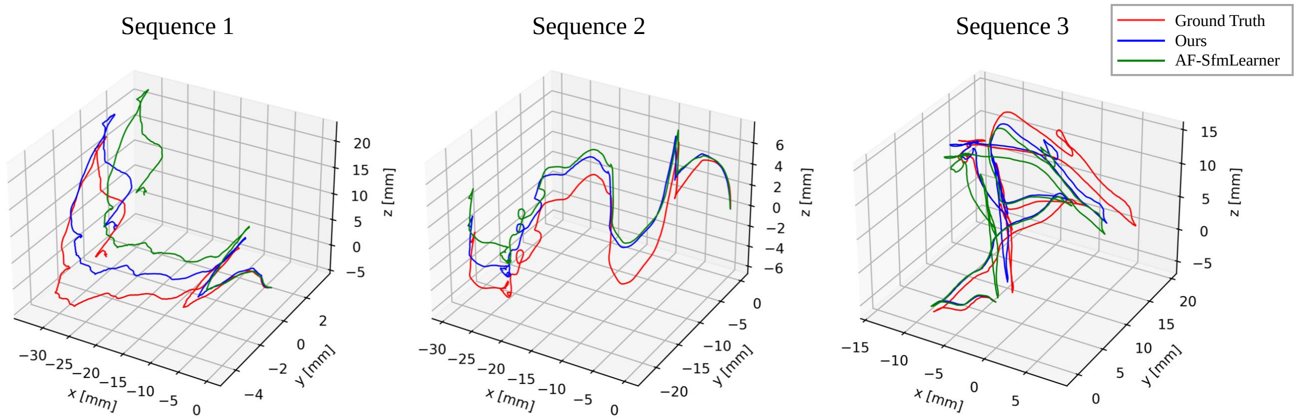


Fig. 7. The visualization of camera motion trajectories.

TABLE VI
THE RESULTS OF TRAJECTORY PREDICTION.

Method	Sequence 1	Sequence 2	Sequence 3
AF-SfmLearner [17]	0.0495	0.0579	0.0642
Ours	**0.0462**	**0.0559**	**0.0638**

sequence 1, 406 frames from sequence 2, and 693 frames from sequence 3. We only compared with the most competitive method (AF-SfmLearner). As shown in Table VI, our method outperforms the AF-SfmLearner on all test sequences. Fig. 7 presents the visualization of camera motion trajectories. It is shown that our approach exhibits reduced trajectory drift and aligns more closely with the ground truth. These favorable trajectory prediction results further validate the effectiveness of our proposed approach.

V. DISCUSSIONS AND CONCLUSION

In this paper, we propose an end-to-end framework based on intrinsic image decomposition for robust depth estimation from monocular endoscopy, which does not rely on any imaging assumptions. In contrary to current methods, our proposed framework employs a decoupling mechanism instead of pre-processing to address the challenge of appearance changes. Specifically, we design intrinsic decomposition module and synthesis reconstruction module to seamlessly integrate image intrinsic decomposition with the framework of unsupervised depth estimation. A Novel framework and dedicated losses are designed to facilitate unsupervised depth estimation. Additionally, we introduce a shading adjustment subnetwork to further mitigate the influence of illumination changes and optimize the consistency constraint between adjacent frames. Extensive experiments on two datasets demonstrate that our method outperforms state-of-the-art methods, showcasing robustness and generalization capabilities. Furthermore, thorough ablation studies validate the effectiveness of the proposed network modules and loss functions. The effectiveness of the proposed method is further illustrated through pose estimation experiments.

Despite the promising results obtained, the employed IID model exhibits certain limitations in effectively handling endoscopic images. As it separates the image into the shading and albedo components without explicitly considering the specular reflection, the intertwined specular reflection poses challenges for accurate decomposition, particularly for methods relying on diffuse reflection priors. While our method operates without such assumptions, it may be beneficial to disentangle the specular reflection from the shading component by applying alternative decomposition model. On the other hand, it should be noted that our basic idea is to decouple the illumination-invariant and illumination-dependent components. We do not necessitate precise decomposition of shading and albedo. As demonstrated by our results, current model may suffice for our specific objective. In our future work, we would like to explore potential advantages by specifically addressing specular reflection.

Overall, the method presented in this article has the potential to enhance the quality of 3D reconstruction in monocular endoscopy, thereby improving the accuracy and robustness of augmented reality navigation technology in MIS. Though it is developed for monocular endoscopy, the idea of separating the original frame into illumination-invariant and illumination-dependent components can also benefit the depth estimation for stereo endoscopy as the addressed problem of illumination variation is a generic challenge for all kinds of endoscopies. Furthermore, since our methodology does not rely on specific organ structures or textures, its applicability extends across a broad spectrum of MIS procedures in neurosurgery, gastroenterology, and orthopedics. Consequently, the advancements we have achieved in endoscopic depth estimation constitute a meaningful contribution to the ongoing progress of MIS.

REFERENCES

[1] T. Bergen and T. Wittenberg, “Stitching and surface reconstruction from endoscopic image sequences: a review of applications and methods,” *IEEE J. Biomed. Health Inform.*, vol. 20, no. 1, pp. 304–321, 2014.

[2] G. A. Puerto-Souza, J. A. Cadeddu, and G.-L. Mariottini, “Toward long-term and accurate augmented-reality for monocular endoscopic videos,” *IEEE Trans. Biomed. Eng.*, vol. 61, no. 10, pp. 2609–2620, Oct. 2014.

[3] B. Yang, C. Liu, W. Zheng, S. Liu, and K. Huang, “Reconstructing a 3d heart surface with stereo-endoscope by learning eigen-shapes,” *Biomed. Opt. Express*, vol. 9, no. 12, pp. 6222–6236, 2018.

[4] J. Lin, N. T. Clancy, J. Qi, Y. Hu, T. Tatla, D. Stoyanov, L. Maier-Hein, and D. S. Elson, "Dual-modality endoscopic probe for tissue surface shape reconstruction and hyperspectral imaging enabled by deep neural networks," *Med. Image Anal.*, vol. 48, pp. 162–176, 2018.

[5] A. R. Widya, Y. Monno, M. Okutomi, S. Suzuki, T. Gotoda, and K. Miki, "Learning-based depth and pose estimation for monocular endoscope with loss generalization," in *Proc. 2021 Annu. Int. Conf. IEEE Eng. Med. Biol. Soc*, 2021, pp. 3547–3552.

[6] R. Wang, M. Zhang, X. Meng, Z. Geng, and F.-Y. Wang, "3-d tracking for augmented reality using combined region and dense cues in endoscopic surgery," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 5, pp. 1540–1551, 2017.

[7] Y. Gu, C. Gu, J. Yang, J. Sun, and G.-Z. Yang, "Vision–kinematics interaction for robotic-assisted bronchoscopy navigation," *IEEE Trans. Med. Imag.*, vol. 41, no. 12, pp. 3600–3610, Dec. 2022.

[8] S. Leonard, A. Sinha, A. Reiter, M. Ishii, G. L. Gallia, R. H. Taylor, and G. D. Hager, "Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data," *IEEE Trans. Med. Imag.*, vol. 37, no. 10, pp. 2185–2195, Oct. 2018.

[9] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "Slam-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality," *Comput. Meth. Programs Biomed.*, vol. 158, pp. 135–146, 2018.

[10] Z. Ren, T. He, L. Peng, S. Liu, S. Zhu, and B. Zeng, "Shape recovery of endoscopic videos by shape from shading using mesh regularization," in *Proc. ICIG*, 2017, pp. 204–213.

[11] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1851–1858.

[12] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1983–1992.

[13] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 3828–3838.

[14] R. Li, J. Pan, Y. Si, B. Yan, Y. Hu, and H. Qin, "Specular reflections removal for endoscopic image sequences with adaptive-rpca decomposition," *IEEE Trans. Med. Imag.*, vol. 39, no. 2, pp. 328–340, Feb. 2020.

[15] X. Liu, A. Sinha, M. Ishii, G. D. Hager, A. Reiter, R. H. Taylor, and M. Unberath, "Dense depth estimation in monocular endoscopy with self-supervised learning methods," *IEEE Trans. Med. Imag.*, vol. 39, no. 5, pp. 1438–1447, May 2020.

[16] K. B. Ozyoruk, G. I. Gokceler, T. L. Bobrow, G. Coskun, K. Incetan, Y. Almalioglu, F. Mahmood, E. Curto, L. Perdigoto, M. Oliveira *et al.*, "Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos," *Med. Image Anal.*, vol. 71, p. 102058, 2021.

[17] 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," *Med. Image Anal.*, vol. 77, p. 102338, 2022.

[18] H. Barrow, J. Tenenbaum, A. Hanson, and E. Riseman, "Recovering intrinsic scene characteristics," *Comput. vis. syst*, vol. 2, no. 3-26, p. 2, 1978.

[19] E. H. Land and J. J. McCann, "Lightness and retinex theory," *Josa*, vol. 61, no. 1, pp. 1–11, 1971.

[20] E. H. Land, "The retinex theory of color vision," *Scientific american*, vol. 237, no. 6, pp. 108–129, 1977.

[21] B. Cai, X. Xu, K. Guo, K. Jia, B. Hu, and D. Tao, "A joint intrinsic-extrinsic prior model for retinex," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 4000–4009.

[22] J. Xu, Y. Hou, D. Ren, L. Liu, F. Zhu, M. Yu, H. Wang, and L. Shao, "Star: A structure and texture aware retinex model," *IEEE Trans. Image Process.*, vol. 29, pp. 5022–5037, 2020.

[23] Z. Wang, Y. Liu, and F. Lu, "Discriminative feature encoding for intrinsic image decomposition," *Comput. Vis. Media*, pp. 1–22, 2023.

[24] A. S. Baslamisli, T. T. Groenestege, P. Das, H.-A. Le, S. Karaoglu, and T. Gevers, "Joint learning of intrinsic images and semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*, 2018, pp. 286–302.

[25] L. Luo, X. Chen, J. Yang, J. Liu, and Z.-X. Yang, "Unsupervised lighting reflectance estimation for robot monitoring under poor illuminance," *IEEE Sens. J.*, 2023.

[26] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Adv. Neural Inf. Process. Syst.(NeurIPS)*, vol. 27, 2014.

[27] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3174–3182, Nov. 2018.

[28] H. Zhang, C. Shen, Y. Li, Y. Cao, Y. Liu, and Y. Yan, "Exploiting temporal consistency for real-time video depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1725–1734.

[29] Z. Li, Z. Chen, X. Liu, and J. Jiang, "Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation," *Mach. Intell. Res.*, vol. 20, no. 6, pp. 837–854, 2023.

[30] F. Mahmood and N. J. Durr, "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy," *Med. Image Anal.*, vol. 48, pp. 230–243, 2018.

[31] A. Rau, P. E. Edwards, O. F. Ahmad, P. Riordan, M. Janatka, L. B. Lovat, and D. Stoyanov, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 14, pp. 1167–1176, 2019.

[32] Y. Yang, S. Shao, T. Yang, P. Wang, Z. Yang, C. Wu, and H. Liu, "A geometry-aware deep network for depth estimation in monocular endoscopy," *Eng. Appl. Artif. Intell.*, vol. 122, p. 105989, 2023.

[33] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 270–279.

[34] Y. Zheng, C. Zhong, P. Li, H.-a. Gao, Y. Zheng, B. Jin, L. Wang, H. Zhao, G. Zhou, Q. Zhang, and D. Zhao, "Steps: Joint self-supervised nighttime image enhancement and depth estimation," in *Proc. IEEE Int. Conf. Robot. Autom.(ICRA)*, 2023, pp. 4916–4923.

[35] J. Zhou, Y. Wang, K. Qin, and W. Zeng, "Moving indoor: Unsupervised video depth learning in challenging environments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 8618–8627.

[36] Z. Yu, L. Jin, and S. Gao, "P 2 net: patch-match and plane-regularization for unsupervised indoor depth estimation," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. Springer, 2020, pp. 206–222.

[37] L. Li, X. Li, S. Yang, S. Ding, A. Jolfaei, and X. Zheng, "Unsupervised-learning-based continuous depth and motion estimation with monocular endoscopy for virtual reality minimally invasive surgery," *IEEE Trans. Ind. Inform.*, vol. 17, no. 6, pp. 3920–3928, June 2021.

[38] Y. Zhang, X. Guo, J. Ma, W. Liu, and J. Zhang, "Beyond brightening low-light images," *Int. J. Comput. Vis.*, vol. 129, pp. 1013–1037, 2021.

[39] Z. Li and N. Snavely, "Learning intrinsic image decomposition from watching the world," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 9039–9048.

[40] W.-C. Ma, H. Chu, B. Zhou, R. Urtasun, and A. Torralba, "Single image intrinsic decomposition without a single intrinsic image," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*, 2018, pp. 201–217.

[41] L. Lettry, K. Vanhoey, and L. Van Gool, "Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences," in *Comput. Graph. Forum.*, vol. 37, no. 7, 2018, pp. 409–419.

[42] E. Garces, C. Rodriguez-Pardo, D. Casas, and J. Lopez-Moreno, "A survey on intrinsic images: Delving deep into lambert and beyond," *Int. J. Comput. Vis.*, vol. 130, no. 3, pp. 836–868, 2022.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[44] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 32, 2019.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[46] J. Spencer, R. Bowden, and S. Hadfield, "Defeat-net: General monocular depth via simultaneous unsupervised representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14 402–14 413.

[47] 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.

[48] D. Recasens, J. Lamarca, J. M. Fácil, J. Montiel, and J. Civera, "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints," *IEEE Robot. Autom. Lett.*, vol. 6, no. 4, pp. 7225–7232, Oct. 2021.