International Journal of Computer Assisted Radiology and Surgery (2025) 20:1255–1263
https://doi.org/10.1007/s11548-025-03371-8

ORIGINAL ARTICLE

SHADeS: self-supervised monocular depth estimation through non-Lambertian image decomposition

Rema Daher[1] · Francisco Vasconcelos[1] · Danail Stoyanov[1]

Received: 18 February 2025 / Accepted: 26 March 2025 / Published online: 13 May 2025

Abstract

Purpose Visual 3D scene reconstruction can support colonoscopy navigation. It can help in recognising which portions of the colon have been visualised and characterising the size and shape of polyps. This is still a very challenging problem due to complex illumination variations, including abundant specular reflections. We investigate how to effectively decouple light and depth in this problem.

Methods We introduce a self-supervised model that simultaneously characterises the shape and lighting of the visualised colonoscopy scene. Our model estimates shading, albedo, depth, and specularities (SHADeS) from single images. Unlike previous approaches (IID (Li et al. IEEE J Biomed Health Inform https://doi.org/10.1109/JBHI.2024.3400804, 2024)), we use a non-Lambertian model that treats specular reflections as a separate light component. The implementation of our method is available at https://github.com/RemaDaher/SHADeS.

Results We demonstrate on real colonoscopy images (Hyper Kvasir) that previous models for light decomposition (IID) and depth estimation (MonoViT, ModoDepth2) are negatively affected by specularities. In contrast, SHADeS can simultaneously produce light decomposition and depth maps that are robust to specular regions. We also perform a quantitative comparison on phantom data (C3VD) where we further demonstrate the robustness of our model.

Conclusion Modelling specular reflections improves depth estimation in colonoscopy. We propose an effective self-supervised approach that uses this insight to jointly estimate light decomposition and depth. Light decomposition has the potential to help with other problems, such as place recognition within the colon.

Keywords Monocular depth · Self-supervision · Specular highlights

Introduction

Colorectal cancer is the third most common cancer worldwide with a 47% fatality rate [2]. Early diagnosis of colorectal cancer plays a key role in improving survival rates [3]. However, only 40% of colorectal cancers are detected early on [4]. One main reason is the difficult visibility conditions in colonoscopy. Computer vision can assist surgeons with visibility through 3D reconstruction, navigation, and polyp detection. In particular, 3D reconstruction could aid in identifying missed regions, characterising polyps, comparing screenings, training endoscopists, and autonomous navigation.

We focus on monocular depth estimation, a crucial part of endoscopic 3D reconstruction and navigation. State-of-the-art (SOTA) methods such as MonoViT [5] have achieved impressive results on non-medical images. However, they still struggle with visibility challenges in endoscopy such as light variations and reflections due to the close-range scene with frequent motion blur and sub-optimal focus. While some non-learning methods [6] have been proposed to tackle this problem, deep learning is still the predominant approach in recent research. Deep networks can be trained either in a supervised manner using virtual or phantom simulated data, or in a self-supervised manner with real endoscopy data. Self-supervised approaches are currently the SOTA in monocular

✉ Rema Daher
rema.daher.20@ucl.ac.uk

Francisco Vasconcelos
f.vasconcelos@ucl.ac.uk

Danail Stoyanov
danail.stoyanov@ucl.ac.uk

[1] Department of Computer Science, UCL Hawkes Institute, University College London, Gower Street, London WC1E 6BT, UK

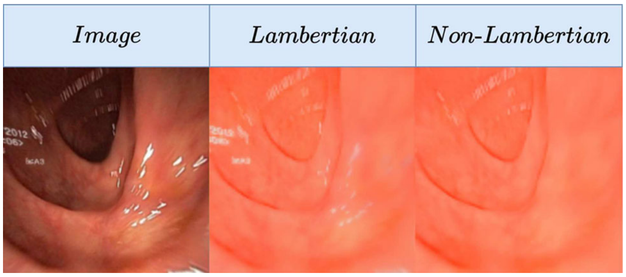

Fig. 1 Extracted albedo from a Lambertian (IID) vs. our non-Lambertian model (SHADeS). The specular reflections produce significantly fewer artefacts with our model

depth estimation for endoscopy, yet they still struggle with challenging illumination conditions, such as abundant specular reflections in endoscopic images.

This paper proposes a novel self-supervised approach that jointly estimates depth and decomposes an image into different light components. We take as inspiration the model in [1] (IID; short for IID-SfMLearner), which simultaneously estimates depth and decomposes the image (I) into albedo (A) and shading (S) following a Lambertian model assumption ($I = AS$). In contrast, we consider specular masks (M) as a third image component, following the relation $I = AS + M$. We do so because handling specular highlights has improved many computer vision tasks [7, 8]. Figure 1 shows that the Lambertian model cannot distinguish between specular reflections and the underlying albedo (A), while our model can extract the albedo free of artefacts. Beyond the raw outputs of our model (depth, light components), it can also implicitly perform semantic segmentation of specular reflections by binarising M as well as specularity removal through image inpainting ($I - M = AS$). In summary, our contributions are as follows:

1. We propose a novel self-supervised monocular depth estimation framework that is more robust to specular reflections than the SOTA (IID, Monodepth2, MonoViT) as demonstrated on real (Hyper Kvasir) and phantom colon data (C3VD).
2. Our model jointly estimates depth, albedo, shading, and specular reflections. This is a direct upgrade from IID, which only estimates depth, shading and albedo. We demonstrate that our model can effectively decouple albedo from specular reflections, while IID extracts albedo with specular artefacts.
3. We can combine the different outputs of our model to implicitly estimate specularity segmentation masks as well as inpainted images without specular reflections.

Related work

In recent years, monocular depth estimation has been dominated both in terms of popularity and performance by self-supervised approaches, and therefore we focus this section on these. SfMLearner [9] was one of the pioneering methods of this kind. It introduced the popular concept of jointly training depth and camera pose regression networks using a loss that measures re-projected photometric consistency on pairs of overlapping views. Most of the more recent self-supervised approaches all follow a similar training methodology. Monodepth2 [10] adds a multi-scale appearance matching loss to address occluded pixels as well as an auto-masking technique to ignore static pixels that generate infinity depth values. In parallel, SC-SfMLearner [11] introduced a constraint for scale consistency and added a self-discovered mask to address dynamic scenes and occlusions. Many works have built upon these methods, with MonoViT [5] being a notable example with state-of-the-art performance that adopts the Monodepth2 methodology while using a transformer-based depth network.

However, these methods have sub-optimal performance when applied to endoscopy data. One of the reasons is that they all assume the visualised scene is approximately a Lambertian surface, i.e. any 3D location is viewed with the same colour and light intensity from any viewpoint. However, in endoscopy, this is not true due to the moving light source and the visualised wet tissue being highly reflective and deformable.

In the endoscopic domain, some methods have incorporated model-free learning based models to estimate an offset that compensates small light changes in different viewpoints. One of the first solutions of this kind proposed a linear affine brightness transformer that was added to the photometric loss [12]. This was extended in [13] by applying domain adaptation so that both real and synthetic data can be combined during training. To further incorporate the appearance changes in endoscopy, AF-SfMLearner [14] added appearance flow and correspondence networks. In [15], a confidence-based colour offset penalty is added to the appearance flow network to improve low-texture and drastic illumination fluctuations. Some have also introduced temporal information to AF-SfMLearner [16, 17].

A different type of methods attempt to filter out regions likely to be inconsistent, such as specular reflections. This can be achieved with a separate specularity detection algorithm that either masks out regions during loss computation [18, 19] or is utilised to learn how to reconstruct surface texture underneath specular regions [20]. In [21], a multitask PoseNet is incorporated to generate pose and two types of masks: one for photometric loss focused on specularities and another for geometric consistency loss focused on deformations. In [22], specular highlights are implicitly incorporated by minimising uncertainty estimated through Bayesian or deep ensemble learning.

Finally, other methods try to model light reflection properties more explicitly. In [23], light intensity is made dependent on its direction. In LightDepth [24], LD for short, a light decline model, coupled with estimated albedo and shading, is utilised as a supervision signal instead of the standard pose estimation network. They account for non-Lambertian properties by adding a specular loss term. The most closely related method to ours, IID-SfMLearner (IID for short) [1], uses an intrinsic decomposition network to simultaneously estimate depth, albedo and shading. To compensate for non-Lambertian properties, they incorporate a shading adjustment network. However, the models described in [1] and [24] can only compensate for small light changes and are still not capable of fully handling saturated specularities. In this paper, we improve on [1] by explicitly modelling a non-Lambertian image decomposition (albedo, shading, and specularities) instead of utilising an adjustment network.

Methodology

Training

Basic monocular depth model

Monocular depth estimation aims at estimating the scene depth of every pixel in a single frame. Self-supervision in monocular depth estimation relies on reconstructing a source image from the viewpoint of a target image.

Consider source and target images I_s, I_t that visualise the same scene under different viewpoints. These images are fed into networks ϕ_{Depth} and ϕ_{Pose} that, respectively, estimate the scene depth maps D_t, D_s and the relative pose $T_{t\rightarrow s}$ between I_t and I_s. Estimated pose ($T_{t\rightarrow s}$), depth (D_t), and known camera intrinsics K are used to reconstruct the target from the source image $I_{s\rightarrow t}$ following the pixel relation in Eq. (1). The supervision signal comes from encouraging the reconstructed image $I_{s\rightarrow t}$ to be closer to the target image I_t using a photoconsistency loss (Eq. (2)), such that the weighting factor $\alpha = 0.85$ [1, 10].

$$p_s \approx K T_{t\rightarrow s} D_t(p_t) K^{-1} p_t \tag{1}$$

$$L_r(I_{s\rightarrow t}, I_t) = \alpha \frac{1 - SSIM(I_{s\rightarrow t}, I_t)}{2} + (1-\alpha)\, \|I_{s\rightarrow t} - I_t\|_1 \tag{2}$$

IID

IID [1] (Fig. 2) extends this basic approach with an additional network $\phi_{\text{Decompose}}$ that decomposes the source image into albedo (A) and shading (S). The photometric loss described above is then computed by comparing a reconstructed source image AS_s (instead of I_s) against target I_t. IID also uses an adjustment network ϕ_{adjust} to learn small light offsets.

Proposed method

Extending [1] and introducing the insights from [25], we consider a more complete image decomposition (Fig. 2) that includes albedo, shading, and specularities (M). For the photometric loss, we compare a reconstructed source image warped to target $AS_{s\rightarrow t}$ against an inpainted target with removed specularities ($I_{t,\text{rem}} = I_t - M$). The specularity component M is effectively a replacement for IID's offset network, ϕ_{adjust}, that more explicitly considers that the dominant light changes are specularities.

Our complete model, ***SHADeS***, which stands for **SH**ading, **A**lbedo, **D**epth and **S**pecularities, has the following components: inpainting module, intrinsic decomposition module, warping module, and auto-masking as shown in Fig. 3.

Inpainting module uses a pre-trained inpainting model P_{Inp} [7]. This model uses a non-learning method [26] to segment specularities before inpainting them. The inpainted images $I_{s,\text{rem}}$, $I_{t,\text{rem}}$ are used in photoconsistency and decomposition losses.

The intrinsic decomposition module uses a U-shaped network, $\phi_{\text{Decompose}}$, adopted from [1], that decomposes the input images into Albedo A and shading S (without specularities). This model is guided by the decomposition loss (Eq. (3)) making sure the reconstructed image from albedo and shading is similar to I_{rem}. Unlike the Lambertian image decomposition assumption ($I = AS$) used in [1] where specular highlights are ignored, we use I_{rem} instead of I since a more accurate non-Lambertian model is $I = AS + M \implies AS = I - M \implies AS \approx I_{\text{rem}}$. The albedo loss in Eq. (4) also guides the intrinsic decomposition model [1]. We apply this loss to ensure that the albedo is influenced solely by warping.

$$L_d(AS, I_{\text{rem}}) = \alpha \frac{1 - SSIM(AS, I_{\text{rem}})}{2} + (1-\alpha)\, \|AS - I_{\text{rem}}\|_1 \tag{3}$$

$$L_a = \|A_t - A_{s\rightarrow t}\|_1 \tag{4}$$

The warping module consists of pose and depth estimation networks, ϕ_{Depth} and ϕ_{Pose}, following the basic self-supervision strategy described in the first paragraph of Sect. Basic monocular depth model. We introduce the reconstruction loss (Eq. (5)) adapted from [1] by replacing I with I_{rem}. We use their edge-aware smoothness loss to ensure smoothness along the depth gradient (Eq. (6)). We also omit the shading adjustment network ϕ_{Adjust} proposed in [1] because it does not impact the results empirically.

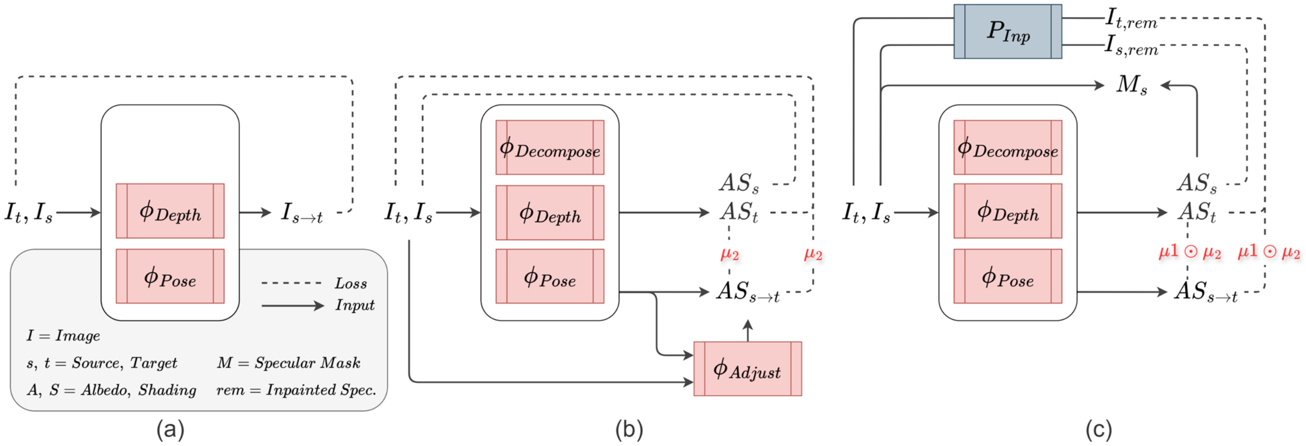

Fig. 2 A high-level representation of depth estimation training. **a** The basic self-supervision relies on reconstructing a source image from the viewpoint of a target image ($I_{s\to t}$). **b** The system proposed in [1] (IID) extends the basic approach with Lambertian decomposition ($\Phi_{Decompose} \to$ I=AS), auto-masking (μ_2), and a light adjustment network, Φ_{Adjust}. **c** Our proposed system extends IID with non-Lambertian decomposition (I=AS+M) through a pre-trained inpainting network (P_{Inp}) and two auto-masking techniques ($\mu_1 \odot \mu_2$) without the need for an adjustment network

$$L_r(AS_{s\to t}, I_{t,\text{rem}}) = \alpha \frac{1 - SSIM(AS_{s\to t}, I_{t,\text{rem}})}{2} + (1-\alpha) \left\| AS_{s\to t} - I_{t,\text{rem}} \right\|_1 \tag{5}$$

$$L_{es}(D_t, I_t) = |\partial x D_t| \, e^{-|\partial x I_t|} + |\partial y D_t| \, e^{-|\partial y I_t|} \tag{6}$$

Two auto-masking techniques were adopted and applied to the L_a and L_r losses. The first auto-masking technique of Eq. (7) from Monodepth2 [10] reduces the problem of infinite depth with objects that move with the camera such as overlaid text and shapes from the endoscopic system. The second auto-masking technique from [1] tackles the problem of missing regions between frames due to camera movement (Eq. (8)). The final mask is their element-wise multiplication $\mu = \mu_1 \odot \mu_2$.

$$\mu_1 = \min_s L_r(I_t, I_{s\to t}) < \min_s L_r(I_t, I_s) \tag{7}$$

$$\mu_2 = I_{s\to t} > 0 \tag{8}$$

The final loss in Eq. (9) is composed of the decomposition, albedo, reconstruction, and smoothness losses. Here, λ_d, λ_a, λ_r, and λ_{es} are set to 0.2, 0.2, 1, and 0.01 as advised in [1].

$$L = \lambda_d(L_d(AS_t, I_{t,\text{rem}}) + L_d(AS_s, I_{s,\text{rem}})) + \lambda_a L_a \odot \mu + \lambda_r L_r \odot \mu + \lambda_{es} L_{es} \tag{9}$$

Our modifications and the difference between the proposed system training and [1] include the removal of their adjustment module, the incorporation of the inpainting module affecting L_d and L_r, and the addition of the auto-masking technique for static pixels proposed in [10]. These modifications are highlighted in orange in Fig. 3 and visually summarised in Fig. 2 with a high-level comparison of methods. We also train on real colonoscopy data as opposed to ex vivo data used in [1].

Inference

At inference time, a single frame is used to generate pose, depth, albedo, shading, inpainted image (AS), and specularity mask ($M = binarize(I - AS),\ threshold = 50$).

Experiments

Data

The following datasets were used in our experiments:

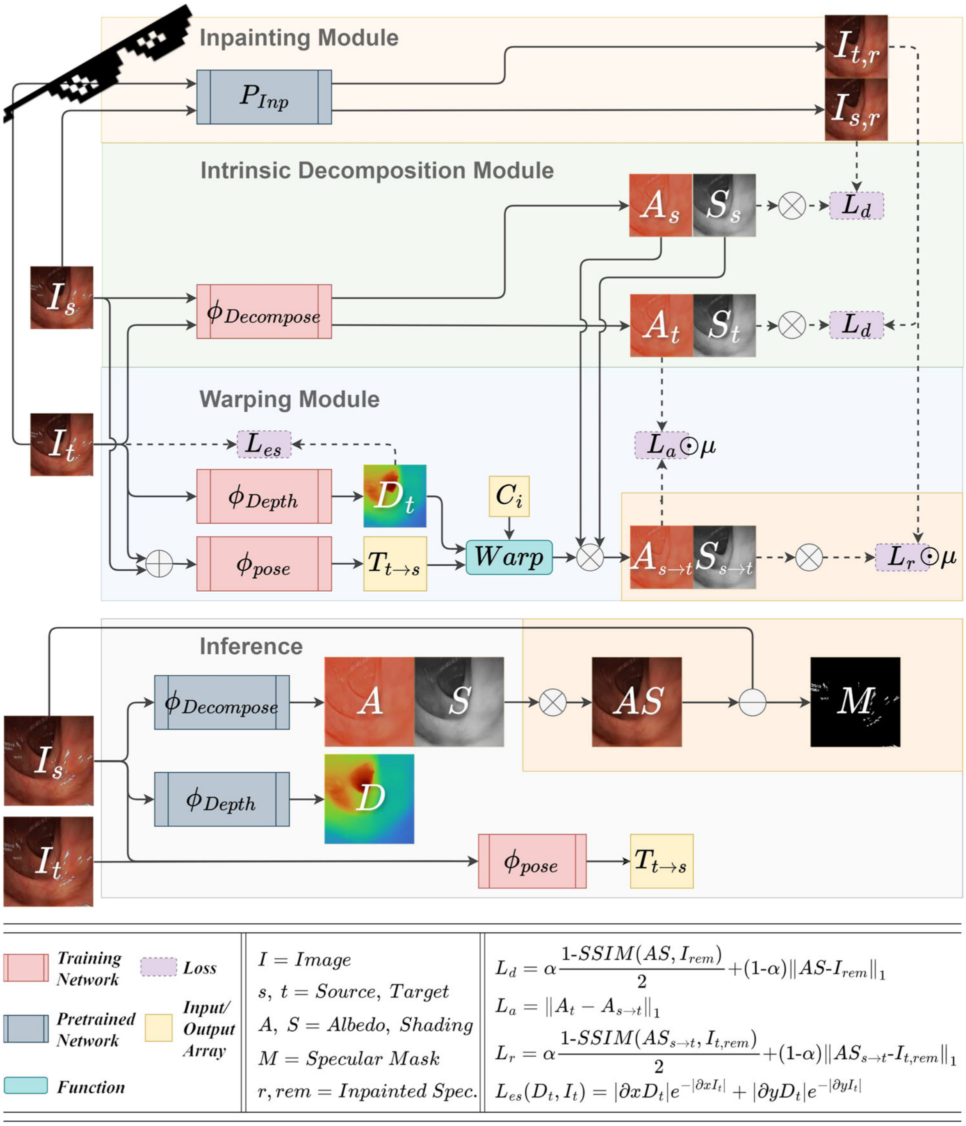
Inpainting Module
P_{Inp}
$I_{t,r}$
$I_{s,r}$
Intrinsic Decomposition Module
I_s
$\phi_{Decompose}$
A_s S_s
L_d
A_t S_t
L_d
Warping Module
I_t
L_{es}
$L_a \odot \mu$
ϕ_{Depth}
D_t
C_i
ϕ_{pose}
$T_{t\to s}$
$Warp$
$A_{s\to t}$ $S_{s\to t}$
$L_r \odot \mu$
Inference
$\phi_{Decompose}$
A S
AS
M
I_s
ϕ_{Depth}
D
I_t
ϕ_{pose}
$T_{t\to s}$
Training Network
Loss
Pretrained Network
Input/ Output Array
Function
I = Image
$s,\ t$ = Source, Target
$A,\ S$ = Albedo, Shading
M = Specular Mask
r, rem = Inpainted Spec.
$L_d = \alpha \frac{1\text{-}SSIM(AS, I_{rem})}{2} + (1\text{-}\alpha)\|AS\text{-}I_{rem}\|_1$
$L_a = \|A_t - A_{s\to t}\|_1$
$L_r = \alpha \frac{1\text{-}SSIM(AS_{s\to t}, I_{t,rem})}{2} + (1\text{-}\alpha)\|AS_{s\to t}\text{-}I_{t,rem}\|_1$
$L_{es}(D_t, I_t) = |\partial x D_t| e^{-|\partial x I_t|} + |\partial y D_t| e^{-|\partial y I_t|}$

Fig. 3 Flowchart of the proposed system. During training we compare a reconstructed source image warped to target $AS_{s\to t}$ against an inpainted target with removed specularities ($I_{t,\text{rem}}$) through the loss L_r, while making sure the depth is smooth (L_{es}) and the decomposition is self-supervised through L_d and L_a. At inference time albedo, shading, pose, and depth are estimated (A, S, T, D) and from those a reconstructed specular free image (AS) and a specular mask (M) are also generated. Our contributions are highlighted in orange

- Data$_{\text{real}}$—A colonoscopy dataset from Hyper Kvasir [27] with a Boston Bowel Preparation Scale of 2 or 3, which indicates high-quality mucosal views. We used 16,976 frames for training and 786 for testing. A cap of 926 frames per video was set.
- Data$_{\text{phantom}}$—A phantom dataset from C3VD [28] with 22 video sequences (10,015 images). This dataset was used for testing generalisability.

All images were first cropped to square and then resized to 288 × 288. Next, all datasets were undistorted using the camera intrinsics and distortion coefficients from Data$_{\text{phantom}}$ [28]. These parameters were applied to both Data$_{\text{phantom}}$ and Data$_{\text{real}}$, as the latter did not provide its own intrinsics, and the Data$_{\text{phantom}}$ parameters provided the most reasonable approximation. We observed that applying this undistortion yielded better results than leaving the data uncorrected. This approach aligns with common practices for datasets lacking camera intrinsics, where parameters are estimated when unavailable [10].

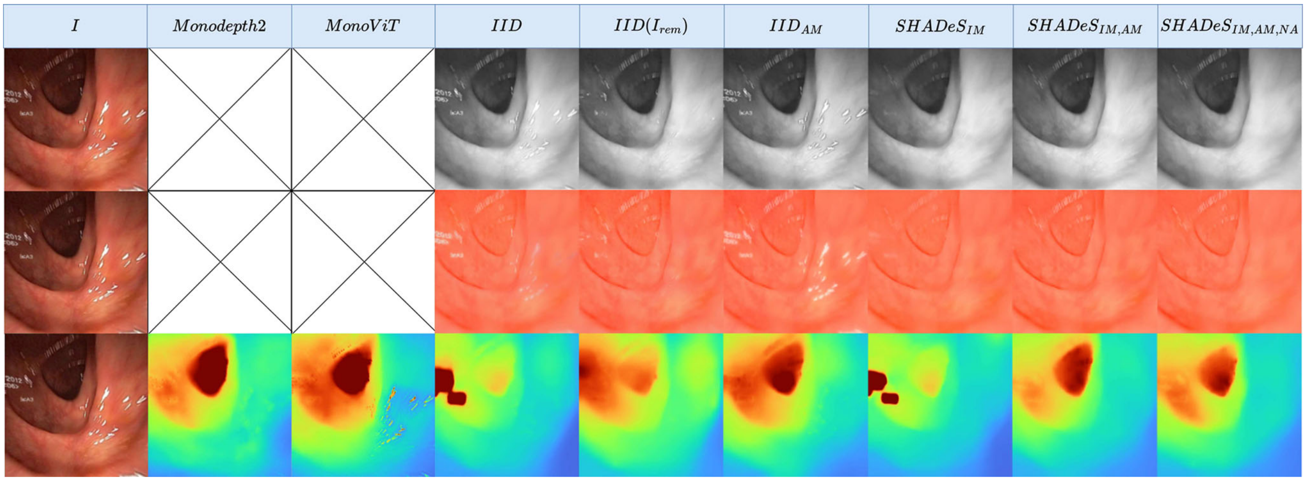

Fig. 4 Visual results of estimated shading, albedo, and depth on $\text{Data}_{\text{real}}$. For visual clarity, we clip the depth at 0.8

Table 1 Specularity surrounding metric (SSM) results on $\text{Data}_{\text{real}}$. SSM evaluates the percentage of smooth depth specular regions by comparing the depth in those regions to their surroundings. The best results are in bold and second-best are underlined

Methods	Monodepth2	MonoViT	IID	$IID(I_{\text{rem}})$	IID_{AM}	$SHADeS_{IM}$	$SHADeS_{IM,AM}$	$SHADeS_{IM,AM,NA}$
SSM (%)	39.1	41.8	63.7	62.8	43.2	68.3	**70.6**	<u>70.0</u>

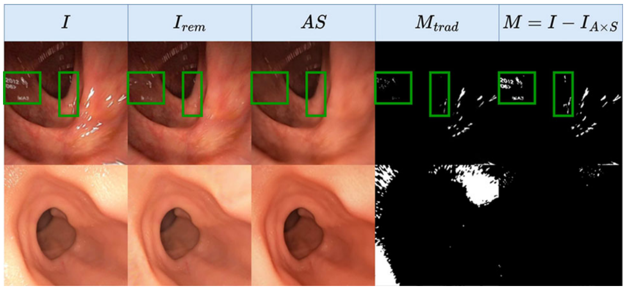

Fig. 5 Results on (row 1) $\text{Data}_{\text{real}}$ and (row 2) $\text{Data}_{\text{phantom}}$ showing estimated reconstructed images AS and specularity masks M vs. their counterparts (I_{rem}, M_{trad}) from [7]

Models

For comparison, we train $Monodepth2$ [10], $MonoViT$ [5], IID [1], and our method, $SHADeS$, which adds an inpainting module (IM), μ_1 auto-masking (AM), and removes the adjustment network (no adjustment: NA) from IID and thus we also refer to it as $SHADeS_{IM,AM,NA}$ for clarity. To analyse the importance of our modifications, we perform ablation studies by training IID with the added μ_1 auto-masking, IID_{AM}. We also train the proposed model without the adjustment network, $SHADeS_{IM,AM}$, and without μ_1 auto-masking, $SHADeS_{IM}$.

Setup

All experiments were performed on an NVIDIA V100-DGXS. For training, we follow the parameters and implementation of each method. However, we remove flipping since the camera centre is not in the image centre. The number of training epochs was also changed from 30 to 20 for IID. For training $SHADeS$, we followed the same parameters of IID with the changes described.

We initialise $SHADeS$ and IID with their pre-trained depth model [1]. For $Monodepth2$ [10] and $MonoViT$ [5], we also used their pre-trained models (mono_640x192) for initialisation. However, both IID and $MonoViT$ did not provide a pre-trained model for the pose network, thus we used the Monodepth2's pre-trained pose model to initialise them and $SHADeS$.

Evaluations

We calculate metrics for each image and then compute the average across all images. The metrics we rely on are:

1. **Specularity surrounding metric (SSM):** Since $\text{Data}_{\text{real}}$ lacks ground truth, we evaluate performance in specular regions segmented using [26]. Note that this segmentation method was also used during training, introducing a potential bias. We calculate the percentage of specular regions whose mean depth ($\text{Mean}_{\text{spec}}$) is close to their surrounding mean depth ($\text{Mean}_{\text{surr}}$) within a bounding box.

Table 2 Depth estimation quantitative results (in mm) on $\text{Data}_{\text{phantom}}$ with best results in bold

Methods	Monodepth2	MonoViT	IID	$IID(I_{\text{rem}})$	IID_{AM}	$SHADeS_{IM}$	$SHADeS_{IM,AM}$	$SHADeS_{IM,AM,NA}$
$MAE \downarrow$	4.6	5.0	4.6	4.6	4.7	4.8	4.5	**4.4**
$MedAE \downarrow$	3.3	3.1	3.2	3.3	3.3	3.4	**3.0**	3.1
$RMSE \downarrow$	**6.3**	7.4	6.8	6.6	6.7	7.0	6.4	**6.3**
$RMSE_{\log} \downarrow$	0.1694	0.1667	0.1856	0.1855	0.1714	0.1972	**0.1607**	0.1609
$Abs_{\text{Rel}} \downarrow$	0.1384	0.1391	0.1476	0.1481	0.1420	0.1590	0.1314	**0.1312**
$Sq_{\text{Rel}} \downarrow$	1.0198	1.2823	1.2793	1.1903	1.1288	1.5685	0.9879	**0.9599**
$\delta < 1.25 \uparrow$	0.8114	0.8230	0.8053	0.8031	0.8162	0.7817	0.8396	**0.8397**
$\delta < 1.25^2 \uparrow$	0.9839	0.9839	0.9613	0.9612	0.9763	0.9547	0.9855	**0.9858**
$\delta < 1.25^3 \uparrow$	0.9987	**0.9991**	0.9926	0.9920	0.9971	0.9887	0.9984	0.9985

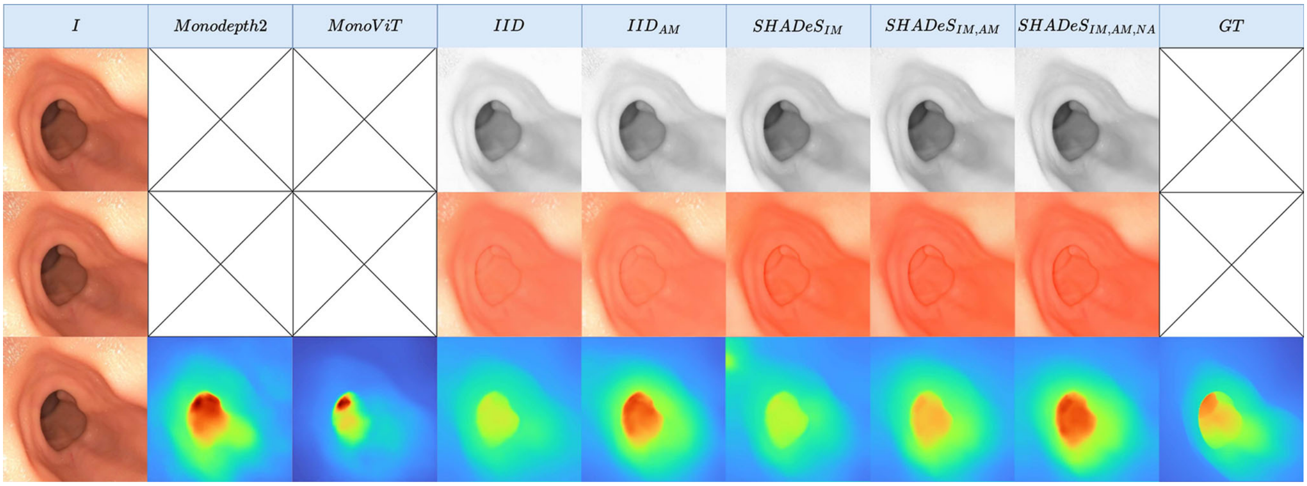

Fig. 6 Visual results of estimated shading, albedo, and depth on $\text{Data}_{\text{phantom}}$

This portrays the method's ability to generate smooth depth maps along specularities. More details can be found in the supplementary material.

2. **Direct error metrics following LD** [24] (MAE, $MedAE$, $RMSE$, $RMSE_{\log}$, Abs_{Rel}, Sq_{Rel}, $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$): To evaluate the generalisability of the models, we calculate standard depth metrics between $\text{Data}_{\text{phantom}}$ and the ground truth. To scale the predicted depth maps, we use median scaling [24] where the ratio between the median ground truth and the median prediction is applied.

Results & discussion

Qualitative and quantitative depth estimation results on $\text{Data}_{\text{real}}$ are shown in Fig. 4 row 3 and Table 1. From these results, we can see that Monodepth2 and MonoViT, which are not tailored for the medical field, perform the worst; This shows the importance of image decomposition. Furthermore, inpainting images as a preprocessing step (IID_{inp}) does not positively affect IID showing that one-step solutions without preprocessing can perform as well and even significantly better (e.g. all $SHADeS$ variations). We also find that methods with an inpainting module (IM) perform better than others, particularly in specular regions. We also notice that μ_1 auto-masking (AM) degrades IID while improving our more realistic non-Lambertian model. We also note the importance of auto-masking in removing the infinite depth effect of static pixels such as text overlays on the image. Finally, removing the adjustment module (NA) did not impact results significantly, which suggests that the adjustment module is unnecessary with our method.

The same conclusions can be made for albedo and shading when looking at visual results in rows 1 and 2 of Fig. 4 with even more obvious improvements in specular regions. We also notice that the albedo and shading with IM methods are even better than the albedo and shading of $IID(I_{\text{rem}})$. This suggests that the model does not only learn to inpaint these specularities but also learns to detect and inpaint specularities not detected by the inpainting pipeline's segmented maps M_{trad} [26].

This learnt specularity knowledge can also be seen in the reconstructed image (AS) and specularity mask (M) in row 1 Fig. 5, where both improved from the traditional methods used within the inpainting module in training [7, 26].

To analyse our model's generalisability, we evaluate on Data$_{\text{phantom}}$. Quantitatively (Table 2), $SHADeS_{IM,AM,NA}$ slightly outperforms other methods. Qualitatively (row 3 Fig. 6), depth estimates are similar across methods, likely due to all being trained on real data, making generalisation to phantom data challenging. In conclusion, our method $SHADeS_{IM,AM,NA}$ generalises on par with SOTA methods while still improving albedo and shading in specular regions (rows 1, 2 Fig. 6).

Conclusion

This paper introduces a non-Lambertian self-supervised model that decomposes a single image into its intrinsic components, shading, albedo, depth, and specularity map (SHADeS). Our model improves over Lambertian methods by generating and utilising an additional specular component. In comparison with state-of-the-art methods, results on real data (Hyper Kvasir) show the robustness of our method to specularities visually and using a specularity smoothness depth metric. Our model can also generalise to phantom data (C3VD) as demonstrated visually and quantitatively (RMSE).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11548-025-03371-8.

Acknowledgements This research was funded, in part, by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145/Z/16/Z]; the Engineering and Physical Sciences Research Council (EPSRC) [EP/P027938/1, EP/R004080/1, EP/P012841/1]; the Royal Academy of Engineering Chair in Emerging Technologies Scheme; H2020 FET (GA863146); and the UCL Centre for Digital Innovation through the Amazon Web Services (AWS) Doctoral Scholarship in Digital Innovation. For the purpose of open access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

Funding Wellcome/EPSRC Centre for Interventional and Surgical Sciences, Amazon Web Services, H2020 Future and Emerging Technologies, Royal Academy of Engineering.

References

1. Li B, Liu B, Zhu M, Luo X, Zhou F (2024) Image intrinsic-based unsupervised monocular depth estimation in endoscopy. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2024.3400804
2. World Cancer Research Fund International: Colorectal Cancer Statistics (2022). www.wcrf.org/cancer-trends/colorectal-cancer-statistics
3. World Health Organization: Colorectal Cancer (2023). www.who.int/news-room/fact-sheets/detail/colorectal-cancer
4. Society AC (2024) Can colorectal cancer be found early? www.cancer.org/cancer/types/colon-rectal-cancer/detection-diagnosis-staging/detection.html
5. Zhao C, Zhang Y, Poggi M, Tosi F, Guo X, Zhu Z, Huang G, Tang Y, Mattoccia S (2022)Monovit: self-supervised monocular depth estimation with a vision transformer. In: 2022 international conference on 3D vision (3DV). IEEE, pp 668–678
6. Liu S, Fan J, Yang Y, Xiao D, Ai D, Song H, Wang Y, Yang J (2024) Monocular endoscopy images depth estimation with multi-scale residual fusion. Comput Biol Med 169:107850
7. Daher R, Vasconcelos F, Stoyanov D (2023) A temporal learning approach to inpainting endoscopic specularities and its effect on image correspondence. Med Image Anal 90:102994
8. Daher R, Barbed OL, Murillo AC, Vasconcelos F, Stoyanov D (2023) Cyclesttn: a learning-based temporal model for specular augmentation in endoscopy. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 570–580
9. Zhou T, Brown M, Snavely N, Lowe DG (2017) Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1851–1858
10. Godard C, Mac Aodha O, Firman M, Brostow GJ (2019) Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3828–3838
11. Bian J, Li Z, Wang N, Zhan H, Shen Cu, Cheng M-M, Reid I (2019) Unsupervised scale-consistent depth and ego-motion learning from monocular video. In Advances in neural information processing systems, vol 32
12. Ozyoruk KB, Gokceler GI, Bobrow TL, Coskun G, Incetan K, Almalioglu Y, Mahmood F, Curto E, Perdigoto L, Oliveira M et al (2021) Endoslam dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos. Med Image Anal 71:102058
13. Rau A, Bhattarai B, Agapito L, Stoyanov D (2023) Task-guided domain gap reduction for monocular depth prediction in endoscopy. In: MICCAI workshop on data engineering in medical imaging. Springer, pp 111–122
14. Shao S, Pei Z, Chen W, Zhu W, Wu X, Sun D, Zhang B (2022) Self-supervised monocular depth and ego-motion estimation in endoscopy: appearance flow to the rescue. Med Image Anal 77:102338
15. Zhou L, Luo J, Wang H, Zhao S, Han Y, Li W (2023) Tackling challenges of low-texture and illumination variations for endoscopy self-supervised monocular depth estimation. In: 2023 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 2427–2432
16. Lou A, Noble J (2024) Ws-sfmlearner: self-supervised monocular depth and ego-motion estimation on surgical videos with unknown camera parameters. In: Medical imaging 2024: image-guided procedures, robotic interventions, and modeling, vol. 12928. SPIE, pp 119–127

17. Shi X, Cui B, Clarkson MJ, Islam M (2024) Long-term reprojection loss for self-supervised monocular depth estimation in endoscopic surgery. Artif Intell Surg 4(3):247–257
18. Li Y (2023) Endodepthl: lightweight endoscopic monocular depth estimation with CNN-transformer. In: 2023 IEEE international conference on bioinformatics and biomedicine (BIBM). IEEE, pp 4344–4351
19. Yue H, Gu Y (2023) Tcl: triplet consistent learning for odometry estimation of monocular endoscope. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 144–153
20. Wu J, Gu Y (2023) Unleashing the power of depth and pose estimation neural networks by designing compatible endoscopic images. arXiv preprint arXiv:2309.07390
21. Liao C, Wang C, Wang P, Wu H, Wang H (2024) Self-supervised learning of monocular depth and ego-motion estimation for non-rigid scenes in wireless capsule endoscopy videos. Biomed Signal Process Control 91:105978
22. Rodriguez-Puigvert J, Recasens D, Civera J, Martinez-Cantin R (2022) On the uncertain single-view depths in colonoscopies. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 130–140
23. Wang S, Zhang Y, McGill SK, Rosenman JG, Frahm J-M, Sengupta S, Pizer SM (2023) A surface-normal based neural framework for colonoscopy reconstruction. In: International conference on information processing in medical imaging. Springer, pp 797–809
24. Rodríguez-Puigvert J, Batlle VM, Montiel J, Martinez-Cantin R, Fua P, Tardós JD, Civera J (2023) Lightdepth: single-view depth self-supervision from illumination decline. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 21273–21283
25. Shi J, Dong Y, Su H, Yu SX (2017) Learning non-Lambertian object intrinsics across shapenet categories. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1685–1694
26. El Meslouhi O, Kardouchi M, Allali H, Gadi T, Benkaddour YA (2011) Automatic detection and inpainting of specular reflections for colposcopic images. Cent Eur J Comput Sci 1(3):341–354
27. Borgli H, Thambawita V, Smedsrud PH, Hicks S, Jha D, Eskeland SL, Randel KR, Pogorelov K, Lux M, Nguyen DTD et al (2020) Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Sci Data 7(1):1–14
28. Bobrow TL, Golhar M, Vijayan R, Akshintala VS, Garcia JR, Durr NJ (2023) Colonoscopy 3D video dataset with paired depth from 2D–3D registration. Med Image Anal 90:102956