

Depth-Aware Endoscopic Video Inpainting

Francis Xiatian Zhang¹[0000–0003–0228–6359], Shuang
Chen¹[0000–0002–6879–7285], Xianghua Xie²[0000–0002–2701–8660], and
Hubert P. H. Shum¹(✉)[0000–0001–5651–6039]

¹ Department of Computer Science, Durham University, Durham, United Kingdom
{xiatian.zhang, shuang.chen, hubert.shum}@durham.ac.uk

² Department of Computer Science, Swansea University, Swansea, United Kingdom
x.xie@swansea.ac.uk

Abstract. Video inpainting fills in corrupted video content with plausible replacements. While recent advances in endoscopic video inpainting have shown potential for enhancing the quality of endoscopic videos, they mainly repair 2D visual information without effectively preserving crucial 3D spatial details for clinical reference. Depth-aware inpainting methods attempt to preserve these details by incorporating depth information. Still, in endoscopic contexts, they face challenges including reliance on pre-acquired depth maps, less effective fusion designs, and ignorance of the fidelity of 3D spatial details. To address them, we introduce a novel Depth-aware Endoscopic Video Inpainting (DAEVI) framework. It features a Spatial-Temporal Guided Depth Estimation module for direct depth estimation from visual features, a Bi-Modal Paired Channel Fusion module for effective channel-by-channel fusion of visual and depth information, and a Depth Enhanced Discriminator to assess the fidelity of the RGB-D sequence comprised of the inpainted frames and estimated depth images. Experimental evaluations on established benchmarks demonstrate our framework’s superiority, achieving a 2% improvement in PSNR and a 6% reduction in MSE compared to state-of-the-art methods. Qualitative analyses further validate its enhanced ability to inpaint fine details, highlighting the benefits of integrating depth information into endoscopic inpainting.

Keywords: Endoscopy · Video Inpainting · Deep Learning

1 Introduction

In endoscopic videos, occlusions or artifacts caused by reflections or instrument shadows significantly degrade the visual quality. This issue is commonly known as *corruptions*, hiding critical anatomical details required for endoscopy examinations and surgeries, affecting clinical decision significantly [16]. As a technique to improve video quality by reconstructing the corrupted regions based on the uncorrupted information, video inpainting is introduced by [20,21,1,6] into the endoscopic scenario to mitigate the corruptions, known as endoscopic video inpainting. While these methods could reconstruct 2D visual information in corrupted endoscopic videos, they suffer from preserving vital 3D spatial details,

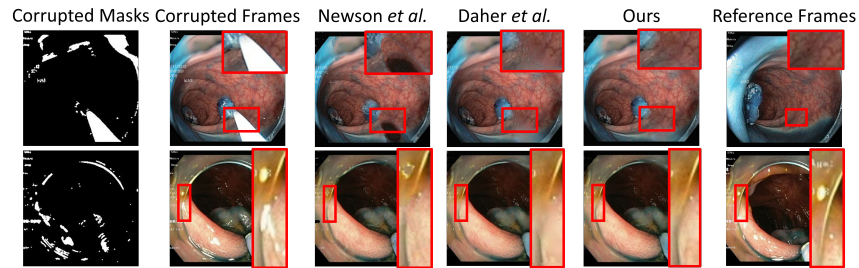


Fig. 1. Comparison with previous methods by Newson *et al.* [17] and Daher *et al.* [6] on corrupted frames from the HyperKvasir dataset [4]. Red boxes highlight significant differences. Reference frames are near frames with less corruption. Our inpainted content is not only visually plausible but also contextually realistic.

resulting in the artifact and spatial inconsistency at the inpainted regions, such low-fidelity performance limits their reliability for clinical applications.

Employing depth map to complement 3D spatial information is widely applied in general video painting [15,22,14], which offers a promising solution to preserve 3D spatial awareness for endoscopic video inpainting. Nevertheless, applying this solution is hindered by three significant challenges: First, it is not feasible to pre-acquire endoscopic depth maps, as the depth sensor is not available in standard monocular endoscopic cameras [9]. Second, given the learned depth features from deep-learning-based methods, simply concatenating visual features channel-wise and using vanilla convolution for fusion [14] fails to effectively exploit the depth representation, as they tend to capture redundant representations from visual features [13], losing the 3D spatial details complementing by depth maps. Third, none of these methods [15,22,14] assess the 3D spatial fidelity in RGB inpainted outputs, which compromises the reliability of inpainted content.

To address these challenges, we propose the Depth-Aware Endoscopic Video Inpainting (DAEVI) framework that provides more reliable inpainted details for improved clinical reference. It consists of a Spatial-Temporal Guided Depth Estimation (STGDE) module, a Bi-Modal Paired Channel Fusion (BMPCF) module, and a Depth-Enhanced Discriminator (DED), each designed to overcome the respective challenge. First, our STGDE module extracts depth information during visual feature learning to provide 3D spatial information, thus avoiding the requirement for pre-acquired depth maps as input. Second, the BMPCF module conducts a tailor-made feature fusion algorithm to better correlate the 3D spatial relevancy between visual and depth features by pair-wise fusing each visual and depth feature. Third, our DED assesses the 3D spatial fidelity of the RGB-D sequence formed by the inpainted frames and estimated depths, promoting realistic outputs with plausible 3D spatial details.

We evaluate our method on the HyperKvasir endoscopic video dataset [4] and compare it with the corresponding benchmark [23,6]. The quantitative experiments demonstrate that our proposed DAEVI outperforms state-of-the-art

approaches [6], achieving approximately 2% better Peak Signal-to-Noise Ratio (PSNR) and 6% lower Mean Squared Error (MSE). Our qualitative results show that DAEVI inpaints more fine-grained details, such as microvessels and the boundary of instruments, as depicted in Fig. 1. Furthermore, we directly apply our DAEVI trained on HyperKvasir to the SERV-CT datasets [8], demonstrating our method’s generalizability in endoscopic video inpainting. Our source code is available on <https://github.com/FrancisXZhang/DAEVI>.

Our work contributes in several ways, as outlined below:

1. To the best of our knowledge, DAEVI is the first endoscopic video inpainting framework to incorporate depth information. The effectiveness is demonstrated by comprehensive experiments.
2. We propose a Spatial-Temporal Guided Depth Estimation module, to translate depth representation directly from latent visual features, hence circumventing the challenge of acquiring depth maps during endoscopic surgery.
3. We design a Bi-Modal Paired Channel Fusion module that fuses each pair of channels from visual and depth features, which effectively leverages 3D spatial details in endoscopic inpainting.
4. We introduce a Depth-Enhanced Discriminator within our end-to-end optimization, which assesses the fidelity of the inpainted RGB-D sequence, promoting realistic outputs with more plausible 3D spatial details.

2 Methodology

Given the input endoscopic video frames $X \in \mathbb{R}^{T \times H \times W \times 3}$, we leverage the binary mask $M \in \mathbb{R}^{T \times H \times W \times 1}$, which identifies the corrupted regions, to get the input $X_M = X \odot M$. After processing by our DAEVI to generate the uncorrupted output $\hat{Y} \in \mathbb{R}^{T \times H \times W \times 3}$. \odot is the element-wise product, $H \times W$ is the spatial dimension. The whole formulation is denoted as: $\hat{Y} = DAEVI(X_M)$.

The overall architecture is shown in Fig. 2. First, DAEVI employs a convolutional encoder to embed X_M into compact visual features $F \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 4C}$ to effectively represent local visual features. Following this, our Spatial-Temporal Guided Depth Estimation (STGDE) module learns multi-level visual features and translates them into depth maps. Subsequently, the proposed Bi-Modal Paired Channel Fusion (BMPCF) module fuses visual and depth features to obtain an integrated representation with enhanced 3D spatial details. After that, a convolutional decoder reconstructs the final inpainted frames \hat{Y} . During training, our Depth-Enhanced Discriminator (DED) assesses the visual and spatial fidelities of the inpainted RGB-D sequence.

2.1 Spatial-Temporal Guided Depth Estimation (STGDE)

Depth-aware endoscopic video inpainting faces a unique challenge in acquiring depth data, as the standard endoscopic cameras are technically unable to provide the raw depth [9]. To address this challenge, we propose a STGDE module to

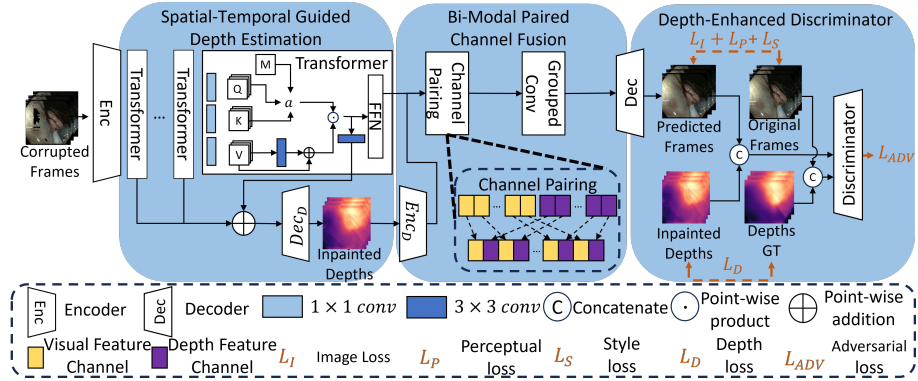


Fig. 2. The overview of our framework. First, our Spatial-Temporal Guided Depth Estimation module translates depth information from corrupted frames (See 2.1). Second, our Bi-Modal Paired Channel Fusion module effectively fuses visual features with depth features (See 2.2). Third, our Depth Enhanced Discriminator assesses the fidelity of the inpainted RGB-D sequence (See 2.3).

translate depth features from latent visual features. STGDE involves multiple transformer blocks TB and a depth decoder Dec_D . Each TB is a spatial-temporal transformer block [23] with a spatial enhancement to enhance the local feature learning, thereby improving the representation ability. Dec_D aims to gather the latent visual feature across all TBs for effective depth estimation.

Specifically, after the encoder, corrupted frames X_M are embedded as visual features F , which are fed into the first transformer block TB_1 . The output from TB_i is subsequently fed into the following TB_{i+1} , where $i \in (1, N_s - 1)$ and N_s is the number of TBs . The F^{i-1} , as the input for each TB_i , are linearly transformed into query Q^i , key K^i , and value V^i separately. We introduce a 3×3 depth-wise convolution $P_V^i(\cdot)$ to enhance the V^i for better spatial feature learning:

$$Q^i, K^i, V^i = P_Q^i(F^{i-1}), P_K^i(F^{i-1}), P_V^i(F^{i-1}) + P_V^i(F^{i-1}), \quad (1)$$

where $P_{Q,K,V}^i(\cdot)$ denote 1×1 2D convolutions. After that, we split each of Q^i , K^i , and V^i into $n = r_1 \times r_2$ smaller patches Q_p^i , K_p^i , and $V_p^i \in \mathbb{R}^{Tn \times c \times h/r_1 \times w/r_2}$, where $h/r_1 \times w/r_2$ is the spatial dimension of patches. Then we utilize the patched Q_p^i , K_p^i , and V_p^i to get the attention output F_{att}^i :

$$S^i = Q_p^i (K_p^i)^\top / \sqrt{r_1 \times r_2 \times c}, \quad (2)$$

$$F_{att}^i = \text{softmax}(S^i \odot M) V_p^i, \quad (3)$$

where S^i is the attention score, and M denotes a resized binary mask matrix indicating the corrupted regions [23]. Then, after a convolutional projection P_F^i

followed by a feed-forward network FFN^i , we obtain the output F^i of TB_i :

$$F^i = FFN^i (P_F^i (F_{att}^i)). \quad (4)$$

To translate depth maps \hat{D} from latent visual features, we aggregate F_{att} across all TB_s to gather the multi-layer visual representation. After a convolutional projection P_D^i , the depth decoder generate the depth maps \hat{D} :

$$\hat{D} = Dec_D \left(\sum_{i=1}^{N_s} P_D^i (F_{att}^i) \right). \quad (5)$$

2.2 Bi-Modal Paired Channel Fusion (BMPCF)

Endoscopic depth-aware inpainting encounters a challenge in effectively integrating depth with visual information, as the simple channel-wise concatenation followed by a vanilla convolution is unable to fully exploit the correlation between depth and visual feature, especially in endoscopic scenes such complex nature involving varied spatial structures [19].

To address this challenge and effectively enhance visual information with 3D spatial details, we design a BMPCF module to correlate each visual and depth feature by a tailor-made pair-wise fusion algorithm.

Specifically, given the depth maps \hat{D} translated by Dec_D , we first enlarge the channel capacity of \hat{D} with a depth encoder Enc_D to obtain the embedded depth feature F_D , which has the same number of channels as the STGDE’s output F^{N_s} . Then, to ensure each depth correlates precisely to the corresponding visual feature, we sequentially interleave the sliced F^{N_s} and F_D in channel-wise:

$$F_{pair}[:, 2i] = \begin{cases} F^{N_s}[:, i] & \text{for } i = 0, 2, \dots, c-2, \\ F_D[:, i] & \text{for } i = 1, 3, \dots, c-1, \end{cases} \quad (6)$$

where c is the channel number of F^{N_s} and F_D , also indicates the number of pairs. After that, the group-wise convolution $G(\cdot)$ [12] is employed on F_{pair} to fuse each pair of visual and depth features, producing the fused output $F_f \in \mathbb{R}^{T \times c \times h \times w}$:

$$F_f = G(F_{pair}). \quad (7)$$

In this way, each convolutional kernel facilitates the fusion between every two adjacent channels consisting of one visual feature and one depth feature. Subsequently, a convolutional decoder reconstructs inpainted frames \hat{Y} from F_f .

2.3 Depth-Enhanced Discriminator (DED)

Effectively assessing the spatial fidelity is critical in endoscopic depth-aware inpainting, as minor inaccuracies, such as incorrect anatomical details, could significantly impact clinical decisions [7]. While Generative Adversarial Network (GAN) strategies [10] in previous depth-aware inpainting methods [22,14]

only enhance the fidelity of RGB content and ignore the fidelity in 3D spatial details, resulting in unreliable outputs for clinical reference [14]. To this end, we introduce DED during training, to comprehensively assess the RGB-D inpainted endoscopic frames across spatial, temporal and depth dimensions.

Specifically, we followed [5] to build our DED with 6 convolutional blocks. The inpainted frames and depth are concatenated as the RGB-D data to be the input of DED. The adversarial loss $L_{ADV} = L_{GEN} + L_{DED}$ is adopted from [10]:

$$L_{DED} = \mathbb{E}_{(D,Y) \sim P_{Data}} [Relu(1 - DED([D, Y]))] + \mathbb{E}_{(\hat{D}, \hat{Y}) \sim P_G} [Relu(DED([\hat{D}, \hat{Y}]))], \quad (8)$$

$$L_{GEN} = -\mathbb{E}_{(\hat{D}, \hat{Y}) \sim P_G} [DED([\hat{D}, \hat{Y}])], \quad (9)$$

To enhance parameter optimization efficiency, we adopt an end-to-end optimization strategy for our DAEVI. Our full loss function is as follows:

$$L = \lambda_D L_D + \lambda_I L_I + \lambda_{GEN} L_{GEN} + \lambda_P L_P + \lambda_S L_S, \quad (10)$$

where λ denotes the weight of each loss term. L_D and L_I are the L1 reconstruction loss [24] for translated depth and inpainted frames, respectively. L_P and L_S denote the perceptual loss and style loss, respectively [11] (details can be found in our supplementary material). In each iteration, L and L_{DED} optimize our inpainting network and our DED, respectively.

3 Experiment

3.1 Experimental Setting

We evaluate our method against the existing benchmark established [6] on the HyperKvasir endoscopic video dataset [4]. This dataset includes 373 videos with a total of 889,372 frames, 343 for training and 30 for testing. Depth ground truth is derived from a pre-trained endoscopic depth estimator [19] on unmasked frames, which is needed only in training. Following the benchmark setting [6], we employ their provided masks to identify corrupted regions in frames and calculate metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Mean Square Error (MSE) specifically for corrupted regions.

We configure the block number $N_s = 8$ and set the weights for λ_D , λ_P , λ_S , λ_I , and λ_{GEN} to [0.1, 0.1, 250, 1, 0.01]. We employ Adam optimizer with learning rate = 1e-4, β_1 : 0, β_2 : 0.99. All experiments are trained on an NVIDIA TITAN RTX 24G GPU with a batch size of 4 for 200k iterations. Training iterations alternate between selecting 5 random or consecutive frames resized to 288 x 288 pixels. For inference, the model processes every 5 corrupted frames alongside 10 nearby corrupted frames sampled for reference, reassembling them into the full video with a real-time processing speed of approximately 0.03 seconds per frame.

Table 1. Inpainting Performance Comparison and Ablation Study. w/o STGDE: A pre-trained depth estimator [19] is leveraged for depth estimation instead of STGDE; w/o BMPCF: Simple concatenation is used for fusion instead of BMPCF; w/o DED: A standard RGB discriminator [5] is used in GAN training instead of DED.

Methods	$PSNR_{Crop} \uparrow$	$SSIM_{Crop} \uparrow$	$MSE_{Crop} \downarrow$
Arnold <i>et al.</i> [2]	19.909	0.559	895.222
Newson <i>et al.</i> [17]	22.27	0.650	543.636
STTN [23]	28.683	0.793	119.541
Daher <i>et al.</i> [6]	29.542	0.785	104.719
DAEVI (Full Framework)	30.126	0.797	97.873
w/o STGDE	29.801	0.788	105.150
w/o BMPCF	29.695	0.791	103.861
w/o DED	29.286	0.797	108.903

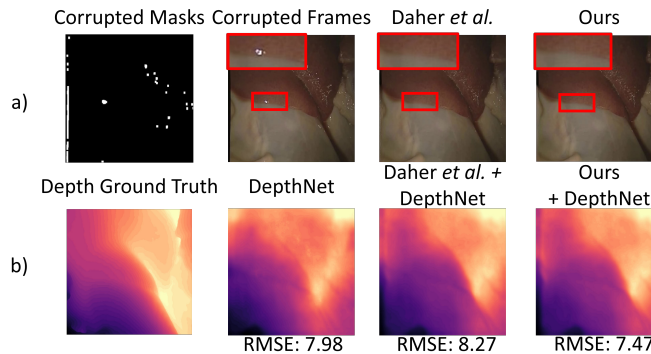


Fig. 3. Comparison of deep learning-based inpainting performance on the SERV-CT dataset: a) Generalization Capability, and b) Depth Information Preservation.

3.2 Results

Comparison with Existing Methods In Table 1, we benchmark our DAEVI against Arnold *et al.* [2], a diffusion-based approach; Newson *et al.* [17], employing temporal patches; STTN [23], a transformer-based method; and Daher *et al.* [6], the state-of-the-art (SOTA) method in endoscopic video inpainting. Our method outperforms existing methods across all metrics. These results highlight the benefits of integrating depth information into endoscopic video inpainting. Clinically, these improvements enhance the visibility of endoscopic videos [21], crucial for accurate diagnostics and surgical planning.

Qualitative Results To demonstrate our inpainting performance on corrupted regions, we remove specular reflections and high reflection from instruments in the HyperKvasir dataset [4] and the SERV-CT dataset, an external-body endoscopic dataset with depth ground truth provided [8]. Figure 1 from the HyperKvasir dataset illustrates our method more effectively restoring details such as microvessels and the interface between instruments and organs. These enhancements are crucial for safer and more efficient endoscopic operations [3]. More qualitative results can be found in our supplementary material.

Table 2. Performance Analysis Across Depth Estimation Block Configurations.

Methods	$PSNR_{Crop} \uparrow$	$SSIM_{Crop} \uparrow$	$MSE_{Crop} \downarrow$
DAEVI (First 4 Blocks)	29.921	0.792	100.851
DAEVI (Last 4 Blocks)	29.900	0.796	101.313
DAEVI (All 8 Blocks)	30.126	0.797	97.873

Table 3. Online Inference Performance Analysis

Methods	$PSNR_{Crop} \uparrow$	$SSIM_{Crop} \uparrow$	$MSE_{Crop} \downarrow$
Daher <i>et al.</i> [6]	29.542	0.785	104.719
DAEVI	30.126	0.797	97.873
DAEVI (Online)	30.117	0.797	98.147

Additionally, Fig. 3 a) shows our method’s ability to inpaint highly plausible content on the SERV-CT dataset without specific fine-tuning, underscoring our approach’s strong generalization capabilities. By applying the pre-trained DepthNet [19] to our inpainted frames, Fig. 3 b) reveals that our method improves depth estimation, achieving the lowest Root Mean Squared Error (RMSE) between the ground truth depth and the depth estimation after inpainting. This further underscores our method’s superior effectiveness in preserving 3D spatial details compared to existing methods.

Ablation Study Our ablation study in Table 1 shows that our full framework yields the best results, highlighting the importance of each module. Notably, removing DED lowers MSE and PSNR but doesn’t affect SSIM, which remains comparable to the complete framework. This may be because DED significantly enhances 3D spatial fidelity, which does not directly correspond to the features assessed by SSIM [18], such as contrast and luminance.

Depth Estimation Block Configuration Analysis Table 2 evaluates inpainting performance using different configurations of the first 4, last 4, and all 8 blocks of STGCN for depth estimation within the DAEVI framework. This analysis reveals that employing all 8 blocks, which combines both low-level and high-level features, achieves optimal performance, highlighting the effectiveness of our current STGDE design.

Online Inference Performance Analysis Table 3 assesses our method’s online inference ability, which employs only past frames for reference. Our online performance still outperforms the SOTA method [6] that uses both past and future frames as reference frames, demonstrating its potential for live endoscopic videos.

4 Conclusion and Discussions

In this paper, we propose the DAEVI framework, the first endoscopic video inpainting framework designed to incorporate depth information to achieve reliable 3D spatial details. This work offers a potential solution to enhance the quality of endoscopic videos, facilitating informed clinical decision-making. Our DAEVI leverages depth information from latent visual features of corrupted endoscopic frames with STGDE, effectively fuses visual and depth information

with BMPCF, and assesses the fidelity of the RGB-D content with DED. Experimental evaluations demonstrate its significant superiority compared to existing methods.

While our framework demonstrates improvements in inpainting performance, it is not without limitations. As our work’s focus is on corruption inpainting, the real-world applicability of DAEVI could still be influenced by the effectiveness of any external corruption detection backbone. Future work could integrate existing endoscopic corruption detection methods such as semantic segmentation [1] into our framework, optimizing both corruption detection and inpainting tasks.

Acknowledgments. This research is supported in part by the EPSRC NorthFutures project (ref: EP/X031012/1).

Disclosure of Interests. The authors declare that there are no conflicts of interest related to this paper.

References

1. Ali, S., Zhou, F., Bailey, A., Braden, B., East, J.E., Lu, X., Rittscher, J.: A deep learning framework for quality assessment and restoration in video endoscopy. *Medical Image Analysis* **68**, 101900 (2021)
2. Arnold, M., Ghosh, A., Ameling, S., Lacey, G.: Automatic segmentation and inpainting of specular highlights for endoscopic imaging. *EURASIP Journal on Image and Video Processing* **2010**, 1–12 (2010)
3. Ben-Menachem, T., Decker, G.A., Early, D.S., Evans, J., Fanelli, R.D., Fisher, D.A., Fisher, L., Fukami, N., Hwang, J.H., Ikenberry, S.O., et al.: Adverse events of upper gi endoscopy. *Gastrointestinal Endoscopy* **76**(4), 707–718 (2012)
4. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Scientific data* **7**(1), 283 (2020)
5. Chang, Y.L., Liu, Z.Y., Lee, K.Y., Hsu, W.: Free-form video inpainting with 3d gated convolution and temporal patchgan. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 9066–9075 (2019)
6. Daher, R., Vasconcelos, F., Stoyanov, D.: A temporal learning approach to inpainting endoscopic specularities and its effect on image correspondence. *Medical Image Analysis* p. 102994 (2023)
7. Dray, X., Histace, A., Robertson, A., Segui, S.: Artificial intelligence for protruding lesions. In: *Artificial Intelligence in Capsule Endoscopy*, pp. 121–148. Elsevier (2023)
8. Edwards, P.E., Psychogyios, D., Speidel, S., Maier-Hein, L., Stoyanov, D.: Serv-ct: A disparity dataset from cone-beam ct for validation of endoscopic 3d reconstruction. *Medical Image Analysis* **76**, 102302 (2022)
9. Fu, Z., Jin, Z., Zhang, C., He, Z., Zha, Z., Hu, C., Gan, T., Yan, Q., Wang, P., Ye, X.: The future of endoscopic navigation: A review of advanced endoscopic vision technology. *IEEE Access* **9**, 41144–41167 (2021)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. *Communications of the ACM* **63**(11), 139–144 (2020)

11. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II* 14. pp. 694–711. Springer (2016)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012)
13. Li, J., Wen, Y., He, L.: Seconv: Spatial and channel reconstruction convolution for feature redundancy. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6153–6162 (2023)
14. Li, S., Zhu, S., Ge, Y., Zeng, B., Imran, M.A., Abbasi, Q.H., Cooper, J.: Depth-guided deep video inpainting. *IEEE Transactions on Multimedia* (2023)
15. Liao, M., Lu, F., Zhou, D., Zhang, S., Li, W., Yang, R.: Dvi: Depth guided video inpainting for autonomous driving. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI* 16. pp. 1–17. Springer (2020)
16. Monkam, P., Wu, J., Lu, W., Shan, W., Chen, H., Zhai, Y.: Easyspec: Automatic specular reflection detection and suppression from endoscopic images. *IEEE Transactions on Computational Imaging* **7**, 1031–1043 (2021)
17. Newson, A., Almansa, A., Fradet, M., Gousseau, Y., Pérez, P.: Video inpainting of complex scenes. *Siam Journal on Imaging Sciences* **7**(4), 1993–2019 (2014)
18. Pambrun, J.F., Noumeir, R.: Limitations of the ssim quality metric in the context of diagnostic imaging. In: *2015 IEEE International Conference on Image Processing*. pp. 2960–2963. IEEE (2015)
19. Shao, S., Pei, Z., Chen, W., Zhu, W., Wu, X., Sun, D., Zhang, B.: Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue. *Medical Image Analysis* **77**, 102338 (2022)
20. Shen, D.F., Guo, J.J., Lin, G.S., Lin, J.Y.: Content-aware specular reflection suppression based on adaptive image inpainting and neural network for endoscopic images. *Computer Methods and Programs in Biomedicine* **192**, 105414 (2020)
21. Tukra, S., Marcus, H.J., Giannarou, S.: See-through vision with unsupervised scene occlusion reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(7), 3779–3790 (2021)
22. Yamashita, Y., Shimosato, K., Ukita, N.: Boundary-aware image inpainting with multiple auxiliary cues. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 619–629 (2022)
23. Zeng, Y., Fu, J., Chao, H.: Learning joint spatial-temporal transformations for video inpainting. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI* 16. pp. 528–543. Springer (2020)
24. Zhou, S., Li, C., Chan, K.C., Loy, C.C.: Propainter: Improving propagation and transformer for video inpainting. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10477–10486 (2023)

Depth-Aware Endoscopic Video Inpainting – Supplementary Material

Reconstruction Loss The details for L_D and L_I are as follows:

$$L_D = \|\hat{D} - D\|, \quad (1)$$

$$L_I = \|\hat{Y} - Y\|, \quad (2)$$

where $\|\cdot\|$ denotes the L1 Norm, D and \hat{D} denote the ground truth depth map and the translated depth map, respectively, and Y and \hat{Y} denote the ground truth frames and the inpainted frames, respectively.

Perceptron and Style Loss The details for L_P and L_S are as follows:

$$L_P = \sum_{l \in \text{Layers}} \frac{1}{N_l} \|F_l(\hat{Y}) - F_l(Y)\|_2^2, \quad (3)$$

$$L_S = \sum_{l \in \text{Layers}} \|G_l(\hat{Y}) - G_l(Y)\|_F^2, \quad (4)$$

where $F_l(\cdot)$ denotes the feature map extracted from layer l of a pre-trained network given frames as input, and $G_l(\cdot)$ represents the Gram matrix of the feature map from layer l , capturing the style information. $\|\cdot\|_2$ denotes the squared Euclidean (L_2) norm, and $\|\cdot\|_F$ denotes the squared Frobenius norm.

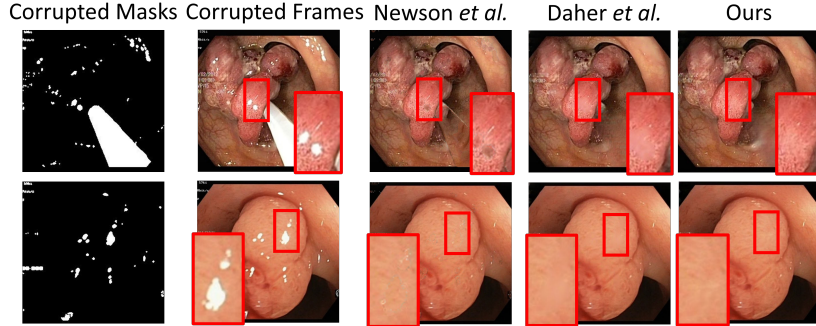


Fig. 1. More Cases from the HyperKvasir Dataset: These cases further demonstrate that our method outperforms others, especially in generating fewer artifacts and more plausible details during endoscopic inpainting. This underscores our approach’s superior corruption removal capability.

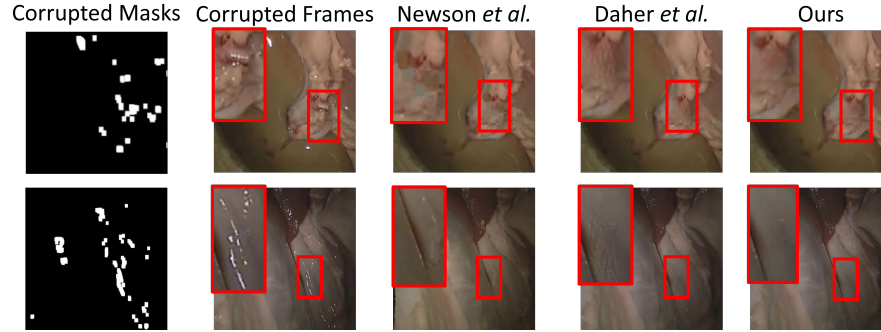


Fig. 2. More Cases from the SERV-CT Dataset: These cases further demonstrate that our method outperforms others without the need for any fine-tuning, especially in generating fewer artifacts during inpainting. This underscores our approach’s superior generalization capability.

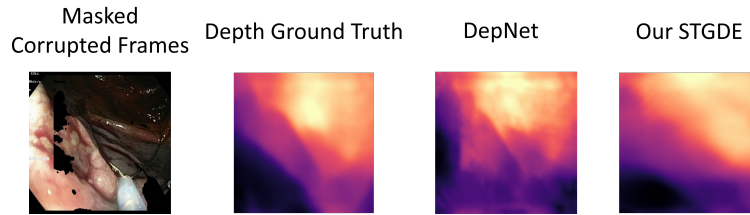


Fig. 3. Depth Estimation Performance Analysis of Our Spatial-Temporal Guided Depth Estimation (STGDE) Module. This analysis compares the performance of our STGDE module against a pre-trained endoscopic depth estimator DepthNet, on masked corrupted frames. The ground truth is derived from depth estimation on unmasked frames. It is observed that our STGDE module estimates depth more accurately and closer to the ground truth compared to the direct application of the pre-trained model on masked frames.