Depth-Aware Endoscopic Video Inpainting – Supplementary Material

Francis Xiatian Zhang[1][0000-0003-0228-6359], Shuang Chen[1][0000-0002-6879-7285], Xianghua Xie[2][0000-0002-2701-8660], and Hubert P. H. Shum[1]()[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`

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

$$L_D = \left|\hat{D} - D\right|, \tag{1}$$

$$L_I = \left|\hat{Y} - Y\right|, \tag{2}$$

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 [2]:

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

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

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 ($L2$) norm, and $|\cdot|_F$ denotes the squared Frobenius norm.

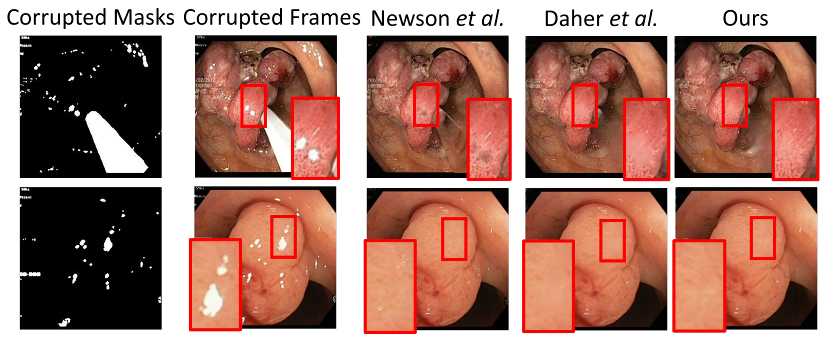

Fig. 1. More Cases from the HyperKvasir Dataset [1]: These cases further demonstrate that our method outperforms others, especially in generating fewer artifacts and more plausible details during endoscopic inpainting. This underscores our approach's superior corruption removal capability.

References

1. 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)
2. 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)
3. 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)

Fig. 2. More Cases from the SERV-CT Dataset [1]: 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.

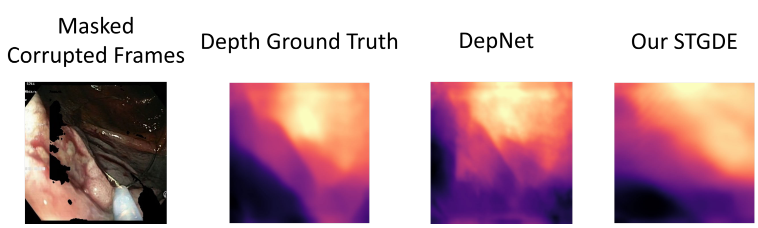

Fig. 3. Depth Estimation Performance Analysis of Our Spatial-Temporal Guided Depth Estimation (STGDE) Module. This analysis compares the performance of our STGDE module against a pre-trained endoscopic depth estimator DepthNet [3], on masked corrupted frames. The ground truth is derived from depth estimation 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.