A. Feature Visualization

In this section, we make a visual demonstration of features from different layers to show how CASCADE identifies salient (important) features and refines them to produce accurate segmentation maps. Figure 1 in this Supplementary shows that CASCADE adopts the features (i.e., X1, X2, X3, X4) from the hierarchical encoder backbone network (i.e., ImageNet pretrained PVTv2 [10]) and progressively groups pixels (of lesions) and enhances them. More specifically, Attention Gates (AGs) add additional (sometimes missing) features using the attention-based fusion of upsampled features and the features from skip connections. Afterward, Channel Attention (CA) and Spatial Attention (SA) identify important features and enhance them which is evident in Figure 1. It is also visible that CASCADE has a better feature map after aggregating features from all the layers of the encoder.

B. Qualitative Results

In this section, we compare the qualitative results of our proposed CASCADE decoder with state-of-the-art methods. We use some challenging examples from unseen ColonDB [8] and ETIS-LaribDB [7]. To train our model, we use the combined CVC-ClinicDB and Kvasir dataset, and keep experimental settings the same as described earlier. The results are described next.

B.1. Qualitative results on ColonDB

Figure 2 in this Supplementary shows visual outputs of different methods on five challenging images from ColonDB testset. In Figure 2(a), most of the methods are confused by the artifacts due to illumination and thus produce false positive results. However, CASCADE effectively ignores the artifact and correctly segments the lesion area having no false positives. Although CASCADE misses the part with low illumination in Figure 2(b), it effectively segments the right part of the lesion that is ignored by all other methods. This illumination issue hopefully can be resolved if we use some relevant data augmentations. CASCADE correctly segments the low contrast lesion region in Figure 2(c), where all other methods fail even with false positives. In Figure 2(d), CASCADE and PraNet [3] produce better results; PolypPVT [2] and PVT-CUP produce false positives distracting by the illumination effect in the low illumination part. Most of the methods fail to segment the lesion in Figure 2(e) due to illumination effects and the noisy texture of the lesion. However, CASCADE effectively overcomes such challenges and produces a high-quality segmentation map. From the visual results presented in Figure 2, we can conclude that due to using attention mechanisms in a sophisticated way CASCADE overcomes challenges that exist in images, and produce high-quality segmentation maps.

B.2. Qualitative results on ETIS-LaribDB

In this Supplementary Figure 3, we present the qualitative results of CASCADE along with state-of-the-art methods on another set (five) of images taken from the unseen ETIS-LaribDB testset. In this case, most of the lesions (polyp) are small and of low contrast (i.e., lesions are hardly separable from the background), however, we observe similar trends in results as ColonDB. In Figure 3(a-e), CNN-based methods, i.e., UNet [6], UNet++ [11], PraNet [3], UACANet [5] fail (except PraNet in Figure 3(c)) to segment the lesions rather lead to false positives. Among the transformer-based methods, SSFormerPVT [9] shows poor performance in small lesion segmentation, PolypPVT [2] fails to segment the lesion in Figure 3(e), and PVT-CUP fails to segment the lesions in Figure 3(b, e). However, our CASCADE segments the lesions well in all five images. Therefore, we can conclude that qualitative results demonstrate the superior performance of our proposed CASCADE decoder over state-of-the-art methods.

References

[1] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015. 2

[2] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *arXiv preprint arXiv:2108.06932*, 2021. 1

[3] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020. 1

[4] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020. 2

[5] Taehun Kim, Hyemin Lee, and Daijin Kim. Uacanet: Uncertainty augmented context attention for polyp segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2167–2175, 2021. 1

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[7] Juan Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Toward embedded detection of

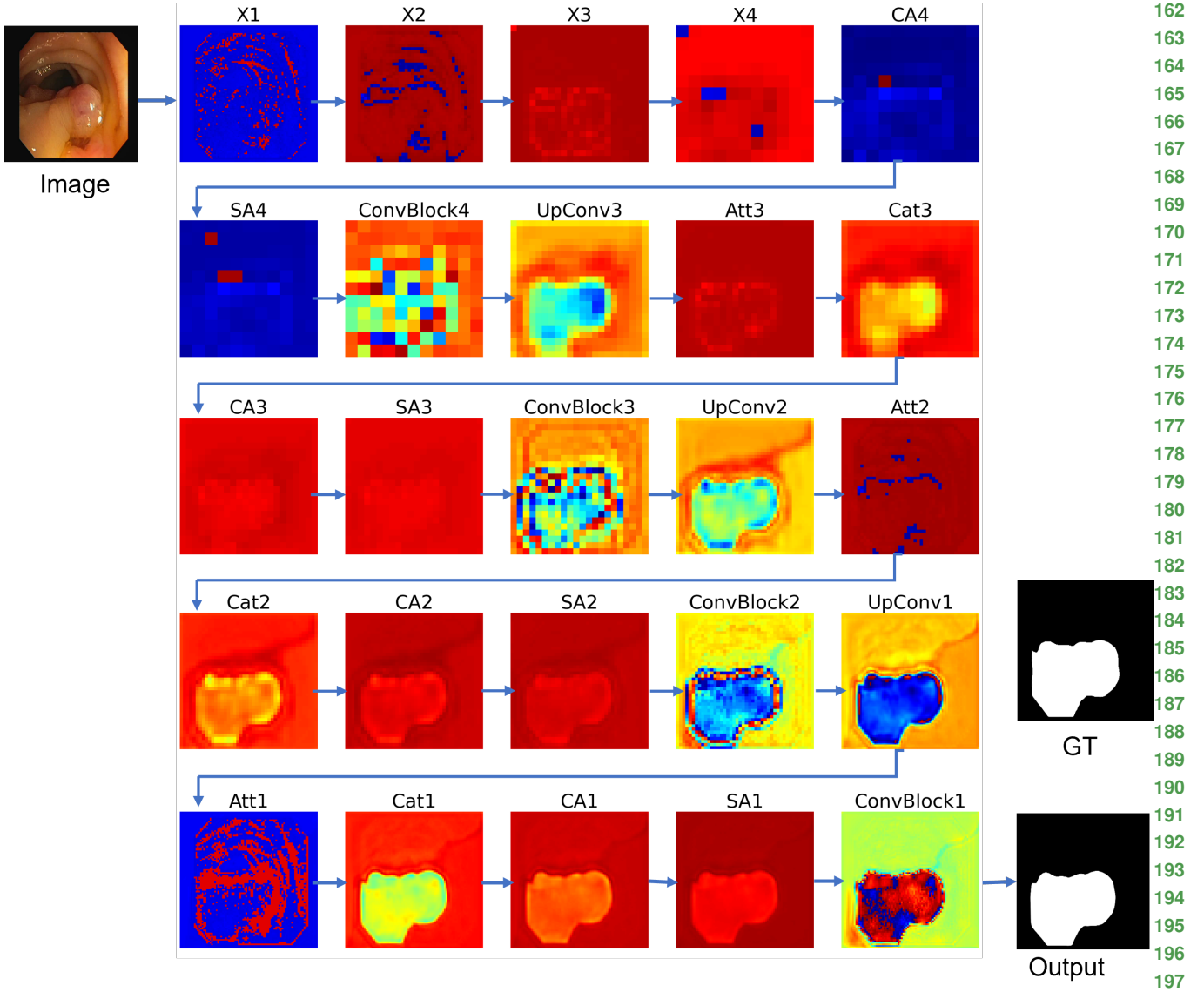

Figure 1. Visual demonstration of how CASCADE refines the feature maps (i.e., X1, X2, X3, X4) and produces the accurate segmentation map. The image is taken from CVC-ClinicDB testset. The features are captured during model evaluation. The model is trained on the combined CVC-ClinicDB [1] and Kvasir [4] dataset (with the same experimental setup discussed earlier).

polyps in wce images for early diagnosis of colorectal cancer. *International journal of computer assisted radiology and surgery*, 9(2):283–293, 2014. 1

[8] Nima Tajbakhsh, Suryakanth R Gurudu, and Jianming Liang. Automated polyp detection in colonoscopy videos using shape and context information. *IEEE transactions on medical imaging*, 35(2):630–644, 2015. 1

[9] Jinfeng Wang, Qiming Huang, Feilong Tang, Jia Meng, Jionglong Su, and Sifan Song. Stepwise feature fusion: Local guides global. *arXiv preprint arXiv:2203.03635*, 2022. 1

[10] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 1

[11] Z Zhou, MMR Siddiquee, N Tajbakhsh, and J Liang. A nested u-net architecture for medical image segmentation. arxiv 2018. *arXiv preprint arXiv:1807.10165*. 1

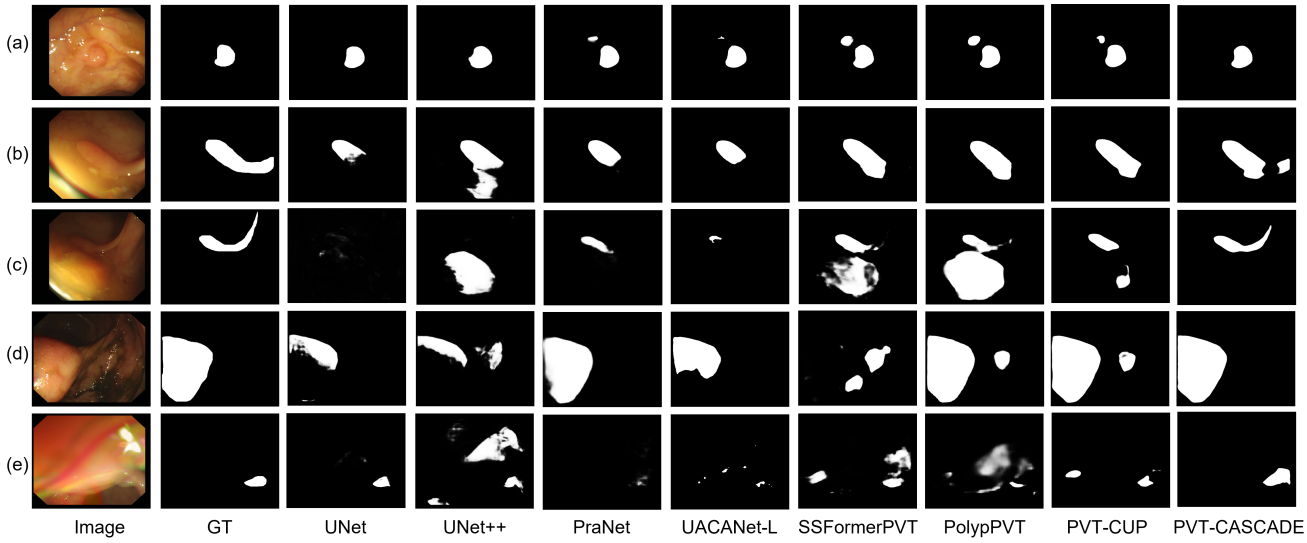

Figure 2. Qualitative results of polyp segmentation on unseen ColonDB. Five challenging images are selected from the ColonDB testset. As it can be seen, the segmentation maps generated by PVT-CASCADE (our) have strong similarity with the GroundTruth (GT).

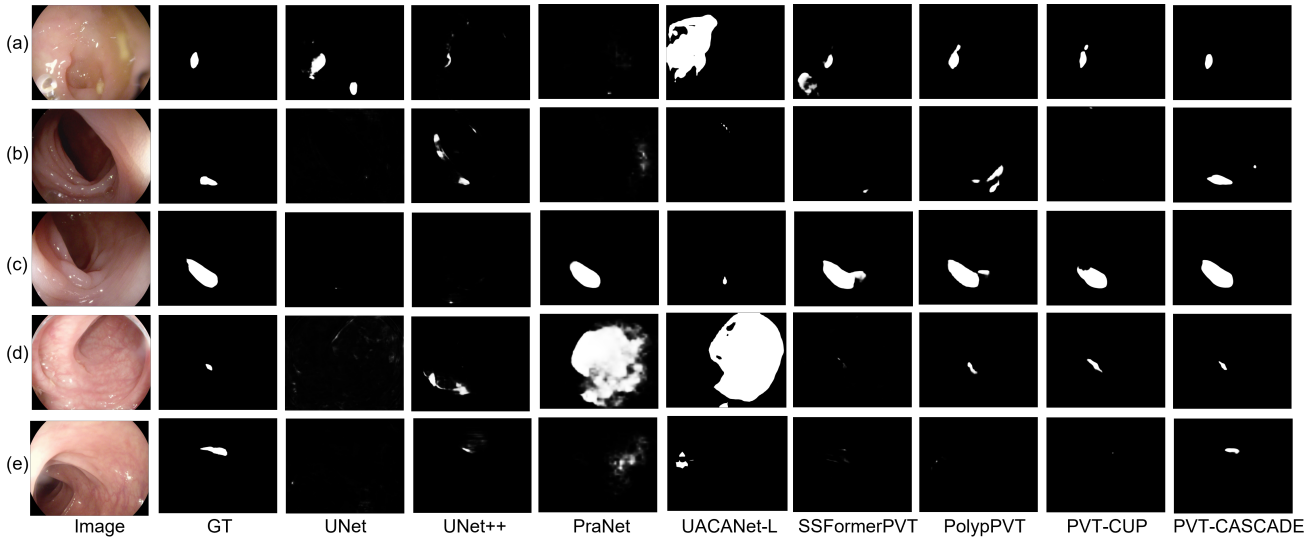

Figure 3. Qualitative results of polyp segmentation on unseen ETIS-LaribDB. Five challenging images with small lesion (polyp) regions are selected from the ETIS-LaribDB testset. As it can be seen, PVT-CASCADE (our) effectively segments the small lesions and outperforms all the state-of-the-art methods.