— Supplementary Material — Masked Supervised Learning for Semantic Segmentation

Hasib Zunair
hasibzunair@gmail.com
A. Ben Hamza
hamza@ciise.concordia.ca
Concordia University – CIISE
Montreal, QC, Canada

1 Implementation Details

Data Preprocessing. All images and masks are of size 224×224, and normalized to have values in $[-1, 1]$.

Architecture. In all our experiments for gland and polyp segmentation tasks, LeViT-UNet-384 [17] is used as the base network in the proposed MaskSup method. For indoor scene segmentation experiments, we use U-Net++ [19]. We found that using LeViTs [17], which is a transformer-based method, for indoor scene segmentation yields poor performance compared to convolutional-based methods.

Model Training. MaskSup is trained in a single step, requiring a dataset of images and segmentation masks. For gland and polyp segmentation, we use Adam optimizer with an initial learning rate of 0.0001, and we set the batch size to 16. For indoor scene segmentation, we use SGD optimizer with an initial learning rate of 0.1 with momentum 0.9, and the batch size is set to 8. Training is performed for 200 epochs and the best weights according to the mIoU score are retained.

Model Testing. After training, given an image, the model makes pixel-level predictions by assigning a pixel to a semantic class (i.e. gland, polyp, chair, wall, etc.)

Hardware and software details. Experiments are conducted on a Linux workstation running 4.8Hz and 64GB RAM and a single NVIDIA RTX 3080 GPU. All algorithms are implemented in Python programming language and PyTorch deep learning framework.

2 Description of Datasets

Datasets. We demonstrate and analyze the performance of our method on binary and multi-class image segmentation tasks. The summary descriptions of the benchmark datasets used in our experiments are as follows:

- **GLaS:** Gland segmentation from histology images is an important task for morphological analysis and is an effective tool in cancer diagnosis. This is difficult due to the diversity in size and texture of glands. The Gland Segmentation (GLaS) [14] dataset consists of a total of 165 images, where 85 are taken for training and 80 for testing following [15, 16].

- **Kvasir & CVC-ClinicDB:** Segmenting polyps from colonoscopy images is a crucial task, as it provides valuable information for diagnosis and surgery. However, this is challenging because the same types of polyps may have various sizes, colors and the boundary between polyp and background is usually ambiguous. Following ResU-Net++ [8], we use the same training and test splits of the Kvasir [9] and CVC-ClinicDB [2] datasets, which consist of 1000 and 612 images and labels, respectively.

- **NYUDv2:** Scene understanding aims to parse the scene into objects, surfaces and their relations with applications in robotics [10]. Scene segmentation is a challenging task, as objects and surfaces are close together and there also exists an imbalance in the semantic classes. The NYUDv2 dataset [13] consists of 1499 RGB-D images with 40 segmentation class labels. Following previous work [11, 12], we use 795 images and labels for training and 654 for testing. It is important to mention that unlike HybridNet A2 [11], we do not leverage the depth information.

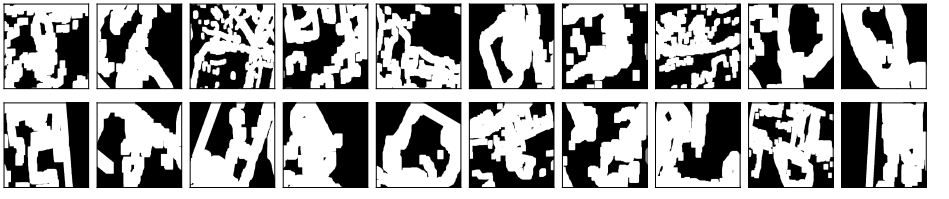

Figure 1: Visual of masks during the training process of MaskSup. Some masks cover more than 50% of the image.

3 Relation to Self-Supervised Learning via Dual Networks

Self-supervised learning (SSL) is generally used as a pre-training task, typically known as prextext task, whose goal is to learn the relationship between input and the label, which can easily be generated from the input itself. A specific class of these SSL methods [1, 4, 5, 6, 7, 18] are based on Siamese networks [3] with the goal of making the two outputs similar, demonstrating the effectiveness of self-supervised pre-training in target tasks such as image classification, segmentation and detection [1]. While this idea resembles the proposed Task Similarity Constraint, the type of output is different in the sense that all these methods output latent representations (i.e. feature vectors), whereas our MaskSup approach outputs prediction labels (i.e. segmentation map).

MaskSup differs from these SSL methods in a number of aspects. First, SSL methods require an initial stage of pre-training, typically using large volumes of unlabeled data and then a fine-tuning stage for the target task. Our approach is generic and directly applies to the target task (i.e. image segmentation). Second, these methods operate on two randomly

augmented samples of the image by applying image rotation, cropping, translation or blurring [1, 5, 13]. In contrast, our approach operates on the image and its masked version. Third, our approach differs from BYOL [7], SwAV [4] and SimSiam [6] in that these methods use two independent networks that operate on the two samples. Our method, however, is a Siamese network with shared weights, where the two networks are identical. Fourth, MaskSup employs a completely different objective function as SimCLR [5] uses a contrastive loss requiring both positive and negative pairs, Barlow Twins [13] and VICReg [1] use a cross-correlation matrix based loss, while BYOL [7], SwAV [4] and SimSiam [6] use a cosine similarity loss. Our approach can be regarded as an attempt to unify self-supervised pre-training and fine-tuning, potentially avoiding discrepancies involved in pre-training and fine-tuning. In summary, MaskSup aims to unify the concept of self-supervised pretraining [1, 5, 6, 13] on unlabeled data and fine-tuning on a target task into a single framework.

4 Additional Experimental Results

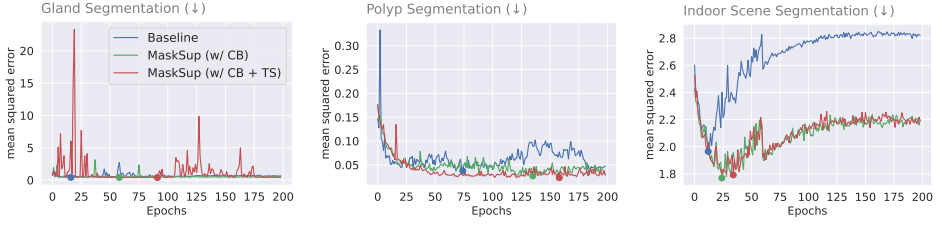

Figure 2: Comparison of mean squared error during training. MaskSup achieves the lowest error across multiple baselines across all three tasks.

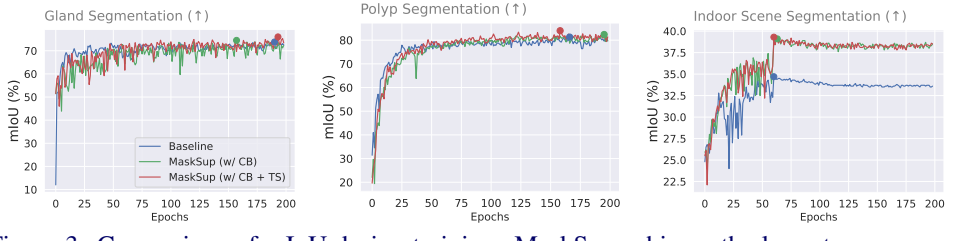

Figure 3: Comparison of mIoU during training. MaskSup achieves the lowest error across multiple baselines across all three tasks.

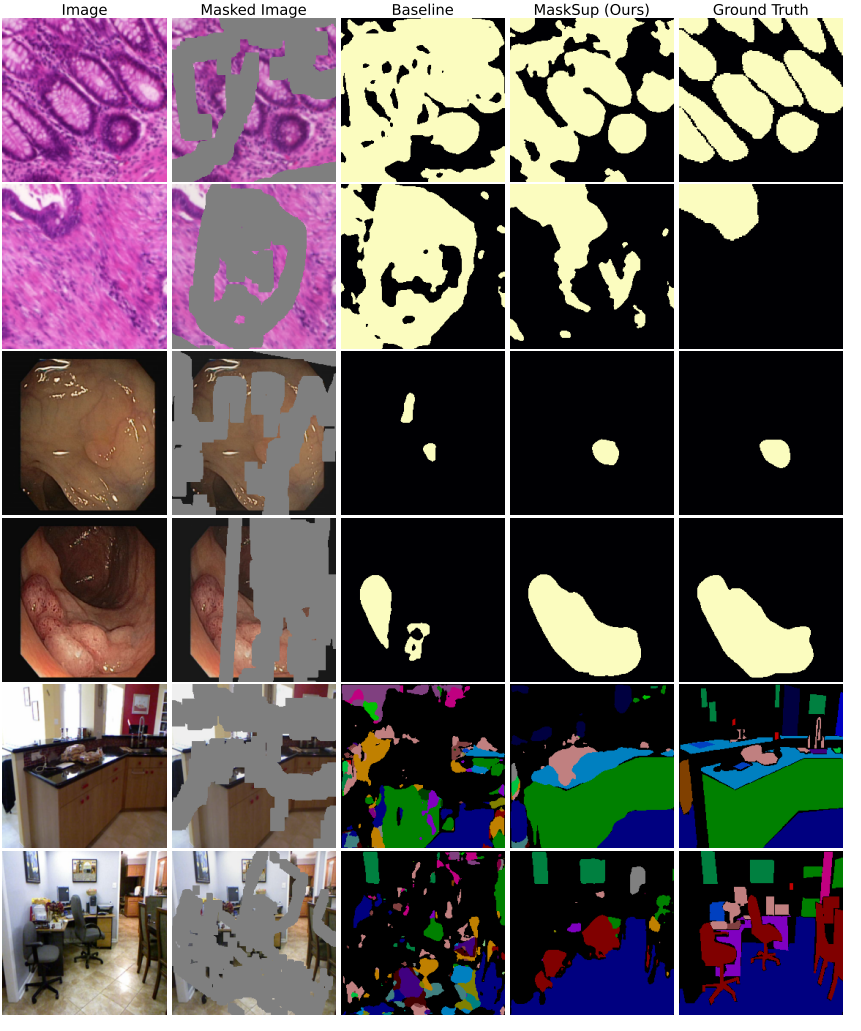

Figure 4: Visual comparison of predictions of MaskSup against baselines when provided **masked regions** as input. MaskSup is robust against masked corrptions.

Figure 5: Visual comparison of MaskSup against baselines on NYUDv2 test set. MaskSup can better segment minority class instances compared to baselines.

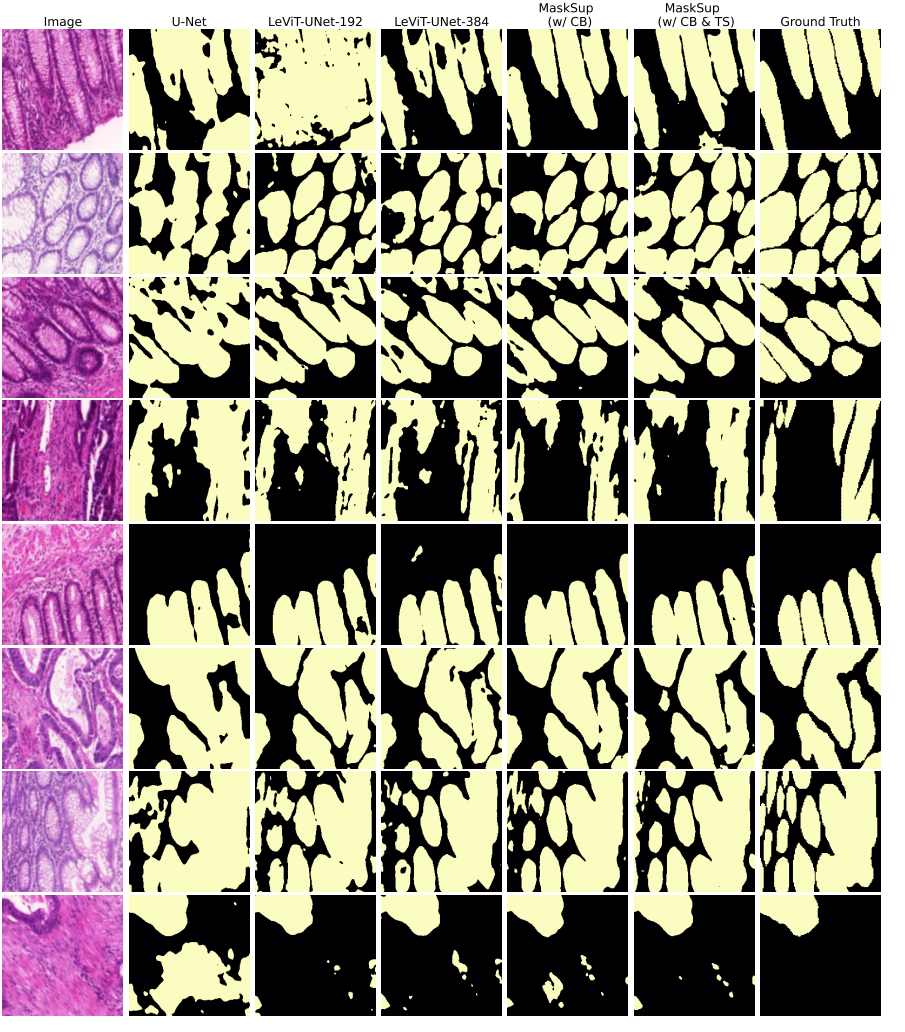

Figure 6: Visual comparison of MaskSup against baselines on GLaS test set. MaskSup can better distinguish between ROI and background and output less discontinuous predictions compared to baselines. MaskSup better tackles ambiguous ROIs and segment very small regions which baselines tend to miss.

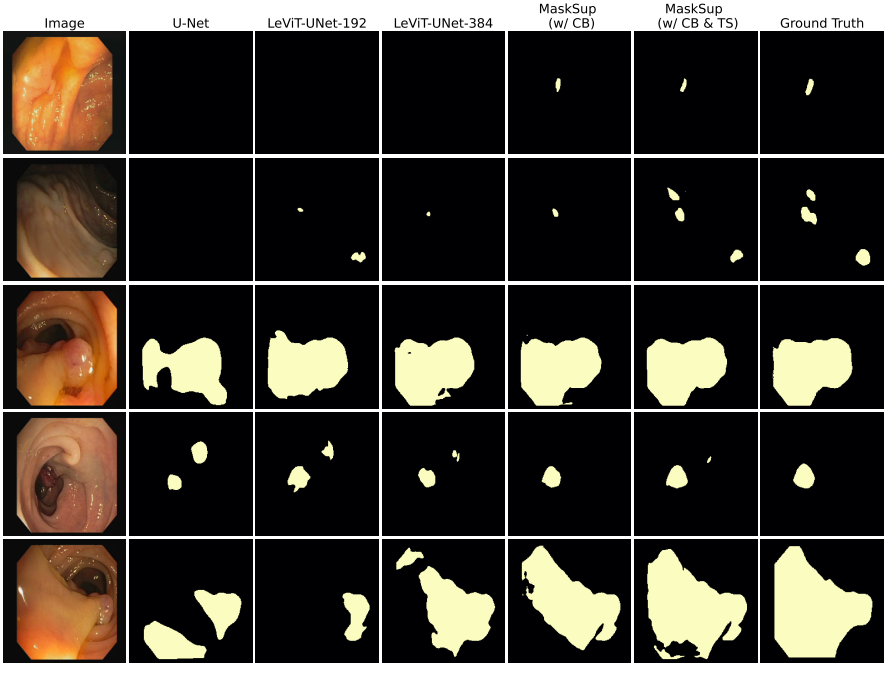

Figure 7: Visual comparison of MaskSup against baselines on Kvasir & CVC-ClinicDB test set.

References

[1] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2021.

[2] 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.

[3] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, 1993.

[4] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924, 2020.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[6] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[7] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.

[8] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. ResUNet++: An advanced architecture for medical image segmentation. In *Proc. IEEE International Symposium on Multimedia*, pages 225–2255. IEEE, 2019.

[9] 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.

[10] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *Proc. IEEE Conference on Robotics and Automation*, pages 1817–1824, 2011.

[11] Xiao Lin, Dalila Sánchez-Escobedo, Josep R Casas, and Montse Pardàs. Depth estimation and semantic segmentation from a single RGB image using a hybrid convolutional neural network. *Sensors*, 19(8):1795, 2019.

[12] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[13] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Proc. European Conference on Computer Vision*, 2012.

[14] Korsuk Sirinukunwattana, Josien PW Pluim, Hao Chen, Xiaojuan Qi, Pheng-Ann Heng, Yun Bo Guo, Li Yang Wang, Bogdan J Matuszewski, Elia Bruni, Urko Sanchez, et al. Gland segmentation in colon histology images: The GLAS challenge contest. *Medical Image Analysis*, 35:489–502, 2017.

[15] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2021.

[16] Jeya Maria Jose Valanarasu, Vishwanath A Sindagi, Ilker Hacihaliloglu, and Vishal M Patel. KiU-Net: Overcomplete convolutional architectures for biomedical image and volumetric segmentation. *IEEE Transactions on Medical Imaging*, 41(4):965–976, 2021.

[17] Guoping Xu, Xingrong Wu, Xuan Zhang, and Xinwei He. LeViT-UNet: Make faster encoders with transformer for medical image segmentation. *arXiv preprint arXiv:2107.08623*, 2021.

[18] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow Twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[19] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6):1856–1867, 2019.