# Supplementary Material for Region-of-interest Attentive Heteromodal Variational Encoder-Decoder for Segmentation with Missing Modalities

Seung-wan Jeong[1,2], Hwan-ho Cho[3], Junmo Kwon[1,2], and Hyunjin Park[1,2*]

[1] Sungkyunkwan University, Suwon, Republic of Korea
[2] Center for Neuroscience Imaging Research, Suwon, Republic of Korea
[3] Konyang University, Daejon, Republic of Korea
{jsw93,skenfn1231,hyunjinp}@skku.edu
nara9313@gmail.com

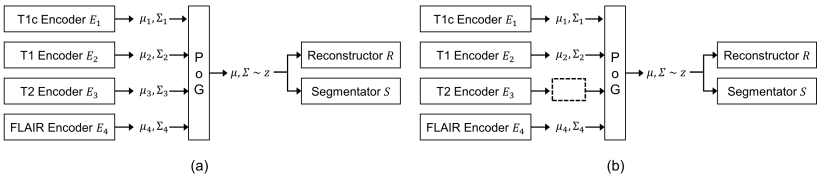


**Fig. 1.** MVAE [5] architecture on U-HVED [2]. (a) Full modalities scenario. (b) Missing modalities scenarios.

**Network Structure of RA-HVED.** The encoder consists of multiple encoders for each input modality. Each encoder consists of 9 convolutional layers. We denote Ck as a convolutional layer with k channels. M stands for max-pooling layer. The architecture of the encoder is defined as C16, M, 2 × C16, M, 2 × C32, M, 2 × C64, M, 2 × C128. All convolutional layers, except the first convolutional layer with a dropout layer, follow instance normalization and leaky ReLU (slope of 0.02). In addition, a multi-scale latent representation is created by adding DRB to all the connections between encoders and decoders. The decoders are divided into the reconstructor and the segmentor. Each decoder consists of 7 convolutional layers. The decoder has an upsampling layer between layers to restore the resolution of the original image. Let Uk denotes convolutional layer with k channels and trilinear upsampling layer. The architecture of the decoder is defined as U128, 2 × C128, U64, 2 × C64, U32, 2 × C32, C3 or C4. All convolutional layers, except the output convolutional layer, follow instance normalization and leaky ReLU with a slope of 0.02. Except for

* Corresponding author.

the size of the output channel in the final $1 \times 1 \times 1$ convolutional layer, the two networks have the same structure. The discriminator has a similar structure to patchGAN [4]. The discriminator consists of 5 convolutional layers with $4 \times 4 \times 4$ kernels. Except for the first and last layers whose stride is 1, all other layers have a stride of 2. The number of feature maps in each layer is 64, 128, 256, 512, 1. In the first four layers, except the last one, each convolutional layer is followed by instance normalization and a leaky ReLU with a slope of 0.2.

**Table 1.** Inference time on other methods. Except for Chen et al. [1], there is no significant difference in inference time for other methods.

| **Methods** | U-HeMIS | U-HVED | Chen et al. | Ours |
|---|---|---|---|---|
| **Inference time (ms)** | **113.45** | 151.52 | 752.81 | 168.31 |

**Table 2.** Comparison of segmentation performance with respect to all 15 missing modality scenarios on the BraTS 2018 dataset. The presence of modality is denoted by •, and the missing of modality is denoted by ◦. All results are evaluated with a dice score (%) and bold represents the best score in each comparison.

| **Available Modalities** | | | | **Dice score (%)** | | | |
|---|---|---|---|---|---|---|---|
| T2 | T1 | F | DWI | U-HeMIS | U-HVED | Chen et al. | Ours |
| ◦ | ◦ | ◦ | • | 38.39 | 38.51 | 36.95 | **47.65** |
| ◦ | ◦ | • | ◦ | 38.11 | 43.73 | 39.11 | **46.83** |
| ◦ | • | ◦ | ◦ | 12.76 | 27.47 | 24.35 | **27.92** |
| • | ◦ | ◦ | ◦ | 36.79 | **38.92** | 36.90 | 38.17 |
| ◦ | ◦ | • | • | 44.76 | 42.94 | 41.05 | **49.76** |
| ◦ | • | • | ◦ | 38.68 | 43.80 | 42.06 | **48.62** |
| • | • | ◦ | ◦ | 37.18 | 39.80 | 38.59 | **43.26** |
| ◦ | • | ◦ | • | 41.70 | 39.57 | 41.30 | **48.74** |
| • | ◦ | ◦ | • | 45.89 | 46.29 | 44.72 | **51.47** |
| • | ◦ | • | ◦ | 42.38 | 47.67 | 43.22 | **50.16** |
| • | • | • | ◦ | 42.96 | 44.97 | 44.51 | **46.51** |
| • | • | ◦ | • | 45.17 | 49.11 | 47.26 | **56.89** |
| • | ◦ | • | • | 45.92 | 47.91 | 45.97 | **56.24** |
| ◦ | • | • | • | 44.63 | 43.88 | 44.46 | **55.25** |
| • | • | • | • | 46.08 | 49.21 | 46.95 | **56.92** |
| Average | | | | 40.09 | 42.92 | 41.16 | **48.29** |

**Results of Segmentation with missing modalities on ISLES.** The segmentation results for the ISLES are shown in Table 2. In most scenarios of missing modality, our method shows better segmentation performance than previous methods. The DWI and FLAIR modalities provide meaningful information about stroke lesions. Compared with other methods, our method achieves high accuracy in both FLAIR and DWI. The T1 modality has less information about stroke lesions than the other modalities. In contrast to Chen et al. [1], our method achieves similar accuracy to U-HVED [2] even when only the T1 modality is available. In Fig 2, our method obtains more accurate segmentation results than other methods, even when we only have the T1 modality, which contains has relatively little information

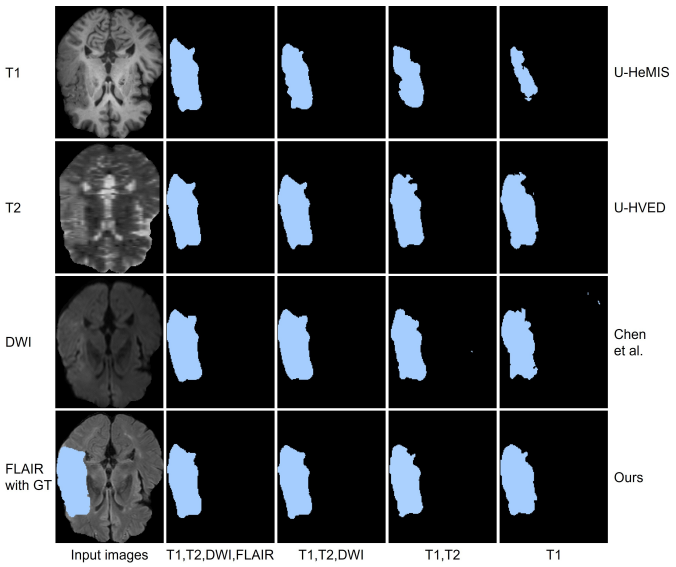


**Fig. 2.** Results of Stroke lesions segmentation produced by using different methods on the ISLES dataset. The first column is the input image of each modality, and each row shows segmentation results with missing modalities of comparison methods.

**Results of Reconstruction with missing modalities on ISLES.** Fig. 3 shows the results of FLAIR image reconstruction for ISLES when the modalities are missing. U-HeMIS [3] and U-HVED produce images that are most like the corresponding images. However, as the number of missing modalities increases, the internal information of the stroke lesions in U-HVED is lost. U-HeMIS and Chen's approach performs poorly in generating not only the inside of the stroke lesion but also the boundary as the number of missing modalities increases. Compared with other methods, our method can generate stroke lesions well

because it imposes constraints on the ROI in image reconstruction even though the number of available modalities is reduced.

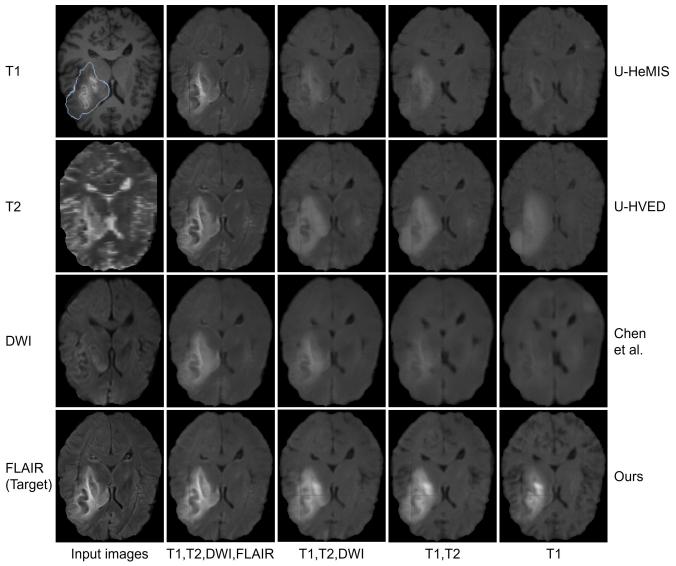


**Fig. 3.** Image reconstruction results generated by different methods on the ISLES dataset. The first column is the input image of each modality, and each row shows the reconstruction results with missing modalities of the comparison methods. Ground truth for segmentation is overlaid on T1.

## References

1. Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.A.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 447–456. Springer (2019)
2. Dorent, R., Joutard, S., Modat, M., Ourselin, S., Vercauteren, T.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 74–82. Springer (2019)
3. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: Hemis: Hetero-modal image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 469–477. Springer (2016)
4. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
5. Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. Advances in Neural Information Processing Systems **31** (2018)