Supplementary Material for
M^{3}AE: Multimodal Representation Learning for Brain Tumor Segmentation with Missing Modalities

Anonymous submission

Network Architecture Details: We use the same backbone network as Wang et al. (2021b), which is essentially a 3D U-Net comprising asymmetric encoder and decoder with skip connections (Fig. S1). The network employs residual blocks (He et al. 2016), where each block consists of two $3 \times 3 \times 3$ convolutions with group normalization (Wu and He 2018) and ReLU, followed by additive identity skip connection. The encoder progressively downsizes image/feature dimensions with strided convolutions while simultaneously increasing the feature numbers by a factor of 2, resulting in an eventual combined factor of 8. Given our input of size $4 \times 128 \times 128 \times 128$ (channel, depth, height, width) and 16 filters in the initial residual block, the size of the bottleneck output is $128 \times 16 \times 16 \times 16$, which is used for the proposed self distillation (Eqn. (2)). The decoder is similar to the encoder, but with only one residual block per spatial level. Each decoder level begins with upsizing, which reduces the feature numbers (with $1 \times 1 \times 1$ convolutions) and increases the spatial dimensions (with 3D bilinear upsampling) by a factor of 2, followed by the addition of the encoder output at the same spatial level, and lastly a residual block. In the end, the decoder outputs 16 feature maps of the original input size, or in other words, a tensor of size $16 \times 128 \times 128 \times 128$. For deep supervision, the output of the first two decoder blocks (of spatial dimensions $32 \times 32 \times 32$ and $64 \times 64 \times 64$, respectively) are used. For reproducible research, our exact implementation is available at: https://github.com/ccarliu/m3ae.

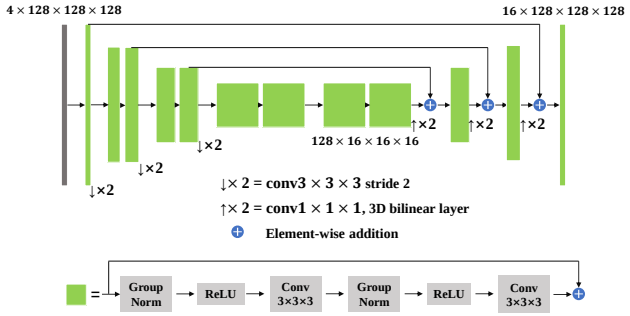
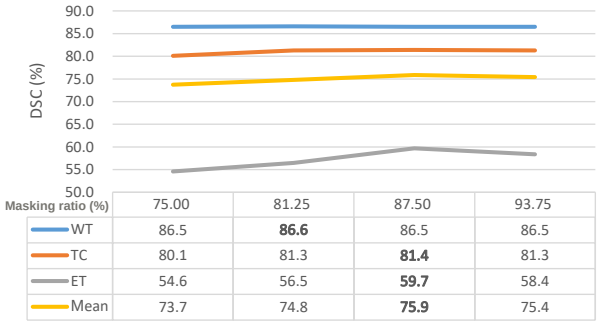


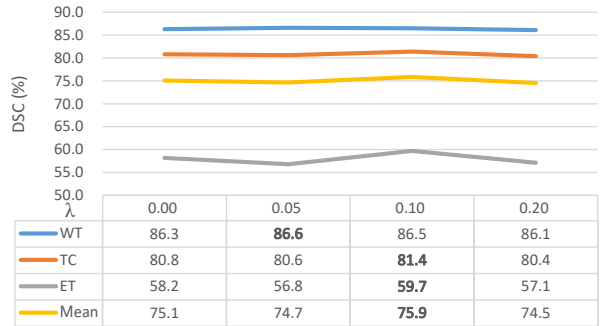
Figure S1: Backbone network used in our M^{3}AE.

Sensitivity Analysis with Respect to Hyperparameters: Figs. S2(a) and (b) show the sensitivity analysis of our M^{3}AE framework's performance with respect to two hyperparameters: the combined masking ratio during pretraining and the weight λ of the consistency loss (Eqn. (4)) during self distillation, on the validation split of BraTS 2018. As we can see, the performance is relatively stable across different masking ratios and λ values, with most variations for the enhancing tumor. The highest average performance across tissue types is achieved with the masking ratio of 0.875 and $\lambda = 0.1$, respectively.



Masking ratio (%)	75.00	81.25	87.50	93.75
WT	86.5	**86.6**	86.5	86.5
TC	80.1	81.3	**81.4**	81.3
ET	54.6	56.5	**59.7**	58.4
Mean	73.7	74.8	**75.9**	75.4

(a) Masking ratio



λ	0.00	0.05	0.10	0.20
WT	86.3	**86.6**	86.5	86.1
TC	80.8	80.6	**81.4**	80.4
ET	58.2	56.8	**59.7**	57.1
Mean	75.1	74.7	**75.9**	74.5

(b) Weight λ

Figure S2: Sensitivity analysis of performance with respect to varying values of (a) masking ratio and (b) weight λ on the testing split of BraTS2018, using mean DSCs of all modal combinations. WT: whole tumor; TC: necrotic and non-enhancing tumor core; ET: enhancing tumor.

Comparison to General-Purpose Multimodal Pretraining Methods: Poklukar et al. (2022) presented a novel Geometric Multimodal Contrastive (GMC) representation learning method where full-modal and modality-specific representations were contrasted for alignment. The method only pretrained the encoder, and only for full and single modality. In contrast, our M^{3}AE pretrains both the encoder and decoder for all possible modal combinations from single, double, etc., to full modalities. A concurrent work (Geng et al. 2022) on vision-language data (denoted by M3AE-VL) partially overlaps with ours in the concept of multimodal patch/token masked autoencoders. The concurrency demonstrates the very recent emergence of such idea and emphasizes its novelty. However, M3AE-VL missed the modality dropout in our method, and excluded the masked content from being encoded, which we utilize effectively for the model inversion based modality completion. In addition,

we propose self-distillation between heterogenous missing-modal situations.

We implement GMC (Poklukar et al. 2022) for the one full- and four single-modal situations and the mean DSCs (%) on the testing split of BraTS 2018 are 78.0/68.4/46.4 (whole/core/enhancing), inferior to our 82.8/73.9/50.8. We also implement M3AE-VL (Geng et al. 2022) for all the 15 modal combinations and the mean DSCs (%) on the testing split of BraTS 2018 are 85.1/76.5/58.6, again inferior to our 85.8/77.4/59.9.

Full-Modal Results on BraTS 2020: The full-model performance on the online validation set of BraTS 2020 is shown in Table S1. Compared to other non-challenge methods, our M^{3}AE achieves the best performance in five of six evaluated metrics, and the second best for the last metric (HD95 of the enhancing tumor; with a marginal difference of less than 1% from the best while outperforming others by at least 22%). Meanwhile, it is also mostly comparable with the challenge winner (Isensee et al. 2020), which was an ensemble of five models trained from cross-validation and involved heavy engineering.

Method	DSC (%) ↑			HD95 (mm) ↓		
	Whole	Core	Enhancing	Whole	Core	Enhancing
U-HVED†	87.8±14.1*	80.6±19.6*	71.8±31.1*	8.8±14.3*	11.6±34.1*	42.7±109.4*
ACN†	88.7±9.2	81.1±19.6*	70.7±31.0*	8.1±13.0*	11.3±35.7*	43.6±113.3*
SMU-Net†	89.8±7.6	83.3±15.1*	73.6±30.0*	5.8±10.1	7.2±11.3*	**29.2**±91.2
RFNet†	89.8±7.2	83.3±16.6*	71.9±29.8*	7.0±14.4	9.6±35.1*	37.6±105.2
ModGen†	89.1±8.1	83.7±14.5*	73.2±29.8*	7.6±14.9*	7.7±23.7*	40.0±109.5*
M^{3}AE	**90.3**±6.1	**85.4**±13.8	**79.1**±25.5	**4.6**±5.8	**5.6**±10.1	29.4±96.2
Challenge‡	91.2	85.7	79.9	3.7	5.6	26.4

Table S1: Full-modal performance comparison on the *online* validation set of BraTS 2020, including U-HVED (Dorent et al. 2019), ACN (Wang et al. 2021b), SMU-Net (Azad, Khosravi, and Merhof 2022), RFNet (Ding, Yu, and Yang 2021), and Models Genesis (ModGen; Zhou et al. 2019). Performance of the challenge winner (Isensee et al. 2020) is also included for reference. Best numbers (excluding the challenge entry) are bolded. †: reproduced based on the authors' codes; ‡: provided by the authors; *: $p < 0.05$ by Wilcoxon signed rank test for pairwise comparison with our method; Format: mean±std., if available.

Gain in Space and Time Efficiency: We compare the space and time efficiency in training and inference of our model, to those of ACN (Wang et al. 2021b) and RFNet (Ding, Yu, and Yang 2021) (the existing best performing dedicated and catch-all methods, respectively). As shown in Table S2, our method achieves the best performance with an overall best efficiency (the numbers of parameters are also visualized in Fig. 1).

Method	ACN	RFNet	M^{3}AE
DSC (%)	70.3	72.7	74.4
No. parameters (million)	71.5	8.4	4.7
Memory (GB)	18.5	18.7	9.1
Training time (h)	450	22	22
Inference time (s)	3	6	2.5

Table S2: Comparison of space and time efficiency in training and inference with the existing best performing dedicated and catch-all methods ACN (Wang et al. 2021b) and RFNet (Ding, Yu, and Yang 2021), respectively. The DSC is the mean value averaged over all modal combinations and three tumor regions on the testing split of BraTS 2018.

Detailed Missing-Modal Results on BraTS 2020: The detailed results on the testing split of BraTS 2020, including per region performance for each modal combination, are presented in Table S3.

Modality				Whole tumor					Tumor core					Enhancing tumor				
FLAIR	T1	T1c	T2	U-HVED	ACN	SMU-Net	RFNet	M^{3}AE	U-HVED	ACN	SMU-Net	RFNet	M^{3}AE	U-HVED	ACN	SMU-Net	RFNet	M^{3}AE
○	○	○	●	81.5±15.0*	84.4±10.9*	85.5±11.5	86.0±8.3	**86.1**±8.5	59.4±24.4*	**72.7**±20.1	72.4±20.8	70.6±24.0	71.8±21.4	27.7±18.9*	46.6±26.8	46.6±27.6*	46.4±27.2	**47.1**±26.8
○	○	●	○	62.5±21.1*	80.0±11.3	**79.4**±14.5	76.2±17.1*	78.9±14.2	68.7±30.5*	**85.1**±14.7	83.8±19.9	82.4±20.7	83.6±19.3	63.8±31.7*	**74.2**±29.3*	72.5±31.1*	71.2±31.8	73.6±29.7
○	●	○	○	56.3±21.6*	77.8±15.7	78.2±14.9	76.1±17.6*	**79.0**±14.4	40.4±25.2*	67.0±24.4	**71.1**±21.2*	63.5±26.5*	69.4±22.0	11.0±9.6*	41.3±25.7	**42.4**±25.1	35.2±24.9*	40.4±26.2
●	○	○	○	79.9±16.7*	87.4±11.5	86.9±11.8*	87.3±12.3	**88.0**±8.8	45.8±20.7*	71.4±19.7*	**72.3**±20.4*	68.9±23.8	68.7±20.6	16.3±11.5*	**44.8**±26.1*	44.6±26.6*	38.9±26.8	40.2±25.5
○	○	●	●	84.1±9.9*	87.2±7.5	87.0±8.1	**87.5**±7.7	87.1±11.5	76.4±22.9*	85.2±16.8	82.3±21.2*	85.0±18.5	**85.6**±16.9	66.8±31.0*	72.7±30.2	73.4±29.1	74.0±30.0	**76.0**±27.6
○	●	●	○	69.6±15.8*	79.5±14.3	80.2±14.5	**80.4**±13.8	80.1±13.7	75.4±21.3*	83.1±18.8	83.6±19.0	82.9±21.6	**83.8**±19.0	66.2±31.3*	72.1±30.6	73.6±29.8	74.2±29.5*	**75.3**±28.1
●	●	○	○	84.2±11.7*	86.0±11.7*	85.5±12.7*	**89.7**±6.7	89.6±7.8	52.1±20.6*	71.1±21.4*	69.1±23.2*	71.7±25.8	**72.8**±21.8	18.9±13.0*	44.8±26.8	42.0±27.3	42.3±29.1	**43.7**±27.1
○	●	○	●	83.1±12.5*	86.4±10.0*	86.2±10.3	**87.9**±7.2	87.3±8.6	61.5±21.5*	73.1±20.7	**73.9**±20.1	72.5±24.0	72.9±21.1	31.1±20.3*	**49.2**±26.5	48.2±27.2	47.7±28.2	48.7±27.4
●	○	○	●	87.7±9.4*	87.8±9.7*	86.6±12.8*	89.6±9.0	**90.1**±6.6	62.6±20.6*	72.3±21.4	70.8±21.4*	73.2±23.2	**74.3**±19.5	29.8±18.7*	**48.1**±27.8	46.6±26.9	47.9±28.2	47.1±26.7
●	○	●	○	84.4±13.7*	88.0±12.5*	87.8±12.0*	**89.6**±10.8	89.5±8.5	74.9±22.9*	83.5±19.4	82.3±19.9*	85.4±17.8	**85.5**±16.9	65.9±30.9*	71.5±32.0	72.1±30.7	72.4±31.3	**75.9**±27.4
●	●	●	○	86.4±10.0*	86.9±11.4*	87.7±9.6*	**90.5**±6.6*	89.6±8.0	79.1±17.9*	81.8±20.8*	83.8±18.0	**86.0**±16.1	85.6±16.8	67.8±30.1*	71.8±30.9	73.1±30.2	74.4±29.7	**76.3**±27.3
●	●	○	●	88.4±7.4*	87.0±10.1*	86.3±12.1*	**90.4**±6.4	90.2±6.8	64.3±19.6*	72.5±21.8*	68.6±22.8*	74.1±23.8	**74.4**±20.6	31.3±19.1*	44.6±26.6*	44.6±26.3*	**48.5**±29.0	48.2±27.8
●	○	●	●	88.5±9.4*	88.4±11.6*	88.2±11.7*	90.4±8.0	**90.5**±7.4	78.0±20.7*	82.8±20.7*	83.3±19.3*	85.4±17.5	**85.8**±16.8	67.2±30.9*	71.2±31.8	72.0±30.2*	72.8±31.1	**77.4**±26.3
○	●	●	●	84.9±8.4*	85.7±11.4*	86.8±10.3	**88.2**±7.4	87.4±11.3	79.0±19.1*	**86.4**±14.1	84.9±15.8*	85.3±18.4	85.8±28.2	68.1±30.3*	73.0±30.3	72.4±30.5	75.0±29.4	**78.0**±25.6
●	●	●	●	89.0±7.3*	87.9±12.0*	87.4±12.5*	**90.9**±6.0	90.4±7.8	80.0±17.8*	85.8±20.0*	83.2±22.2*	85.9±16.8	**86.2**±16.4	68.2±30.1*	73.0±30.5	71.0±31.2*	74.9±29.4	**77.5**±26.4
Mean				80.7±9.9*	85.4±3.4*	85.3±3.2*	86.7±5.0	**86.9**±4.2	66.5±12.8*	77.9±7.0	77.7±6.5	78.2±7.7*	**79.1**±7.2	46.7±22.8*	59.9±14.0*	59.7±14.3*	59.7±15.8*	**61.7**±16.3

Table S3: Performance comparison (DSC % in mean±std.) with SOTA methods, including U-HVED (Dorent et al. 2019), ACN (Wang et al. 2021b), SMU-Net (Azad, Khosravi, and Merhof 2022), and RFNet (Ding, Yu, and Yang 2021), on the testing split of BraTS 2020. Present and missing modalities are denoted by ● and ○, respectively. $*$: $p < 0.05$ by Wilcoxon signed rank test for pairwise comparison with our method.