Multi-Modal Contrastive Masked Autoencoders: A Two-Stage Progressive Pre-training Approach for RGBD Datasets

Supplementary Material

6. More Results

6.1. Additional Baselines Results

Table 11 highlights the performance of additional baselines on ScanNet semantic segmentation.

Method	Pretrain	mIoU
MultiMAE (RGB + Depth)	IN + ScanNet	65.1
Denoising only	IN + ScanNet	66.6
RGB + Depth Reconstruction only (no stage-1)	ScanNet	40.8
Ours w/o stage-1	ScanNet	42.9

Table 11. Performance of Additional Baselines on ScanNet Semantic Segmentation.

6.2. Effect of σ_{max}

The figure 4 shows the performance on ScanNet 2D semantic segmentation for different σ_{max} values.

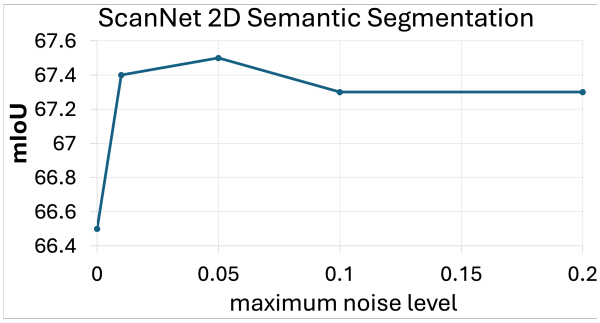

Figure 4. Effect of different values of σ_{max} pn ScanNet 2D semantic segmentation.

6.3. NYUv2 Semantic Segmentation

We show the generalizibility of the approach by fine-tuning the pre-trained model on NYUv2. More specifically, we use the model where the stage-2 pre-training is conducted using the SUN RGB-D dataset. It contains 1449 RGB-D images and we use the official split of 795 training images and 654 testing images. Table 12 shows the superiority of our approach compared to pre-training using Mask3D and other baselines for NYUv2 2D semantic segmentation.

7. Implementation Details

7.1. Pre-training and Fine-tuning Details

In the stage-1, we pre-train the encoders for 300 epochs using ImageNet [22] dataset using learning rate of 1.5e-4 with adamw optimizer. We report the stage-2 pre-training details on ScanNet [21] and SUN RGB-D [74] in Table 14. Furthermore, we report the fine-tuning details for semantic segmentation task and depth estimation in Table 15 and Table 13 respectively. Finally, we follow Mask3D [45] for instance segmentation and fine-tune the model using Detectron2 with 1x schedule.

7.2. Model Architecture

The pre-training model consists of modality-specific encoders and a shared decoder following MAE [38]. The modality specific encoders are ViT-B while the decoder consists of 8 blocks with 16 multi-head attentions. The dimension is set to 512. We add a single fully connected layer on the top of the decoder for the depth reconstruction. For noise embedding, we use 2 Fully Connected (FC) layers and 1 ReLU. For fine-tuning, we follow MultiMAE [6] for downstream task head. For 2D semantic segmentation, we use ConvNeXt [58] based decoder while for depth estimation, we use DPT [70]. Lastly, we use MaskRCNN [37] for instance segmentation.

Methods	Reconstruction task	Backbone	Pre-train	Fine-tune Modality	mIoU
MAE [38]	RGB	ViT-B	ImageNet	RGB	46.9
MAE [38]	RGB	ViT-B	ImageNet+SUN RGB-D	RGB	47.5
Mask3D [45]	Depth	ViT-B	ImageNet+SUN RGB-D	RGB	47.9
MultiMAE [6]	RGB + Depth	ViT-B	ImageNet+SUN RGB-D	RGB	47.4
Ours	Depth	ViT-B	ImageNet+SUN RGB-D	RGB	**49.2**

Table 12. We report the mean IoU of ViT-B + ConvNeXt based segmentation header on **NYUv2 Semantic Segmentation**.

Configuration	NYUv2 [62]
Optimizer	AdamW
Optimizer betas	{0.9, 0.999}
Base learning rate	1e-4
Weight decay	1e-4
Learning rate schedule	cosine decay
Warmup epochs	100
Warmup learning rate	1e-6
Epochs	2000
Batch Size	128
Layer-wise lr decay	0.75
Input resolution	256 x 256
Augmentation	RandomCrop, Color jitter

Table 13. Fine-tune setting for NYUv2 [62] **depth estimation**.

Configuration	ScanNet [21]	SUN RGB-D [74]
Optimizer	AdamW	
Optimizer betas	{0.9, 0.95}	
Base learning rate	1e-4	
Weight decay	5e-2	
Learning rate schedule	cosine decay	
Stage-2 epochs	100	250
Augmentation	Gaussian Blur, ColorJitter	
α	1.0	
β	0.01	0.1
γ	1.0	
Masking ratio	0.8	

Table 14. Stage-2 pre-training setting on ScanNet [21] and SUN RGB-D [74].

Configuration	ScanNet [21]	NYUv2 [62]	SUN RGB-D [74]
Optimizer	AdamW		
Optimizer betas	{0.9, 0.999}		
Base learning rate	1e-4		
Layer-wise lr decay	0.75		
Weight decay	5e-2		
Learning rate schedule	cosine decay		
Warmup epochs	1		
Warmup learning rate	1e-6		
Drop path	0.1		
Epochs	60	200	60
Input resolution	240 x 320	640 x 640	640 x 640
Color jitter	✗	✓	✗
RandomGaussianBlur	✓	✗	✓
RandomHorizontalFlip	✓	✗	✓

Table 15. Fine-tune setting on ScanNet [21], NYUv2 [62] and SUN RGB-D [74] for **2D semantic segmentation**.