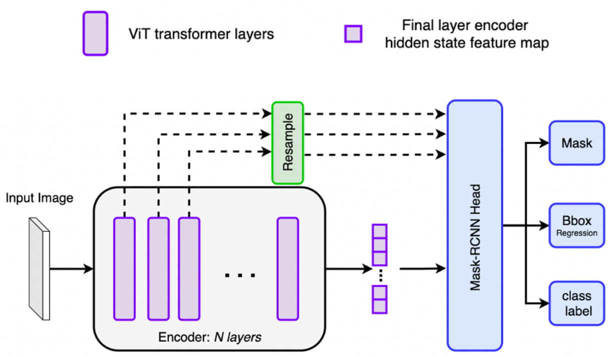
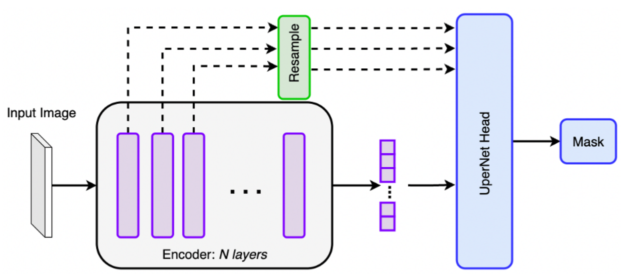
Supplementary Material for Improving Visual Representation Learning through Perceptual Understanding

Samyakh Tukra   Frederick Hoffman   Ken Chatfield
Tractable AI
{samyakh.tukra,frederick.hoffman,ken}@tractable.ai



(a) Architecture for object detection / segmentation on MS-COCO



(b) Architecture for semantic segmentation on ADE20K

Figure 1. the augmentation performed on the MAE pretrained encoder architecture for downstream fine-tuning

A. Implementation Details

A.1. Pre-training

The encoder model architecture follows closely that of the MAE paper [1]. The full configuration of the model and pre-training settings are shown in Table 1. For the ViT-B encoder, the width is set to 768 dimensions and comprises 12 layers each with 12 self-attention heads. For the ViT-L encoder, the width is 1024 dimensions and comprises 24 layers each with 16 self-attention heads. The decoder comprises in both cases 8 layers each with 16 self-attention heads and a width of 512 dimensions.

A.2. Fine-tuning

The full configuration of the model and fine-tuning settings are shown in Table 2. For MS-COCO, we use a Mask-RCNN backbone as shown in Figure 1a. This uses Feature Pyramid Networks (FPNs) [2] and we adapt the ViT encoder model accordingly. Specifically, as the encoder is composed of multiple ViT transformer layers outputting feature maps at a single scale (unlike convolutional layers), we extract feature maps at 4 layer intervals *i.e.* $[0, 4, 8, 12]$. Feature maps are resampled to the respective size required by the original Mask-RCNN head. For upsampling, bilinear interpolation is used with a scale factor of two (if required) followed by a 3x3 convolution. For downsampling, the features are reshaped to a square matrix followed by a 3x3 convolution. For segmentation on ADE20K [8], we adopt the UperNet model [5] as our decoder. This follows a similar strategy as Mask-RCNN, necessitating a FPN backbone and the same processing steps described above are applied, as illustrated in Figure 1b

B. Additional Reconstruction Results

Figure 2 shows additional randomly sampled reconstruction results from the ImageNet 1K validation set. Of note is how as additional perceptual supervision is increased, high-frequency detail such as fur in samples 1 and 3 are reconstructed more faithfully. More interestingly, in sample 2 the eyes are reconstructed properly despite being masked out entirely in the input to the decoder. This suggests that the model learns to capture higher-level semantic information better than when using plain vanilla MSE.

Table 1. Hyperparameters used for pre-training MAE and MSG-MAE on ImageNet 1K data.

Hyperparameter	ViT-B	ViT-L
Image patch size	16x16	
Hidden size	768	1024
No. of layers	12	24
Attention heads	12	16
FFN hidden size	3072	4096
Decoder hidden size	512	
Decoder No. of layers	8	
Decoder attention heads	16	
Training epochs	300	1200
Batch size	32	
Optimizer	Weighted Adam [3]	
learning rate	1.5e-4	
Weight decay	0.05	
Adam β	(0.9, 0.999)	
Learning rate schedule	Cosine	
Warmup epochs	40	
Data augmentations	RandomResizedCrop	
Input resolution	224x224	
Colour jitter	0.4	
Masking ratio	75%	

Table 2. Hyperparameters for linear probing and fine-tuning pre-trained MAE and MSG-MAE on downstream datasets.

Hyperparameter	Value
Training epochs	130
Batch size	32
Optimizer	Weighted Adam [3]
learning rate	0.001
Weight decay	0.05
Adam β	(0.9, 0.999)
Learning rate schedule	Cosine
label smoothing ϵ [4]	0.1
mixup [7]	0.8
cutmix [6]	1.0
Warmup epochs	10
Data augmentations	RandAug(9, 0.5)

References

[1] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022. 1, 3

[2] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 936–944, Los Alamitos, CA, USA, jul 2017. 1

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*. OpenReview.net, 2019. 2

[4] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 2

[5] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*. Springer, 2018. 1

[6] S. Yun, D. Han, S. Chun, S. Oh, Y. Yoo, and J. Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6022–6031, Los Alamitos, CA, USA, nov 2019. IEEE Computer Society. 2

[7] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*. OpenReview.net, 2018. 2

[8] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127(3):302–321, 2019. 1

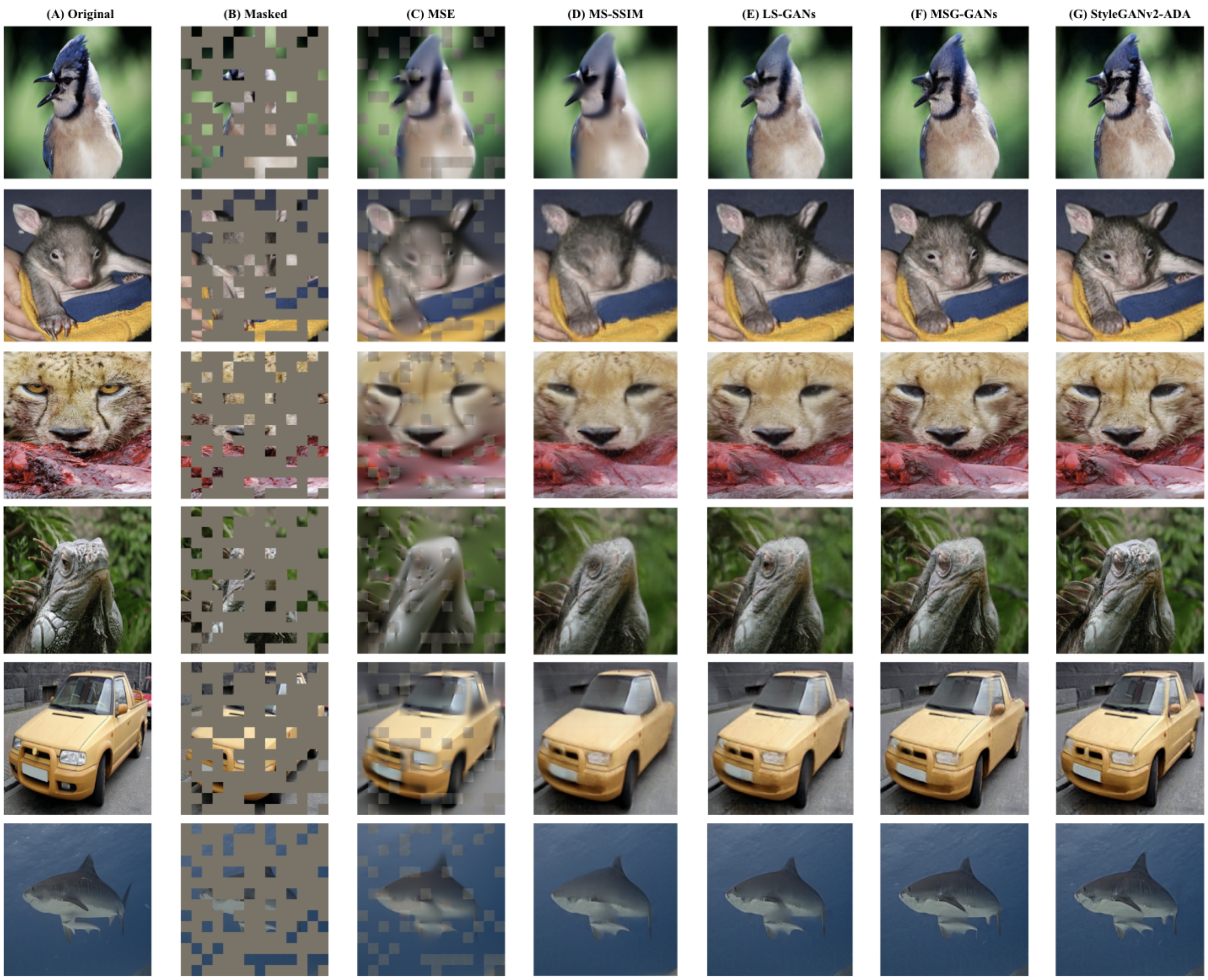


Figure 2. Randomly sampled reconstructions from the ImageNet-1K validation set. Columns are: (A) the original ground truth, (B) the masked input (MIR 75%), (C-G) are the reconstructed outputs generated MAE model trained with: MSE [1], SSIM+L1, LS-GAN-P, MSG-GAN-P, StyleGANv2-ADA-P losses respectively.