Correlational Image Modeling for Self-Supervised Visual Pre-Training —Supplementary Material

A. Appendix

In the supplementary material, we provide the detailed pre-training and fine-tuning recipes in Section A.1. Section A.2 provides more qualitative visualization for *exemplar-context* images and predicted correlation maps.

A.1. Implementation Details

Pre-training. Table 1 summarizes the pre-training settings for vanilla ViT and ResNet-50 models. All experiments are conducted on 8 A100 GPUs for both ViT and ResNet-50 models. Our `CIM` is *general* across architectures that the configurations are *shared* by different architectures, without specialized tuning.

Fine-tuning. Table 2 and Table 3 summarize the fine-tuning settings for vanilla ViT and ResNet-50 models, respectively. The configurations for ViT are *shared* across models. The configurations for ResNet-50 basically follow [16], using the AdamW optimizer following [8].

Semantic segmentation on ADE20K. Following the configurations in BEiT [1], we fine-tune UperNet [17] using AdamW as the optimizer for 160K iterations with a batch size of 16. The input resolution is 512×512, and we use single-scale inference. Following the common practice of BERT [6] fine-tuning in NLP [12], we initialize all segmentation models using model weights after supervised fine-tuning on ImageNet-1K as suggested in BEiT [1].

Table 1. **Pre-training settings for vanilla ViT-S/16, ViT-B/16 and ResNet-50 models on ImageNet-200 and ImageNet-1K.** Note that we adopt the *same* pre-training configurations across different architectures without further parameter tuning.

Configuration	Value
Optimizer	AdamW [11]
Pre-training epochs	300
Peak learning rate	2.4e-3
Batch size	4096
Weight decay	0.05
Optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$ [3]
Learning rate schedule	Cosine decay
Warmup epochs	40
Gradient clipping	1.0
Dropout [13]	✗
Stochastic depth [10]	✗
LayerScale [15]	✗
Data augmentation	RandomResizedCrop
Pos. emb. in Transformer layers	1-D absolute pos. emb. [7]
Patch size	16
Pre-training resolution of *context* image	160
Pre-training resolution of *exemplar* image	64
Number of *exemplars*	6

Table 2. **Fine-tuning settings for vanilla ViT-S/16 and ViT-B/16 on ImageNet-200 and ImageNet-1K.** We fine-tune ViT-S/16 for 200 epochs, and ViT-B/16 for 100 epochs. All other hyper-parameters are the same.

Configuration	Value
Optimizer	AdamW [11]
Fine-tuning epochs	200 (S), 100 (B)
Peak learning rate	9.6e-3
Layer-wise learning rate decay [1]	0.8 [4]
Batch size	2048
Weight decay	0.05
Optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.999$
Learning rate schedule	Cosine decay
Warmup epochs	5
Loss function	Cross-entropy loss
Gradient clipping	✗
Dropout [13]	✗
Stochastic depth [10]	0.1
Mixup [19]	0.8
Cutmix [18]	1.0
Label smoothing [14]	0.1
Random augmentation [5]	9 / 0.5
Patch size	16
Fine-tuning resolution	224
Test resolution	224

Table 3. **Fine-tuning settings for vanilla ResNet-50 on ImageNet-1K.** The hyper-parameters generally follow [16], except that we adopt the AdamW optimizer following [8].

Configuration	100 epoch FT	300 epoch FT
Optimizer	AdamW [11]	
Peak learning rate	12e-3	
Layer-wise learning rate decay [1]	✗	
Batch size	2048	
Weight decay	0.02	
Learning rate schedule	Cosine decay	
Warmup epochs	5	
Loss function	Binary cross-entropy loss	
Gradient clipping	✗	
Dropout [13]	✗	
Stochastic depth [10]	✗	
Mixup [19]	0.1	
Cutmix [18]	1.0	
Label smoothing [14]	0.1	✗
Repeated augmentation [2, 9]	✗	✓
Random augmentation [5]	6 / 0.5	7 / 0.5
Fine-tuning resolution	160	224
Test resolution	224	
Test crop ratio	0.95	

A.2. More Visualization

We provide more qualitative visualization of *exemplar-context* images together with both ground-truth and predicted correlation maps for `CIM` in Figure 1, using unseen ImageNet-1K *validation* images.

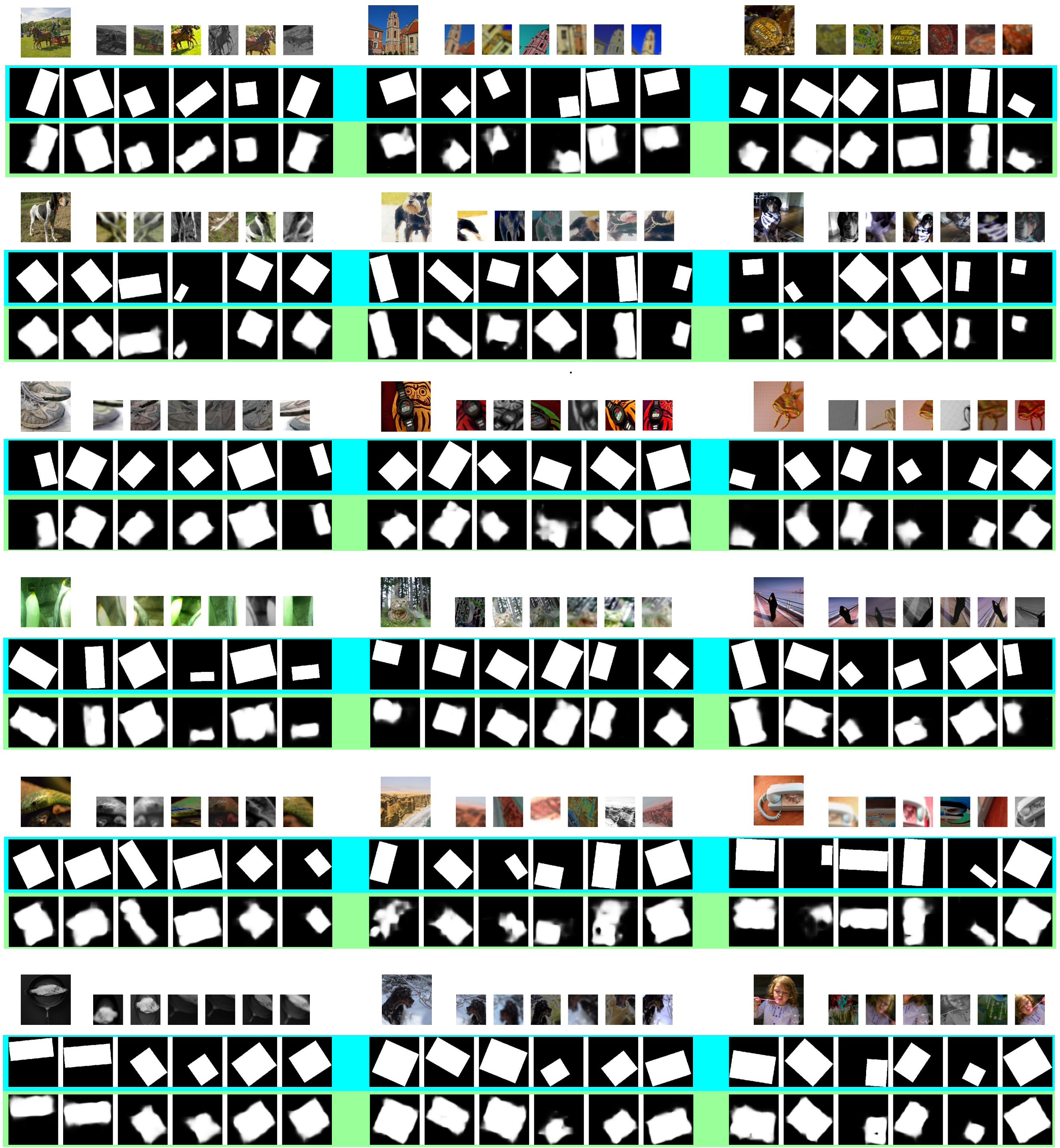

Figure 1. **Visualization** of *exemplar-context* images in company with both ground-truth and predicted correlation maps for `CIM`.

References

[1] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 1, 2

[2] Maxim Berman, Hervé Jégou, Andrea Vedaldi, Iasonas Kokkinos, and Matthijs Douze. Multigrain: a unified image embedding for classes and instances. *arXiv preprint arXiv:1902.05509*, 2019. 2

[3] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *ICML*, 2020. 1

[4] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020. 2

[5] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPRW*, 2020. 2

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 1

[8] Yuxin Fang, Li Dong, Hangbo Bao, Xinggang Wang, and Furu Wei. Corrupted image modeling for self-supervised visual pre-training. *arXiv preprint arXiv:2202.03382*, 2022. 1, 2

[9] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: better training with larger batches. *arXiv preprint arXiv:1901.09335*, 2019. 2

[10] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 1, 2

[11] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1, 2

[12] Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R Bowman. Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work? In *ACL*, 2020. 1

[13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014. 1, 2

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

[15] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *ICCV*, 2021. 1

[16] Ross Wightman, Hugo Touvron, and Hervé Jégou. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021. 1, 2

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

[18] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 2

[19] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2