Appendix of

Unleashing Vanilla Vision Transformer with Masked Image Modeling for Object Detection

Yuxin Fang[1*] Shusheng Yang[1*] Shijie Wang[1*] Yixiao Ge[2,3] Ying Shan[2,3] Xinggang Wang[1†]

[1]School of EIC, Huazhong University of Science & Technology
[2]Tencent AI Lab [3]ARC Lab, Tencent PCG

method	COCO `val`	COCO `test-dev`
MIMDET-B	51.7 AP^{box}/ 46.1 AP^{mask}	51.8 AP^{box}/ 46.3 AP^{mask}
MIMDET-L	54.3 AP^{box}/ 48.2 AP^{mask}	54.5 AP^{box}/ 48.7 AP^{mask}

Table 1: **COCO object detection and instance segmentation results using Mask R-CNN on COCO `val` & `test-dev` set respectively.** Their results are consistent.

backbone	params (M)	FLOPs (T)	ft epochs	AP^{box}	AP^{mask}
Li et al.-B	111	0.8	100	50.3	44.9
MIMDET-B	128	0.9	36	**51.7**	**46.1**
Li et al.-L	331	1.9	100	53.3	47.2
MIMDET-L	349	2.1	36	**54.3**	**48.2**

Table 2: **Params, FLOPs & ft epochs** comparisons with Li et al. [5] using Mask R-CNN.

A. Appendix

Architecture of ConvStem. We adopt a minimalist ConvStem design, *i.e.*, by simply stacking 3×3 regular convolutions with a stride of 2 and doubled feature dimensions. Each convolutional layer is followed by a layer normalization [1] and a GELU activation [4]. The detailed configurations are given in Architecture 1.

Hyper-parameters and Model Configurations. Hyper-parameters and model configurations for fine-tuning on the COCO dataset are shown in Table 3. Since the vanilla ViT encoder is already pre-trained while the task layer is trained from scratch, the learning rate of the ViT encoder part is divided by a "lr multiplier" and the learning rate for the task layer is multiplied by a "lr multiplier".

Compared with Li et al., MIMDET achieves better results with a much shorter training schedule & similar complexity, as shown in Table 2.

Optimization. The loss function of MIMDET keeps the *same* as the canonical Mask R-CNN [3, 5], *i.e.*, explicit reconstruction loss for ViT encoder is *not* needed during the fine-tuning, even though the encoder only receive partial observations. The implicit reconstruction process of ViT encoder is driven by the supervision from the Mask R-CNN detector.

Architecture 1 -ConvStem for ViT-Base (PyTorch Style), which can help preserve low-level details, produce higher resolution hierarchical features for FPN, and introduce 2D inductive biases for the ViT encoder & detector.

```
# Number of Parameters: 4.1M.
ConvStem(
  ModuleList(
    (0): Sequential(
      (0): Conv2d(3, 96, kernel_size=(3, 3), stride=(2,
           2), padding=(1, 1), bias=False)
      (1): LayerNorm2d(96, eps=1e-06, affine=True) &
           GELU()
    )
    (1): Sequential(
      (0): Conv2d(96, 192, kernel_size=(3, 3), stride
           =(2, 2), padding=(1, 1), bias=False)
      (1): LayerNorm2d(192, eps=1e-06, affine=True) &
           GELU() # Input for FPN P2.
    )
    (2): Sequential(
      (0): Conv2d(192, 384, kernel_size=(3, 3), stride
           =(2, 2), padding=(1, 1), bias=False)
      (1): LayerNorm2d(384, eps=1e-06, affine=True) &
           GELU() # Input for FPN P3.
    )
    (3): Sequential(
      (0): Conv2d(384, 768, kernel_size=(3, 3), stride
           =(2, 2), padding=(1, 1), bias=False)
      (1): LayerNorm2d(768, eps=1e-06, affine=True) &
           GELU()
      (2): Conv2d(768, 768, kernel_size=(1, 1), stride
           =(1, 1)) # Input for ViT-Base Enc.
    )))
```

Results on COCO `test-dev` set and comparisons with COCO `val` set results are shown in Table 1, which imply that our models & settings are not biased towards `val` set.

Feature Visualizations Figure 1 and 2 visualizes some backbone & FPN feature maps with a stride of 4 for both [5] and our MIMDET. The stride-4 backbone feature of [5] is obtained from a stride-16 ViT encoder feature via upsampling

*Equal contribution. † Xinggang Wang (`xgwang@hust.edu.cn`) is the corresponding author. This work was done when Shusheng Yang was interning at ARC Lab, Tencent PCG.

backbone	hyper-parameters					model configs		
	lr	lr multiplier	weight decay	drop path	ft epochs	params (M)	FLOPs (G)	inf. time (s)
MIMDET-Base	$8e^{-5}$	2	0.1	0.1	36	128	933	0.29
MIMDET-Large	$8e^{-5}$	3.5	0.1	0.1	36	349	2082	0.58

Table 3: **Hyper-parameters and model configurations for COCO fine-tuning with Mask R-CNN.** We report the average number of FLOPs and inference time for the first 100 images in the COCO `val` set following [2] on a V100 GPU. Hyper-parameters for Cascade Mask R-CNN and RetinaNet are same as Mask R-CNN.

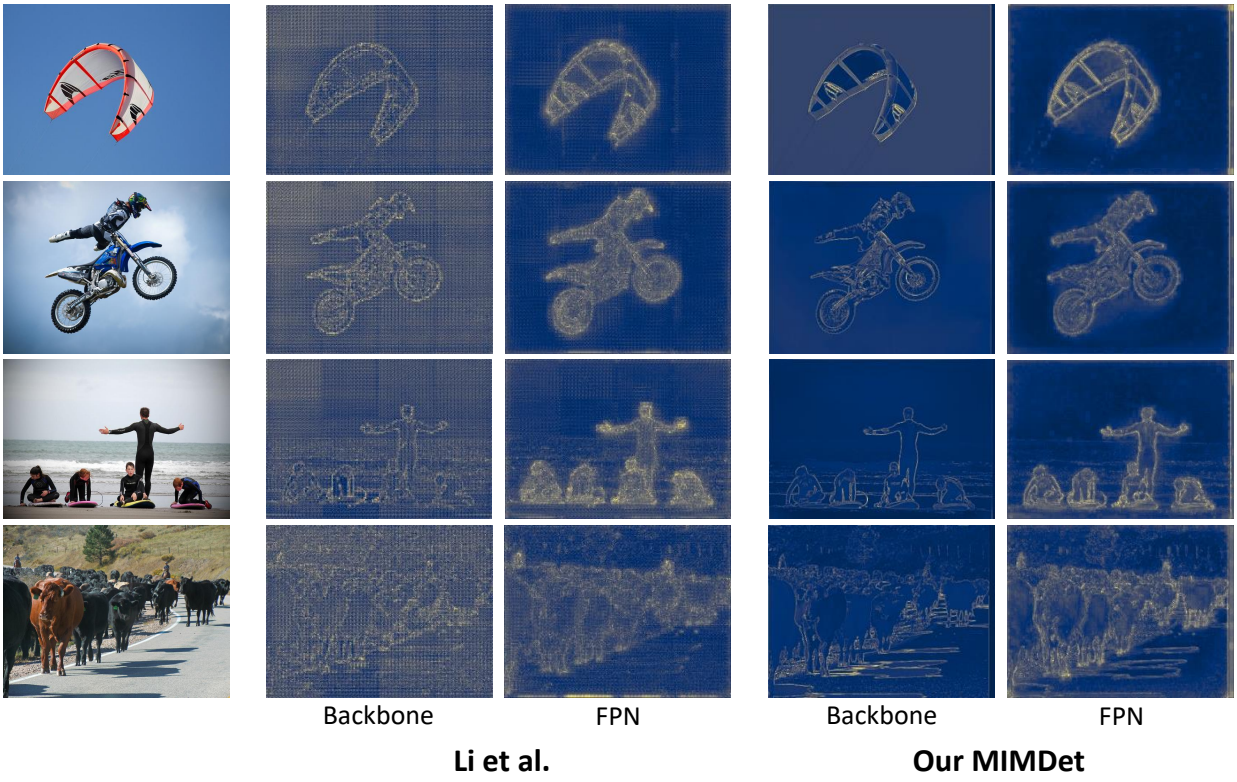

Figure 1: **Feature visualizations and comparisons of some stride-4 backbone and FPN feature maps.** The feature maps of [5] is obtained from our re-implementation which successfully reproduces its reported results.

using two stride-2 transposed convolutions with 2×2 kernel. The resulting features suffer from very strong "checkerboard artifacts [6]". If we look closer, the evidence of ViT attention's window partition emerges. Thanks to FPN, the noise can be mitigated to some extent. However, many low-level details are still fuzzy. On the other hand, our ConvStem in MIMDET can always produce clear and tidy features, which is beneficial to both the ViT encoder as well as the Mask R-CNN detector.

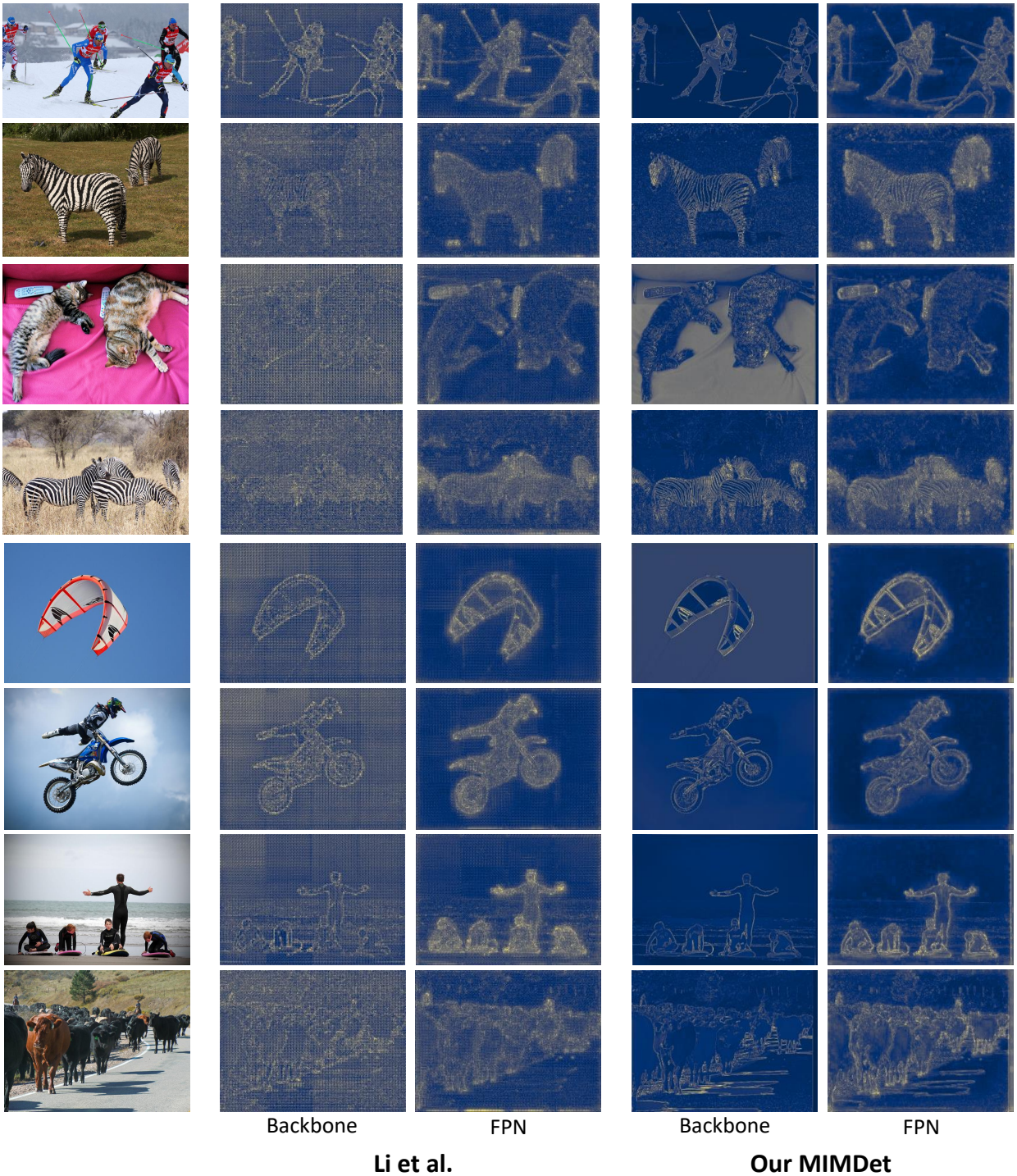

Figure 2: **Feature visualizations and comparisons of some stride-4 backbone and FPN feature maps.** The feature maps of [5] is obtained from our re-implementation which successfully reproduces its reported results.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[3] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. In *ICCV*, 2017. 1

[4] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 1

[5] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 1, 2, 3

[6] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. 2