SimPLR: A Simple and Plain Transformer for Object Detection and Segmentation —Supplementary Material—

Anonymous CVPR submission

Paper ID 7071

This supplementary material provides additional implementation details, further information for better reproducibility, additional quantitative and qualitative results as well as license information.

Contents

A. Implementation Details

Masked Instance-Attention. The masked instance-attention follows the grid sampling strategy of the box-attention in [9], but differs in the computation of attention scores to better capture objects of different shapes. To be specific, the region of interest r'_i is divided into 4 bins of 2×2 grid, each of which contains a $\frac{m}{2} \times \frac{m}{2}$ grid features sampled using bilinear interpolation. Instead of assigning an attention weight to each feature vector, a linear projection ($\mathbb{R}^d \rightarrow \mathbb{R}^{2\times2}$) is adopted to generate the 2×2 attention scores for 4 bins. The $\frac{m}{2} \times \frac{m}{2}$ feature vectors within the same bin share the same attention weight. This is equivalent to the *average* aggregation of feature values covered by each bin, which shows to reduce misalignments in RoIAlign [3]:

$$\text{head}_i = \sum_{k=0}^{2\times2} \sum_{j=0}^{\frac{m}{2}\times\frac{m}{2}} \frac{\alpha_k}{\frac{m}{2} \cdot \frac{m}{2}} v_{i_{k,j}}, \tag{1}$$

where a_k is the attention weight corresponding to k-th bin and $v_{i_{k,j}}$ is the j-th feature vector inside k-th bin.

Inspired by [1], we utilize the mask prediction of the previous decoder layer $\mathcal{M}_q \in \mathbb{R}^{H_m \times W_m}$ corresponding to the object query q. Given the coordinates of grid features within the region of interest r'_i, we sample the corresponding mask scores using bilinear interpolation. The sampled mask scores are binarized with the 0.5 threshold before $\mathrm{softmax}$ in the attention computation. Note that in masked instance-attention, we sample the feature grid of 14×14.

Fig. 1 shows the difference between box-attention [9] and masked instance-attention. By utilizing the mask prediction from previous decoder layer, masked instance-attention can effectively capture object of different shapes.

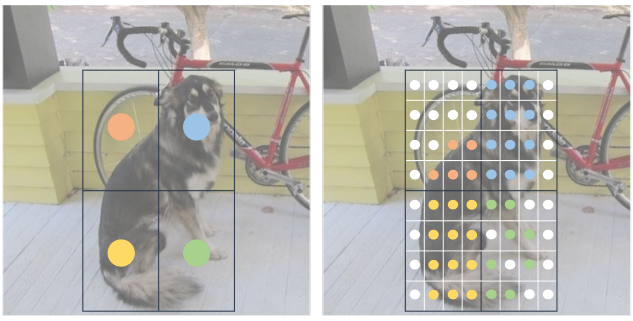

Figure 1. **Masked Instance-Attention. Left:** The box-attention [9] which samples 2×2 grid features in the region of interest. **Right:** Our masked instance-attention for dense grid sampling that employs masking strategy to capture object boundary. The 2×2 attention scores are denoted in four colours and the masked attention score is shown in white.

The creation of input features. In Fig. 2, we compare the creation of input features to detection head between SimpleFPN and our method. In [5], the multi-scale feature maps are created by different sets of convolution layers. Instead, SimPLR simply applies a deconvolution layer following by a GroupNorm layer [13].

Losses in training of SimPLR. We use focal loss [7] and dice loss [8] for the mask loss: $\mathcal{L}_{\text{mask}} = \lambda_{\text{focal}}\mathcal{L}_{\text{focal}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}$ with $\lambda_{\text{focal}} = \lambda_{\text{dice}} = 5.0$. The box loss is the combination of ℓ_1 loss and GIoU loss [10], $\mathcal{L}_{\text{box}} = \lambda_{\ell_1}\mathcal{L}_{\ell_1} + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}}$, with $\lambda_{\ell_1} = 5.0$ and $\lambda_{\text{giou}} = 2.0$. The focal loss is also used for our classification loss, $\mathcal{L}_{\text{cls}}$. Our final loss is formulated as: $\mathcal{L} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{box}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}$

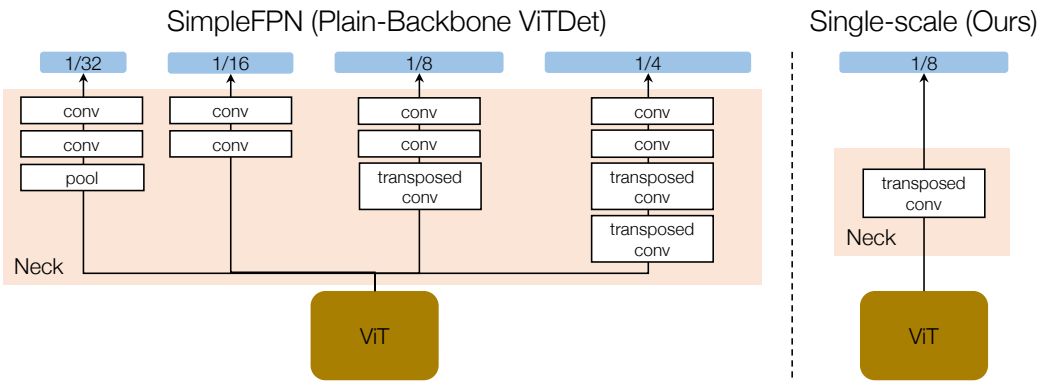

Figure 2. **The creation of input features. Left:** The creation of feature pyramids from the last feature of the plain backbone, ViT, in SimpleFPN [5] where different stacks of convolutional layers are used to create features at different scales. **Right:** The design of our single-scale feature map with only one layer.

($\lambda_{\text{cls}} = 2.0$ for object detection and instance segmentation, $\lambda_{\text{cls}} = 4.0$ for panoptic segmentation).

Hyper-parameters of SimPLR. SimPLR contains 6 encoder and decoder layers. The adaptive-scale attention in SimPLR encoder samples 2×2 grid features per region of interest. In the decoder, we compute attention on a grid of 14×14 features within regions of interest. The dimension ratio of feed-forward sub-layers to 4. The number of object queries is 300 in the decoder as suggested in [9]. The size of input image is 1024×1024 in both training and inference. Note that we also use this setting for the baseline (*i.e.*, BoxeR with ViT backbone).

In Tab. 2d, we show that the *decouple* between feature scale and dimension of the ViT backbone and the detection head helps to boost the performance of our plain detector while keeping the efficiency. This comes from the fact that the complexity of global self-attention in the ViT backbone increase quadraticaly w.r.t. the feature scale and the detection head enjoys the high-resolution input for object prediction. Note that with ViT-H as the backbone, we follow [5] to interpolate the kernel of patch projection into 16×16. The hyper-parameters for each SimPLR size (Base, Large, and Huge) are in Tab. 1.

model size		Base	Large	Huge
backbone	dim	768	1024	1280
	# head	12	16	16
	feat. scale	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
detection head	enc. dim	384	768	960
	dec. dim	256	384	384
	# head	12	16	16
	feat. scale	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

Table 1. Hyper-parameters of backbone and detection head for different sizes of SimPLR (base – large – huge models). Note that these settings are the same for all three tasks.

B. Additional Results

method	backbone	pre-train	Panoptic Segmentation PQ	PQ$^{\text{th}}$	PQ$^{\text{st}}$	FPS
MaskFormer	Swin-B	sup-1K	51.1	56.3	43.2	-
Mask2Former	Swin-B	sup-1K	55.1	61.0	46.1	-
SimPLR	ViT-B	sup-1K	**55.2**	**61.2**	**46.2**	**13**

Table 2. **More panoptic segmentation comparison** between SimPLR with ViT-B backbone and other methods with Swin-B backbone. All backbones are pre-trained on ImageNet-1K with supervised pre-training. SimPLR still shows competitive results when using only single-scale input.

More panoptic segmentation comparison. Here, we provide more results of SimPLR with ViT-B backbone and other methods with Swin-B backbone using supervised pre-training on COCO panoptic segmentation in Tab. 2. SimPLR continues to show strong segmentation performance when using only single-scale input.

Ablation on pre-training strategies. Tab. 3 compares the ViT backbone when pre-trained using different strategies with different sizes of pre-training data. SimPLR with the ViT backbone benefits from better pre-training methods even with supervised approaches. Among supervised pre-training methods, DEiTv3 [12] shows better results than DEiT [11], and the pre-training on ImageNet-21K further improves the performance of DEiTv3.

However, self-supervised methods like MAE [4] provides strong pre-trained backbones when only pre-trained on ImageNet-1K. This further confirms that our plain detector, SimPLR, enjoys the significant progress of self-supervised learning and scaling ViTs. A similar observation is also pointed out in ViTDet [5] where the plain ViT backbone initialized with MAE shows better improvement over hierarchical backbones.

pre-train	Object Detection				Instance Segmentation			
	AP^b	AP^b_S	AP^b_M	AP^b_L	AP^m	AP^m_S	AP^m_M	AP^m_L
IN-1K, DEiT	53.6	33.7	58.1	71.5	46.1	24.5	50.4	67.2
IN-1K, DEiTv3	54.0	34.3	58.8	70.5	46.4	24.8	51.1	66.7
IN-21K, DEiTv3	54.8	35.4	59.0	**72.4**	47.1	25.8	51.2	**68.5**
IN-1K, MAE	**55.4**	**36.1**	**59.1**	70.9	**47.6**	**26.8**	**51.4**	67.1

Table 3. **Ablation on pre-training strategies** of the plain ViT backbone using SimPLR evaluated on COCO object detection and instance segmentation. We compare the ViT backbone pre-trained using supervised methods (*top* row) *vs*. self-supervised methods (*bottom* row) with different sizes of pre-training dataset (ImageNet-1K *vs*. ImageNet-21K). Here, we use the 5× schedule as in [9]. It can be seen that SimPLR with the plain ViT backbone benefits from better pre-training approaches and with more pre-training data.

Figure 3. **Qualitative results** for object detection, instance segmentation, and panoptic segmentation generated by SimPLR using ViT-B as backbone on the COCO `val` set. In each pair, the left image shows the visualization of object detection and instance segmentation, while the right one indicates the panoptic segmentation prediction.

C. Qualitative results

We provide qualitative results of the SimPLR prediction with ViT-B backbone on three tasks: COCO object detection, instance segmentation, and panoptic segmentation in Fig. 3.

D. Asset Licenses

Dataset	License
ImageNet [2]	https://image-net.org/download.php
COCO [6]	Creative Commons Attribution 4.0 License

References

[1] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Mask2former: Masked-

attention mask transformer for universal image segmentation. In *CVPR*, 2022. 1

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 3

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

[4] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2

[5] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, 2022. 1, 2

[6] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014. 3

[7] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1

[8] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 1

[9] Duy-Kien Nguyen, Jihong Ju, Olaf Booij, Martin R. Oswald, and Cees G. M. Snoek. Boxer: Box-attention for 2d and 3d transformers. In *CVPR*, 2022. 1, 2, 3

[10] Seyed Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and A loss for bounding box regression. In *CVPR*, 2019. 1

[11] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers and distillation through attention. In *International Conference on Machine Learning*, 2021. 2

[12] Hugo Touvron, Matthieu Cord, and Herve Jegou. Deit iii: Revenge of the vit. *arXiv preprint arXiv:2204.07118*, 2022. 2

[13] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 1