Vision Transformers Are Good Mask Auto-Labelers
Supplementary Material

Shiyi Lan[1] Xitong Yang[2] Zhiding Yu[1] Zuxuan Wu[3] Jose M. Alvarez[1] Anima Anandkumar[1,4]
[1]NVIDIA [2]Meta AI, FAIR [3]Fudan University [4]Caltech
https://github.com/NVlabs/mask-auto-labeler

1. Additional details of CRF

In the main paper, we define the energy terms of CRF but skip the details on how we use the Mean Field algorithm to minimize the energy. Here, we provide more details on how we use the Mean Field algorithm [1].

We define $\boldsymbol{l} = \{\boldsymbol{l}_1, ..., \boldsymbol{l}_N\}$ as the label being inferred, where $N = H \times W$ is the size of the input image and $\boldsymbol{x}_i$ is the label of the i-th pixel in $\boldsymbol{I}$. We also assume that the network predicts a mask $\boldsymbol{m} = \{\boldsymbol{m}_1, ..., \boldsymbol{m}_N\}$ is where $\boldsymbol{m}_i$ is the unary mask score of the i-th pixel in $\boldsymbol{I}$. The pseudo-code to obtain $\boldsymbol{l}$ using mean field is attached in Alg. 1:

Algorithm 1 Mean field algorithm for CRFs.

procedure MEANFIELD($\boldsymbol{m}, \boldsymbol{I}$)
  $\boldsymbol{K}_{i,j} \leftarrow \omega \exp(-\frac{|\boldsymbol{I}_i - \boldsymbol{I}_j|}{2\zeta^2})$
  ▷ Initialize the Gaussian kernels
  $\boldsymbol{l} \leftarrow \boldsymbol{m}$ ▷ Initialize $\boldsymbol{l}$ using $\boldsymbol{m}$
  while not converge **do** ▷ Iterate until convergence
    for $i \leftarrow 1$ **to** $|\boldsymbol{l}|$ **do**
      $\hat{\boldsymbol{l}}_i \leftarrow \boldsymbol{l}_i$
      for $j \in \mathcal{N}(i)$ **do**
        $\hat{\boldsymbol{l}}_i \leftarrow \hat{\boldsymbol{l}}_i + \boldsymbol{K}_j * \boldsymbol{l}_j$
        ▷ Message passing
      end for
    end for
    $\boldsymbol{l} \leftarrow \varphi(\hat{\boldsymbol{l}})$ ▷ φ is a clamp function
  end while
  return $\lambda(\boldsymbol{l})$ ▷ λ is a threshold function
end procedure

2. Additional implementation details

We use the same hyper-parameters on all benchmarks for all image encoders (Standard ViTs [5–7], Swin Transformers [8], and ConvNeXts [4]) and mask decoders (fully connected decoder, fully convolutional decoder, attention-based decoder,), including batch size, optimization hyper-parameters. We observe a performance drop when we add parametric layers or multi-scale lateral/skip connections [3, 9] between the image encoder (Standard ViTs, Swin Transformers, ConvNeXts) and the mask decoder (attention-based decoder). We insert a couple of the bilinear interpolation layers to resize the feature map between the image encoder and the mask decoder and resize the segmentation score map. Specifically, we resize the feature map produced by the image encoder to 1/16 (small), 1/8 (medium), 1/4 (large) size of the raw input according to the size of the objects. We divide the objects into three scales regarding to the area of their bound boxes. We use the area ranges of [0, 32^2), [32^2, 96^2), [96^2, ∞) to cover small, medium, and large objects, respectively. We resize the mask prediction map to 512 × 512 to reach the original resolution of the input images.

Moreover, we also try three naive ways to add classification loss, but it does not work well with MAL. First, we add another fully connected layer as the classification decoder, which takes the feature map of the first fully connected layer of the instance-aware head K. With this design, the classification causes a significant performance drop. Secondly, we use two extra fully connected layers or the original classification decoder of standard ViTs as the classification decoder, which directly takes the feature map of the image encoder. However, the classification loss does not provide performance improvement or loss in this scenario.

3. Benefits for detection

The supervised object detection models benefit from the extra mask supervision [10], which improves detection results. Specifically, we follow the settings in Mask R-CNN [10]. First, we use RoI Align, the box branch, and the box supervision without mask supervision. Second, we add the mask branch and ground-truth mask supervision on top of the first baseline. The second baseline is the original Mask R-CNN. Thirdly, we replace the ground-truth masks with the mask pseudo-labels generated by MAL on top of the second baseline. It turns out that using MAL-generated mask pseudo-labels for mask supervision brings in an im-

Figure 1. The qualitative comparison between Mask2Former trained with GT mask and Mask2Former trained with MAL-generated mask pseudo-labels. Note that we use ViT-MAE-Base as the image encoder of MAL and Swin-Small as the backbone of the Mask2Former.

InstSeg Backbone	Dataset	Mask Labels	(%)AP	(%)AP_{50}	(%)AP_{75}	(%)AP_S	(%)AP_M	(%)AP_L
ResNet-50-DCN [2]	LVIS v1	None	22.0	36.4	22.9	16.8	29.1	33.4
ResNet-50-DCN [2]	LVIS v1	GT mask	22.5	36.9	23.8	16.8	29.7	35.0
ResNet-50-DCN [2]	LVIS v1	MAL mask	22.6	37.2	23.8	17.3	29.8	34.6
ResNet-101-DCN [2]	LVIS v1	None	24.4	39.5	26.1	17.9	32.2	36.7
ResNet-101-DCN [2]	LVIS v1	GT mask	24.6	39.7	26.1	18.3	32.1	38.3
ResNet-101-DCN [2]	LVIS v1	MAL mask	25.1	40.0	26.7	18.4	32.5	37.8
ResNeXt-101-32x4d-FPN [2, 3]	LVIS v1	None	25.5	41.0	27.1	18.8	33.7	38.0
ResNeXt-101-32x4d-FPN [2, 3]	LVIS v1	GT mask	26.7	42.1	28.6	19.7	34.7	39.4
ResNeXt-101-32x4d-FPN [2, 3]	LVIS v1	MAL mask	26.3	41.5	28.3	19.5	34.5	39.6
ResNeXt-101-64x4d-FPN [2, 3]	LVIS v1	None	26.6	42.0	28.3	19.8	34.7	39.9
ResNeXt-101-64x4d-FPN [2, 3]	LVIS v1	GT mask	27.2	42.8	29.2	20.2	35.7	41.0
ResNeXt-101-64x4d-FPN [2, 3]	LVIS v1	MAL mask	27.2	42.7	29.1	19.8	35.9	40.7
ConvNeXt-Small [4]	COCO	None	51.5	70.6	56.1	34.8	55.2	66.9
ConvNeXt-Small [4]	COCO	GT mask	51.8	70.6	56.3	34.5	55.9	66.6
ConvNeXt-Small [4]	COCO	MAL mask	51.7	70.5	56.2	35.2	55.7	66.8

Table 1. Results of detection by adding different mask supervision. The models are evaluated on COCO val2017 and LVIS v1. By adding mask supervision using ground-truth masks or mask pseudo-labels, we can get around 1% improvement on different AP metrics on LVIS v1. On COCO val2017, the detection performance also benefits from mask pseudo-labels. Although the improvement is less than COCO's, the improvement is consistent over different random seeds.

provement similar to ground-truth masks on detection. We show the results in Tab. 1.

4. Additional qualitative results

We also visualize the prediction results produced by the instance segmentation models trained with ground-truth masks and mask pseudo-labels in Fig. 1. In most cases, we argue that humans cannot tell which results are produced by the models supervised by human-annotated labels.

References

[1] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1, 2

[4] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 1, 2

[5] 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. *arXiv preprint arXiv:2010.11929*, 2020. 1

[6] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 1

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[9] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 1

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1