TinyWT: A Large-Scale Wind Turbine Dataset of Satellite Images for Tiny Object Detection (Supplementary Materials)

Mingye Zhu[1,2,*], Zhicheng Yang[3,†], Hang Zhou[4,‡], Chen Du[3], Andy Wong[3], Yibing Wei[3,5,*], Zhuo Deng[2,6,*], Mei Han[3], Jui-Hsin Lai[3,†]
[1]USTC, China [2]Ping An Technology, China [3]PAII Inc., USA [4]Alchemy Insight, USA
[5]University of Wisconsin - Madison, USA [6]Tsinghua SIGS, China

A. Impact of Hyper-Parameters

In the training stage, two important hyper-parameters are the decoder feature dimension C and the feature embedding D' in Eq. 7 during the computation of supervised contrastive loss. Next we conduct two ablation studies to investigate the impact of these two hyper-parameters.

Selection of different feature dimensions. In Sec. 5.2, we simply use the last single-layer feature representation for SCL. Here the impact of different feature map layers is investigated. We evaluate our proposed framework with modifications of feature map selection and demonstrate the results in Table A.1. We observe that the proposed SCL is not obviously sensitive to different feature map selections.

Feat. Dim. C	96	192	384	768	1440[§]
Val mIoU(%)	81.11	81.03	80.64	81.07	81.39

Table A.1. Selection of different feature dimensions. [§]Resizing and concatenating all the feature maps.

Impact of different feature embeddings in supervised contrastive loss. In SCL, the feature representations of both positive and negative samples need to be normalized into an embedding dimension of D' to calculate the contrastive loss. Previously we set D'=128 and now we experiment with different embeddings. The results are displayed in Table A.2 and we see that within a certain range, the results are not sensitive to the selection of embedding dimensions within a certain range.

Feat. Embedding D'	64	128	256	512
Val mIoU(%)	80.96	80.64	80.70	79.52

Table A.2. Impact of different feature embeddings in the supervised contrastive loss.

B. Experiment Results of Detection Framework on TinyWT

In the main text, we apply Transformer-based segmentation methods on TinyWT. To accomplish more generalization of our dataset, we also provide reference detection results in this section. The detection model we adopt here is the recently advanced Transformer-based detection framework DINO [6], which has achieved state-of-the-art and outstanding performance for various detection tasks [3]. We re-annotate every one-dot label to the format of a bounding box of 5×5 pixels, and make it compatible with the state-of-the-art detection models. Table B.1 lists the results of DINO on TinyWT with different backbones. The experiment results of the standard detection metric Average Precision (AP) are displayed for image patches. Similarly, we employ the same evaluation protocol as in Sec. 5.1 and merge the patch-level inference results back to the original image size and calculate the overall precision, recall, and accuracy results for the whole TinyWT. As we can see, DINO achieves comparable precision but worse recall values compared with Transformer-based segmentation methods, resulting in lower accuracies as well. This suggests that when the problem is treated as a conventional detection task, tiny objects are more likely to be omitted by detectors. This degradation in performance also justifies our endeavor to reposition the tiny object detection problem and exploit segmentation methods to localize and count the wind turbine instances in the first place.

*This work was done when Mingye Zhu and Zhuo Deng were interns at Ping An Technology, and when Yibing Wei was an intern at PAII Inc..
†Correspondence: Zhicheng Yang (zcyangpingan@gmail.com); Jui-Hsin Lai (juihsin.lai@gmail.com)
‡This work was done when Hang Zhou was at PAII Inc., USA.

Method	AP	AP_{50}	AP_{75}	Precision(%)	Recall(%)	Accuracy(%)
	(on image patches)			*(on whole images)*		
DINO w/ ResNet50	0.583	0.826	0.448	97.60	94.22	92.09
DINO w/ Swin-T	0.605	0.836	0.472	97.97	94.52	92.70

Table B.1. Detection results for TinyWT using DINO.

C. Geographic Distribution of TinyWT

Fig. C.1 depicts the overall geographic distribution of TinyWT. We can see that TinyWT is endowed with an extensive longitude and latitude layout. Moreover, TinyWT displays the rich distribution of land cover land use, covering rangelands, croplands, shorelines, and many other areas. This diversity of landscape included in our dataset makes TinyWT an exemplar of wind turbine detection on a large scale.

D. Visualization Results

Next we exhibit visualization results on TinyWT from three example regions, which are displayed in Fig. C.1 with red boxes. These three regions cover the top four land use land cover (LULC) categories of TinyWT (rangeland (53%), crop (19%), bareground (9%), and tree (8%)).

Region 1. Fig. D.1 provides visualization results of various models in Region 1 covered by *rangeland*, the most dominant land use land cover (LULC) category of TinyWT. As shown in Fig. D.1a and D.1b, DeepLabv3 [1] and PSPNet [7] fail to recognize some wind turbines in rangeland, which have slightly different blocky background textures. On the other hand, SegFormer MiT-B2 [5] and UperNet [4] with Swin-T [2] mistake certain background objects with wind turbine-like features as wind turbines. However, they can successfully spot all ground-truth samples without false positives when incorporated with our design.

Region 2. The visualization results in Region 2 are presented in Fig. D.2. This region's LULC category is a mixture of *rangeland* and *bareground* (the 3rd top category of LULC). As we can see, all baseline models have several false negatives or false positives, resulting in less desirable detection results. However, when taking advantage of the proposed CSC and SCL modules, stronger constraints are imposed and more robust and distinguishable patterns are learned, thus the models become more acute and robust detectors. In Fig. D.2f, all false positives in Fig. D.2e with confusing appearance but not displayed in the regular arrangements of normal wind turbines are successfully suppressed. These visualization results directly demonstrate that we can extract more discriminative features and achieve more satisfying results for tiny object detection.

Region 3. In Fig. D.3, the visualization results of Region 3 are demonstrated, whose LULC category is a combination of *crop* and *tree*. As is clearly observed, CNN-based models (DeepLabv3 and PSPNet) have a tendency to omit true positives, which is also the case in Region 1 and 2. We thus conclude that crucial features of tiny objects tend to degrade during deep convolution operations. On the contrary, with the multi-head attention mechanism, Transformer-based models do not usually omit true positives but make additional false positives. However, when combined with our design, they can achieve more precise and accurate results, as illustrated in Fig. D.3d and Fig. D.3f.

References

[1] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 2, 4, 5, 6

[2] 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. 2, 4, 5, 6

[3] Papers with Code. Dino: Detr with improved denoising anchor boxes for end-to-end object detection — papers with code. https://paperswithcode.com/paper/dino-detr-with-improved-denoising-anchor-1, 2022. Accessed: 2022-11-08. 1

[4] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 2, 4, 5, 6

[5] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2, 4, 5, 6

[6] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 1

[7] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2, 4, 5, 6

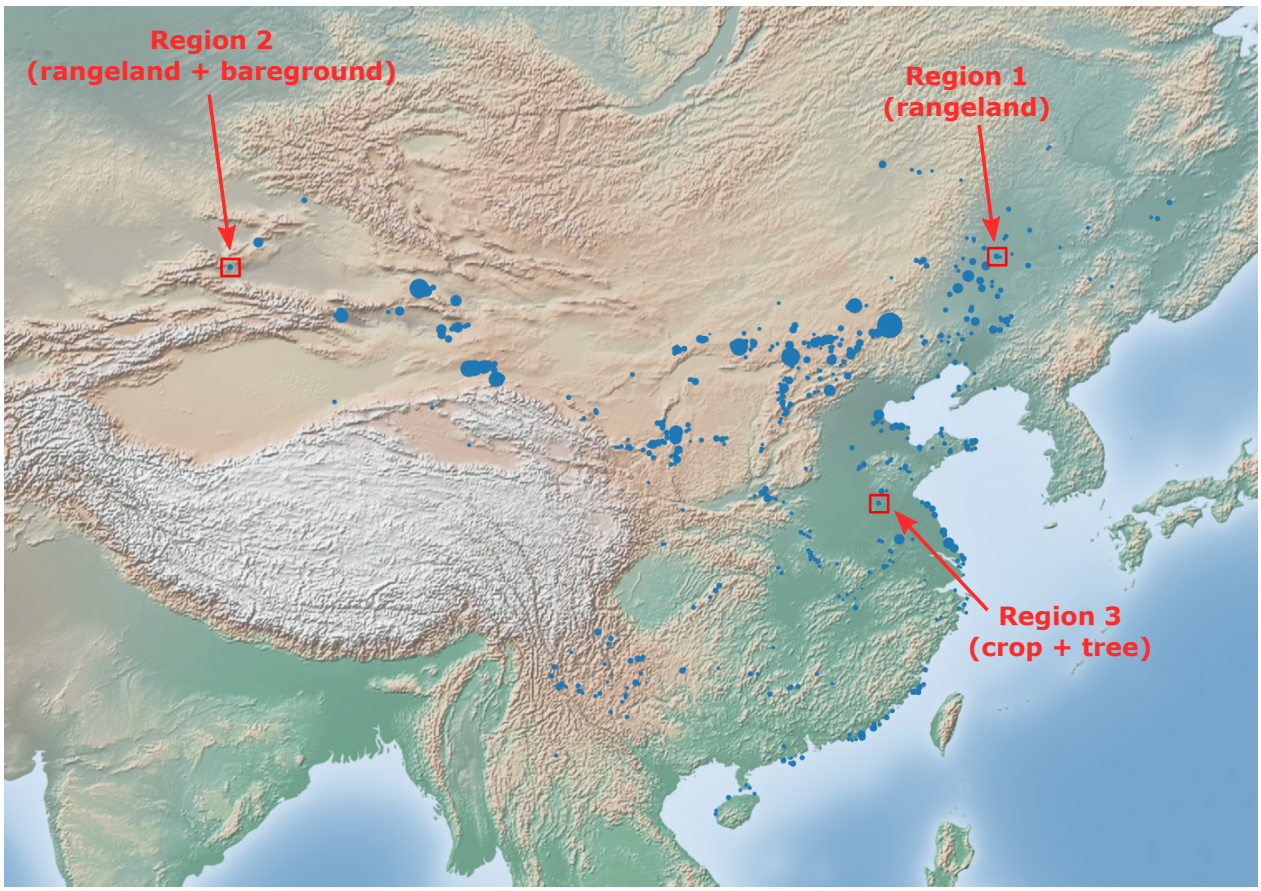
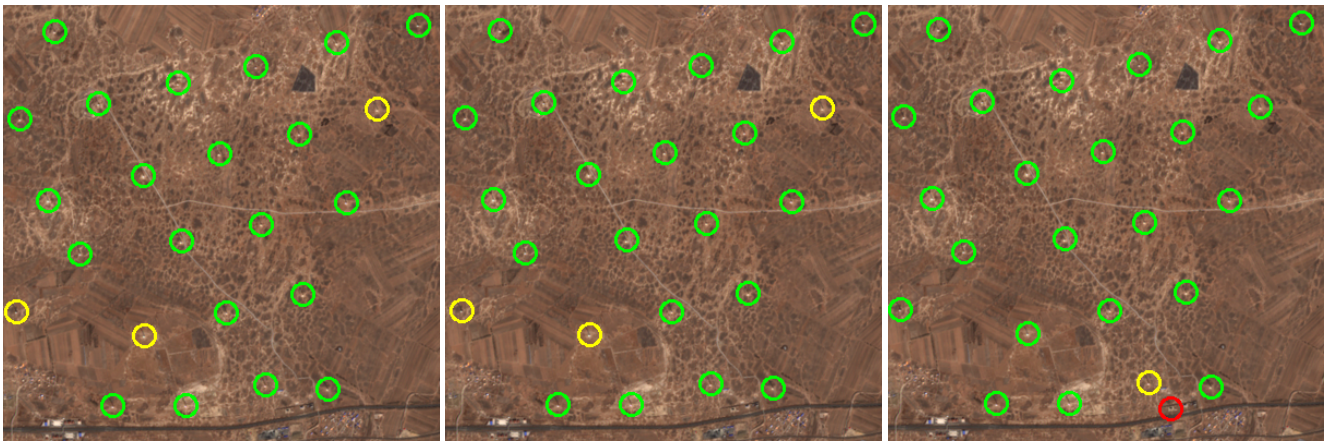


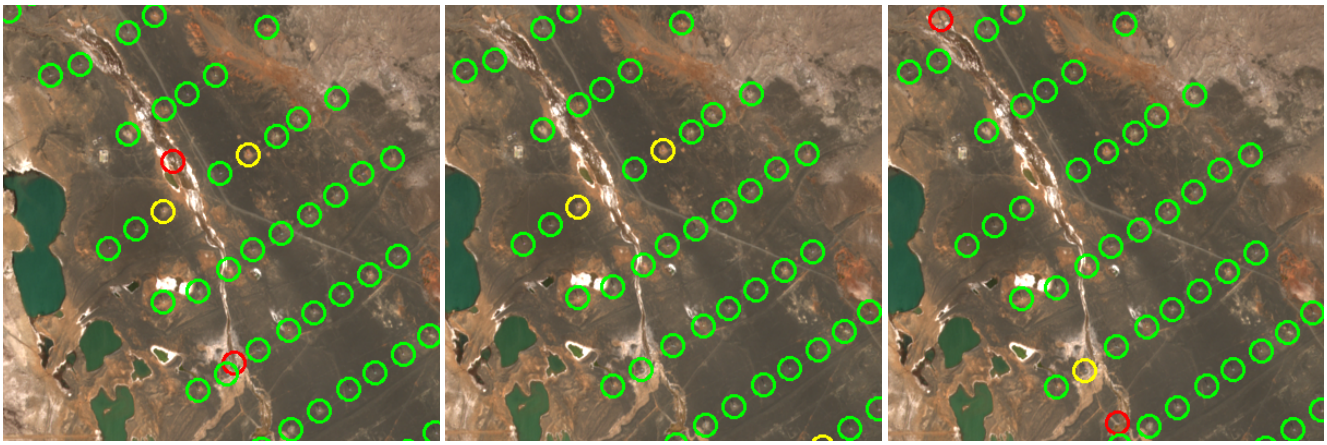
Figure C.1. An overview of data distribution for TinyWT. Regions dotted in blue refer to the data collection location of TinyWT. The diameter of each dotted point indicates the number of wind turbines in each region. The red boxes refer to three example regions for visualization results in Sec. D.



(a) DeepLabv3 [1] (b) PSPNet [7] (c) SegFormer MiT-B2 [5]

(d) SegFormer MiT-B2 [5] + **Our design** (e) Swin-T [2] + UperNet [4] (f) Swin-T [2] + UperNet [4] + **Our design**

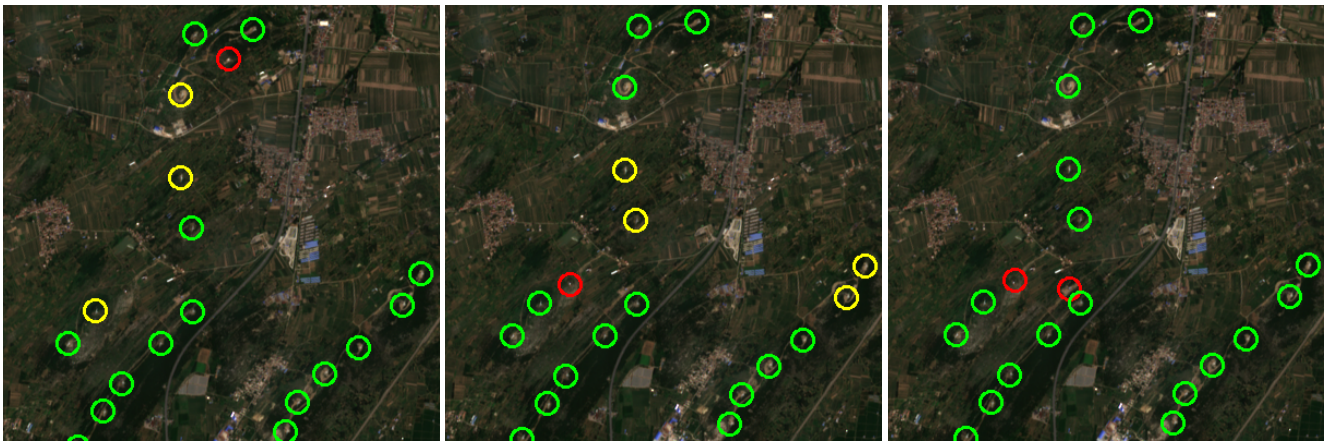
Figure D.1. Example visualization results of different baseline models and our proposed ones in Region 1. *Green circle*: True Positives; *Yellow circle*: False Negatives; *Red circle*: False Positives. The center of a circle denotes the centroid of a detected wind turbine blob.



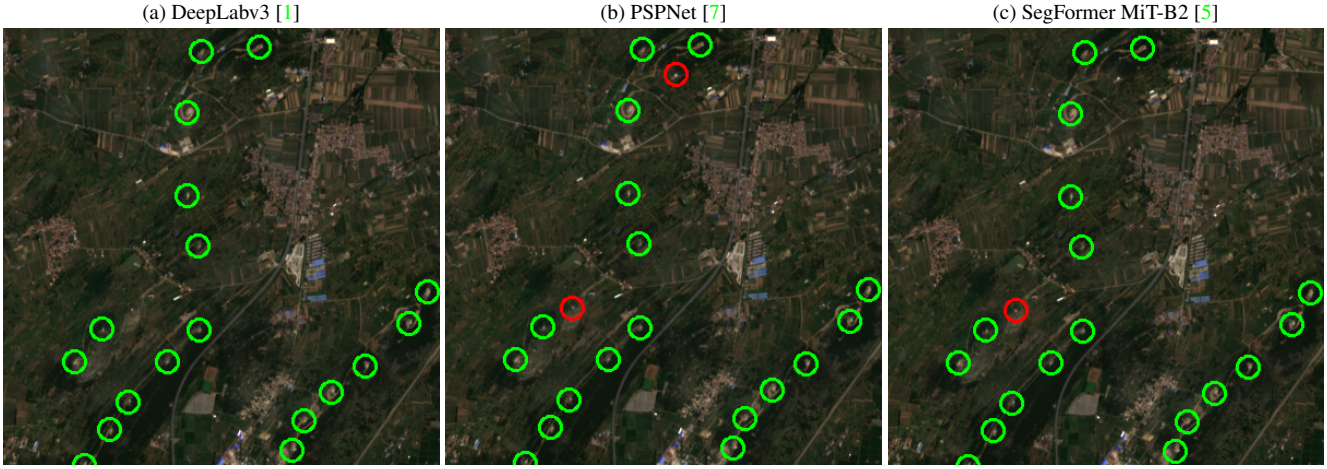
(a) DeepLabv3 [1] (b) PSPNet [7] (c) SegFormer MiT-B2 [5]

(d) SegFormer MiT-B2 [5] + **Our design** (e) Swin-T [2] + UperNet [4] (f) Swin-T [2] + UperNet [4] + **Our design**

Figure D.2. Example visualization results of different baseline models and our proposed ones in Region 2. *Green circle*: True Positives; *Yellow circle*: False Negatives; *Red circle*: False Positives. The center of a circle denotes the centroid of a detected wind turbine blob.



(a) DeepLabv3 [1] (b) PSPNet [7] (c) SegFormer MiT-B2 [5]



(d) SegFormer MiT-B2 [5] + **Our design** (e) Swin-T [2] + UperNet [4] (f) Swin-T [2] + UperNet [4] + **Our design**

Figure D.3. Example visualization results of different baseline models and our proposed ones in Region 3. *Green circle*: True Positives; *Yellow circle*: False Negatives; *Red circle*: False Positives. The center of a circle denotes the centroid of a detected wind turbine blob.