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

Anonymous CV4EO Algorithms Track submission

Paper ID 10

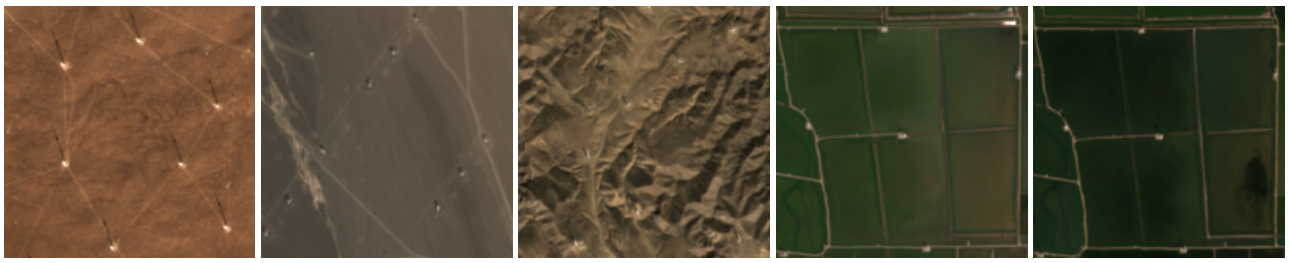

Figure 1. Wind turbine examples in different regions and seasons of our dataset. *First:* Scattered on rangeland in spring. *Second:* Regularly spaced in a Gobi desert region in summer. *Third:* Deployed alongside the ridgelines in autumn. *Fourth and Fifth:* Arranged on agricultural land in summer and winter, respectively.

Abstract

Tiny object detection is a challenging task. Many datasets for this task are released in past years, spanning from natural scene to remote sensing images. However, wind turbines in satellite images, a significant category of tiny objects, have not been well included. Aiming at completing the tiny object datasets, we release TinyWT, a large-scale year-round tiny wind turbine dataset of satellite images. It has 8k+ images, a very tiny object size of 3–6 pixels, and 700k+ annotations in total with the extensive effort of human correction. Unlike other tiny object datasets of aerial/satellite images that are limited to academic research only, our dataset is free for commercial use. Every pixel's geographic coordinates are also explicitly extracted for researchers without related domain knowledge. Meanwhile, we reposition the tiny object detection task as a localizing-and-counting problem and incorporate segmentation techniques, and propose a novel design to exploit the strengths of contextual similarity constraint and supervised contrastive learning. The experiment results of both baseline models (CNN-based and Transformer-based models) as well as our special design are presented. Without bells and whistles, our design effectively improves the baseline models' performance, achieving a maximum of 4.94% mIoU gain where 21.15% of false negatives are recalled and 22.02% of false positives are removed.

1. Introduction

Object detection is one of the most challenging and crucial tasks in Computer Vision (CV). Specifically, small or tiny object analysis is a critical and indispensable branch owing to its wide application in real-life scenarios. Many datasets of natural scene images have been publicly available to accelerate the progress of this research direction. Their primary purposes include crowd counting for public monitoring and surveillance [22, 48, 52, 57, 69, 71], driving assistance [68], and emergency rescue [66].

Compared with natural scene images, remote sensing images inherently present abundant occurrences of small objects. Recent years have witnessed various datasets of aerial and satellite images for small object detection, covering the applications of vehicle monitoring [11, 20, 44], hub throughput management [38, 74], and more general purposes [33, 54, 59, 60]. However, one significant category, *wind turbine* in satellite images, has been long-neglected.

Admittedly, several global remote sensing-based wind turbine data resources [12, 19, 70] are recently available, however, due to the highly professional annotation formats (GeoJSON [2] or Shapefile [13]), they are not directly usable for researchers from other disciplines. To this end, a large-scale off-the-shelf wind turbine dataset is needed

to substantially complete the tiny object datasets for researchers with different knowledge backgrounds. Furthermore, wind turbines serve an integral role in the hot topics related to earth observation, such as green power and carbon neutrality [51]. These tasks are economically profit-oriented or highly associated with commercialization. Most of the existing datasets are composed of high-resolution images or obtained from Google Earth [15], limiting their use for commercial purposes. Unlike the aerial image sources for academic research only, the Sentinel-2 (S2) satellite imagery [10] with a medium spatial resolution of 10m is an ideal image source for dataset construction with commercial use permission.

Along with releasing diverse small/tiny object datasets, scholars also explore effective approaches to solve this specific problem. The *detection*-based frameworks are commonly adopted to achieve the baseline and more advanced performance [17, 32, 63]. The overall detection accuracy is acceptable because the great majority of those objects are not tiny enough but still with a recognizable bounding box annotation. Nevertheless, the performance of very tiny objects is still not satisfying using the conventional detection frameworks and evaluation metrics [26, 55, 66]. Another methodology to analyze tiny objects exploits *segmentation* techniques. For example in the task of crowd counting aforementioned, a bounding box annotation is not the optimal option because human heads are highly tiny and dense in surveillance images. Instead, a one-dot annotation is placed at a human head and the problem is primarily reformulated as calculating the density similarity of the predicted segmentation map and ground truth to evaluate the pixel-level estimation [14, 18, 56].

In this paper, we introduce a public dataset **TinyWT** and propose a novel design to enhance the overall performance of wind turbine detection. To the best of our knowledge, TinyWT is the first large-scale dataset for tiny wind turbine detection of satellite images. It covers 474 wind turbine farm regions in one entire year, containing a total of 8,871 images and 736,426 annotations. The annotated area has a *very tiny* object size of 3–6 pixels. Fig. 1 provides several examples of our TinyWT in different regions and seasons. To investigate an appropriate solution for tiny object detection based on this dataset, we rethink the problem, develop the corresponding evaluation protocols, and propose a novel design to boost the model's performance. Our key contributions are summarized as follows.

- Unlike normally transferring CV knowledge to other application sectors, TinyWT *reversely nourishes* the CV community by fundamentally completing the entire tiny object detection dataset family. First, the wind turbine is a tiny object category with long-term insufficient attention. Second, we clear the domain knowledge hurdle by extracting the professional geographic annotations and converting them to local pixel-level annotations on S2 satellite images. Third, TinyWT is completely free for commercial use, distinct from the vast majority of public tiny object datasets.

- We perform an extensive manual examination of every image to rectify the missing or incorrect labels in [12]. The number of corrected images occupies 40.86% of the total amount of images. Moreover, we provide two types of corrected annotations: one-dot annotation at a wind turbine hub center as well as the original format, and an expanded binary mask to cover the hub area.

- We reposition the tiny object detection task as a *localizing-and-counting* problem and incorporate segmentation techniques. Besides, two novel plug-and-play modules are proposed to emphasize contextual similarity constraint and supervise contrastive learning, respectively. Our design effectively improves the baseline model's performance by a maximum of 4.94% mIoU gain and 1.22% accuracy improvement on our updated evaluation metrics, without introducing extra computational overhead during inference.

2. Related Work

2.1. Datasets for Tiny Object Analysis

Tiny object analysis has long been a challenging task in natural images. Many dedicated public datasets contribute to the progress of this research domain. Tiny ImageNet [31] is constructed for tiny object classification. The datasets such as Wider Face [64], WiderPerson [69], CityPersons [68], TinyPerson [66] are benchmark datasets for detecting person's face or figure. Moreover, many datasets related to crowd counting [21, 22, 40, 52, 57, 71] aim at counting people's heads in images for public monitoring and surveillance. Datasets for other purposes include tiny object tracking [76], few-shot learning [48], etc.

Remote sensing imagery naturally contains small objects. Many datasets are released for detecting single-class objects, such as car [20, 44] and ship [38], or multi-class detection tasks [33, 60, 74]. However, the overall object size of those datasets is not tiny enough. Recently, a dataset AI-TOD [55] is built upon the public aerial image datasets of DOTA v1.5 [60], xView [27], VisDrone2018-Det [75], Airbus-Ship [24] and DIOR [33]. It exacts the objects from those datasets only with an object size of no more than 64 pixels and an average object size of 12.8 pixels.

Nonetheless, those remote sensing datasets neither contain sufficient wind turbine instances nor are allowed for commercial use. In Table 2, we comprehensively compare TinyWT with those public datasets of remote sensing images mentioned above, and detail the uniqueness of our dataset in Sec. 3.1.

Type	#Images	%(of total)	#Instances	%(of total)
Mislabeled	529	5.96	6197	0.84
Unlabeled	3096	34.90	38913	5.28

Table 1. Number of mislabeled and unlabeled instances in the original dataset before correction.

2.2. Methods for Tiny Object Analysis

For the datasets with the bounding box annotations, detection-based approaches are widely applied. Several works [33, 60] evaluate the multiple classical CNN-based detection baselines [8, 29, 34–36, 49, 50]. More specific designs are proposed for better tiny object representation such as $\mathcal{R}^2$-CNN [46], Scale Match [66], M-CenterNet [55], C3Det [32]. In recent years, contrastive learning has proved its powerful efficacy in self-supervised learning (SSL) [6] and even supervised tasks [23, 25], also in remote sensing-based SSL [1, 58] and supervised contrastive learning tasks [3, 30]. Overall, the aforementioned technologies try to incorporate stronger context information or improve the representation of features to better distinguish between positive and negative samples.

Segmentation techniques are also widely exploited to analyze tiny objects. For example, researchers have leveraged segmentation approaches to address the crowd counting problem, including density map estimation [14, 18, 56], localization [22], and counting with blobs [28]. Authors in [65] adopt DeepLabv3 [5] and PSPNet [72] as the segmentation baselines for validating their designs for small objects augmentation. In recent years, the impressive performance of Transformer-based approaches has been witnessed in segmentation tasks [4, 37, 53, 62, 73]. Segformer [62] is equipped with a hierarchical Transformer encoder and a lightweight all-MLP decoder design, achieving impressive results. ViT [9] and Swin Transformer [37] are also utilized as feature extractors for other segmentation methods [53, 61]. In this work, we design two modules to explicitly incorporate contextual information and impose similarity constraint, and adopt supervised contrastive learning, to further boost performance of Transformer-based baselines.

3. The TinyWT Dataset

3.1. Dataset Construction

Original form of annotations. Authors in [12] extract the worldwide locations of wind turbines based on the open-source global mapping project OpenStreetMap (OSM) [16]. Nonetheless, those locations with geospatial coordinates are stored in professional GeoJSON file format, which is a geographic information system (GIS) vector format, requiring special software or libraries for further processing. This format is thus not for the wider research community lacking related domain knowledge.

Figure 2. Varying visual appearance of the same wind turbine during four quarters.

Image collection. The original data form described above has only geo-location annotations without images. To build an annotated image dataset, we leverage S2 image source to extract corresponding images from the annotated regions. Due to the moderate spatial resolution, the locations of some adjacent wind turbines are close. Rather than cropping a small image patch to contain one single wind turbine sample only, we thus leverage a union-find algorithm to cluster the wind turbine label points based on their geo-locations. Specifically, a label point and a distance between every two points are regarded as a node and an edge, respectively. As a result, nearby points within a distance threshold are included in the same group to occupy a large region. We also set a distance buffer extended from the outermost wind turbines to avoid partially cropping a sample on the image margin. Finally, a rectangle region is cropped to contain the grouped wind turbines plus the buffered distance.

Since the annotations were established using the OSM in 2018 [12], we extract the S2 images during the same year. Considering the overall dataset size, this version of TinyWT contains the entire onshore territory of China. Aside from eliminating the images with cloud cover (white regions) or no data (abnormal black regions), we also remove images with snow where the hubs are unable to be recognized. As a result, a total of 8,871 3-channel PNG images from 474 regions across China with an average image size of 956×1011 are achieved. The geographic location information of every region is separately stored in a NumPy format [45].

Annotation correction. As the OSM is an open-source platform with limited quality control, the original annotations suffer from a considerable number of mislabeled and omitted instances. We thus amend the ground truth labels with extensive human scrutiny based on the protocol below.

1) All year-round images are grouped into 474 sets in the QGIS software [47], each sharing the same location but different time stamps. Since it is possible that the visual information on a single image is difficult to be identified, we cross-check the time-series image sequence to recognize the wind turbines. Fig. 2 illustrates that the example visual appearance of wind turbines varies during different quarters.

2) For any instance that is labeled as a wind turbine by OSM but is vacant on the image, we record the geographic coordinates in QGIS and remove those misla-

Dataset	Source	#Img.	Img. Size(px.)	#Cat.	#Anno.	Target	Anno. Way	Spat. Res.	Object Size(px.)
UCAS-AOD [74]	aer./sat.	2420	659–1372	2	14596	plane/car	bbox	/	65.5±24.5
COWC [44]	aerial	53	2k–19k	1	32716	car	one dot	0.15m	24–48
CARPK [20]	aerial	1448	1280×720	1	89777	car	bbox	∼0.017m	58.4±12.2
HRSC2016 [38]	aer./sat.	1061	300–1.5k	19	2976	ship	bbox/seg	0.4–2m	140.6±67.9
DOTA [60]	aerial	2806	800–13k	15	118282	common	bbox	∼0.4m	55.3±63.1
DIOR [33]	aer./sat.	23463	800×800	20	192472	common	bbox	0.5–30m	65.7±91.8
TinyPerson [66]	aerial	1610	∼500–∼5k	1	72651	person	bbox	/	18±17.4
AI-TOD [55]	aerial	28306	800×800	8	700621	common	bbox	∼0.4–30m⋆	12.8±5.9
FAIR1M [54]	aer./sat.	15266	1k–10k	37	∼1.02m	common	bbox	0.3–0.8m	⩾16
TinyWT(Ours)‡	satellite	8871	513–6731	1	736426	wind turbine	one dot	10m	∼3–6

Table 2. Comparison of our dataset and the existing public benchmarks and datasets of remote sensing imagery for small/tiny object detection. ⋆It is estimated based on the datasets used for AI-TOD dataset construction since the spatial resolution of this dataset is not explicitly given. ‡Free for commercial use.

beled instances from the ground truth. For wind turbines that are observable from the image but not annotated in ground truth labels, we supplement the annotations with the new coordinates.

3) Such steps are repeated by three human annotators for every image to avoid possible mistakes. Each annotator approximately takes one hour for scrutinizing all year-round images of one region, leading to a total of ∼450 hours to screen the whole dataset.

Table 1 lists the results of our annotation correction. Once the entire annotations have been rectified, we compute the data statistics and obtain an overview of data distribution, as shown in Sec. 3.2.

Licenses for commercial use. Our dataset construction leverages three existing resources: 1) the QGIS software has `CC BY-SA 3.0` license; 2) the original annotation dataset [12] has `CC0` license; 3) the image data are acquired from the S2 satellite. All of them are free for commercial use [7].

3.2. Dataset Overview

Dataset comparison. Table 2 comprehensively compares our dataset with other public ones related to our research topic. As we can see, our outstanding properties mainly contain the following aspects. First, the object type is the overlooked category in remote sensing images, compared with other common tiny objects in the community such as planes, vehicles, and ships. AI-TOD [66] covers the wind turbine category, but its instance percentage is very limited (∼0.076%). Second, our source is S2 satellite with a medium spatial resolution (10m) eligible for commercial use. Third, due to the various annotation ways, we define the "object size" as the square root of the area of the region of interest (RoI). Our object size has the smallest pixels, belonging to the descriptive *very tiny* category that has an object size of [2, 8] pixels [55, 66].

Moreover beyond Table 2, our dataset uniquely provides each pixel's geographic coordinates and year-round image sequence at each location. A typical public dataset of remote sensing images simply releases the PNG/JPG images, or the raw professional TIFF/JP2 images with embedded geographic coordinates. However, explicitly leveraging such geospatial information is non-trivial for CV researchers. To this end, we extract the commonly used WGS84 EPSG:4326 coordinates as Google Earth adopts for every pixel location.

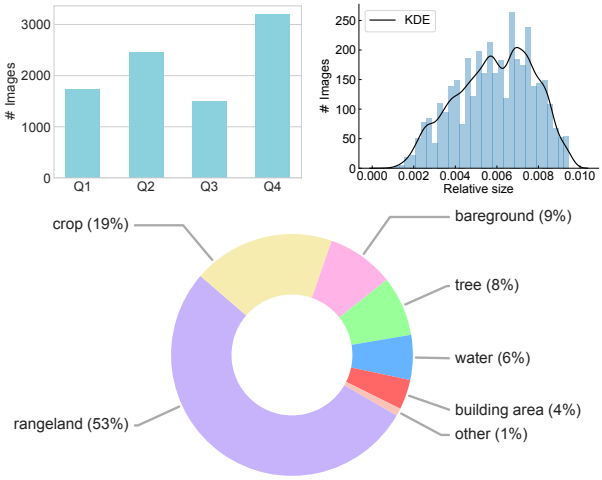

Figure 3. Statistical analysis of our TinyWT dataset. *Upper left:* Number of images in different quarters. *Upper right:* Distribution of relative size. *Bottom:* LULC distribution.

Statistical analysis. We show detailed statistical information in Fig. 3. From the upper left plot, we observe that a large number of images in TinyWT are taken in the second and fourth quarters. This is mainly due to the fewer available images during the snowy season (Q1) and the rainy/cloudy season (Q3). The upper right plot provides the distribution and the kernel density estimation (KDE) of "relative size" of our dataset. The bottom plot unveils the distribution of land cover land use (LULC) of our dataset.

Dataset attributes. One typical attribute of TinyWT is the lower object resolution. A tiny wind turbine instance oc-

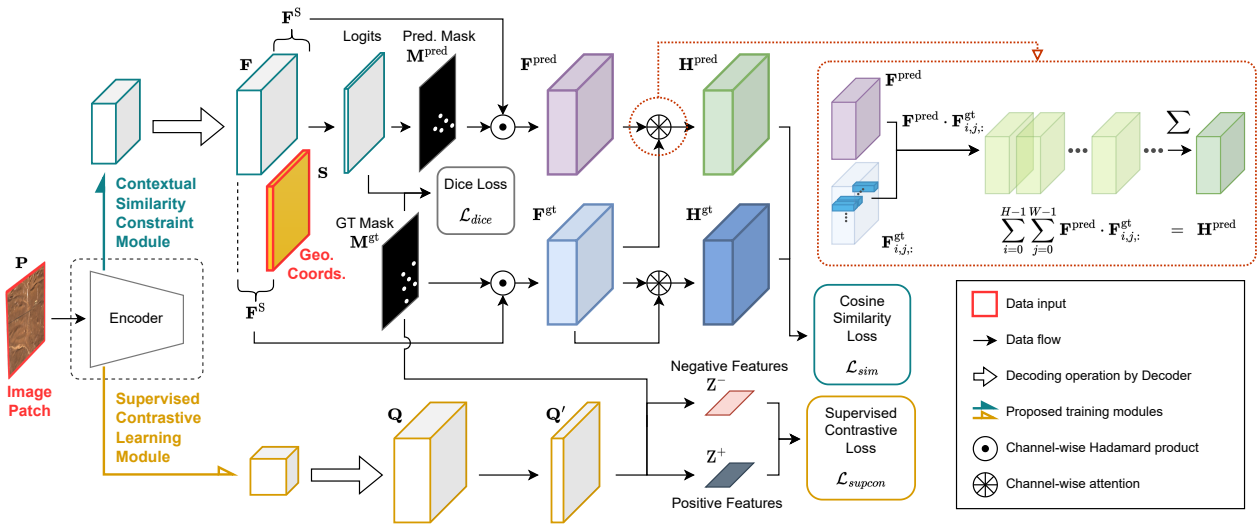

Figure 4. Framework of the proposed training scheme. Given an image patch $\mathbf{P}$, an encoder is adopted to extract multi-scale feature representations. The *Contextual Similarity Constraint* (CSC) module is utilized to incorporate spatial information into feature representation $\mathbf{F}$ and compute the cosine similarity loss $\mathcal{L}_{sim}$. The *Supervised Contrastive Learning* (SCL) module leverages feature representation $\mathbf{Q}$ to form positive and negative pairs and result in $\mathcal{L}_{supcon}$. Besides, the dice loss $\mathcal{L}_{dice}$ is calculated from the 2-dimensional feature logits and ground-truth mask. By combining these three losses, we get our final objective function.

cupies only several pixels of an S2 image, presenting very limited appearance information (shown in Fig. 2). More prominently, a clear shadow of the turbine blade can be observed at an instance in AI-TOD [55], but it is not distinguishable in our dataset. Similar to other tiny objects in public datasets, a wind turbine is also susceptible to background interference. For example, a transmission tower or a water tower has a very similar size of RoI.

4. Methodology

4.1. Motivation

Regarding our very tiny object detection, predicting a perfect shape is challenging due to the object resolution. This enables us to rethink the problem. As mentioned in Sec. 1 and 2, the crowd counting task has similar difficulty detecting crowded people in surveillance images. Authors in [22] define three tasks to appropriately tackle this problem, which are counting, localization, and density map estimation. The density map estimation is not necessary when the objects do not have much occlusion effects as in our case [28]. Therefore, the more suitable description of our problem is "*localizing and counting*" the tiny objects.

To address the original crowd counting problem, valid solutions using segmentation methods have been reviewed in Sec. 2.2. Inspired by that methodology, we convert the geographic one-dot hub annotation into very tiny segmentation masks. Even though wind turbines have different heights, the annotated hub area in an S2 image has an approximate object size of 3 to 6 pixels (shown in Table 2). As the original one-point annotations correspond to the coordinates of the centers of wind turbine hubs, we expand the annotated point into a binary mask of 5×5 pixels to cover the foreground hub area. This scheme of annotation conversion was also endorsed in a recent pilot study of detecting wind turbines in S2 images using image processing approaches [41, 42]. Accordingly, our methods contain the segmentation framework to predict the location and size of wind turbine regions and an appropriate measure to count them. We elaborate on our proposed approach in the following sub-sections. The design of corresponding evaluation protocols is introduced in Sec. 5.1.

4.2. Overview of the Proposed Approach

In this section, we briefly describe our entire framework as shown in Fig. 4. Due to the various size of raw images, we crop each image $\mathbf{I} \in \mathbb{R}^{H_I \times W_I \times 3}$ into patches $\mathbf{P} \in \mathbb{R}^{H \times W \times 3}$ with a fixed height H and width W. During the training flow, we design two special learning modules to boost the performance of segmentation methods for tiny objects without introducing any overhead during inference, which are *Contextual Similarity Constraint* (CSC) module and *Supervised Contrastive Learning* (SCL) module. These two modules leverage the off-the-shelf spatial information provided in our dataset, and can be considered as a plug-and-play integration to any encoder-decoder-based segmentation models. The respective losses can be added to any original loss options in gradient backward computation. Note that no additional operations or computation is needed during inference.

4.3. Contextual Similarity Constraint

Heuristically, the inherent spatial patterns of wind turbine layout greatly help discriminate wind turbines, which is also verified during the process of label correction by human annotators. Inspired by this observation, we design CSC to explicitly utilize geospatial information for imposing location-aware bias, and combine it with the resized feature representations from the decoder to create a *visual and spatial similarity* constraint.

We first normalize the longitude and latitude values of each pixel's location available in our dataset, resulting in an array of the normalized geo-coordinates $\mathbf{S} \in \mathbb{R}^{H\times W\times K}$, where the dimension K relies on the choice of normalization manner. $\mathbf{S}$ is next concatenated with the resized decoder feature map $\mathbf{F} \in \mathbb{R}^{H\times W\times C}$ to obtain $\mathbf{F}^{\mathrm{S}} \in \mathbb{R}^{H\times W\times(C+K)}$.

The binary prediction mask generated by $\mathbf{F}$ is given by:

$$\mathbf{M}^{\mathrm{pred}} = \underset{\mathrm{dim}=-1}{\arg\max}\, \mathbf{F} \in \{0,1\}^{H\times W}. \tag{1}$$

We let $\odot$ denote the channel-wise Hadamard product operation. The masked feature maps $\mathbf{F}^{\mathrm{pred}}$ and $\mathbf{F}^{\mathrm{gt}}$ are thus calculated by

$$\mathbf{F}^{\mathrm{pred}} = \mathbf{M}^{\mathrm{pred}} \odot \mathbf{F}^{\mathrm{S}} \in \mathbb{R}^{H\times W\times(C+K)}, \tag{2}$$

$$\mathbf{F}^{\mathrm{gt}} = \mathbf{M}^{\mathrm{gt}} \odot \mathbf{F}^{\mathrm{S}} \in \mathbb{R}^{H\times W\times(C+K)}. \tag{3}$$

Now that $\mathbf{F}_{\mathrm{pred}}$ represents the model's faith in the semantic information of all the predicted instances, we implement channel-wise attention to emphasize the similarity between any predicted and ground truth instances in the embedding space within the entire patch to reach a consensus about the fidelity of the prediction.

$$\mathbf{H}^{\mathrm{pred}} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} \mathbf{F}^{\mathrm{pred}} \cdot \mathbf{F}^{\mathrm{gt}}_{i,j,:} \in \mathbb{R}^{H\times W\times(C+K)}, \tag{4}$$

$$\mathbf{H}^{\mathrm{gt}} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} \mathbf{F}^{\mathrm{gt}} \cdot \mathbf{F}^{\mathrm{gt}}_{i,j,:} \in \mathbb{R}^{H\times W\times(C+K)}. \tag{5}$$

Then we flatten $\mathbf{H}^{\mathrm{gt}}$ and $\mathbf{H}^{\mathrm{pred}}$ into $\hat{\mathrm{H}}^{\mathrm{gt}} \in \mathbb{R}^{N\times(C+K)}$ and $\hat{\mathrm{H}}^{\mathrm{pred}} \in \mathbb{R}^{N\times(C+K)}$, respectively, where $N = HW$. Finally, the cosine similarity is computed:

$$\mathcal{L}_{sim} = 1 - \cos\varphi = 1 - \frac{\hat{\mathrm{H}}^{\mathrm{pred}}\,\hat{\mathrm{H}}^{\mathrm{gt}}}{\left|\hat{\mathrm{H}}^{\mathrm{pred}}\right|\left|\hat{\mathrm{H}}^{\mathrm{gt}}\right|}. \tag{6}$$

This contextual constraint reasons about the visual and geometrical affinity of the predicted instances with the global context to alleviate local misalignment and avoid outliers, learning latent patterns in wind turbine distribution and thus guiding semantic propagation.

It is worth noticing that our design is also effective in the more general scenario where the geo-coordinates are not available. The geo-coordinates can be otherwise replaced by the pixel's relative location of the patch. We present the evaluation results in Sec. 5.3 and Table 4.

4.4. Supervised Contrastive Learning

Along with the success of contrastive learning in SSL tasks in recent years, variants of contrastive learning methods have been employed in SSL of remote sensing images [1, 3]. Furthermore, researchers also investigate contrastive learning in supervised tasks [23, 25], directly enhancing the learned representations guided by ground truth labels.

Inspired by the supervised contrastive learning paradigm, we develop SCL to emphasize object representations. Instead of using different views of the same image as image-level positive pairs, we let all the wind turbines in the same image form positive pairs and contrast against the background areas, which are negative pairs. Such a design enables the learned representations to be more robust and invariant to subtle changes in location. Once resizing the last-layer feature representation, we get the feature map $\mathbf{Q} \in \mathbb{R}^{H\times W\times D}$. Next we utilize a projection network $\phi(\cdot)$, to map the vectors into a fixed feature dimension D' to obtain $\mathbf{Q}'$:

$$\mathbf{Q}' = \phi(\mathbf{Q}) \in \mathbb{R}^{H\times W\times D'}. \tag{7}$$

Vectors of clustered pixels of wind turbines from $\mathbf{Q}'$ are regarded as positive pairs and flattened into $\mathrm{Z}^{+} \in \mathbb{R}^{N\times D'}$, where $N = HW$. $\mathrm{Z}^{-} \in \mathbb{R}^{N\times D'}$ are randomly sampled negative pairs from the background. A batch of positive and negative sample/label pairs can be thus denoted as $\mathrm{Z} = \{\boldsymbol{z}_\ell, \boldsymbol{y}_\ell\}_{\ell=1,\ldots,2N}$, where $\boldsymbol{z}_k \in \mathrm{Z}^{+}, k = 1,\ldots,N$; $\boldsymbol{z}_m \in \mathrm{Z}^{-}, m = (N+1),\ldots,2N$, and $\boldsymbol{y}_\ell \in \{0,1\}_\ell$ denotes the corresponding labels. Let $n \in \Gamma = \{1,\ldots,2N\}$ be the index of an arbitrary sample in Z, $\Theta(n) = \Gamma\backslash\{n\}$, and $\Omega(n) = \{\omega \in \Theta(n) : \boldsymbol{y}_\omega = \boldsymbol{y}_n\}$ is the set of indices of all positives in Z excluding n. Next the supervised contrastive loss is constructed as:

$$\mathcal{L}_{supcon} = \sum_{n\in\Gamma} \frac{-1}{|\Omega(n)|} \sum_{\omega\in\Omega(n)} \log \frac{\exp\left(\boldsymbol{z}_n \boldsymbol{z}_\omega/\tau\right)}{\sum\limits_{\theta\in\Theta(n)} \exp\left(\boldsymbol{z}_n \boldsymbol{z}_\theta/\tau\right)}, \tag{8}$$

where $\tau \in \mathbb{R}^{+}$ is a temperature parameter and $|\Omega(n)|$ is the cardinality of $\Omega(n)$.

As pointed out in [25], the loss in Eq. 8 has an intrinsic ability to perform hard positive/negative mining because the gradient contributions from hard positives/negatives are large while those for easy positives/negatives are small. As a result, the tiny objects are endowed with strong feature representations which pull the disturbing candidates (e.g.,

	Method	Encoder	#Param.(M)	GFLOPs	mIoU(%) *(on image patches)*	Precision(%) *(on whole images)*	Recall(%) *(on whole images)*	Accuracy(%) *(on whole images)*
CNN	DeepLabv3 [5]	ResNet50	68.1	67.4	75.52	97.73	90.80	88.91
	PSPNet [72]	ResNet50	49.0	44.7	75.74	98.29	91.04	89.62
Transformer	SegFormer [62]	MiT-B2	24.7	4.48	74.98	97.09	97.17	94.42
			28.3	4.48	**77.41** (+2.43)	**97.28** (+0.19)	**97.52** (+0.35)	**94.93** (+0.51)
		MiT-B3	56.6	6.18	79.46	97.71	97.81	95.62
			60.2	6.18	**82.34** (+2.88)	**97.79** (+0.08)	**98.07** (+0.26)	**95.95** (+0.33)
		MiT-B5	82.0	13.0	82.96	**98.36**	98.10	**96.52**
			85.5	13.0	**83.81** (+0.85)	98.14 (−0.22)	**98.20** (+0.10)	96.41 (−0.11)
	UperNet [61]	Swin-T	59.8	59.6	76.73	96.23	97.92	94.30
			66.6	59.6	**81.07** (+4.94)	**97.06** (+0.83)	**98.36** (+0.44)	**95.52** (+1.22)
		Swin-S	81.1	66.4	84.25	97.44	98.35	95.88
			88.0	66.4	**87.31** (+3.06)	**97.75** (+0.31)	**98.78** (+0.43)	**96.58** (+0.70)
		Swin-B	121.2	77.3	89.33	98.00	98.72	96.76
			129.5	77.3	**89.71** (+0.38)	**98.15** (+0.15)	**98.96** (+0.24)	**97.15** (+0.39)

Table 3. Experiment results of baseline methods and our proposed framework on the validation set. The patch size is set as 256×256. Rows in gray color indicate the results of baseline models equipped with our special design.

false positives caused by background clutter and occlusion) away from true positives. The final objective function for the proposed framework is a linear combination of the conventional dice loss [43] $\mathcal{L}_{dice}$, the cosine similarity loss $\mathcal{L}_{sim}$ and the supervised contrastive loss $\mathcal{L}_{supcon}$ with weight coefficient α, β, and γ:

$$\mathcal{L} = \alpha\mathcal{L}_{dice} + \beta\mathcal{L}_{sim} + \gamma\mathcal{L}_{supcon}. \tag{9}$$

By assimilating CSC and SCL into the framework, we learn the stronger feature representations which are critical in tiny object recognition in versatile environments.

5. Experiment Results

In this section, we demonstrate the usage of the TinyWT dataset and present several baselines that can be used as reference results for future research. Apart from that, we investigate the effects of our proposed framework for tiny object detection.

5.1. Implementation Details

Dataset splitting. To evaluate our proposed framework, we split the entire TinyWT dataset into training, validation, and test sets by a ratio of 3:1:1, resulting in 5,322, 1,774, and 1,775 images without any overlapped regions, respectively. We conduct inference on the validation set only, and the ground truth labels for the test set will be reserved for future challenge uses.

Parameter settings. The encoders are pretrained on the ImageNet-1K dataset and the decoders are randomly initialized. Firstly the satellite images of different sizes are cropped into patches with a fixed size of 256×256, which are then fed into the model. We apply normalization to data and train the models using 4 Tesla V100 SXM2 GPUs (32G RAM) for 160k iterations with synchronized batch normalization. The batch size is set to 8 for all segmentation tasks. The optimizer is AdamW [39] with an initial learning rate of 0.00006, and a weight decay of 0.01.

During training, we simply leverage sine and cosine functions to normalize longitude and latitude in Sec. 4.3 separately, resulting in $\mathbf{S} \in \mathbb{R}^{256\times256\times4}$. In Sec. 4.4, we take the last-layer feature representation from the encoder, so the number of feature dimensions is up to the choice of different encoder settings. $\phi(\cdot)$ is a 1×1 convolutional layer to reduce channel numbers to $D' = 128$, and the temperature parameter τ is 0.07 thorough all experiments in contrastive loss. We simply set the weight parameters α, β, γ to 1 in Eq. 9. Note that the inference procedure remains unchanged as the original segmentation setting, where no additional design is needed.

Evaluation protocols. Since we reposition our problem as "localizing and counting" the tiny objects, relying on the canonical indicator Mean Intersection over Union (mIoU) to simply evaluate the semantic segmentation performance is not sufficient. To this end, we merge the patch-level inference mask back to the original image size and calculate the overall precision, recall, and accuracy results for the whole TinyWT. As the annotation is a blob of 5×5 pixels to cover the hub area, we define that a *true positive* is counted when 1) the area of the predicted blob is no more than 6×6 pixels, and 2) the distance between the centers of the predicted blob and the ground truth blob is no more than 7. These two rules are designed to jointly screen all predicted areas for their size and location. The former rule considers the

CSC (Rel.)	CSC (Geo.)	SCL	Val mIoU(%)
-	-	-	76.73
✓	✗	-	78.73 (2.60↑)
✗	✓	-	79.26 (3.13↑)
-	-	✓	79.42 (3.29↑)
✓	✗	✓	80.54 (3.81↑)
✗	✓	✓	81.07 (4.94↑)

Table 4. Effectiveness of CSC and SCL. "Rel." refers to a pixel's relative coordinates in an image patch. "Geo." represents a pixel's geographic coordinates.

diversity of hub areas (i.e. object size of TinyWT in Table 2) and thus adopts a loose criterion for predicted blob area. The latter rule is based on the fact that the shortest distance between the centers of the two ground truth wind turbines is ∼14 pixels, so a predicted blob with the location error of at most 7 pixels can be accepted. We calculate the true/false positives/negatives using a greedy one-to-one matching strategy.

5.2. Performance Comparison

Table 3 presents the results of both CNN-based and Transformer-based baselines on TinyWT. With the prosperous development of Transformer-based approaches, we here select the recently popular frameworks SegFormer [62] and UperNet [61] with Swin Transformer [37] as our baselines. For SegFormer, we experiment with the backbones MiT-B2, MiT-B3, and MiT-B5. For UperNet, Swin-T, Swin-S, and Swin-B are adopted. As we can see, the simplest Transformer-based models (MiT-B2 and Swin-T) have similar mIoU values as those of CNN-based baselines, but their precision, recall, and accuracy results outperform those of them significantly. This observation also indicates the necessity of including additional metrics to comprehensively evaluate models for the tiny object detection task. Furthermore, we incorporate the Transformer-based baselines with CSC and SCL to demonstrate the efficacy of our proposed modules (at the gray rows). It is obvious that our special design robustly improves the performance of tiny object detection, without adding any computational cost during inference. Specifically, the significance of our design is more salient when the base model size is smaller.

5.3. Ablation Studies

Next we conduct ablation studies to further validate our proposed framework using TinyWT. For the sake of both memory usage conservation as well as performance illustration, we use Swin-T as the encoder for UperNet throughout the rest of the experiments.

Effectiveness of a single module. Table 4 shows the performance improvement by applying CSC and SCL individually or collaboratively. Our modified pixel-level contrastive learning in SCL substantially contributes to performance improvement. For a better generalization without available geographic coordinates, we also take local pixel locations within a patch as the spatial information in CSC. Even though geographic coordinates enable the model to achieve a better learning capacity across the patches, we observe that local pixel locations are still able to enrich the overall performance, demonstrating the universality and effectiveness of CSC.

Train	Q1	Q2	Q3	Q4	Q1∼Q2	Q1∼Q3	Q1∼Q4
mIoU	63.11	68.38	64.89	68.15	73.61	77.64	81.07

Table 5. Effects of partial training. "Train" refers to the partial training data (from different quarters), and "mIoU" is calculated on the whole-year validation set.

Effects of partial-training. To probe into the relationship between training data distribution and model performance, i.e., how the model infers if trained on partial data, we explicitly train our model on part of the four quarters and test on a whole-year scope. From Table 5 we see that the result experiences a monotonic increment accordingly when trained on data varying from only a single quarter to the entire four quarters. Aside from this observation, we see better performance accomplished when trained on Q2 and Q4 because there are more images collected in these two quarters according to Fig. 3a.

6. Conclusions and Discussion

In this paper, we present TinyWT, a large-scale and year-round tiny wind turbine dataset of satellite images. It covers 474 wind turbine farm regions and explicitly extracts each region's pixel-level geographic coordinates for researchers without domain knowledge. Furthermore, TinyWT is readily available for academic as well as commercial purposes. In order to appropriately tackle this tiny object detection task, we reposition it as a localizing-and-counting problem and leverage segmentation methods to achieve benchmark performance. To pursue further improvement, we propose a novel design with two modules to emphasize contextual similarity constraints and supervised contrastive learning in tandem. The experiment results demonstrate the robust effectiveness of our design on different Transformer-based segmentation models.

For more generalization of our dataset, we also provide reference detection results in the supplemental material using the recent Transformer-based detection framework DINO [67]. Compared with this version of TinyWT, we future work will include more countries and regions to cover broader geographical distributions. Regarding the negative societal impact, if TinyWT is used for estimating the power generation of wind farms, any false predictions (false positives or false negatives) might affect the accuracy of such estimations.

References

[1] Kumar Ayush, Burak Uzkent, Chenlin Meng, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Geography-aware self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10181–10190, 2021. 3, 6

[2] Howard Butler, Martin Daly, Allan Doyle, Sean Gillies, Stefan Hagen, and Tim Schaub. The geojson format. Technical report, 2016. 1

[3] Jianqi Chen, Keyan Chen, Hao Chen, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Contrastive learning for fine-grained ship classification in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022. 3, 6

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 3

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3, 7

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[7] Creative Commons. About the licenses. https://creativecommons.org/licenses/, 2014. Accessed: 2022-11-01. 4

[8] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. *Advances in neural information processing systems*, 29, 2016. 3

[9] 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. 3

[10] Matthias Drusch, Umberto Del Bello, Sébastien Carlier, Olivier Colin, Veronica Fernandez, Ferran Gascon, Bianca Hoersch, Claudia Isola, Paolo Laberinti, Philippe Martimort, et al. Sentinel-2: Esa's optical high-resolution mission for gmes operational services. *Remote sensing of Environment*, 120:25–36, 2012. 2

[11] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 1

[12] Sebastian Dunnett, Alessandro Sorichetta, Gail Taylor, and Felix Eigenbrod. Harmonised global datasets of wind and solar farm locations and power. *Scientific data*, 7(1):1–12, 2020. 1, 2, 3, 4

[13] ESRI. Esri shapefile technical description. https://www.esri.com/content/dam/esrisites/sitecore-archive/Files/Pdfs/library/whitepapers/pdfs/shapefile.pdf, 1998. Accessed: 2022-11-01. 1

[14] Junyu Gao, Qi Wang, and Xuelong Li. Pcc net: Perspective crowd counting via spatial convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3486–3498, 2019. 2, 3

[15] Noel Gorelick, Matt Hancher, Mike Dixon, Simon Ilyushchenko, David Thau, and Rebecca Moore. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote sensing of Environment*, 202:18–27, 2017. 2

[16] Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive computing*, 7(4):12–18, 2008. 3

[17] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2786–2795, 2021. 2

[18] Tao Han, Junyu Gao, Yuan Yuan, and Qi Wang. Focus on semantic consistency for cross-domain crowd understanding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1848–1852. IEEE, 2020. 2, 3

[19] Thorsten Hoeser, Stefanie Feuerstein, and Claudia Kuenzer. Deepowt: A global offshore wind turbine data set derived with deep learning from sentinel-1 data. *Earth System Science Data*, 14(9):4251–4270, 2022. 1

[20] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. In *Proceedings of the IEEE international conference on computer vision*, pages 4145–4153, 2017. 1, 2, 4

[21] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013. 2

[22] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–546, 2018. 1, 2, 3, 5

[23] Heechul Jung, Yoonju Oh, Seongho Jeong, Chaehyeon Lee, and Taegyun Jeon. Contrastive self-supervised learning with smoothed representation for remote sensing. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 3, 6

[24] Kaggle. Airbus ship detection challenge. https://www.kaggle.com/c/airbus-ship-detection, 2018. Accessed: 2022-11-01. 2

[25] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 3, 6

[26] Mate Kisantal, Zbigniew Wojna, Jakub Murawski, Jacek Naruniec, and Kyunghyun Cho. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296*, 2019. 2

[27] Darius Lam, Richard Kuzma, Kevin McGee, Samuel Dooley, Michael Laielli, Matthew Klaric, Yaroslav Bulatov, and Brendan McCord. xview: Objects in context in overhead imagery. *arXiv preprint arXiv:1802.07856*, 2018. 2

[28] Issam H Laradji, Negar Rostamzadeh, Pedro O Pinheiro, David Vazquez, and Mark Schmidt. Where are the blobs: Counting by localization with point supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 547–562, 2018. 3, 5

[29] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 3

[30] Hoàng-Ân Lê, Heng Zhang, Minh-Tan Pham, and Sébastien Lefèvre. Mutual guidance meets supervised contrastive learning: Vehicle detection in remote sensing images. *Remote Sensing*, 14(15):3689, 2022. 3

[31] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 2

[32] Chunggi Lee, Seonwook Park, Heon Song, Jeongun Ryu, Sanghoon Kim, Haejoon Kim, Sérgio Pereira, and Donggeun Yoo. Interactive multi-class tiny-object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14136–14145, 2022. 2, 3

[33] Ke Li, Gang Wan, Gong Cheng, Liqiu Meng, and Junwei Han. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159:296–307, 2020. 1, 2, 3, 4

[34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[35] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018. 3

[36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 3

[37] 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. 3, 8

[38] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *International conference on pattern recognition applications and methods*, volume 2, pages 324–331. SciTePress, 2017. 1, 2, 4

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[40] Zheng Ma, Lei Yu, and Antoni B Chan. Small instance detection by integer programming on object density maps. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3689–3697, 2015. 2

[41] Nicolas Mandroux, Tristan Dagobert, Sébastien Drouyer, and R Grompone Von Gioi. Wind turbine detection on sentinel-2 images. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 4888–4891. IEEE, 2021. 5

[42] Nicolas Mandroux, Tristan Dagobert, Sébastien Drouyer, and Rafael Grompone von Gioi. Single date wind turbine detection on sentinel-2 optical images. *Image Processing On Line*, 12:198–217, 2022. 5

[43] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 7

[44] T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *European conference on computer vision*, pages 785–800. Springer, 2016. 1, 2, 4

[45] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006. 3

[46] J Pang, C Li, J Shi, Z Xu, and H Feng. R2-cnn: Fast tiny object detection in large-scale remote sensing images. arxiv 2019. *arXiv preprint arXiv:1902.06042*. 3

[47] QGIS Development Team. *QGIS Geographic Information System*. Open Source Geospatial Foundation, 2009. 3

[48] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021. 1, 2

[49] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 3

[50] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[51] David Rolnick, Priya L Donti, Lynn H Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, Nikola Milojevic-Dupont, Natasha Jaques, Anna Waldman-Brown, et al. Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55(2):1–96, 2022. 2

[52] Vishwanath Sindagi, Rajeev Yasarla, and Vishal MM Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2

[53] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021. 3

[54] Xian Sun, Peijin Wang, Zhiyuan Yan, Feng Xu, Ruiping Wang, Wenhui Diao, Jin Chen, Jihao Li, Yingchao Feng, Tao Xu, et al. Fair1m: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:116–130, 2022. 1, 4

[55] Jinwang Wang, Wen Yang, Haowen Guo, Ruixiang Zhang, and Gui-Song Xia. Tiny object detection in aerial images. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3791–3798. IEEE, 2021. 2, 3, 4, 5

[56] Qian Wang and Toby P Breckon. Crowd counting via segmentation guided attention networks and curriculum loss. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 2, 3

[57] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2141–2149, 2020. 1, 2

[58] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. Self-supervised learning in remote sensing: A review. *arXiv preprint arXiv:2206.13188*, 2022. 3

[59] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 1

[60] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3974–3983, 2018. 1, 2, 3, 4

[61] 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. 3, 7, 8

[62] 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. 3, 7, 8

[63] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Querydet: Cascaded sparse query for accelerating high-resolution small object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13668–13677, 2022. 2

[64] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 2

[65] Zhengeng Yang, Hongshan Yu, Mingtao Feng, Wei Sun, Xuefei Lin, Mingui Sun, Zhi-Hong Mao, and Ajmal Mian. Small object augmentation of urban scenes for real-time semantic segmentation. *IEEE Transactions on Image Processing*, 29:5175–5190, 2020. 3

[66] Xuehui Yu, Yuqi Gong, Nan Jiang, Qixiang Ye, and Zhenjun Han. Scale match for tiny person detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1257–1265, 2020. 1, 2, 3, 4

[67] 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. 8

[68] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017. 1, 2

[69] Shifeng Zhang, Yiliang Xie, Jun Wan, Hansheng Xia, Stan Z Li, and Guodong Guo. Widerperson: A diverse dataset for dense pedestrian detection in the wild. *IEEE Transactions on Multimedia*, 22(2):380–393, 2019. 1, 2

[70] Ting Zhang, Bo Tian, Dhritiraj Sengupta, Lei Zhang, and Yali Si. Global offshore wind turbine dataset. *Scientific Data*, 8(1):1–12, 2021. 1

[71] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016. 1, 2

[72] 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. 3, 7

[73] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 3

[74] Haigang Zhu, Xiaogang Chen, Weiqun Dai, Kun Fu, Qixiang Ye, and Jianbin Jiao. Orientation robust object detection in aerial images using deep convolutional neural network. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 3735–3739. IEEE, 2015. 1, 2, 4

[75] Pengfei Zhu, Longyin Wen, Xiao Bian, Haibin Ling, and Qinghua Hu. Vision meets drones: A challenge. *arXiv preprint arXiv:1804.07437*, 2018. 2

[76] Yabin Zhu, Chenglong Li, Yao Liu, Xiao Wang, Jin Tang, Bin Luo, and Zhixiang Huang. Tiny object tracking: A large-scale dataset and a baseline. *arXiv preprint arXiv:2202.05659*, 2022. 2