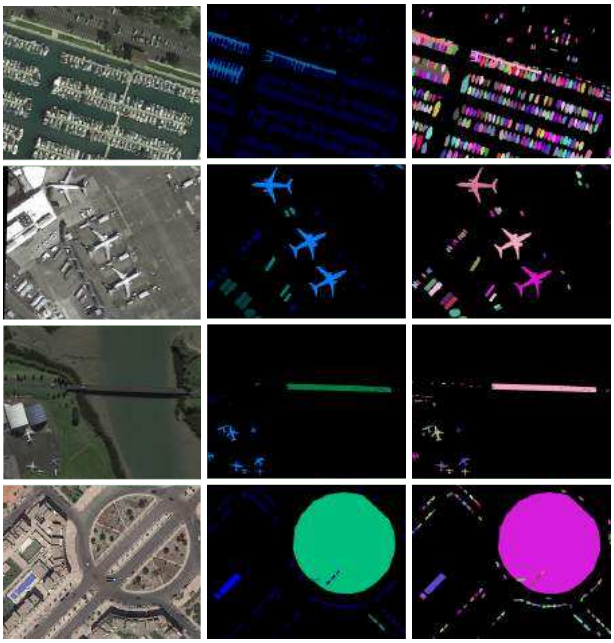
iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images

Syed Waqas Zamir[1,*] Aditya Arora[1,*] Akshita Gupta[1] Salman Khan[1] Guolei Sun[1]
Fahad Shahbaz Khan[1] Fan Zhu[1] Ling Shao[1] Gui-Song Xia[2] Xiang Bai[3]
[1]Inception Institute of Artificial Intelligence, UAE, [2]Wuhan University, China
[3]Huazhong University of Science and Technology, China
[1]firstname.lastname@inceptioniai.org
[2]guisong.xia@whu.edu.cn, [3]xbai@hust.edu.cn

Abstract

Existing Earth Vision datasets are either suitable for semantic segmentation or object detection. In this work, we introduce the first benchmark dataset for instance segmentation in aerial imagery that combines instance-level object detection and pixel-level segmentation tasks. In comparison to instance segmentation in natural scenes, aerial images present unique challenges e.g., a huge number of instances per image, large object-scale variations and abundant tiny objects. Our large-scale and densely annotated Instance Segmentation in Aerial Images Dataset (iSAID) comes with 655,451 object instances for 15 categories across 2,806 high-resolution images. Such precise per-pixel annotations for each instance ensure accurate localization that is essential for detailed scene analysis. Compared to existing small-scale aerial image based instance segmentation datasets, iSAID contains 15× the number of object categories and 5× the number of instances. We benchmark our dataset using two popular instance segmentation approaches for natural images, namely Mask R-CNN and PANet. In our experiments we show that direct application of off-the-shelf Mask R-CNN and PANet on aerial images provide suboptimal instance segmentation results, thus requiring specialized solutions from the research community.



(a) Original image (b) SS maps (c) IS maps

Figure 1: Some typical examples from iSAID containing objects with high density, arbitrary shapes and orientation, large aspect ratios and huge scale variation. SS and IS denote semantic segmentation and instance segmentation, respectively.

1. Introduction

Given an image, the aim of instance segmentation is to predict category labels of all objects of interest and localize them using pixel-level masks. Large-scale datasets such as ImageNet [7], PASCAL-VOC [8], MSCOCO [17], Cityscapes [6] and ADE20K [34] contain natural scenes in which objects appear with upward orientation. These datasets enabled deep convolutional neural networks (CNN), that are data hungry in nature [14], to show unprecedented performance in scene understanding tasks such as image classification [29, 11, 30], object detection [27, 26, 19], semantic labeling and instance segmentation [10, 18, 4]. However, the algorithms developed to solve these tasks in regular images do not transfer well to overhead (aerial) imagery. In aerial images, objects occur in high density (Fig. 1, row 1), arbitrary shapes and orientation (Fig. 1, row 2), large aspect ratios (Fig. 1, row 3), and with huge scale variation (Fig. 1, row 4). To accurately address

*Equal contribution

the challenges of aerial images for high-level vision tasks, tailor-made solutions on appropriate datasets are desired.

To encourage the advancements in aerial imagery for earth observation, a few well-annotated datasets for object detection [16, 32] and semantic labeling [21, 9] have recently been introduced. However, they do not provide per-pixel accurate labelings for each object instance in an aerial image and are therefore unsuitable for instance segmentation task (see Table 1). Publicly available instance segmentation datasets [1, 31] typically focus on a single object category; for example, [31] only contains building footprints and [1] only has labelings for ships. To address the shortcomings of these existing datasets, we introduce a large-scale Instance Segmentation in Aerial Images Dataset (iSAID). Our dataset contains annotations for an enormous 655,451 instances of 15 categories in 2,806 high-resolution images. Having such large number of instances and class count makes iSAID suitable for real-world applications in complicated aerial scenes.

Compared to other aerial datasets for instance segmentation, iSAID is far more diverse, comprehensive and challenging. It exhibits the following distinctive characteristics: **(a)** large number of images with high spatial resolution, **(b)** fifteen important and commonly occurring categories, **(c)** large number of instances per category, **(d)** large count of labelled instances per image, which might help in learning contextual information, **(e)** huge object scale variation, containing small, medium and large objects, often within the same image, **(f)** Imbalanced and uneven distribution of objects with varying orientation within images, depicting real-life aerial conditions, **(g)** several small size objects, with ambiguous appearance, can only be resolved with contextual reasoning, **(h)** precise instance-level annotations carried out by professional annotators, cross-checked and validated by expert annotators complying with well-defined guidelines.

2. Related Work

Both in terms of historical context and recent times, large-scale datasets have played a key role in progressing the state-of-the-art for scene understanding tasks such as image classification [29, 11, 30], scene recognition [33], object detection [27, 26, 19] and segmentation [10, 18, 4]. For instance, ImageNet [7] is one of the most popular large-scale dataset for image classification task, on which the state-of-the-art methods [29, 11, 30] are able to reach human-level performance. Similarly, large-scale annotated datasets, such as MSCOCO [17], Cityscapes [6] and ADE20K [34] for object detection, semantic and instance segmentation have driven the development of exciting new solutions for natural scenes. Introduction of datasets, that are larger in scale and diversity, not only provide room for new applications but also set new research directions. Moreover, challenging datasets push research community to develop more sophisticated and robust algorithms; thus enabling their application in real-world scenarios.

There are numerous lucrative application areas of Earth Vision research, including security and surveillance [28], urban planning [22], precision agriculture, land type classification [2] and change detection [15]. In general, deep-learning based algorithms show excellent performance when provided with large-scale datasets, as demonstrated for several high-level vision tasks [11, 10, 26] involving conventional large-scale image datasets. A key limitation towards building solutions for the Earth Vision applications is the unavailability of aerial datasets resembling the scale and diversity of natural-scene datasets (*e.g.* ImageNet [7] and MSCOCO [17]). Specifically, existing overhead imagery datasets are significantly lagging in terms of category count, instance count and the quality of annotations. The advanced off-the-shelf methods trained on conventional datasets when applied on aerial image datasets, fail to provide satisfactory results due to large domain shift, high density objects with large variations in orientation and scale. As an example, an otherwise robust object detector SSD [19] yields an mAP of just 17.84 on the dataset for object detection in aerial images (DOTA) [32]. Recently, large-scale aerial image datasets (DOTA [32] and xView [16]) have been introduced to make advancement in object detection research for earth observation and remote sensing. Both of these datasets [32, 16] are more diverse, complex, and suitable for real-world applications than previously existing aerial datasets for object detection [5, 3, 23, 35, 20]. On the down side, these datasets do not provide pixel-level masks for the annotated object instances.

Instance segmentation is a challenging problem that goes one step ahead than regular object detection as it aims to achieve precise per-pixel localization for each object instance. Unlike aerial object detection, there exist no large-scale annotated dataset for instance segmentation in aerial images. A few publicly available datasets in this domain only contain instances of just a single category (e.g., ships [1] and buildings [31]). Owing to the significance of precise localization of each instance in aerial imagery, we introduce a novel dataset, iSAID, that is significantly large, challenging, well-annotated, and offers 15$\times$ the number of object categories and 5$\times$ the number of instances than existing datasets [1, 31].

3. Dataset Details

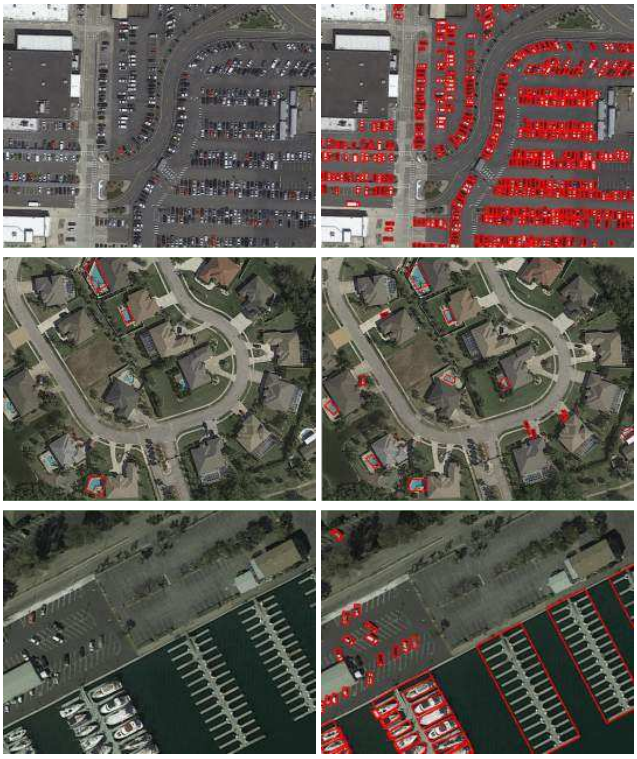
3.1. Images, Classes and Dataset Splits

In order to create a dataset for instance segmentation task, we build on the large-scale aerial image dataset: DOTA [32], that contains 2,806 images. The images are collected from multiple sensors and platforms to reduce bias.

Note that the original DOTA dataset only contains bounding box annotations for object detection, thus cannot be used for accurate instance segmentation. Furthermore, DOTA [32] suffers with several aberrations such as incorrect labels, missing instance annotations, and inaccurate bounding boxes. To avoid these issues, our dataset for instance segmentation is **independently annotated from scratch**, leading to 655,451 instances compared to 188,282 instances provided originally in DOTA [32] (a $\sim$ 250% relative increase, see Fig. 2 for examples).

It is important to note that the our instance segmentation dataset in aerial images has unique challenges compared to regular image datasets (e.g., less object details, small size and different viewpoints- see Fig. 3). On the other hand, as summarized in Table 1, most of the existing aerial image datasets are annotated with bounding boxes or point-labels that only coarsely localize the object instances. Furthermore, these datasets are often limited to a small scale with only a few object categories. In comparison, our proposed iSAID dataset provides a large number of instances, precisely marked with masks denoting their exact location in an image (Fig. 6). The two existing instance segmentation datasets for aerial imagery only comprise of a single object category (e.g., ships [1] or buildings [31]). In contrast, iSAID has a diverse range of 15 categories and much larger scale ($\sim$5$\times$ more instances).

In order to select object categories we follow the experts in overhead satellite imagery interpretation [32] and provide annotations for the following 15 classes: *plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, swimming pool* and *soccer ball field*. Objects from these categories occur frequently and are important for various real-world applications [5, 23, 35]. For dataset splits, we use half of the original images to form train set, $1/6$ images for validation set and $1/3$ for test set. Both images and ground-truth annotations for the train and validation sets will be released publicly. In the case of test set, we will publicly provide images without annotations. The test set annotations will be used to set up an evaluation server for fair comparison between the developed techniques.



(a) DOTA (b) iSAID

Figure 2: Visualization of missing annotations from DOTA [32] as compared to iSAID.

Figure 3: Ships, buses and cars from MSCOCO [17] (odd columns) and iSAID (even columns). Notice the size variation and the angle at which images are taken.

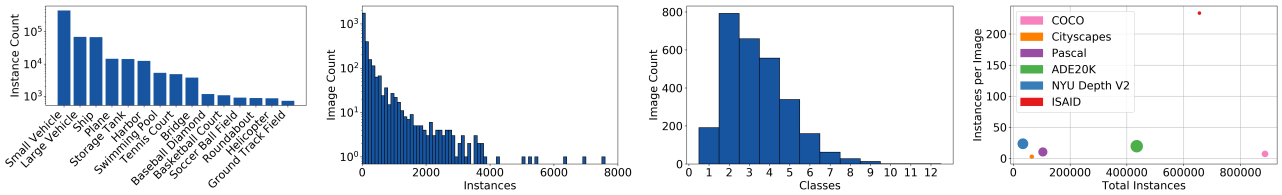
3.2. Annotation Procedure

We design a comprehensive annotation pipeline to ensure that annotations of all images are consistent, accurate and complete. The pipeline includes the following steps: developing annotation guidelines; training annotators; annotating images; quality checks and annotation refinement until satisfaction. For annotation, a high-quality in-house software named *Haibei* was used to draw instance segmentation masks on images.

In order to obtain high-quality annotations, clear and thorough guidelines for annotators are of prime importance. Taking notes from previously proposed datasets [32, 16, 17, 34, 8], we establish the following guidelines: **1)** All clearly visible objects of the above-mentioned 15 categories must be annotated; **2)** Segmentation masks for each instance should match its visual margin in the image; **3)** Images should be zoomed in or out, when necessary, to obtain annotations with refined boundaries; **4)** Cases of un-

Dataset	Bounding box	Segmentation mask	#Main categories	#Fine-grain categories	#Total categories	#Instances	#Images	Image width
NWPU VHR-10 [5]	horizontal	✗	10	✗	10	3,651	800	∼1,000
SZTAKI-INRIA [3]	oriented	✗	1	✗	1	665	9	∼800
TAS [12]	horizontal	✗	1	✗	1	1,319	30	792
COWC [23]	center-point	✗	1	✗	1	32,716	53	2,000 ∼ 19,000
VEDAI [25]	oriented	✗	3	✓	9	3,700	1,200	512, 1,024
UCAS-AOD [35]	horizontal	✗	2	✗	2	6,029	910	∼1,000
HRSC2016 [20]	oriented	✗	1	✗	1	2,976	1,061	300∼1,500
xView [16]	horizontal	✗	16	✓	60	1,000,000	1,127	700∼4,000
DOTA [32]	oriented	✗	14	✓	15	188,282	2,806	800∼13,000
Airbus Ship [1]	polygon	✓	1	✗	1	131,000	192,000	∼800
SpaceNet MVOI [31]	polygon	✓	1	✗	1	126,747	60,000	900
iSAID (Ours)	polygon	✓	14	✓	15	655,451	2,806	800∼13,000

Table 1: Comparison between Aerial Datasets. *Center-point* represents those annotations for which only the center coordinates of the instances are provided.



(a) (b) (c) (d)

Figure 4: Statistics of classes and instances in iSAID. (a) Histogram of the number of instances per class (sorted by frequency). (b) Histogram of number of instances per image. (c) Histogram of number of classes per image. (d) Number of instances vs. instances per image (comparison of our dataset with other large-scale conventional datasets). The size of the circle denotes the number of categories, *e.g.*, big circle represents the presence of large number of object categories.

clear/difficult objects should be reported to the team supervisors and then discussed to get annotations with high confidence; **5)** All work should be done at a single facility using the same software.

The images of proposed iSAID are annotated by the professional annotators. The annotators were trained through multiple sessions, even if they had prior experience in annotating datasets of any kind. During training phase, each annotator was shown both positive and negative examples containing objects from 15 categories. An assessment protocol was developed to shortlist the best annotators in the following manner: annotators were asked to annotate several sample images containing easy and difficult cases while strictly adhering to the established guidelines. The quality of annotations was crossed checked to evaluate their performance. Only those annotators who passed the test were approved to work on this particular project. In general, the selected annotators were given training for approximately 4 hours before assigning them the task of annotating actual aerial image dataset.

At the beginning of the annotation process, the supervisory team distributes different sets of images among annotators. The annotators were asked to annotate all objects belonging to 15 categories appearing in the images. Due to high spatial resolution and large number of instances, it took approximately 3.5 hours for one annotator to finish labelling all objects present in a single image, resulting in 409 man-hours (for 2,806 images) excluding cross checks and refinements.

Once the first round of annotations was completed, a five-stage quality control procedure was put in place to ensure that the annotation quality is good. **1)** The labelers were asked to examine their own annotated images and correct issues like double labels, false labels, missing objects and inaccurate boundaries. **2)** The annotators reviewed the work of other peers on rotational basis. In this stage, object masks for each class were cropped and placed in one specific directory, so that the annotation errors could be easily identified and corrected. **3)** The supervisory team randomly sampled 70% images (around 2000) and analyzed their quality. **4)** A team of experts sampled 20% images (around 500) and ensured the quality of annotations. In case of problems, the annotations were iteratively send back to the annotators for refinement until the experts were satisfied by the labels. **5)** Finally, several statistics (*e.g.*, instance areas, aspect ratios, etc.) were computed. Any outliers were

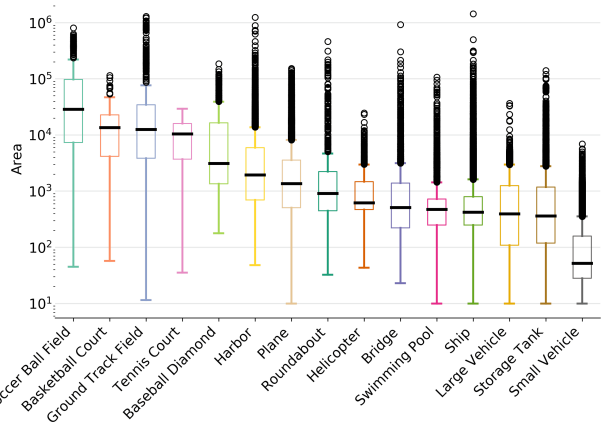


Figure 5: Boxplot depicting the range of areas for each object category. The size of objects varies greatly both among and across classes.

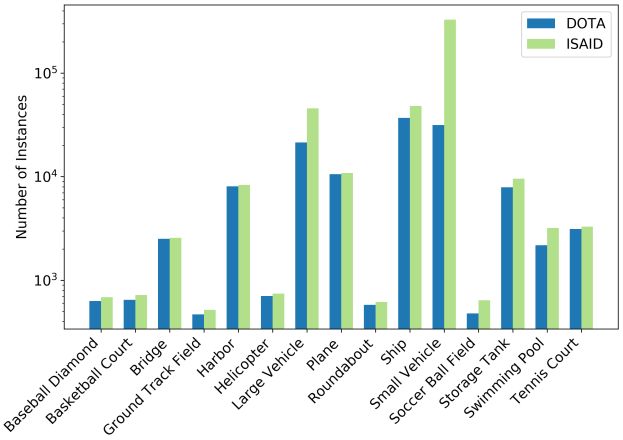


Figure 6: Comparison of DOTA [32] and our dataset (iSAID) in terms of instances per category. iSAID contains, in total, 3.5 times more number of instances than DOTA.

double checked to make sure they are indeed valid and correct annotations.

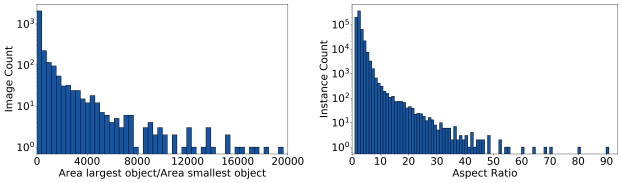
3.3. iSAID Statistics

In this section we analyze the properties of iSAID and compare it with other relevant datasets.

Image resolution. Images in natural datasets (*e.g.*, PASCAL-VOC [8], ImageNet [7]) are generally of limited dimensions, often reaching no more than 1000×1000 pixels. In contrast, aerial images have a very large resolution: for instance the width of some images in COWC [23] dataset is up to 19,000 pixels. In our dataset, the spatial resolution of images ranges from 800 to 13,000 in width. Applying off-the-shelf conventional object detection and instance segmentation methods on such high-resolution aerial images yield suboptimal results, as we shall see in the experiment section.

Instance count. Our dataset comprises 655,451 annotated instances of 15 categories. In Fig. 4a it is shown that there are some infrequent classes with significantly less number of instances than other more frequent classes. For example, small vehicle and ground track field are the most frequent and least frequent classes, respectively. Such a class imbalance usually exists in both natural and aerial imagery datasets and it is important for real-world applications [13]. Fig. 4c illustrates the image histogram in which multiple classes co-exists; on average 3.27 classes appear in each image of iSAID.

Another property, common in all aerial image datasets, is the presence of large number of object instances per image due to a large field of view. As shown in Fig. 4b, the instance count per image in our dataset can reach up to 8,000. Fig. 4d depicts that our dataset contains on average ∼239 instances per image, which is significantly higher compared



(a) (b)

Figure 7: Statistics of images and instances in iSAID. (a) Ratio between areas of largest and smallest object shows the huge variation in scale.(b) shows that instances in iSAID exhibit large variation in aspect ratio.

to traditional large-scale datasets for instance segmentation: MSCOCO [17], Cityscapes [6], PASCAL-VOC [8], ADE20K [34] and NYU Depth V2 [24] contain 7.1, 2.6, 10.3, 19.5, and 23.5 instances per image, respectively. In aerial images, the densely packed instances typically appear in scenes containing parking lots and marina.

Area of categories. In natural as well as aerial images, objects appear in various sizes. Therefore, an instance segmentation method should be flexible and efficient enough to deal with objects of small, medium and large sizes [32]. In our dataset, we consider objects in the range 10 to 144 pixels as small, 144 to 1024 pixels as medium, and 1024 and above as large. The percentage of small, medium and large objects in iSAID is 52.0, 33.7 and 9.7, respectively. The box plot in Fig. 5 presents statistics of area for each class of iSAID. It can be seen that the size of objects varies greatly both among and across classes. For instance, the ship category contains small boats covering area of 10 pixels, as well as, large vessels of sizes upto 1,436,401 pixels, depicting a huge intra-class variation. Similarly, a small vehicle can be as small as 10 pixels and a ground track field can be

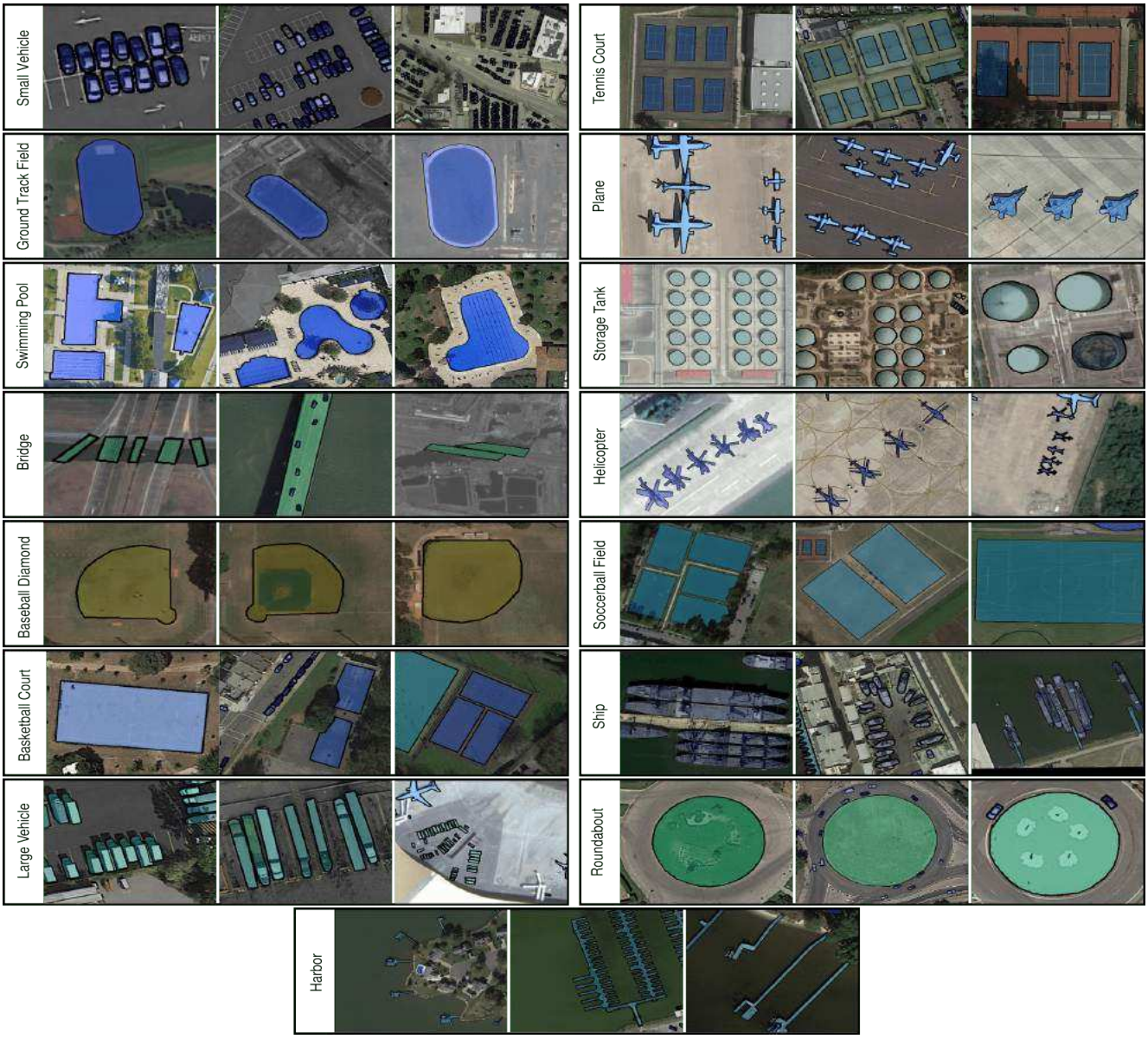


Figure 8: Samples of annotated images in iSAID.

as large as 1,297,121 pixels, illustrating immense inter-class variation. Fig. 7a shows the variation in scale when small and large objects of same or different categories appear together, which is a very common case in aerial imagery. We can notice that the ratio between the area of the largest object and the smallest object can reach up to 20,000. Such enormous scale variation poses an extreme challenge for instance segmentation methods that need to handle both tiny and very large objects, simultaneously.

Aspect ratio. In aerial images many objects occur with unusually large aspect ratios, which is not the case in traditional ground images. Fig. 7b depicts the distribution of aspect ratio for object instances in our proposed dataset. We can notice that instances exhibit huge variation in aspect ratios, reaching up to 90 (with an average of 2.4). Moreover, a large number of instances present in our dataset have a large aspect ratio.

4. Experiments

In this section, we test how general instance segmentation methods, particularly developed for regular scene datasets, perform on our newly developed aerial dataset (some sample images are shown in Fig. 8). To this end, we use MaskR-CNN [10] and PANet [18]: the former for its

Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Mask R-CNN [10]	25.65	51.30	22.72	14.46	31.26	37.71
Mask R-CNN+	33.41	56.77	34.66	35.83	46.50	23.93
PANet [18]	34.17	56.57	35.84	19.56	42.27	**46.62**
PANet+	39.54	63.59	42.22	42.14	53.61	38.50
PANet++	**40.00**	**64.54**	**42.50**	**42.46**	**54.74**	43.16

Table 2: **Instance segmentation** results using mask AP on iSAID test set. PANet [18] and its variants outperform Mask R-CNN [10] and its variants with significant margin. PANet++ with backbone ResNet-152 performs best.

Method	AP^{bb}	$AP^{bb}{}_{50}$	$AP^{bb}{}_{75}$	$AP^{bb}{}_{S}$	$AP^{bb}{}_{M}$	$AP^{bb}{}_{L}$
Mask R-CNN [10]	36.50	59.06	41.27	26.16	43.10	43.32
Mask R-CNN+	37.18	60.79	40.67	39.84	43.72	16.01
PANet [18]	41.66	60.94	46.62	26.92	47.81	**50.95**
PANet+	46.31	66.90	51.68	48.92	53.33	26.52
PANet++	**47.0**	**68.06**	**52.37**	**49.48**	**55.07**	27.97

Table 3: **Object detection** results using bounding box AP on iSAID test set. Similar to instance segmentation case, PANet [18] and its variants generate better results than Mask-RCNN and its variants.

Method	AP	AP_{50}	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC
Mask R-CNN [10]	25.7	51.3	37.7	42.5	13.0	23.6	6.9	7.4	26.6	54.9	34.6	28.0	20.8	35.9	22.5	25.1	5.3
Mask R-CNN+	33.4	56.8	41.7	39.6	15.2	25.9	16.9	30.4	48.8	72.9	43.1	32.0	26.7	36.0	29.6	36.7	5.6
PANet	34.2	56.8	39.2	45.5	15.1	29.3	15.0	28.8	45.9	74.1	47.4	29.6	33.9	36.9	26.3	36.1	9.5
PANet++	**40.0**	**64.6**	**48.7**	**50.3**	**18.9**	**32.5**	**20.4**	**34.4**	**56.5**	**78.4**	**52.3**	**35.4**	**38.8**	**40.2**	**35.8**	**42.5**	**13.7**

Table 4: **Class-wise instance segmentation** results on iSAID test set. Note that short names are used to define categories: BD-Baseball diamond, GTF-Ground field track, SV-Small vehicle, LV-Large vehicle TC-Tennis court, BC-Basketball court, SC-Storage tank, SBF-Soccer-ball field, RA-Roundabout, SP-Swimming pool, and HC-Helicopter.

popularity as a meta algorithm and the latter for its state-of-the-art results. Additionally, we make simple modifications in the baseline models and report the results of these variants. For evaluation, we use the standard COCO metrics: AP (averaged over IoU threshold), AP_{50}, AP_{75}, AP_S, AP_M and AP_L, where S, M and L represent small (area: 10-144 pixels), medium (area:144 to 1024 pixels) and large objects (area:1024 and above), respectively.

Implementation Details. Images with large resolution (*e.g.* 4000 pixels in width) are commonly present in iSAID. The baseline methods [10, 18] cannot handle images with such unusually large spatial dimension. Therefore, we opt to train and test the baseline methods on the patches of size 800×800 extracted from the full resolution images with a stride set to 200. In order to train baseline Mask R-CNN and PANet models, we use the same hyper-parameters as in the original papers [10, 18]. In the training phase, the cropped patches are re-scaled with shorter edges as 800 pixels and longer edges as 1400 pixels. During the cropping process, some objects may get cut. we then generate new annotations for the patches with updated segmentation masks. We use mini-batch size of 16 for training. Our models are trained on 8 GPUs for 180k iterations with an initial learning rate of 0.025, that is decreased by a factor of 10 at 90k iteration. We use weight decay of 0.0001 and momentum of 0.9.

In an effort to benchmark the proposed dataset, we consider the original Mask R-CNN [10] and PANet [18] as our baseline models, both using ResNet101-FPN as backbone. We do not change any hyper-parameter settings in the baseline models. On top of these baselines, we make three minor modifications to develop Mask R-CNN+ and PANet+: **(a)** Since, large number of objects are present per image, we consider the number of detection boxes to be 1000 (instead of 100 considered by default in the baselines) during evaluation. **(b)** As high scale variation exists within aerial images, we use scale augmentations at six scales (1200,1000,800,600,400). In comparison, the baseline considers a single scale of 800 pixels (shorter side). **(c)** An NMS (non-maximal suppression) threshold of 0.6 is used instead of the 0.5 used for baseline. Lastly, for our best model (PANet), we also try a heavier backbone (ResNet-152-FPN) that results in the top performing models for instance segmentation and bounding box detection. We term this model as PANet++. Note that the modifications in baselines are minor, and we expect that more sophisticated algorithmic choices might significantly improve the results.

4.1. Results

In Table 2, we report the results achieved by baselines (Mask R-CNN [10] and PANet [18]) and their variants for the instance segmentation task. It can be seen that the PANet [18] with its default parameters outperforms the Mask R-CNN [10] on iSAID. This trend is similar to the performance of these baselines on the MSCOCO dataset for instance segmentation in regular ground images. Moreover, by making minor modifications in baselines to make them suitable for aerial images, we were able to obtain marginal improvements e.g., an absolute increment of 7.8 AP with Mask R-CNN+ over baseline [10]. The best performance is achieved by PANet++ which uses a stronger ResNet-152-FPN backbone. To study the performance trend for different classes, we also report class-wise AP in Table 4. Notably, in the case of PANet++, we observe a significant performance gain of $\sim$5 points or more in AP_{50} for some categories such as baseball diamond, basketball court and harbour.

In addition to instance segmentation masks, we also

Method	AP^{bb}	AP^{bb}_{50}	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC
Mask R-CNN [10]	36.6	59.1	57.8	44.7	19.7	36.4	17.9	31.7	46.9	70.2	42.7	31.4	25.4	36.4	41.0	36.2	21.9
Mask R-CNN+	37.2	60.8	58.5	38.5	18.6	32.7	20.8	36.8	51.4	72.9	43.1	32.0	26.7	36.0	29.6	48.8	29.6
PANet	41.7	61.0	62.8	47.5	19.3	44.3	18.3	35.0	50.3	77.4	48.5	30.9	35.3	40.4	46.6	40.4	27.9
PANet++	**47.0**	**68.1**	**68.1**	**51.0**	**23.4**	**44.2**	**27.3**	**42.1**	**61.9**	**79.4**	**53.8**	**38.1**	**39.1**	**43.4**	**53.6**	**47.1**	**32.4**

Table 5: **Class-wise object detection** results on iSAID test set. The same short names for categories are used as in Table 4.

(a) Ground Truth (b) Mask R-CNN [10] (c) Mask R-CNN+ (d) PANet [18] (e) PANet++

Figure 9: Visual results on images from test set of iSAID. It can be noticed that the original Mask R-CNN [10] yields the least accurate results, with missing object instances. Whereas, PANet++ produces significantly better results compared to its original counter part [18], as well as Mask R-CNN and Mask R-CNN+.

compute bounding-box object detection results, as reported in Tables 3 and 5. In this experiment, the horizontal bounding-boxes are considered. For object detection, we observe similar trends in methods' ranking as they were for instance segmentation. It is important to note that our results are inferior to those reported in [32], possibly due to the large number of newly introduced object instances in iSAID (655,451 vs 188,282 in DOTA).

Qualitative results for instance segmentation are shown in Fig. 9. The results are shown for Mask R-CNN and PANet baselines and their modified versions. We note that with simple modifcations to these strong baselines, we were able to significantly improve on extreme sized objects (both very small and large objects). As expected from the quantitative results, the PANet++ achieves most convincing qualitative results with accurate instance masks among the other evaluated models.

5. Conclusion

Delineating each object instance in aerial images is a practically significant and a scientifically challenging problem. The progress in this area has been limited due to the lack of large-scale, densely annotated satellite image dataset with accurate instance masks. To bridge this gap, we propose a new instance segmentation dataset which encompasses 15 object categories and 655,451 instances in total. We extensively benchmark the dataset on instance segmentation and object detection tasks. Our results show that the aerial imagery pose new challenges to existing instance segmentation algorithms such as a large number of objects per image, limited appearance details, several small objects, significant scale variations among the different object types and a high class imbalance. We hope that our contribution will lead to new developments on the instance segmentation task in aerial imagery.

References

[1] Dataset for airbus ship dectection challenge. https://www.kaggle.com/c/airbus-ship-detection/, 2018. [Online; accessed 27-May-2019]. 2, 3, 4

[2] R. M. Anwer, F. S. Khan, J. van de Weijer, M. Molinier, and J. Laaksonen. Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS journal of photogrammetry and remote sensing*, 138:74–85, 2018. 2

[3] C. Benedek, X. Descombes, and J. Zerubia. Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics. *TPAMI*, 34(1):33–50, 2012. 2, 4

[4] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1, 2

[5] G. Cheng, P. Zhou, and J. Han. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415, 2016. 2, 3, 4

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 1, 2, 5

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 2, 5

[8] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010. 1, 3, 5

[9] H. Goldberg, M. Brown, and S. Wang. A benchmark for building footprint classification using orthorectified RGB imagery and digital surface models from commercial satellites. In *Proceedings of IEEE Applied Imagery Pattern Recognition Workshop*, 2017. 2

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2, 6, 7, 8

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 2

[12] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *EECV*, 2008. 4

[13] S. Khan, M. Hayat, S. W. Zamir, J. Shen, and L. Shao. Striking the right balance with uncertainty. 2019. 5

[14] S. Khan, H. Rahmani, S. A. A. Shah, and M. Bennamoun. A guide to convolutional neural networks for computer vision. *Synthesis Lectures on Computer Vision*, 8(1):1–207, 2018. 1

[15] S. H. Khan, X. He, F. Porikli, and M. Bennamoun. Forest change detection in incomplete satellite images with deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(9):5407–5423, 2017. 2

[16] D. Lam, R. Kuzma, K. McGee, S. Dooley, M. Laielli, M. Klaric, Y. Bulatov, and B. McCord. xView: Objects in context in overhead imagery. In *arXiv*, 2018. 2, 3, 4

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 2, 3, 5

[18] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018. 1, 2, 6, 7, 8

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016. 1, 2

[20] Z. Liu, H. Wang, L. Weng, and Y. Yang. Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds. *IEEE Geoscience and Remote Sensing Letters*, 13(8):1074–1078, 2016. 2, 4

[21] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2017. 2

[22] D. Marcos, M. Volpi, B. Kellenberger, and D. Tuia. Land cover mapping at very high resolution with rotation equivariant cnns: Towards small yet accurate models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:96–107, 2018. 2

[23] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *ECCV*, 2016. 2, 3, 4, 5

[24] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. 5

[25] S. Razakarivony and F. Jurie. Vehicle detection in aerial imagery: A small target detection benchmark. *Journal of Visual Communication and Image Representation*, 34:187–203, 2016. 4

[26] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. In *CVPR*, 2017. 1, 2

[27] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPs*, 2015. 1, 2

[28] C. Santamaria, M. Alvarez, H. Greidanus, V. Syrris, P. Soille, and P. Argentieri. Mass processing of sentinel-1 images for maritime surveillance. *Remote Sensing*, 9(7):678, 2017. 2

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2014. 1, 2

[30] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *AAAI*, 2017. 1, 2

[31] N. Weir, D. Lindenbaum, A. Bastidas, A. Van Etten, S. McPherson, J. Shermeyer, V. Kumar, and H. Tang. SpaceNet MVOI: a multi-view overhead imagery dataset. *arXiv*, 2019. 2, 3, 4

[32] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang. DOTA: A large-scale dataset for object detection in aerial images. In *CVPR*, 2018. 2, 3, 4, 5, 8

[33] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 2

[34] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 1, 2, 3, 5

[35] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao. Orientation robust object detection in aerial images using deep convolutional neural network. In *ICIP*, 2015. 2, 3, 4