LUAI Challenge 2021 on Learning to Understand Aerial Images

Gui-Song Xia\(^1\), Jian Ding\(^1\), Ming Qian\(^1\), Nan Xue\(^1\), Jiaming Han\(^1\), Xiang Bai\(^2\), Michael Ying Yang\(^3\), Shengyang Li\(^4\), Serge Belongie\(^5\), Jiebo Luo\(^6\), Mihai Datcu\(^7,8\), Marcello Pelillo\(^9\), Liangpei Zhang\(^1\), Qiang Zhou\(^10\), Chao-hui Yu\(^10\), Kaixuan Hu\(^11\), Yingjia Bu\(^11\), Wenning Tan\(^11\), Zhe Yang\(^10\), Wei Li\(^10\), Shang Liu\(^10\), Jiaxuan Zhao\(^12\), Tianzhi Ma\(^12\), Zi-han Gao\(^12\), Lingqi Wang\(^12\), Yi Zuo\(^12\), Licheng Jiao\(^12\), Chang Meng\(^12\), Hao Wang\(^12\), Jiahao Wang\(^12\), Yiming Hui\(^12\), Zhuojun Dong\(^12\), Jie Zhang\(^12\), Qianyue Bao\(^12\), Zixiao Zhang\(^12\), Fang Liu\(^12\)

\(^1\)Wuhan University, China.
\(^2\)Huazhong University of Science and Technology, China.
\(^3\)University of Twente, Netherlands.
\(^4\)Chinese Academy of Sciences, Beijing, China.
\(^5\)Cornell Tech and Cornell University, United States.
\(^6\)University of Rochester, United States.
\(^7\)German Aerospace Center (DLR), Germany.
\(^8\)University POLITEHNICA of Bucharest, Romania.
\(^9\)Computer Science, Ca' Foscari University of Venice, Italy.
\(^10\)Alibaba Group, China.
\(^11\)Hikvision Research Institute
\(^12\)Xidian University, China.

Abstract

This report summarizes the results of Learning to Understand Aerial Images (LUAI) 2021 challenge held on ICCV’2021, which focuses on object detection and semantic segmentation in aerial images. Using DOTA-v2.0\(^7\) and GID-15\(^35\) datasets, this challenge proposes three tasks for oriented object detection, horizontal object detection, and semantic segmentation of common categories in aerial images. This challenge received a total of 146 registrations on the three tasks. Through the challenge, we hope to draw attention from a wide range of communities and call for more efforts on the problems of learning to understand aerial images.

1. Introduction

Earth vision, also known as Earth Observation and Remote Sensing, targets to understand large-scale scenes on the Earth’s surface with aerial images taken from overhead view, which provides a new way to understand our physical world and benefits many applications, e.g., urban management and planning, precise agriculture, emergency rescue and disaster relief.

Recently, the problems in Earth vision span from plane detection\(^27\), ship detection\(^22\), vehicle detection\(^19\), building extraction\(^31\), road extraction\(^5\), and changing detection\(^11\), etc. Most of these applications can be considered as special cases of object detection\(^33, 8, 12, 29, 18, 36\), semantic segmentation\(^35\), and instance segmentation\(^24, 30\) of aerial images.

To advance methods for learning to understand aerial images, we propose a competition that focuses on object detection and semantic segmentation for common categories on an ICCV workshop\(^1\). As known to all, large-scale, well-annotated, and diverse datasets are crucial for learning-based algorithms. Therefore, for object detection in aerial images, this competition uses a new large-scale benchmark database, i.e., DOTA-v2.0\(^7\)\(^2\). The dataset contains 11,268 large-scale images (the maximum size is 20,000), 18 categories, and 1,793,658 instances, each of which is labeled by an arbitrary (8 d.o.f.) oriented bounding box. It is worth mentioning that this database is the same as the DOTA-v1.0 database, but with more annotations and categories.

\(^1\)https://captain-whu.github.io/LUAI2021/challenge.html
\(^2\)https://captain-whu.github.io/DOTA
noticing that the competition tasks on object detection follows two previous object detection in aerial images (ODAI) challenges held on ICPR’2018 [8] and CVPR’2019 [1], using DOTA-v1.0 [33] and DOTA-v1.5 [1], respectively. For semantic segmentation, we propose a new dataset, named as Gaofen Image Dataset with 15 categories (GID-15 [35]\(^3\)). GID-15 contains 150 pixel-level annotated GF-2 images in 15 categories. Based on the two datasets, we propose three competition tasks, namely oriented object detection, horizontal object detection, and semantic segmentation.

Through the dataset and competition tasks, we aim to draw attention from a wide range of communities and call for more future research and efforts on the problems of object detection and semantic segmentation in aerial images. We believe the competition and workshop will not only promote the development of algorithms in Earth vision, but also pose interesting algorithmic questions to the computer vision community.

## 2. Datasets

In this section, we present the statics and properties of the two datasets for object detection and semantic segmentation respectively.

### 2.1. Object Detection

The object detection track is based on DOTA-v2.0 [7], which collects images from Google Earth, GF-2 Satellite, and aerial images. The domain shifts among these different image sources make it possible to develop domain-robust object detectors, which also have practical value. There are 18 common categories, 11,268 images and 1,793,658 instances in DOTA-v2.0. All the objects in DOTA-v2.0 are annotated with oriented bounding boxes. These images and oriented annotations provide a rich resource for researches on rotation-invariant object detection. DOTA-v2.0 are split into training (1,830 images), validation (593 images), test-dev (2792 images), and test-challenge (6053 images). The evaluation in this challenge is on the test-challenge. We do not release the ground truths. The scores can be obtained by submitting results to the evaluation server. The annotated examples for can be found in Fig. 1.

### 2.2. Semantic Segmentation

The semantic segmentation track is based on Gaofen Image Dataset with 15 categories (GID-15) [35], which is a new large-scale land-cover dataset. GID-15 has many superiorities over the existing land-cover dataset owing to its large coverage, wide distribution, and high spatial resolution. The large-scale remote sensing semantic segmentation dataset contains 150 pixel-level annotated GF-2 images in 15 categories. The image sizes of the GF-2 images are 7, 200 × 6, 800. The examples of annotated images in GID-15 can be found in Fig. 2.

## 3. Challenge Tasks

Using the above mentioned DOTA-v2.0 [7] and GID-15, we propose three tasks, namely oriented object detection, horizontal object detection, and semantic segmentation. In what follows, we provide the details of outputs and evaluation metrics of each task.

### 3.1. Task1 - Oriented Object Detection

The target of this task is to locate objects with oriented bounding boxes (OBBs) and give their semantic labels. The OBBs here are quadrilaterals \(\{(x_i, y_i)| i = 1, 2, 3, 4\}\). We adopt the mean average precision (mAP) as the evaluation metric, which follows the PASCAL VOC [10] except that the IoU calculation is performed between OBBs.

### 3.2. Task2 - Horizontal Object Detection

The target of this task is to locate objects with horizontal bounding boxes and give their semantic labels. The HBBs here are rectangles \((x_{min}, y_{min}, x_{max}, y_{max})\), which are generated by calculating axis-aligned bounding boxes of OBBs. We adopt the mean average precision (mAP) as the evaluation metric, which is exactly the same as PASCAL VOC [10].

### 3.3. Task3 - Semantic Segmentation

The target of this task is to predict the semantic labels for all the pixels in aerial images. We adopt the mean intersection over union (mIoU) [10] as the evaluation metric.

## 4. Organization

The registration of this competition starts on July 5, 2021. After registration, the participants can download the images of the train/val/test set and ground truths of the train/val set. Since we do not release the ground truths of these tasks, the participants need to submit their results to our evaluation server for automatic evaluation. If the evaluation process succeeds, the results will be sent to the registered emails. Each team is allowed to submit only once a day to the server during the challenge to avoid overfitting. The deadline for challenge submission is Aug 15, 2021.

There are 84 and 62 teams registered at the DOTA-v2.0 evaluation server and the GID-15 evaluation server, respectively. The teams come from universities, research institutes, and tech companies, such as Alibaba, Tsinghua University, Hikvision Research Institute, Xidian University, Northwestern Polytechnical University (NWPU), and Satrec Initiative.

There are 459, 219, and 374 submissions for task1, task2, and task3, respectively. If one team has multiple sub-

\(^3\)https://captain-whu.github.io/GID15/
missions, only the submission with the highest score is kept. Each team is asked to submit a description of their method. Otherwise, the submissions are not valid.

5. Submissions and Results

The mAP of two object detection tasks and mIoU of semantic segmentation task are calculated by our evaluation server automatically. The Tab. 1, Tab. 2, and Tab. 3 summarize the valid results of each team on three tasks, respec-
5.1. Baseline Results

For reference, we implement and evaluate baseline algorithms for each task. For Task1 and Task2, we adopt RoI Transformer, R50-FPN, 1× schedule [6] and ReDet, ReR50-ReFPN, 1× schedule [13] as our baselines. We follow the single scale default settings in AerialDetection Benchmark [7]. For Task3, we adopt UperNet [34] with Swin-S [21] as our baseline. When training, we only use training data (100 images) and no validation. We change the input size to 1024×1024. Then we ignore the background area when calculating loss. When testing an image, we adopt a slide inference strategy. The patch size is 1024×1024, and stride is 700×700 in this process. Besides, we follow the default setting in Swin Transformer when training on Cityscapes[4] dataset.

5.2. Top 3 Submissions on the Task1

1st Place. The Alibaba_AIEarth, team of Qiang Zhou, Chaohui Yu from Alibaba Group, adopt the ReDet [13] framework as baseline for this challenge. Then they optimize the baseline in mainly two ways, i.e., single model optimization and test-time augmentation (TTA). Specifically, as for single model optimization, they apply Swin [21] transformer as the backbone, Guided Anchor [28] as the RPN, Double-Head [32] as the detection head. In addition, they incorporate RiRoI Align to GRoIE [26] to achieve a better single model. They further use SWA [38] to train each single model for an extra 12 epochs using cyclical learning rates and then average these 12 checkpoints as the final model. Besides, they use random rotation, random crop, random flip (vertical and horizontal), and multi-scale training as data augmentation in the training phase. As for TTA, they apply multi-scale testing by resizing the images by factors of [0.5, 1.0, 1.5] and then crop the images into patches of size 1024×1024. Since the performance of different models in each category may be different, to leverage the advantages of different single models, they further perform model ensembling to achieve better performance. The OBB results and HBB results are shown in Table 4 and Table 5, respectively.

2nd Place. The Tsinghua_cc, team of Chenchen Fan from Tsinghua, used the official ReDet [13] 5 as their baseline and then improve performance based on the baseline. As for the data process, since they use multi-scale training and multi-scale testing, they prepare the train/val dataset, and the test challenge dataset with scales [0.5, 0.75, 1.0, 1.5], and then resize the cropped patch to 1024x1024. As for training, they train multiple single models (backbone includes ReR50, res50, res101) with 12 epochs or 24 epochs. Then they use SWA [38] training to improve the performance of each single model. As for testing, they use multi-scale testing and rotate testing. Finally, they ensemble the results of multiple single models.

3rd Place. The HIK HOW, team of Kaixuan Hu, Yingjia Bu, Wenning Tan from Hikvision Research Institute, adopt Oriented Reppoints [16] and ROI Transformer [6] as baseline-detectors. They only use the competition dataset for training using pre-trained models from ImageNet, and no extra data was added in training. About backbone, they found that Swin Transformer [21] + Path Aggregation Network (PANet) [20] could achieve better performance than ResNet [14]. Another issue with ODAI is that the hyper-parameter setting of conventional object detectors learned from natural images is not appropriate for aerial images due to domain differences. Thus they changed some hyperparameters such as anchor scales and anchor ratios. Besides, tricks like model ensembling, data augmentation, test time augmentation are also adopted to improve the final performance.

5.3. Top 3 Submissions on the Task2

Alibaba_AIEarth, HIK_HOW, and Tsinghua_cc ranked in the first, second, and third place in Task2, respectively. All three teams use the same methods described in Task1 and transfer their OBB results to HBB results for Task2. The ranking order is slightly different from that in Task1.

5.4. Top 3 Submissions on the Task3

1st Place. The Alibaba_AIEarth, team of Zhe Yang, Wei Li, Shang Liu from Alibaba Group adopt segformer [9] and volo [17] as the baseline. They cut the train and validation set to small patches for training. The patch size is 1024 × 1024, while the input image size is 896 × 896 for segformer and 512 × 512 for volo. An ensemble loss of IoU loss and CrossEntropy loss is used. An effective data augmentation pipeline including color jitter, contrast, brightness, and gaussian noise is used. In the training stage, they first use the poly learning rate policy to train the basic models and then use the cyclic learning rate policy to prepare snapshots for SWA [38] average by using basic models as a pre-trained model. Segformer model and volo model are merged to get better performance. They find it is hard for a model to classify the large lake or river area, but the result nearby the land is reliable. Some categories are confusing between each other, for example, arbor woodland and shrub land, natural grassland, and artificial grassland. They design two specific models to distinguish the confusing category pairs.

2nd Place. The ZAIL, team of Jiaxuan Zhao, Tianzhi Ma, Zihan Gao, Lingqi Wang, and Yi Zuo, Licheng Jiao
Table 1: Results of valid submissions in Task1. The abbreviations are the same as in Table 1.

<table>
<thead>
<tr>
<th>Team</th>
<th>mAP</th>
<th>PL</th>
<th>BD</th>
<th>GTF</th>
<th>SV</th>
<th>LV</th>
<th>Ship</th>
<th>TC</th>
<th>BC</th>
<th>ST</th>
<th>SBF</th>
<th>RA</th>
<th>Harbor</th>
<th>SP</th>
<th>CC</th>
<th>Airport</th>
<th>HP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alibaba_Ai_Earth</td>
<td>58.97</td>
<td>87.34</td>
<td>62.82</td>
<td>76.89</td>
<td>68.98</td>
<td>42.64</td>
<td>83.84</td>
<td>63.38</td>
<td>50.01</td>
<td>78.98</td>
<td>51.75</td>
<td>58.81</td>
<td>47.57</td>
<td>62.98</td>
<td>14.02</td>
<td>53.88</td>
<td>88.41</td>
</tr>
<tr>
<td>Tsinghua_tc</td>
<td>56.03</td>
<td>86.39</td>
<td>61.73</td>
<td>45.42</td>
<td>74.61</td>
<td>47.46</td>
<td>29.08</td>
<td>81.41</td>
<td>60.88</td>
<td>45.27</td>
<td>76.61</td>
<td>47.56</td>
<td>57.88</td>
<td>44.8</td>
<td>61.87</td>
<td>11.62</td>
<td>49.27</td>
</tr>
<tr>
<td>HK_HOW</td>
<td>55.93</td>
<td>86</td>
<td>61.88</td>
<td>46.78</td>
<td>76.03</td>
<td>55.19</td>
<td>38.01</td>
<td>77.77</td>
<td>59.42</td>
<td>45.57</td>
<td>74.17</td>
<td>50.48</td>
<td>55.46</td>
<td>45.86</td>
<td>50.80</td>
<td>10.66</td>
<td>53.07</td>
</tr>
<tr>
<td>Beaters</td>
<td>54.59</td>
<td>85.18</td>
<td>58.73</td>
<td>44.44</td>
<td>74.05</td>
<td>56.34</td>
<td>34.75</td>
<td>76.79</td>
<td>55.51</td>
<td>43.96</td>
<td>73.71</td>
<td>49.27</td>
<td>53.56</td>
<td>43.48</td>
<td>58.05</td>
<td>18.29</td>
<td>47.81</td>
</tr>
<tr>
<td>LMY-XiaoXiao</td>
<td>50.08</td>
<td>85.91</td>
<td>57.39</td>
<td>42.55</td>
<td>75.06</td>
<td>46.92</td>
<td>32.87</td>
<td>76.23</td>
<td>57.39</td>
<td>46.21</td>
<td>60.58</td>
<td>45.91</td>
<td>46.15</td>
<td>43.03</td>
<td>53.58</td>
<td>9.75</td>
<td>47.34</td>
</tr>
<tr>
<td>STTU_Pingan_KYZ</td>
<td>47.98</td>
<td>83.81</td>
<td>58.54</td>
<td>38.93</td>
<td>73.73</td>
<td>52.72</td>
<td>66.41</td>
<td>59.72</td>
<td>42.07</td>
<td>42.73</td>
<td>72.53</td>
<td>43.97</td>
<td>54.40</td>
<td>53.56</td>
<td>36.64</td>
<td>13.97</td>
<td>28.63</td>
</tr>
<tr>
<td>pure_white</td>
<td>45.23</td>
<td>79.15</td>
<td>56.89</td>
<td>41.26</td>
<td>73.64</td>
<td>62.73</td>
<td>24.66</td>
<td>58.85</td>
<td>56.82</td>
<td>42.81</td>
<td>45.97</td>
<td>52.64</td>
<td>42.30</td>
<td>46.78</td>
<td>5.5</td>
<td>31.55</td>
<td>70.22</td>
</tr>
<tr>
<td>Im2hard</td>
<td>44.17</td>
<td>77.29</td>
<td>52.17</td>
<td>38.16</td>
<td>70.99</td>
<td>65.93</td>
<td>21.65</td>
<td>65.91</td>
<td>52.22</td>
<td>46.52</td>
<td>63.43</td>
<td>38.71</td>
<td>45.47</td>
<td>34.81</td>
<td>40.15</td>
<td>1.49</td>
<td>43.58</td>
</tr>
<tr>
<td>hakjinlee</td>
<td>42.39</td>
<td>77.72</td>
<td>52.29</td>
<td>39.18</td>
<td>70.86</td>
<td>65.25</td>
<td>20.26</td>
<td>70.63</td>
<td>46.55</td>
<td>36.99</td>
<td>58.86</td>
<td>42.32</td>
<td>45.31</td>
<td>43.97</td>
<td>48.27</td>
<td>1.66</td>
<td>21.02</td>
</tr>
<tr>
<td>xyh</td>
<td>31.24</td>
<td>61.29</td>
<td>49.09</td>
<td>19.58</td>
<td>61.19</td>
<td>25.95</td>
<td>45.80</td>
<td>48.17</td>
<td>5.71</td>
<td>50.22</td>
<td>35.26</td>
<td>9.96</td>
<td>38.05</td>
<td>0.05</td>
<td>10.27</td>
<td>66.86</td>
<td>15.95</td>
</tr>
<tr>
<td>philhe</td>
<td>18.55</td>
<td>51.71</td>
<td>29.22</td>
<td>12.64</td>
<td>45.86</td>
<td>16.75</td>
<td>19.09</td>
<td>32.78</td>
<td>5.05</td>
<td>52.07</td>
<td>29.39</td>
<td>36.46</td>
<td>18.12</td>
<td>38.68</td>
<td>0.04</td>
<td>21.51</td>
<td>68.69</td>
</tr>
</tbody>
</table>

Red Tran Tr [6] (baseline) 38.34 70.38 45.46 32.72 67.71 26.60 21.06 51.22 43.15 30.42 59.46 34.57 32.22 43.42 1.82 62.30 9.86
Red Tran [7] (baseline) 40.05 73.93 48.85 35.58 68.62 26.51 19.24 56.89 48.73 32.58 59.89 35.82 44.36 32.21 37.41 3.20 18.97 69.09 0.03

Table 2: Results of valid submissions in Task2. The abbreviations are the same as in Table 2.

Table 3: Results of valid submissions for categories in Task3. The abbreviations are the same as in Table 3.

Table 4: Ablations of Alibaba_Ai_Earth in Task1.

Table 5: Ablations of Alibaba_Ai_Earth in Task2.
5.5. Summary of Methods and Discussions

In the Task1, the RoI Transformer and ReDet are widely used by all the top submissions, which show the rotation-invariance is crucial in ODAI. Recent transformer based backbone (such as Swin Transformer [21]) is also widely used and shows its advantage in ODAI. The architecture design and data augmentation involving scale and rotation are helpful in ODAI. In the Task2, most of the submissions transfer their OBB results to the HBB results. The possible reason is that the OBB results can be used to perform Rotated NMS (R-NMS) [6, 37], which is better than regular NMS for densely packed object detection. For Task3, the 1st team adopt the transformer-based segmentation network and significantly surpass the other teams. Our baseline algorithm is also a transformer-based method and outperforms most of the participants that use the CNN-based methods. The possible reason is that the transformer can significantly increase the effective receptive field [23] and obtain global information, which is very important for semantic segmentation in aerial images.

For all three tasks, Transformers are widely used, showing that it is also a promising direction in the LUAI. SWA [38] is another widely used method in all three tasks to improve the generalization. However, there still exists much room to conduct research related to generalization in LUAI.

6. Conclusion

We organized a challenge on LUAI with three tasks of oriented object detection, horizontal object detection, and semantic segmentation. In summary, this challenge received a total of 146 registrations from a worldwide range of institutes. Through the results and summary, we find several promising directions in the LUAI. We hope this challenge can draw more attention from vision communities and promote future research on LUAI.

References

[12] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. IEEE Transactions on Geoscience and Remote Sensing, 2021. 1
[13] Jiaming Han, Jian Ding, Nan Xue, and Gui-Song Xia. Redet: A rotation-equivariant detector for aerial object detection. In


[26] Leonardo Rossi, Akbar Karimi, and Andrea Prati. A novel region of interest extraction layer for instance segmentation, 2020. 4


[37] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In CVPR, pages 8232–8241, 2019. 6
