Mask OBB

A Semantic Attention-Based Mask Oriented Bounding Box Representation for Multi-Category Object Detection in Aerial Images

Jinwang Wang, Jian Ding, Haowen Guo, Wensheng Cheng, Ting Pan and Wen Yang

School of Electronic Information, Wuhan University, Wuhan 430072, China

jwwangchn@whu.edu.cn

1. Introduction

Overview

- ✓ We address the influence of ambiguity of regression-based OBB representation methods for oriented bounding box detection, and propose a mask oriented bounding box representation (Mask OBB). As far as we know, we are the first to treat the multi-category oriented object detection in aerial images as a problem of pixel-level classification.
- ✓ We propose an Inception Lateral Connection Feature Pyramid Network (ILC-FPN), which can provide better features to handle huge scale changes of objects in aerial images.
- ✓ We design a Semantic Attention Network (SAN) to distinguish interested objects from cluttered background by providing semantic features when predicting HBBs and OBBs.
- ✓ Some highlighted results on DOTA and HRSC2016 dataset:
  - DOTA OBB task mAP: **75.54**%
  - DOTA HBB task mAP: **76.98**%
  - HRSC2016 OBB task mAP: **96.7**%

Motivation

The representation of Mask-OBB can avoid the problem of ambiguity and obtain better detection results.

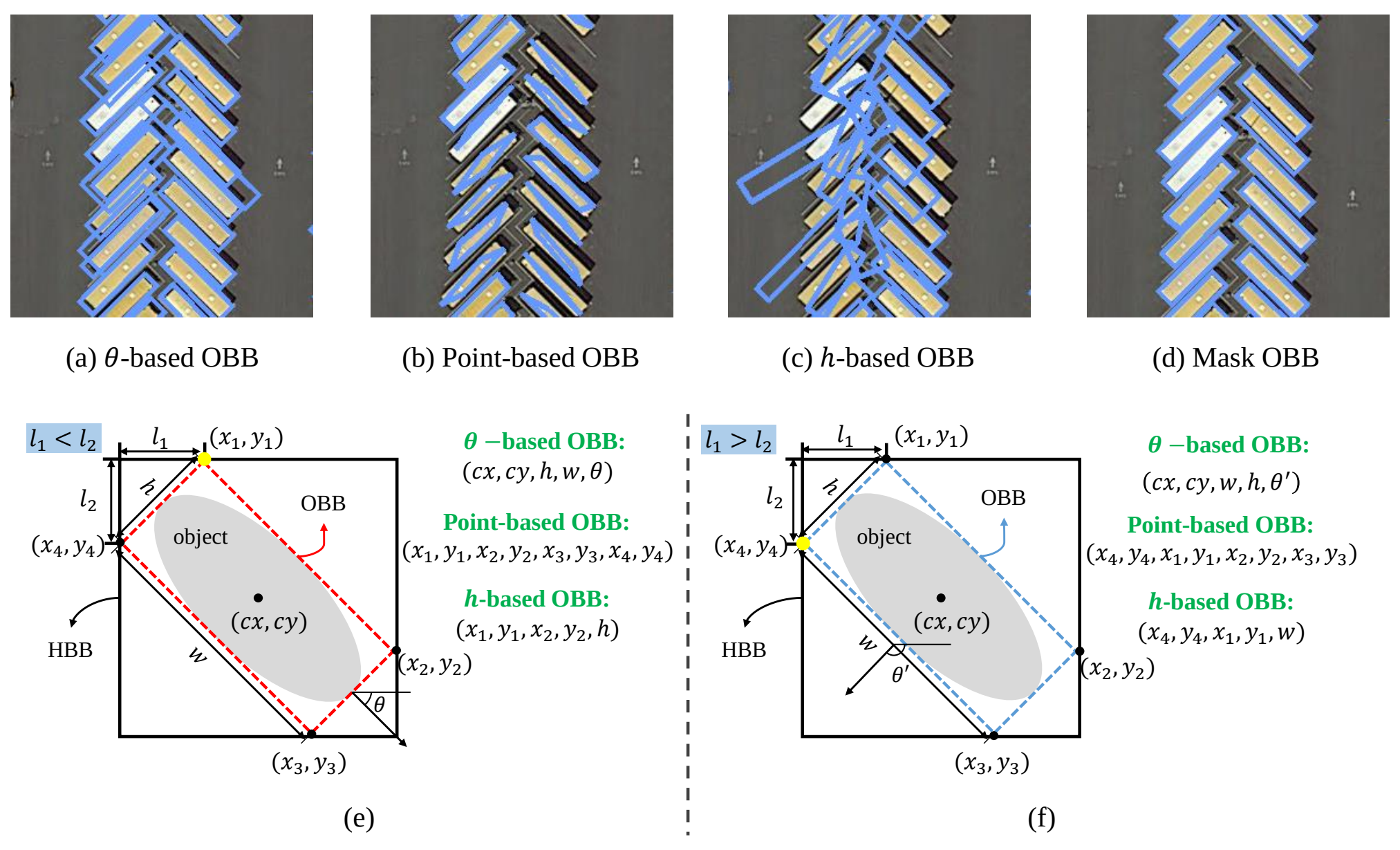


Figure 1: (a-c) Failure modes of (x, y, w, h, θ) (θ-based OBB representation), $\{(x_i, y_i)|i = 1, 2, 3, 4\}$ (point-based OBB representation) and (x_1, y_1, x_2, y_2, h) (h-based OBB representation) respectively. (d) Results from Mask-OBB representation.

2. Methodology

Mask-OBB

Fig. 2 illustrates the point-based OBBs and converted Mask-OBBs on DOTA dataset images. The highlight points are original ground truth, and the highlight regions inside point-based OBBs are new ground truth for pixel-level classification.

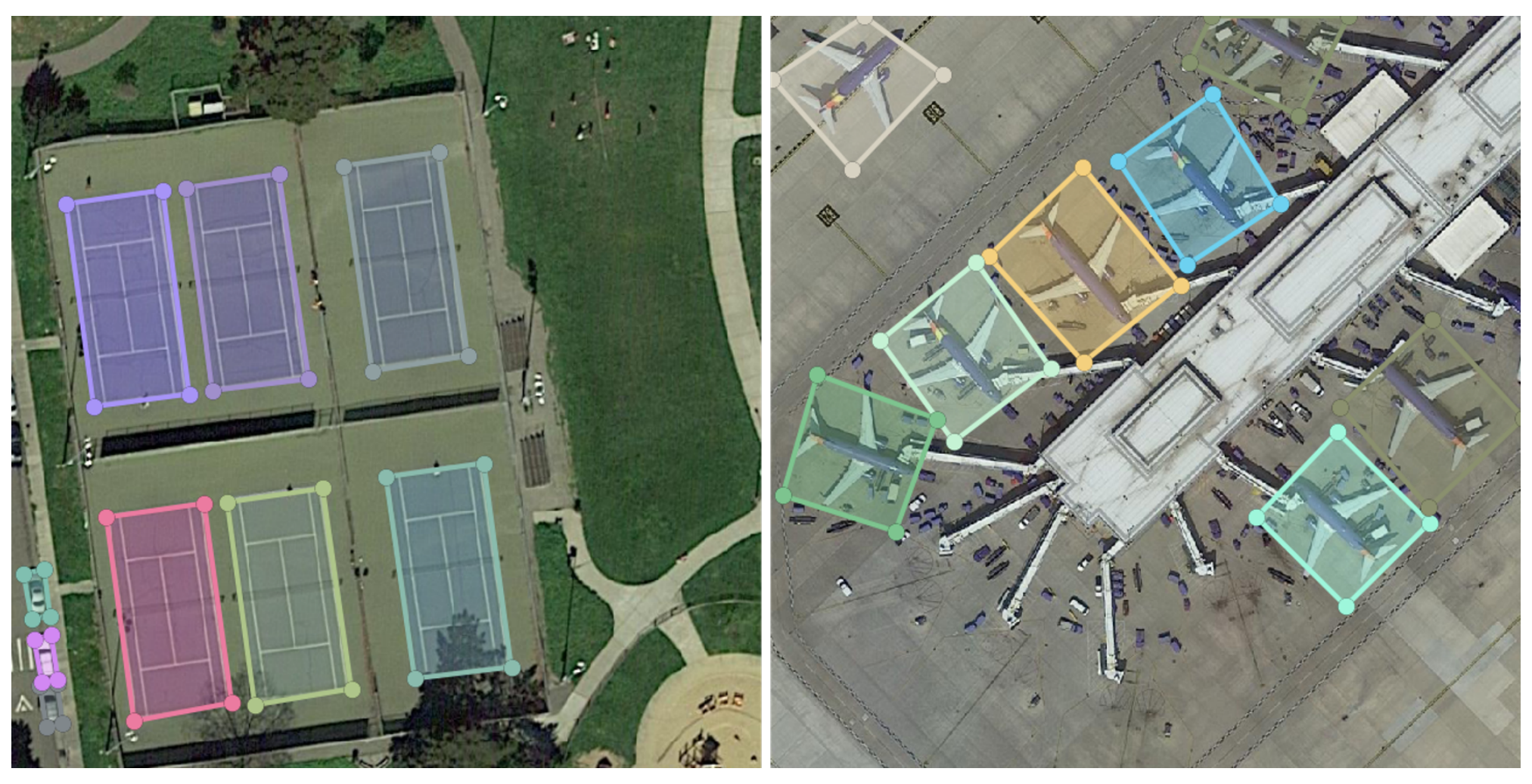


Figure 2: Samples for illustrating Mask-OBB.

Pipeline

1. ILCN-FPN and RPN generate features and region proposals, respectively.
2. RoI Align extracts object features from FPN features and SAN features.
3. HBB and OBB branchs generate horizontal bounding boxes and oriented bounding boxes on region proposals.

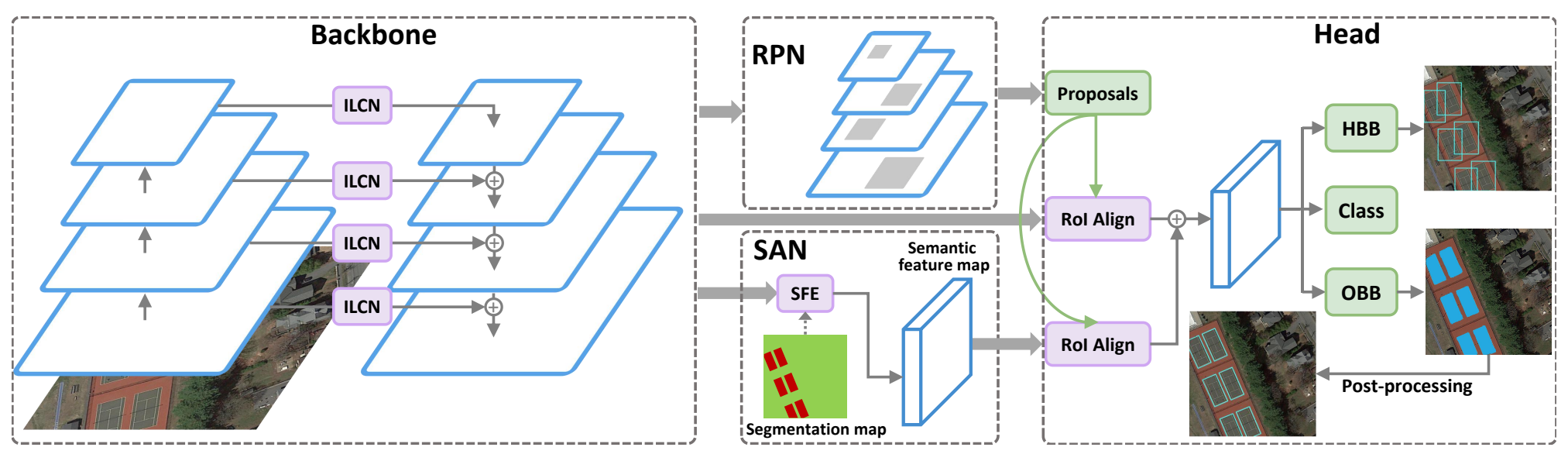


Figure 3: Overview of the pipeline for our method.

Inception Lateral Connection Network (ILCN)

For enhancing the FPN, we design an Inception Lateral Connection Network (ILCN), which uses inception structure to enhance feature propagation. Fig. 4 shows the architecture of ILCN. Different from the original FPN which uses a 1×1 convolutional layer as the lateral connection, ILCN use inception structure as the lateral connection. Besides the original 1×1 convolutional layer, three additional layers are added in the lateral connection. These extra layers include a 5×5 convolutional layer, a 3×3 convolutional layer, and a max pooling layer.

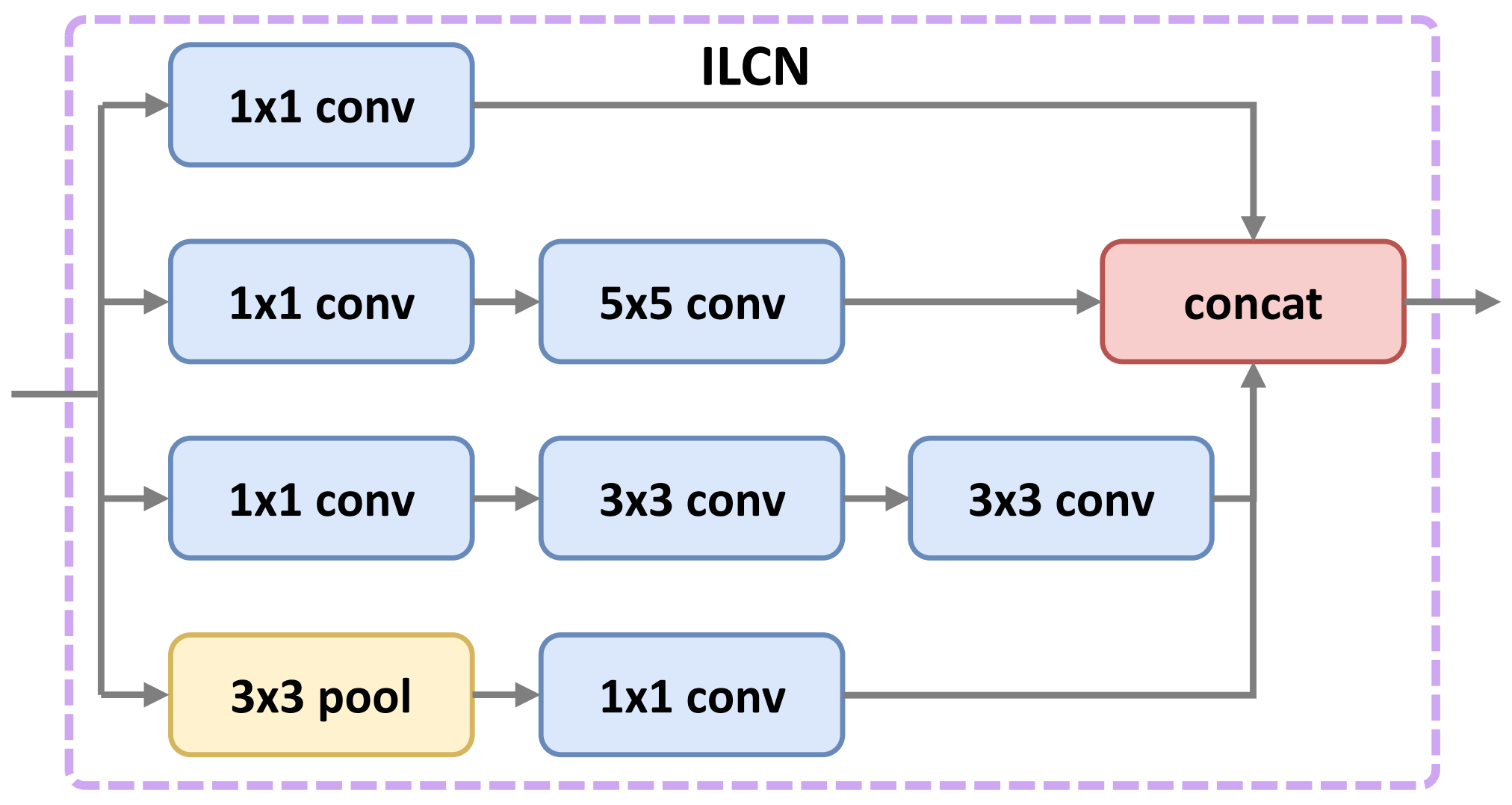


Figure 4: Illustration of the Inception Lateral Connection Network (ILCN).

Semantic Attention Network (SAN)

To generate the semantic feature from the outputs of FPN, we design a Semantic Feature Extraction (SFE) module to incorporate higher-level feature maps with context information and lower-level feature maps with location information for better feature representation. The result is a set of feature maps at the same scale, which are then element-wise summed. Four 3×3 convolutions are then used to obtain the output of SFE.

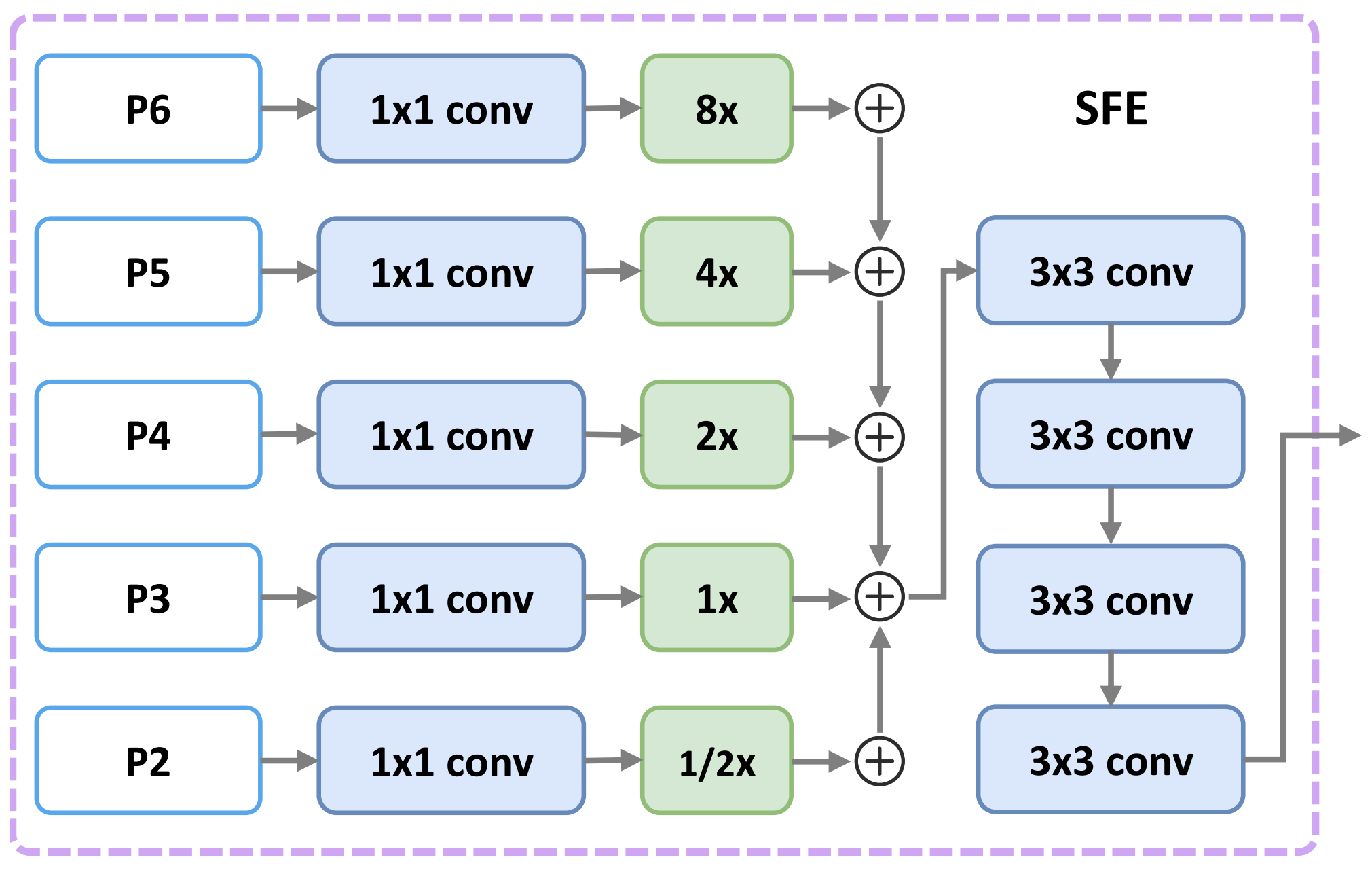


Figure 5: Illustration of the Semantic Feature Extraction (SFE) module.

Aerial image datasets have no semantic segmentation ground truth. For calculating the semantic segmentation loss, we generate the semantic segmentation ground truth by the OBB ground truth.

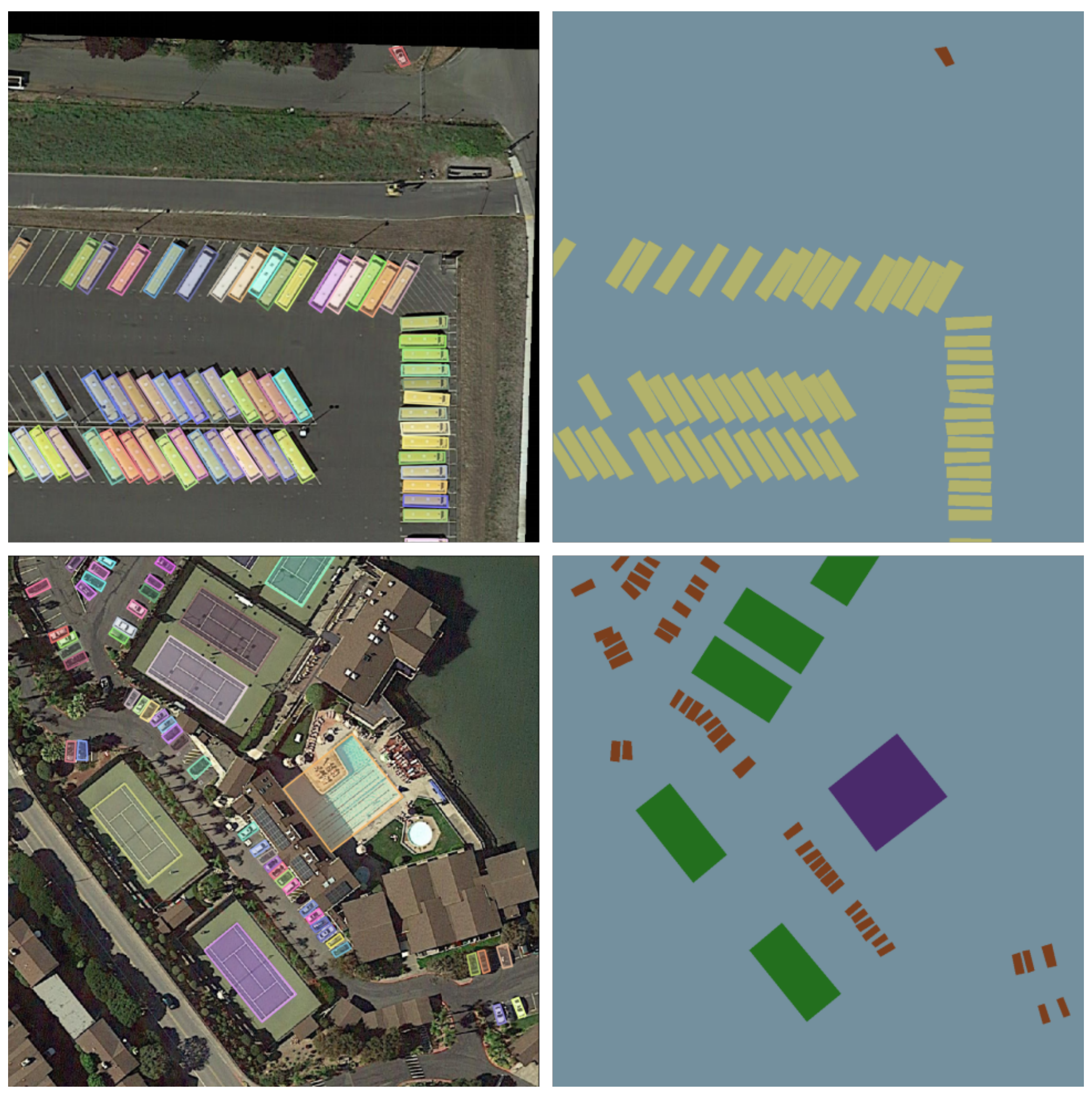


Figure 6: OBB ground truth and corresponding semantic segmentation ground truth. Different colors in semantic segmentation ground truth mean different categories.

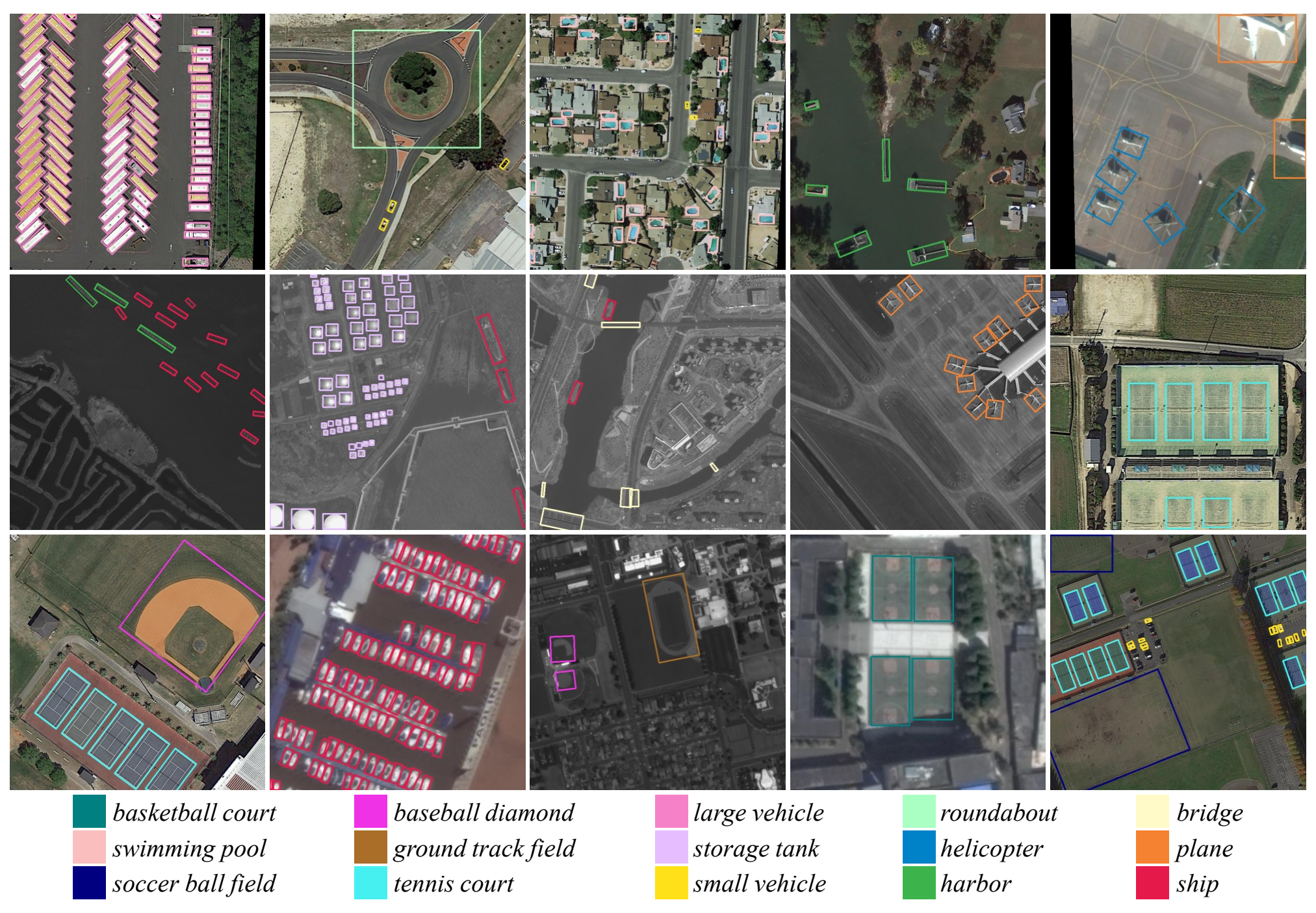
3. Experiments

We demonstrate the performance on DOTA [2] dataset and HRSC2016 [1] dataset with state-of-the-art methods.

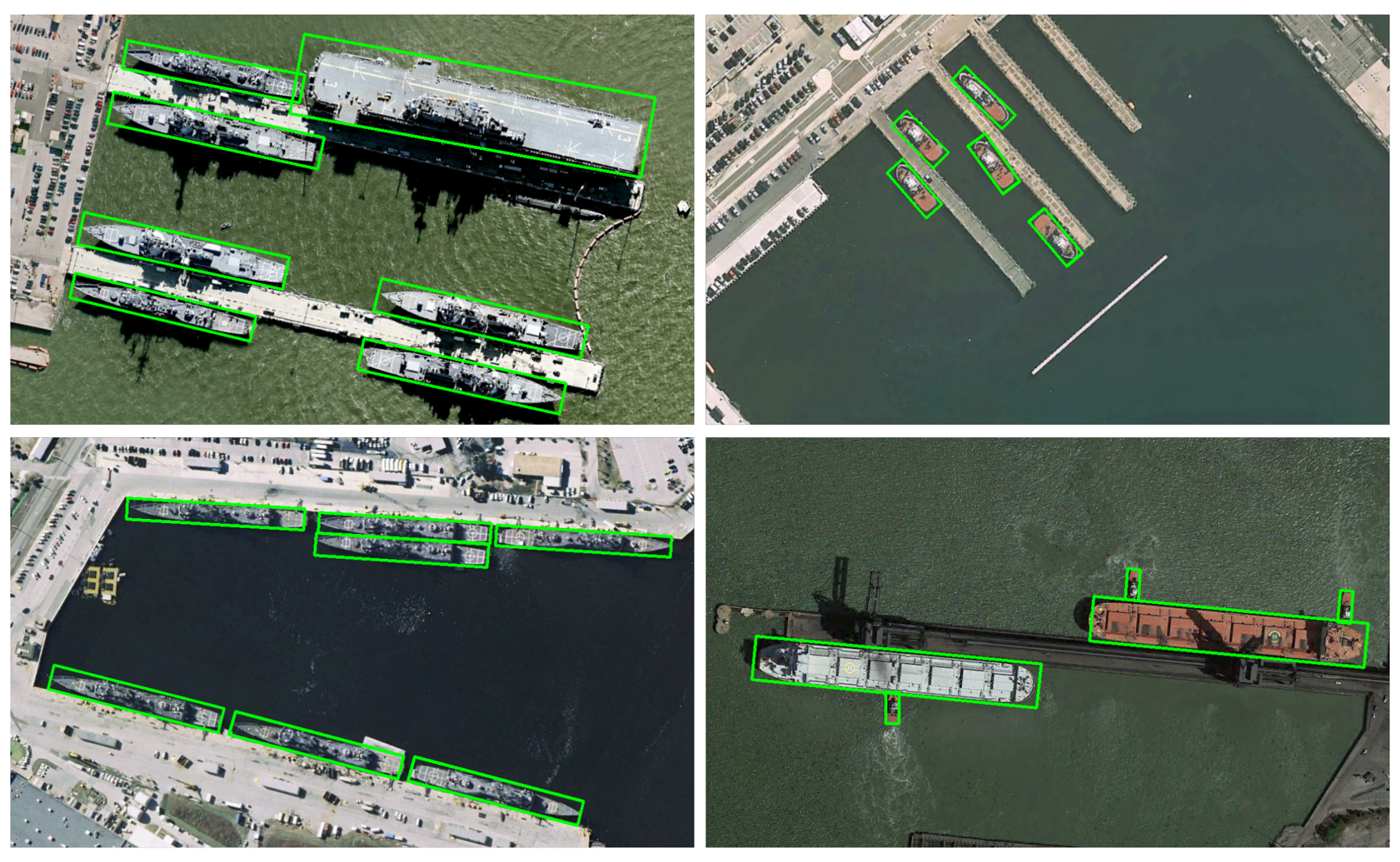
Table 1: Ablation study of each component in our proposed method on DOTA dataset.

model	OBB (%)	HBB (%)
baseline	69.97	70.14
baseline + ILC-FPN	71.09 (↑ 1.12)	71.61 (↑ 1.47)
baseline + SAN	70.69 (↑ 0.72)	72.02 (↑ 0.88)
baseline + ILC-FPN + SAN	71.43 (↑ 1.46)	72.41 (↑ 2.27)

Visualization Results on DOTA dataset



Visualization Results on HRSC2016 dataset



4. Conclusions

We analyzed the influence of different OBB representations for oriented object detection in aerial images, which exposes shortcomings of the typical regression-based OBB representation (θ-based, point-based and h-based OBB). Based on the analysis, the Mask OBB is proposed to tackle the ambiguity in regression-based OBB. With this new representation, we proposed the ILC-FPN which use Inception Lateral Connection (ILCN) to enhance the feature extraction ability of FPN for handling the scale variation problem in aerial images. In addition, we proposed the Semantic Attention Network to extract semantic features to further enhance features of generating HBBs and OBBs. Experimental results on the DOTA and HRSC2016 datasets demonstrated the importance of representations in multi-category arbitrary-oriented object detection. **Notably, our method achieves 75.54% and 76.98% on OBB task and HBB task of DOTA dataset, respectively. At the same time, it achieves 96.7% on HRSC2016 dataset OBB task.**

5. Reference

[1] Zikun Liu, Hongzhen Wang, Lubin Weng, and Yiping 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] 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 *CVPR*, pages 3974–3983, 2018.