Paper ID: 83801
student

IAF EARTH OBSERVATION SYMPOSIUM (B1)
Interactive Presentations - IAF EARTH OBSERVATION SYMPOSIUM (IP)

Author: Dr. Lingyun Gu
School of Aerospace, Tsinghua University, Beijing, China

Dr. Eugene Popov
Peter the Great St. Petersburg Polytechnic University, Russian Federation
Prof. Ge Dong
Tsinghua University, China

TWO-STREAM FEATURE FUSION STRATEGY FOR MULTIMODAL REMOTE SENSING OBJECT DETECTION IN EARTH OBSERVATION

Abstract

Remote sensing object detection is a fundamental task in Earth Observation (EO), and many EO applications require round-the-clock detection of specific areas or objects, such as ocean monitoring, fire warning and traffic direction. Compared to traditional single-modal detection, multimodal object detection is able to overcome challenges such as light variations, adverse weather, and occlusion, making it more adaptable to round-the-clock EO applications. By fusing information from multiple modalities, such as RGB, SAR, and infrared, multimodal object detection approaches can obtain more comprehensive features, thus improving the reliability and robustness of detection. Existing approaches tend to fuse all regions equally, which completes the missing object information in a single modality, but simultaneously introduces background noise, making network learning more difficult. To address this issue, in this paper, a two-stream feature fusion strategy (TSFF) for multimodal remote sensing object detection in EO is proposed, where one stream performs focal fusion on the object regions and another stream performs global fusion on the entire feature map, to achieve more effective multimodal object detection. Specifically, 1) Focal fusion first decouples the foreground and the background, then enhances the key pixels and channels in the foreground by an Object-Aware Module (OAM), and finally fuses the foreground features to eliminate the negative effects of background noise. 2) Global fusion extracts the relationships between long-range information in single-modal images by a Context-Aware Module (CAM), and then fuses them to compensate for the global information missing in the focal fusion. Extensive experiments on VEDAI dataset validate the effectiveness of our TSFF. VEDAI is a large multimodal remote sensing dataset containing 9 vehicle classes with more than 3700 labeled objects in more than 1200 images, each with RGB and IR modalities. With YOLOv5s as the baseline, TSFF achieves 78.8% mean Average Precision (mAP), which is 4.5% mAP improvement over the RGB detector and 4.8% mAP improvement over the IR detector. With Faster RCNN as the baseline, the object detection performance of TSFF is also significantly improved compared with the popular algorithms. Additionally, visualization results demonstrate that our TSFF effectively reduces false positives caused by noisy features. Since TSFF is a generalized fusion strategy, it can benefit both one-stage and two-stage detectors, and thus it has great potential for all-weather, all-day EO missions.