Semi-Supervised Hyperspectral Object Detection Challenge Results - PBVS 2022

Aneesh Rangnekar, Zachary Mulhollan, Anthony Vodacek, Matthew Hoffman, Angel Sappa, Erik Blasch,
Jun Yu, Liwen Zhang, Shenshen Du, Hao Chang, Keda Lu,
Zhong Zhang, Fang Gao, Ye Yu, Feng Shuang, Lei Wang, Qiang Ling,
Pranjay Shyam, Kuk-Jin Yoon, Kyung-Soo Kim

Abstract

This paper summarizes the top contributions to the first semi-supervised hyperspectral object detection (SSHOD) challenge, which was organized as a part of the Perception Beyond the Visible Spectrum (PBVS) 2022 workshop at the Computer Vision and Pattern Recognition (CVPR) conference. The SSHODC challenge is a first-of-its-kind hyperspectral dataset with temporally contiguous frames collected from a university rooftop observing a 4-way vehicle intersection over a period of three days. The dataset contains a total of 2890 frames, captured at an average resolution of 1600 × 192 pixels, with 51 hyperspectral bands from 400nm to 900nm. SSHOD challenge uses 989 images as the training set, 605 images as validation set and 1296 images as the evaluation (test) set. Each set was acquired on a different day to maximize the variance in weather conditions. Labels are provided for 10% of the annotated data, hence formulating a semi-supervised learning task for the participants which is evaluated in terms of average precision over the entire set of classes, as well as individual moving object classes: namely vehicle, bus and bike. The challenge received participation registration from 38 individuals, with 8 participating in the validation phase and 3 participating in the test phase. This paper describes the dataset acquisition, with challenge formulation, proposed methods and qualitative and quantitative results.

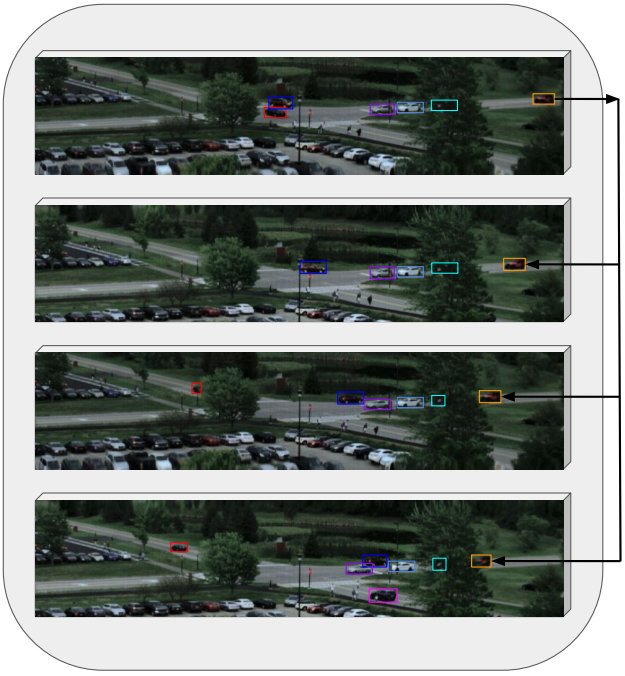


Figure 1. Example composite RGB rendering of four continuous frames from our dataset for the SSHOD challenge at approximately 0.7 frames per second. The operating mechanics of a push-broom sensor camera causes a difference in area observed in consecutive time steps.

1. Introduction

Hyperspectral images (HSI) differ from normal color (RGB) images in that they have roughly 50 - 400 contiguous color bands instead of the conventional three RGB bands. This increase in resolution along the channel dimension provides enhanced detail of the object materials present within the scene, and this has been shown to enhance fine-grained discrimination in deep neural networks for hyperspectral pixel classification, object tracking, and super-resolution [8–10, 13, 14, 21, 27]. Hyperspectral pixel classification, the research area motivating the sensor design

Aneesh Rangnekar[1] (aneesh.rangnekar@mail.rit.edu), Zachary Mulhollan[1], Anthony Vodacek[1], Matthew Hoffman[1], Angel D. Sappa[2,3], and Erik Blasch[4] are the SSHODC - PBVS CVPR 2022 organizers, while the other authors participated in the challenge.
[1]Rochester Institute of Technology, Rochester, NY, USA.
[2]ESPOL Polytechnic University, Guayaquil, Ecuador.
[3]Computer Vision Center, Campus UAB, Barcelona, Spain.
[4]US Air Force Research Lab, Rome, NY.

and data collection, has been primarily studied using three datasets: (1) Indian Pines, (2) Salinas Valley, and (3) University of Pavia. Indian Pines and Salinas Valley contain primarily different types of vegetation and Univ. of Pavia contains classes typically found around a university - for example, trees, soil, and asphalt. In all three cases, the small spatial extent often leads researchers to use Monte-Carlo (MC) cross-validation splits for benchmarking the performance of various deep learning based architectures.

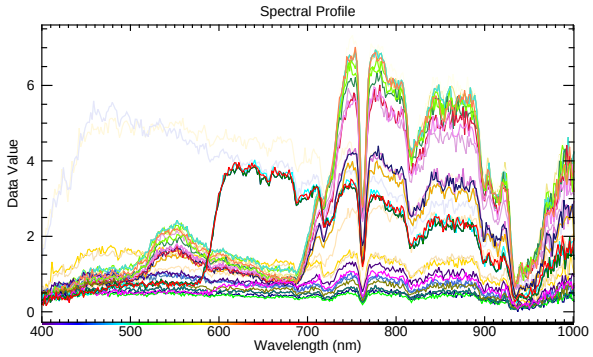
A large body of previous work in the hyperspectral imagery domain [1, 19, 20, 22, 25] has attempted to understand the challenges involved in dynamic scene understanding for spectral images containing a diverse set of materials. However, the current non-synthetic datasets suffer from two major shortcomings from a dynamic application oriented perspective - for example, vehicle object detection. First, they are captured in a static environment, e.g., the flight line presented in AeroRIT, captured at a relatively high ground sampling distance (GSD), but cannot be used for object localization due to significant overlap between small-size pixels [15]. Second, they do not contain rich instances of some major sources of occlusion in spectral imagery, e.g., adjacency effect, glint, and shadows. The need for an annotated dynamic HSI dataset with real-world environmental challenges is essential as neural network approaches have been known to be sensitive to image perturbations and the above-discussed factors with atmospheric variance can significantly alter the image composition, thereby resulting in signatures that may appear to be out-of-the-training distribution for the networks.

With these motivations in mind, we collect a motion dataset (Fig. 1) - our primary goal is providing information to study and solve the challenges that may appear for creating a deployable model that uses spectral signatures, or a multi-modal combination with spectral signatures, for object detection, (future) tracking, and re-identification purposes from the ground and aerial perspectives. The goal of the RooftopHSI dataset is to improve the recently developing collection of datasets in spectral imagery and the robustness of spectral imagery exploitation methods. We mount a hyperspectral imaging system on a university building rooftop overlooking a 4-way moving traffic intersection and gather data over a period three days.

(a) Data acquisition with the Headwall Hyperspec. We focus on a red vehicle of interest to show the details captured via spectral imagery.

(b) The corresponding scene rendered as the RGB-composite with the same car highlighted with a red box.

(c) The mask for the corresponding scene highlighting the area of interest while ignoring regions belonging to the parking lot.



(d) Spectral radiance curves of different objects within the scene. The x-axis denotes the different hyperspectral bands from 390-1000 nm, and the y-axis denotes the corresponding values in radiance.

Figure 2. Visualizing the data acquisition setup, corresponding frame, mask, and spectral signatures plot for a scene instance in the dataset.

2. RooftopHSI dataset

2.1. Data Acquisition

We collect images using a custom built imaging system with a Headwall Micro HE (High Efficiency) Hyperspec E-series camera attached to a high-speed gimbal. In general, there are four types of hyperspectral imaging systems that can be distinguished by their scanning mechanisms: whisk-broom, spectral, snapshot, and push-broom. Of these, snapshot and push-broom systems are the more popular for typical data collections in hyperspectral imaging [1, 5, 21, 24–26]. Snapshot cameras provide high frame rates but lack both spectral and spatial resolution relative to push-broom cameras. For example, Xiong *et al*. use an Imec snapshot camera with a spatial resolution of 512×256 pixels and 16 bands from 470 nm to 620 nm, at 25 frames per second (FPS) [25]. This is a relatively low resolution, both

spatially and spectrally, when it comes to deploying for applications in far-range and high-altitude imaging.

Since the goal is to understand challenges in hyperspectral imagery from both a ground and aerial perspective, we require a system that provides relatively rich spatial and spectral information. Hence, we use a push-broom system which provides temporal, spatial, and spectral data for scene understanding. Push-broom scanners work by collecting one spatial line at a time along with the associated spectra at each pixel. In order to generate a second spatial dimension, the camera must be moved over the imaging area (along-track) and then lines are stitched together to create the hyperspectral cube. The push-broom application obtains along-track (motion) pixels by nodding the pan-tilt unit (Fig. 2a), where the motion is also responsible for the mismatch in consecutive frames as seen in Fig. 1. We use a frame period and exposure of 5 milliseconds, and hence the typical image resolution is between 150-190 pixels vertically, 1600 pixels horizontally and 371 bands (390 nm to 1000 nm) in the spectral (channel) dimension at a modest frame rate of 0.8 $\sim$ 1.2 FPS.

We gathered data over three days (Table 1) and an average duration of 1.5 hours each morning. The camera was mounted at a fixed location on the rooftop (Fig. 2a) overlooking the same intersection spot, at the same relative altitude. We observed changes in the atmosphere (i.e., the weather changed from clear skies to clouds and back), which resulted in images that varied in signal magnitude due to the presence of clouds, variation in illumination, and other environmental interference. Typically, the resulting images are processed by converting the data from digital counts to radiance, then a final conversion to reflectance units through use of calibration panels. However, real-time conversion of hyperspectral cubes from radiance to reflectance is not possible at all times - there may be scenes where deploying a calibration panel is not practical (for example, deploying the camera on a moving unmanned aerial vehicle - plane). Hence, from a real-time usability perspective, we consider the lack of reflectance data as adversary for radiometric remote-sensing.

Figure 2a shows the RooftopHSI setup for data collection. The objectives of our collection are multi-fold:

- to obtain short-time interval hyperspectral imagery that can be used to perform object detection without losing the object's structure,
- to ensure there are sufficient sources of occlusion that cause vehicle misdetections and observe if hyperspectral signatures are helpful for detection under the conditions, and
- to analyze how hyperspectral vehicle detection performance is affected by changes in illumination and weather conditions.

Split	Date	Images	Instances			
			Vehicle	Bus	Bike	Total
Train	09-09-19	989	3299	80	41	3420
Val	08-29-19	605	3088	16	12	3116
Test	09-10-19	1296	3502	44	34	3580

Table 1. Statistics of objects in RooftopHSI over the train, validation and test sets.

2.2. Data Preprocessing

Before recording data from the camera, we closed the shutter and obtained dark current readings at every new set of video captures. These dark current readings were used along with the camera's proprietary processing software to calibrate all images from digital counts to at-sensor spectral radiance in units of $mWm^{-2}sr^{-1}\mu m^{-1}$. Randomly sampled spectra from vehicles, road and vegetation found in the scene are shown in Fig. 2b and plotted them in Fig. 2d. We observed low signal-to-noise ratio (SNR) below 400 nm, and a lot of similar spectra with low SNR in the 900 nm to 1000 nm range, with differences in amplitude. As these bands do not contain discriminative spectral information, we do not use them in our analysis. In addition, we used a sub-sampled band version by sampling at every 10 nm to optimize computation cost versus disk occupancy and reduce adjacent band redundancy. Our final dataset contains 51 bands, from 400 nm to 900 nm, in 10 nm intervals.

2.3. Data Annotation

We used LabelImg to label bounding boxes into three categories (i.e., vehicle, bus, and bike) within the data following a two-step approach: (1) as we are interested in moving vehicles on the road, we created a mask per image that covers the parking lot within the scene (see Fig. 2c), and (2) we then proceed to image labeling the images within the area of interest using a modified version of LabelImg that provides insights into occlusions due to vegetation by using the normalized difference vegetation index (NDVI) algorithm [11]. We annotated every object within the 4-way intersection that is visible or partially occluded to the human eye, as we hypothesized hyperspectral signatures can help compensate for lack of color and edge-based detections. To avoid labeling discrepancies, a team of annotators (scale.ai), the student challenge organizers and two external volunteers further confirmed all labels, for an average of four checks on each labeling instance.

We split the data into train, validation, and test sets based on the days they are captured as shown in Table 1. Our reasoning behind this distribution split is the fact that our camera overlooks the same 4-way intersection over the three days, with minor changes to the observation altitude and

(a) Set of ordered frames from the dataset: the vehicle, along with others in the dataset, appears distorted once the push-broom tilt sweep is complete and all lines stitched together depending on their relative speed to the camera's frame rate and its tilt motion.

(b) Vehicles occluded by trees in the scene.

(c) Vehicles occluded by other vehicles in the scene. The second figure from the left also shows the image perturbations caused by glint in the scene.

Figure 3. Zoomed in composite RGB images of instances from the dataset showing sources of noise and occlusions.

angles. Hence, the data split prevents quick scene generalization, which in turn may cause the convolutional neural networks to overfit quickly. Having different days makes the task relatively difficult as we now have to take into consideration the network's potential to overfit as well as account for changes in atmospheric conditions and surroundings that may cause a shift in the spectral signatures of moving objects. Table 1 also shows a huge imbalance in the number of examples over classes, with the vehicle category dominating the other two classes - which is a realistic scenario over campuses. We do not label the dataset for single or multi-object tracking as the average track length is around 3-4 frames at 0.7 FPS and discussing and developing algorithms for low frame rate tracking is beyond the scope of the current SSHOD challenge.

2.4. Data Exploration

The RooftopHSI dataset contains a total of 2890 manually selected and labeled frames (Table 1). HSI data is relatively expensive to store and process and therefore, we only consider frames that have at least a single-car instance through the intersection in our dataset.

Environment: The camera was mounted on the university rooftop and observed a 4-way intersection as described in Section 2.1. There are multiple trees present throughout this scene and they account for the primary source of occlusion throughout the dataset (Fig. 3b). We also continued to gather data when the environment shifts from clear skies to cloudy weather to replicate aerial data collection settings. This is contrary to common data collection settings in HSI, where images are gathered during particularly clear skies to prevent signal contamination from atmospheric noise. Variety in the atmosphere makes our dataset more challenging from an HSI processing standpoint.

Camera noise: The tilt motion of the Hyperspec camera, to write data in a push-broom setting, introduces motion artifacts in the images as seen in Fig. 3a. The vehicle appears to be deformed right-inclined or left-inclined depending on if the unit is moving from up-down or down-up respectively, which can be considered a low frame-rate mo-

tion blur that occurs due to a mismatch in vehicle speed and camera frame rate. We account for this deformity by interpolating the bounding box throughout the deformed shape and consider it a form of dataset noise. In addition, the most common sensor noise in HSI systems are smile and keystone effects. However, the data did not contain enough distortions as we imaged from ground level at relatively close range and hence, do not consider them in the preprocesing stage (Section 2.2).

Glint: Sun glint, the most common source of occlusion in remote sensing, occurs due to the material's bidirectional reflectance mechanics directly reflecting sunlight into the camera sensor. We observe this only occurs in certain parts of the imagery and is almost always associated with vehicles, and sometimes water (Fig. 3c).

Moving objects: Fig. 3c shows vehicle-to-vehicle occlusion, which most often occurs along the intersection box borders. These are the secondary source of occlusion in the RooftopHSI dataset.

The presence of such variations, coupled with changes in atmospheric settings - clear skies and cloudy weather, make our dataset the first of its kind to tackle object detection with ground hyperspectral imagery.

2.5. SSHOD Challenge 2022

The SSHOD challenge is a semi-supervised object detection problem. Particularly, we provide labels for only 10% of the data from the 989 training images, ensuring the tail classes, namely bus and bike are sufficiently sampled for learning useful features. We also provide a starter code [1] on our baseline MobileNet-v2 Faster-RCNN, which is trained on only the labeled examples within the dataset [4, 16, 17]. The challenge was formulated as a means to encourage participants to develop frameworks for learning meaningful representations that can make up for the loss of labeled examples, while maintaining a backbone complexity not drastically exceeding the MobileNet-v2 architecture. After registration, participants were able to access the links to all the data via CodaLab and submit predictions for automatic evaluation on the competition server. The COCO evaluation metric was used for determining the rankings of all submissions [12]. From 38 registered participants, 8 submitted their predictions in the validation phase and 3 submitted their predictions in the test phase, with corresponding fact sheet and model weights, with only 2 of the entries above our baseline (Table 2). We discuss the approaches and team formations, with a distinct observation that both the teams used a student-teacher framework to generate pseudo labels as a means of compensating for lack of sufficient data.

[1] https://codalab.lisn.upsaclay.fr/

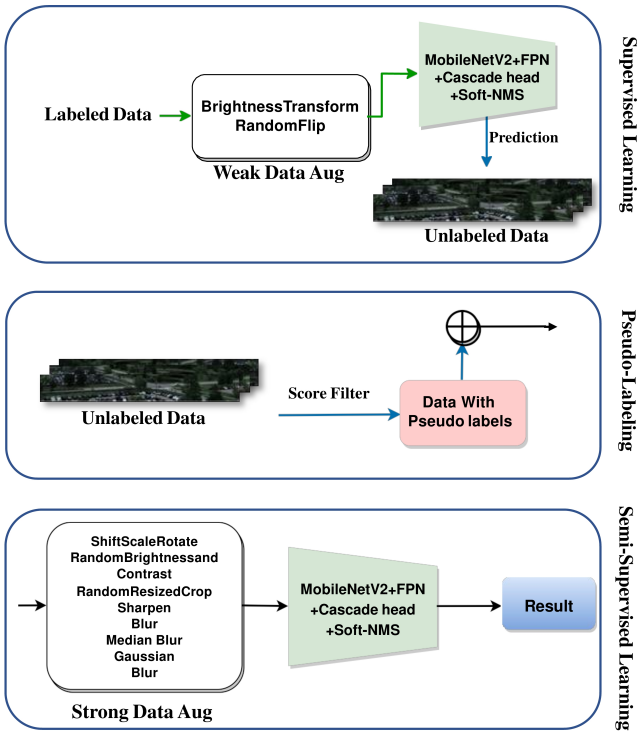


Figure 4. Framework: USTC-IAT-United

3. Proposed Approaches and Teams

This section briefly presents the approaches proposed by the different teams.

3.1. USTC-IAT-United

Figure (4) shows the USTC-IAT-United team's approach for this challenge. The authors modified the standard Faster R-CNN framework with a Cascade R-CNN [3], taking into account the computational complexity of backbone equivalent or lesser than MobileNetv2 [17] (in terms of parameters and GFlops). During training, the team used a multi-scale strategy, setting the scale to [(1600, 188),(1600, 189)]. The second phase of the training included training on pseudo labels obtained on the remaining unlabeled training set, by choosing the predictions from the model trained on the initially labeled set with confidence scores higher than 0.99. In the second stage of training, the team added much stronger data enhancement strategies than before, such as cutout [6], ShiftScaleRotate, RandomBrightnessContrast, and RandomResizedCrop. In the testing phase, the team used a multi-scale testing with Soft-NMS [2] to further improve the accuracy of the model. Table 2 shows that this approach overcomes the baseline resluts on using the entire labeled data, and we believe its primarily due to replacing Faster R-CNN framework with the Cascade R-CNN framework. All models were trained on Nvidia V100 GPUs, with

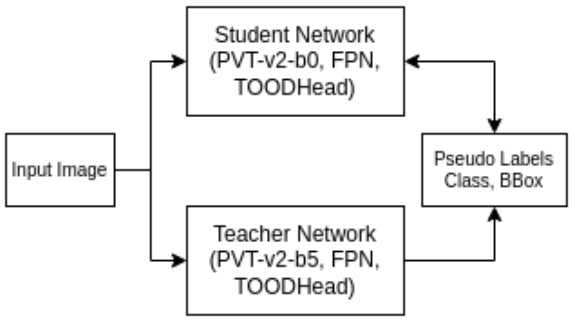


Figure 5. Framework: MSC-1

the MMdetection framework [4].

3.2. MSC-1

Figure (5) shows the MSC-1 team's approach for this challenge. The team trained a teacher-student network, where the teacher network was used to generate pseudo labels on the unlabeled data which is used to train the student network. They also modified the framework to address the class imbalance within the training dataset by adding a sampling technique to increase the frequency of occurrence for rare-classes (bus and bike). The team used a Pyramid Vision Transformer-b5 as the backbone encoder for the teacher network [23], with a modified feature pyramid network [18] and Task-aligned One-stage Object Detection (TOOD) block [7] as the detection head. The team used a PVT-B0 backbone, which is a lightweight version of the PVT family of networks, as the student network for their final submission to meet the computational requirements. All models were trained on Nvidia Titan RTX GPUs, with the MMdetection framework [4].

3.3. Discussion

The winning results are summarized in Table 2 and make two important observations. We observe, for the test set, that the Cascade R-CNN approach (USTC-IAT-United) is able to outperform the performance of a Faster R-CNN network that is trained with the entire set of labeled data, while struggling with the bike class, which is the most infrequently occurring class. The other entry (MSC-1), that uses PVT-B0 backbone, is able to outperform our baseline by using pseudo labels. However, since is not close to the USTC-IAT-United performance, we conclude that the modification of Cascade R-CNN is crucial for better results. We observed that both approaches used some form pseudo-labeling: USTC-IAT-United uses the same set of networks, while MSC-1 uses a relatively expensive backbone PVT-B5 for generating the predictions, and then train a lighter backbone PVT-B0 on the combination of labeled and pseudo-labeled data. Figs. 6, 7, 8 provide some examples of predictions from each of the submissions (USTC-IAT-United and MSC-1), our baseline and our fully-supervised approach for comparison, and discuss the most noticeable points in their captions.

4. Conclusion

In this paper, we introduce the RooftopHSI dataset, a first-of-its-kind dataset to benchmark hyperspectral object detection in realistic scenarios, that includes occlusion, deformations and changes due to weather conditions. We constructed the SSHOD challenge as a semi-supervised learning scenario, by providing labels for only 10% of the training data, and encouraging participants to use algorithms in semi-supervised learning for boosting performance. We hope our dataset and initial survey of methods will boost research in this area of designing frameworks that rely on hyperspectral features for object detection.

Acknowledgements

This work was supported by the Dynamic Data Driven Applications Systems Program, Air Force Office of Scientific Research (AFOSR) under Grant FA9550-19-1-0021. The development of the HSI system was made possible by Grant FA9550-15-1-0444 from the Air Force Office of Scientific Research (AFOSR) Defense University Research Instrumentation Program (DURIP).

Appendix A. Teams Information

SSHODC 2022 organization team:

Members: Aneesh Rangnekar[1] (aneesh.rangnekar@mail.rit.edu), Zachary Mulhollan[1], Anthony Vodacek[1], Matthew Hoffman[1], and Angel D. Sappa[2,3]

Affiliation: [1]Rochester Institute of Technology, Rochester, NY, USA, [2]ESPOL Polytechnic University, Guayaquil, Ecuador, [3]Computer Vision Center, Campus UAB, 08193 Bellaterra, Barcelona, Spain

A.1. USTC-IAT-United:

Members: Jun Yu[1], Liwen Zhang[1], Shenshen Du[1], Hao Chang[1], Keda Lu[1,2], Zhong Zhang[3], Fang Gao[4], Ye Yu[1], Feng Shuang[5], Lei Wang[1], Qiang Ling[1]

Affiliation: [1]University of Science and Technology of China, [2]Ping An Technology Co., Ltd., [3]Hefei ZhanDa Intelligence Technology Co., Ltd, [4]Guangxi University, [5]Hefei University of Technology

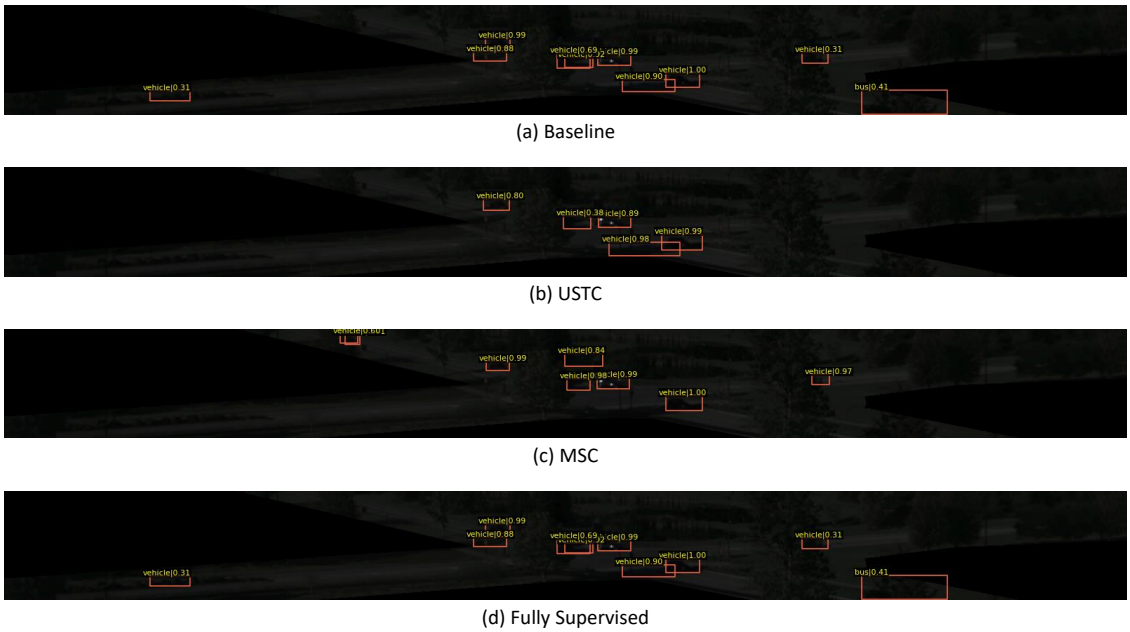
A.2. MSC-1:

Members: Pranjay Shyam, Kuk-Jin Yoon, Kyung-Soo Kim

Affiliation: Department of Mechanical Engineering, Korea

	Framework (No. of Params)	Validation set AP-Vehicle	AP-Bus	AP-Bike	AP	Test set AP-Vehicle	AP-Bus	AP-Bike	AP
Baseline	MobileNetv2 (14.11 M)	48.66	12.85	14.67	25.39	28.00	34.20	0.00	20.70
USTC-IAT -United	MobileNetv2 (31.68 M)	56.80	51.40	52.60	53.60	39.70	61.50	4.30	35.10
MSC-1	PVT-B0 (32.77 M)	49.50	39.80	31.90	40.40	31.80	56.30	0.20	29.40
Fully -Supervised	MobileNetv2 (14.11 M)	58.10	42.60	28.00	42.90	38.50	51.80	6.70	32.30

Table 2. Summary of results, comparing the baseline network (MobileNetv2-Faster-RCNN) trained on 10% data, to the two submissions (USTC-IAT-United and MSC-1), and the fully supervised version of our network trained with 100% of the training examples.



(a) Baseline

(b) USTC

(c) MSC

(d) Fully Supervised

Figure 6. (a), (c) have relatively confident false detections of bus, and only (c) is able to detect a heavily occluded vehicle around the trees, thus indicating a possible advantage of using a vision transformer backbone for hyperspectral object detection.

Advanced Institute of Science and Technology, Daejeon, Republic of Korea

References

[1] Boaz Arad and Ohad Ben-Shahar. Sparse recovery of hyperspectral signal from natural rgb images. In *European Conference on Computer Vision*, pages 19–34. Springer, 2016. 2

[2] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 5

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 5

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 5, 6

[5] Lulu Chen, Yongqiang Zhao, Jiaxin Yao, Jiaxin Chen, Ning Li, Jonathan Cheung-Wai Chan, and Seong G Kong. Object

(a) Baseline

(b) USTC

(c) MSC

(d) Fully Supervised

Figure 7. (b) predicts an object around the intersection that is consistent in its predictions in Fig. 8, and identical to a couple of bounding boxes around the same area in (d). (c) is yet again the only one to have an understanding of the object around the area, thus solidifying our observation from Fig. 6. This indicates a combination of PVT with Cascade R-CNN could provide best of the both circumstance results.

(a) Baseline

(b) USTC

(c) MSC

(d) Fully Supervised

Figure 8. (c) struggles to perfectly understand how a bus looks like, and this can be attributed to the patchwise workings of a vision transformer as the network struggles to grasp the overall shape. (b) has a relatively lower confidence for a vehicle partly occluded by tree as compared to others - this can also be attributed to a possible distribution bias in the number of occlusions present in the labeled examples, which may make training with pseudo labels difficult. This comes slightly as a surprise though, as the network trained with only 10% of the data, is able to detect the vehicle with a high confidence, as compared to the same network when adjusted for a complex framework and pseudo labels.

tracking in hyperspectral-oriented video with fast spatial-spectral features. *Remote Sensing*, 13(10):1922, 2021. 2

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 5

[7] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *ICCV*, 2021. 6

[8] Renlong Hang, Feng Zhou, Qingshan Liu, and Pedram Ghamisi. Classification of hyperspectral images via multitask generative adversarial networks. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 1

[9] Ronald Kemker and Christopher Kanan. Self-taught feature learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2693–2705, 2017. 1

[10] Zhuanfeng Li, Fengchao Xiong, Jun Zhou, Jing Wang, Jianfeng Lu, and Yuntao Qian. Bae-net: A band attention aware ensemble network for hyperspectral object tracking. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2106–2110, 2020. 1

[11] Tzuta Lin. labelimg. https://github.com/tzutalin/labelImg, 2015. 3

[12] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[13] Zhenqi Liu, Xinyu Wang, Meng Shu, Guanzhong Li, Chen Sun, Ziying Liu, and Yanfei Zhong. An anchor-free siamese target tracking network for hyperspectral video. In *2021 11th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2021. 1

[14] Xin Miao, Xin Yuan, Yunchen Pu, and Vassilis Athitsos. l-net: Reconstruct hyperspectral images from a snapshot measurement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4059–4069, 2019. 1

[15] Aneesh Rangnekar, Nilay Mokashi, Emmett J Ientilucci, Christopher Kanan, and Matthew J Hoffman. Aerorit: A new scene for hyperspectral image analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 2020. 2

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 5

[17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 5

[18] Pranjay Shyam, Kuk-Jin Yoon, and Kyung-Soo Kim. Dynamic anchor selection for improving object localization. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9477–9483. IEEE, 2020. 6

[19] Burak Uzkent, Matthew J Hoffman, and Anthony Vodacek. Real-time vehicle tracking in aerial video using hyperspectral features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 36–44, 2016. 2

[20] Burak Uzkent, Aneesh Rangnekar, and Matthew J Hoffman. Aerial vehicle tracking by adaptive fusion of hyperspectral likelihood maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 233–242. IEEE, 2017. 2

[21] Burak Uzkent, Aneesh Rangnekar, and Matthew J Hoffman. Tracking in aerial hyperspectral videos using deep kernelized correlation filters. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):449–461, 2018. 1, 2

[22] Anthony Vodacek, John P Kerekes, and Matthew J Hoffman. Adaptive optical sensing in an object tracking dddas. *Procedia Computer Science*, 9:1159–1166, 2012. 2

[23] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *ArXiv*, abs/2106.13797, 2021. 6

[24] Fengchao Xiong, Jun Zhou, Jocelyn Chanussot, and Yuntao Qian. Dynamic material-aware object tracking in hyperspectral videos. In *2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–6. IEEE, 2019. 2

[25] Fengchao Xiong, Jun Zhou, and Yuntao Qian. Material based object tracking in hyperspectral videos. *IEEE Transactions on Image Processing*, 29:3719–3733, 2020. 2

[26] Longbin Yan, Min Zhao, Xiuheng Wang, Yuge Zhang, and Jie Chen. Object detection in hyperspectral images. *IEEE Signal Processing Letters*, 28:508–512, 2021. 2

[27] Lei Zhang, Jiangtao Nie, Wei Wei, Yanning Zhang, Shengcai Liao, and Ling Shao. Unsupervised adaptation learning for hyperspectral imagery super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3073–3082, 2020. 1