SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Asymmetric Object-Background Latent Pretraining For Domain Adaptation

Arthur Toder[1*], Martin Perez-Izaguirre[2], Adrien Chan Hon Tong[3]

[1] DTIS, ONERA, Palaiseau, France. E-mail: arthur.toder@onera.fr, Orcid: https://orcid.org/0009-0006-4071-3867
[2] MBDA, Le Plessis-Robinson, France. E-mail: martin.perez-izaguirre@mbda-systems.com, Orcid: https://orcid.org/0009-0007-5405-9971
[3] DTIS, ONERA, Palaiseau, France. E-mail: adrien.chan_hon_tong@onera.fr, Orcid: https://orcid.org/0000-0002-7333-2765

Abstract
This article addresses the question of the consistency of unmanned aerial vehicle (UAV) detection under a strong change of environment. It presents a self-supervised pretraining strategy, applied to a state-of-the-art one-stage object detector, that decouples the learning of the objects of interest from the learning of the background. This greatly improves the detection of UAVs on urban backgrounds while requiring only rural UAV datasets.

Keywords: Computer vision, domain adaptation, object detection

I. INTRODUCTION

Unmanned Aerial Vehicles (UAVs) are becoming increasingly ubiquitous, raising new challenges for object detection. The ``drone'' class exhibits high intra-class variability, with large changes in scale, viewpoint, and instance count across images, making reliable detection difficult. An additional central challenge is the domain shift between training and deployment. While UAVs operate in diverse environments, data collection is often biased toward easier settings (e.g., rural areas), leaving urban scenarios underrepresented despite their practical importance. Consequently, models trained on a rural source distribution tend to generalize poorly to different urban target distributions, leading to significant performance degradation.

Such data drift can sometimes be mitigated by domain adaptation (Yaroslav & Victor, 2015), domain generalisation (Gauri, Ritvik, Keshav, & Ramesh, 2024), or data augmentation (Devries & Taylor, 2017). These methods aim to either align the source and target distributions or improve the model's robustness to unseen conditions, thereby enhancing generalisation performance in environments that differ from the training data. There is a large body of state-of-the-art work on this subject, but it mainly concerns transferring performance between general datasets (e.g., from COCO (Tsung-Yi, et al., 2014) to Objects365 (Shao, et al., 2019)).

However, these methods do not precisely address the use case of a UAV in a new background. To our knowledge, no method has considered domain adaptation with specifically different drifts between the object and the background.

The key contribution of this paper is to design a Self-Supervised Learning (SSL) method to deal with the background data drift common in UAV detection, which outperforms the state-of-the-art UAV detector by a large margin (22% to 39% mAP) on widely used UAV datasets in a transfer setting.

II. STATE OF THE ART

A. UAV DETECTION

Object detection aims to localize and classify objects within images. In UAV detection, this task is particularly challenging due to the diversity of UAV types and the large variation in apparent size caused by changes in distance, viewpoint, and resolution. Recent state-of-the-art methods are largely based on DETR (Nicolas, et al., 2020), which combines a CNN backbone with a transformer-based detection head. By replacing region proposal networks with learnable queries, DETR (Nicolas, et al., 2020) provides a unified, end-to-end framework. However, it requires large training datasets and struggles with small objects, although later variants such as Deformable DETR (Xizhou, et al., 2021) and DINO (Hao, et al., 2022) mitigate

these limitations. Multi-modal approaches have also recently emerged, incorporating language information into detection models (e.g., CLIP-based methods (Alec, et al., 2021)).
Despite this progress, UAV detection benchmarks such as the AVSS 2021 and Anti-UAV 2023 challenges show that CNN-based one-stage detectors remain dominant for this use case (Akyon, Eryuksel, Ozfuttu, & Altinuc, 2021), (Jie, Jingshu, Dongdong, & Dong, 2022). In particular, YOLO (Joseph & Ali, 2018) still offers a favourable trade-off between speed and accuracy for real-time applications, compared to slower two-stage methods like Fast R-CNN (Girshick, 2015). However, it is widely known that YOLO generalizes poorly to new settings, requiring domain adaptation techniques for real-life robustness.

B. Domain Adaptation

Domain adaptation was first studied in image classification (e.g. (Ganin, et al., 2016)), where early approaches learn domain-invariant feature representations to improve cross-domain generalization. Extending these ideas to object detection introduces additional challenges due to the need for precise spatial localisation. More recent work addresses domain shift in both classification and detection through a variety of strategies, typically depending on the severity of the shift and the application context. One line of methods introduces intermediate domains to progressively reduce large distribution gaps (Adrian Lopez & Krystian, 2019). While effective for extreme shifts (e.g., real-to-illustration), this approach is less suitable for subtler transitions such as rural-to-urban scenes. Data-level adaptation methods, such as ConfMix (Giulio, Luca, Elisa, & Yiming, 2022), instead blend source and target samples through image compositing and have shown effectiveness in autonomous driving scenarios, although they rely on explicit access to target data and assume relatively limited domain shifts.
Another family of approaches relies on student-teacher frameworks (Tarvainen & Valpola, 2017), where a teacher model trained on the source domain supervises a student adapted to the target domain without requiring labels. The advantage of these approaches is that they are source-free: there is no retraining on the source data. However, the performance relies heavily on the initial guess of the teacher. A large data drift can make these approaches useless. Also, these methods usually deal with background and foreground symmetrically.
Conversely, in this paper we propose to jointly integrate self-supervised learning, object detection, and domain adaptation in a unified framework that takes advantage of foreground-free target data at training time.

C. Self-Supervised Learning

Recent self-supervised learning (SSL) approaches are largely based on contrastive learning, which learns representations by pulling together augmented views of the same image (positive pairs) while pushing apart different images (negative pairs). However, these methods typically require carefully designed negative sampling and large batch sizes to structure the embedding space effectively. An alternative paradigm was introduced with BYOL (Grill, et al., 2020), which removes the need for negative pairs. BYOL (Grill, et al., 2020) employs an online network and a target network updated via exponential moving average. Given two augmented views of the same image, the online network is trained to predict the representation produced by the target network.
SSL has been shown to improve downstream object detection performance when used as pre-training (Dang, Kornblith, Nguyen, Chin, & Khademi, 2022). However, such approaches typically rely on data closely related to the source domain. Other works address domain shift using SSL or related teacher--student frameworks, particularly for adversarial backgrounds or cross-modality transfer (e.g., IR to visible) (Kim, et al., 2022). While related to our setting, where target-like samples may be integrated into training data, these methods differ as they rely on teacher-student consistency rather than direct representation learning.
Common SSL frameworks include MoCo (He, Fan, Wu, Xie, & Girshick, 2020), SimCLR (Ting, Simon, Mohammad, & Geoffrey, 2020), and SimSiam (Chen & He, 2021). In this work, we adopt SimSiam, which avoids negative pairs and instead learns invariance between two augmented views of the same image using a Siamese architecture.
Its design is particularly suitable for our setting, where defining negative pairs is difficult due to the lack of prior structure in the data. By relying solely on positive pairs generated via augmentation, SimSiam provides a stable and efficient representation learning framework without the need for explicit negative sampling.

III. Datasets Presentation

Datasets play a critical role in the performance and robustness of object detection architectures. For example, as shown in (Liu, Gao, Wan, Wang, & Lyu, 2023), a lack of representation for small objects in the training data often leads to poor detection performance on such targets. This issue highlights the importance of dataset design: insufficient coverage of target scales, backgrounds, or environmental conditions can severely limit the generalisation ability of deep models. Conversely, incorporating datasets

that provide diverse examples and adequate representation of small UAVs has been shown to significantly improve detection accuracy and robustness across scenarios (Nan Jiang and Kuiran Wang and Xiaoke, et al., 2021), (Coluccia, et al., 2024).

A. DATASETS PRESENTATION

1) UAV datasets: In this work, we construct a diverse training dataset by combining three publicly available datasets. This aggregation provides varied backgrounds, UAV scales, and sufficient data for robust training. Although all datasets originate from video sequences, we discard temporal information and treat each frame independently, in line with a standard image-based object detection setting. Throughout this paper, as UAV training datasets we use Drone versus Bird (DvB) (Coluccia, et al., 2024) and Anti-UAV (Nan Jiang and Kuiran Wang and Xiaoke, et al., 2021), and as an inference-only dataset, Det-Fly (Zheng, et al., 2021). These three datasets, shown in Figures 1 to 6, ensure a large diversity of backgrounds, UAV models, target sizes, and numbers of targets to be detected.

Figure 1 : EXAMPLE of RURAL in DvB

Figure 2 : FIXED POV from DET-FLY (URBAN)

Figure 3 : EXAMPLE of RURAL IMAGE in

Figure 4 : EXAMPLE of URBAN in DvB

Figure 5 : From ABOVE POV DET-FLY (URBAN)

Figure 6 : EXAMPLE of URBAN IMAGE IMAGE in ANTI-UAV

2) Splitting rural / urban for train and test sets: To evaluate the robustness of the algorithms to a strong change of environment, we created our own test-bench, represented in Figure 7 composed of three datasets based on the same protocol. The goal here is to create a challenging dataset with urban images, as this is the domain we chose to target.

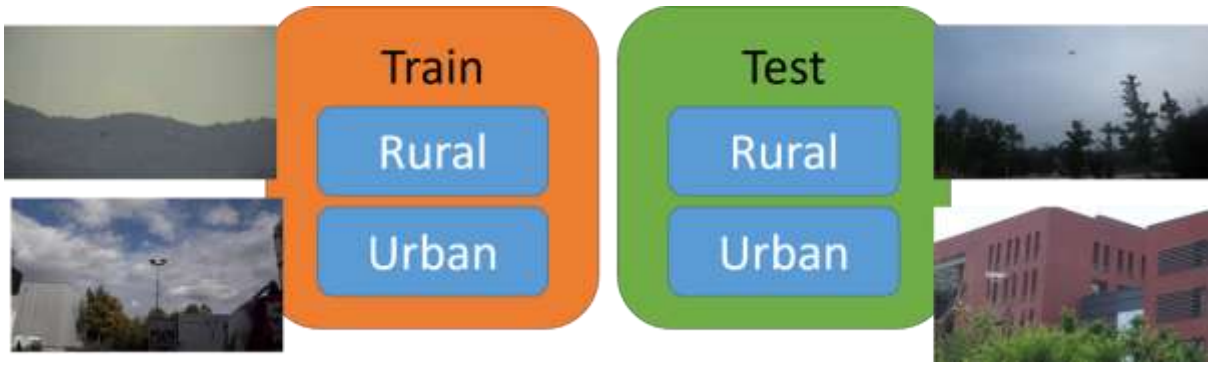

Figure 7 : Division of the UAV datasets in 4 differents sets

B. CITYSCAPES

In addition, we rely on a completely different kind of urban dataset. Cityscapes (Marius, et al., 2016) is a large-scale benchmark dataset widely used for semantic urban scene understanding. It is designed to support tasks such as semantic segmentation, instance segmentation, and object detection in the context of autonomous driving; therefore, it contains no UAVs. As a result, we can use it as background imagery.

To use it as a test set, we cut out drones from the three datasets presented earlier and insert them into the Cityscapes dataset, following Figure 8. Even if this technique can create insertion artefacts in the output datasets, it should not cause any problems, since these images are used only for inference and the network cannot learn these artefacts.

Figure 8 : Exemple of insertion of UAVs (left) and model of inserted UAVs (right)

IV. LATENT BACKGROUND PRETRAINING METHOD

For the proposed architecture, we focus on creating something robust to a change of environment. As mentioned in the introduction, a possible and simple solution would be to introduce empty backgrounds from the target environment, for example Cityscapes, into the training dataset, so that during training the architecture would see the target backgrounds but without any object to detect in them. However, this solution has some problems, which we describe below, and so we opt for another approach.

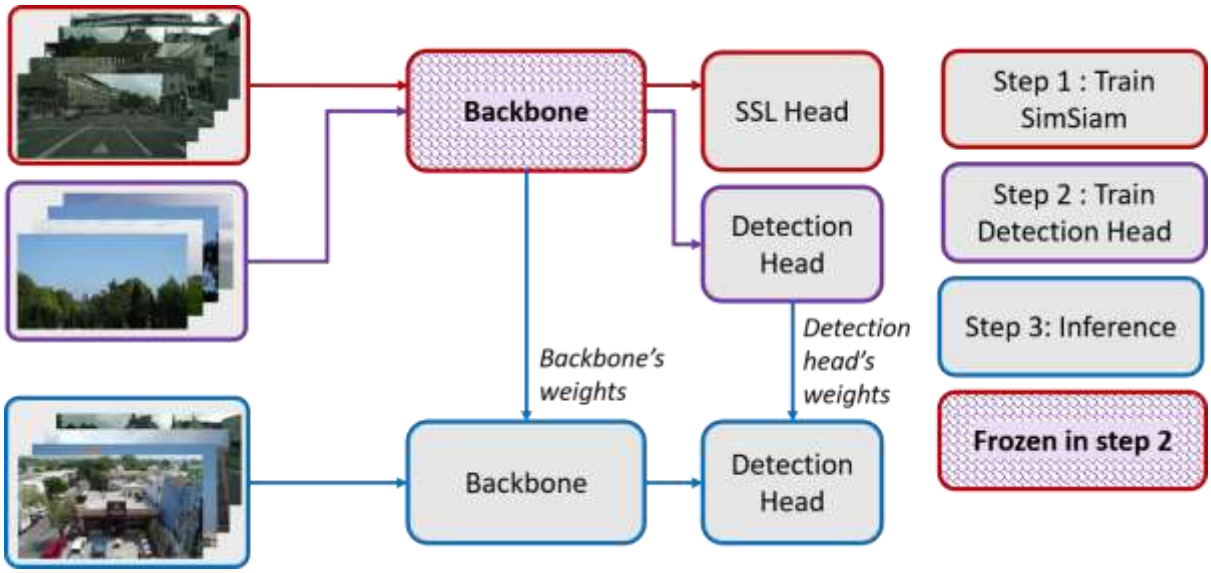

Figure 9 : Scheme of the presented architecture

Rather than introducing a new architecture, our contribution, Figure 9, is a training strategy aimed at improving robustness when a detector trained on a source domain is deployed on a different target distribution. The proposed method consists of three stages: two training phases followed by inference. It relies on two core components: a self-supervised learning (SSL) module and a detection network. We present the approach in a generic manner, independent of the specific SSL method or detector architecture, in order to emphasize its flexibility and compatibility with different backbones and training paradigms.

A. THE SSL TRAINING

The first stage of the proposed method consists of self-supervised learning (SSL), applied to the backbone of the detection network to construct a domain-robust embedding space. Following a SimSiam formulation, each image is augmented twice using standard transformations (e.g., cropping, rotation, colour jittering), producing two correlated views of the same sample. These views are encoded into embeddings by the backbone, and a lightweight prediction head is trained to align one representation with the other.

This stage is performed on a mixture of data drawn from the source distribution and images representative of the target environment (e.g., backgrounds or auxiliary sequences), while strictly excluding any inference images. The objective is not object recognition, but rather the learning of shared visual structure across domains, mainly driven by background information, in order to initialize a more domain-invariant feature space.

B. LEARNING TO DETECT OBJECTS

In the second stage, the detection head is trained while the backbone remains frozen. The detector is fine-tuned on labelled source-domain images containing UAV instances (e.g., rural sequences), enabling it to learn object-specific representations within the previously learned embedding space.

The key idea is that the detection head operates on features that already encode shared structure across domains, reducing sensitivity to background shifts. Consequently, the model learns to separate UAVs from the background using a representation space that is less dependent on the source environment, which is expected to improve generalization to unseen target domains.

Finally, inference is performed on previously unseen target data, including either synthetic composites or held-out datasets, with no overlap with the training stages.

V. EXPERIMENTATION

A. SCENARIO DESCRIPTION

The two main datasets are Drone versus Bird and Anti-UAV, which are divided into rural and urban sequences, as well as into train/test splits. The strong changes of environment are modelled as transferring performance from the rural training sets presented earlier to urban or Cityscapes inference.

As the UAV datasets are based on video sequences, and we do not want to split a single sequence between the train and test sets, the division is made according to complete sequences. Since some sequences could fit into both the rural and urban categories, we simply do not use them. COCO is also used as additional data in SSL training (no UAV).

Table I. PERFORMANCE (PRECISION (%), RECALL (%), mAP@0.5 (%)) of a UAV DETECTOR on THREE URBAN DATASETS CREATED by INSERTING DRONES from DRONE VS BIRD, ANTI-UAV and DET-FLY into CITYSCAPES (RESPECTIVELY C_{DvB}, $C_{Anti-UAV}$ and $C_{Det-Fly}$), TRAINED on RURAL SEQUENCES.

Method	C_{DvB} (known drones)			$C_{Anti-UAV}$ (known drones)			$C_{Det-Fly}$ (unknown drones)		
	P (%)	**R (%)**	**mAP (%)**	**P (%)**	**R (%)**	**mAP (%)**	**P (%)**	**R (%)**	**mAP (%)**
Rural only (1)	46.5	26.4	22.8	45.5	34.4	30.0	45.5	23.5	22.8
Simple adding (2)	70.5	19.0	21.8	49.3	8.9	10.3	54.7	13.5	14.5
With SSL (ours)	57.5	35.2	**34.8**	56.5	**34.3**	34.4	61.4	36.8	**39.4**

B. PERFORMANCE ON CITYSCAPES

The most important result is given in Table I. It shows the results of inference on the three datasets created by inserting a UAV into Cityscapes (cf. Figure 8). Our architecture, presented in the Latent Background Pretraining Method section, is compared to two references: a model trained only on the rural sequences of DvB and Anti-UAV (1), and a model trained on the rural sequences of DvB and Anti-UAV and, more importantly, also on Cityscapes images without any target (2).

It is important to remember that every version of YOLOv5 was pre-trained on COCO. Training only on rural data (1) provides a strong baseline but yields limited generalisation to urban composites. As expected, simply adding target-domain background images without objects (2) does not improve performance and often degrades mAP, despite improving exposure to the target visual domain. This drop is consistent with the observed reduction in recall, suggesting a bias toward predicting empty scenes when no objects are present during training. These results indicate that naive background adaptation is insufficient and may introduce a negative bias in object-presence estimation. In contrast, our SSL-based approach consistently improves performance across all settings, achieving the best mAP in each case. This suggests that SSL helps decouple scene understanding from object presence, enabling better utilisation of target-domain context without reinforcing the ``empty scene'' prior.

Table II. mAP of YOLOv5 TESTED on DATASETS with and without BUILDINGS, for SEVERAL SSL COMBINATIONS.

	wo SSL	**CS**	**CS+rural**	**CS+urban**	**CS+COCO**
Rural – test	66.8%	58.6%	67.1%	62.6%	**71.0%**
Urban - test	41.9%	**44.3%**	40.8%	37.5%	42.6%

C. PERFORMANCE ON URBAN UAV

To further study the influence of data drift, we consider performance in various intra-UAV-dataset settings, using only Cityscapes as a UAV-free background. Table II reports the impact of different SSL pre-training strategies on YOLOv5 performance in urban settings. Overall, SSL consistently improves performance over the baseline, with varying gains across domains and training compositions. Two configurations include UAV data during SSL (Cityscapes + rural/urban UAVs), where UAV instances appear during representation learning. However, this does not really improve the detection, suggesting that SSL mainly shapes the embedding space rather than directly benefiting object-specific recognition. On urban test sequences, SSL trained solely on Cityscapes performs best, while Cityscapes + COCO remains second and still improves over the baseline. This indicates that adding heterogeneous data does not uniformly transfer across domains.

VI. CONCLUSION

We developed a new method to counter the performance loss of common object detectors under strong background changes. We designed a strategy that takes advantage of SSL to create an embedding space based on empty images from the target distribution, which is then used in a second phase to train on images containing objects from the source distribution. Our method outperforms both pure transfers can deal with heavy background changes, contrary to common TTA or pseudo-labelling methods. This result is particularly important for UAV detection, where datasets are easier to collect in a rural context than for applications in an urban environment: in some settings, we outperform the baselines by 20% of mAP.

Références

1. Adrian Lopez, R., & Krystian, M. (2019). Domain Adaptation for Object Detection via Style Consistency.
2. Akyon, F. C., Eryuksel, O., Ozfuttu, K. A., & Altinuc, S. O. (2021). Track Boosting and Synthetic Data Aided Drone Detection. 2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 1–5.
3. Alec, R., Jong Wook, K., Chris, H., Aditya, R., Gabriel, G., Sandhini, A., . . . Ilya, S. (2021). Learning transferable visual models from natural language supervision (CLIP). International Machine Learning Society.
4. Chen, X., & He, K. (2021, June). Exploring Simple Siamese Representation Learning. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 15745-15753.
5. Coluccia, A., Fascista, A., Sommer, L., Schumann, A., Dimou, A., & Zarpalas, D. (2024). The Drone-vs-Bird Detection Grand Challenge at ICASSP 2023: A Review of Methods and Results. IEEE Open Journal of Signal Processing, 5, 766-779.
6. Dang, T., Kornblith, S., Nguyen, H. T., Chin, P., & Khademi, M. (2022, October). A Study on Self-Supervised Object Detection Pretraining. Computer Vision – ECCV 2022 Workshops, 86–99.
7. Devries, T., & Taylor, G. W. (2017). Improved Regularization of Convolutional Neural Networks with Cutout.
8. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., . . . Lempitsky, V. (2016, January). Domain-adversarial training of neural networks. The Journal of Machine Learning Research, 17(1), 2096–2030.
9. Gauri, G., Ritvik, K., Keshav, G., & Ramesh, R. (2024). Domain generalization via robust invariant representations.
10. Girshick, R. (2015). Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV), 1440-1448.
11. Giulio, M., Luca, Z., Elisa, R., & Yiming, W. (2022, Octobre). ConfMix: Unsupervised Domain Adaptation for Object Detection via Confidence-based Mixing.
12. Grill, J.-B., Strub, F. a., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., . . . Valko, M. (2020). Bootstrap your own latent a new approach to self-supervised learning. Proceedings of the 34th International Conference on Neural Information Processing Systems.
13. Hao, Z., Feng, L., Shilong, L., Lei, Z., Hang, S., Jun, Z., . . . Heung-Yeung, S. (2022). DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. International Conference on Learning Representations.
14. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020, June). Momentum Contrast for Unsupervised Visual Representation Learning. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 9726-9735.
15. Jie, Z., Jingshu, Z., Dongdong, L., & Dong, W. (2022). Vision-based Anti-UAV Detection and Tracking.
16. Joseph, R., & Ali, F. (2018). YOLOv3: An Incremental Improvement.

17. Kim, J., Huh, J., Park, I., Bak, J., Kim, D., & Lee, S. (2022). Small Object Detection in Infrared Images: Learning from Imbalanced Cross-Domain Data via Domain Adaptation. MDPI.
18. Liu, Z., Gao, X., Wan, Y., Wang, J., & Lyu, H. (2023). An Improved YOLOv5 Method for Small Object Detection in UAV Capture Scenes. IEEE Access, 11, 14365-14374.
19. Marius, C., Mohamed, O., Sebastian, R., Timo, R., Markus, E., Rodrigo, B., . . . Bernt, S. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3213-3223.
20. Nan Jiang and Kuiran Wang and Xiaoke, P., Xuehui, Y., Qiang, W., Junliang, X., Guorong, L., Jian, Z., . . . Zhenjun, H. (2021). Anti-UAV: A Large Multi-Modal Benchmark for UAV Tracking.
21. Nicolas, C., Francisco, M., Gabriel, S., Nicolas, U., Alexander, K., & Sergey, Z. (2020, August). End-to-end object detection with transformers (DETR). Computer Vision – ECCV: 16th European Conference, Glasgow, UK,, 213 - 229.
22. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., . . . Sun, J. (2019, October). Objects365: A large-scale, high-quality dataset for object detection. International Conference on Computer Vision (ICCV).
23. Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 1195–1204.
24. Ting, C., Simon, K., Mohammad, N., & Geoffrey, H. (2020). A Simple Framework for Contrastive Learning of Visual Representations. Proceedings of the 37th International Conference on Machine Learning(149).
25. Tsung-Yi, L., Michael, M., Serge, B., Lubomir, B., Ross, G., James, H., . . . Piotr, D. (2014). Microsoft COCO: Common objects incontext. Eur. Conf. Computer Vision (ECCV).
26. Xizhou, Z., Weijie, S., Lewei, L., Li, B., Xiaogang, W., & Jifeng, D. (2021). Deformable DETR: Deformable transformers for end-to-end object detection. International Conference on Learning Representations.
27. Yaroslav, G., & Victor, L. (2015). Unsupervised domain adaptation by backpropagation. Int. Conf. Machine Learning (ICML).
28. Zheng, Y., Chen, Z., Lv, D., Li, Z., Lan, Z., & Zhao, S. (2021). Air-to-Air Visual Detection of Micro-UAVs: An Experimental Evaluation of Deep Learning. IEEE Robotics and Automation Letters, 6(2), 1020-1027.