Optimizing Object Detection via Metric-driven Training Data Selection

Changyuan Zhou,* Yumin Guo, Qinxue Lv and Ji Yuan
Onewo Space-Tech Service Co., Ltd.
{zhoucy26, guoym14, lvqx02, yuanj36}@vanke.com

Abstract

*In the realm of object detection, training models with limited, unlabelled data from target domains presents significant challenges. This study focuses on the critical issue of optimizing image dataset selection to enhance object detection performance, especially when dealing with small sample sizes and closely-related target data that lacks predefined labels. Our proposed method adopts an integrated approach that combines data exploration, pseudo labeling, and strategic image selection from varied datasets, e.g. COCO and KITTI. By ranking source images based on their image-wise Average Precision (***AP***) scores followed by mosaic augmentation on selected images, experimental results demonstrate the efficiency of this data selection mechanism, indicating significant advancements in object detection performance and domain adaptability. Our method won the 2nd DataCV Challenge with the* **AP** *of 0.2285, achieving a 0.052* **AP** *increase over the baseline method. This work offers a robust pathway to overcome key challenges in applying object detection models across various domains, particularly in scenarios with limited annotations from target set. Our codes have been available at: https://github.com/welovecv/datacv.*

1. Introduction

Object detection represents for a cornerstone in the field of computer vision, with widespread applications ranging from autonomous driving to surveillance systems. The effectiveness of object detection models, such as RetinaNet [14], heavily relies on the quality and diversity of the datasets used for training. Typically, these models are trained and evaluated on benchmark datasets like COCO [13] and KITTI [6], which are meticulously annotated and curated to represent a wide array of scenarios and object categories. In practice, acquiring annotated data from the target domain is often impractical due to time, privacy, or cost constraints, rendering traditional data labeling approaches unfeasible.

In practical applications, a significant challenge arises when deploying these models in real-world environments: the discrepancy between the data characteristics of the source (training) and target (deployment) domains, often leading to degraded performance. This phenomenon, known as domain shift, has spurred research into domain adaptation techniques, aiming to bridge the gap between source and target datasets without requiring extensive labeling efforts for the latter [4, 12, 24]. Traditionally, domain adaptation has focused on modifying the model or its learning process to better align with the target domain's characteristics. While effective, these methods can be complex and computationally intensive, requiring substantial tuning and expertise.

Semi-supervised learning [9, 23, 26] aims to address the challenges posed by limited labeled data. These strategies leverage a small amount of labelled data alongside a larger volume of unlabelled data to train models, potentially offering a partial solution to the annotation scarcity in real-world applications. However, while semi-supervised learning provides a valuable framework for utilizing unlabelled data, its effectiveness is inherently dependent on the relevance and quality of the available labelled data. In the context of object detection, where the domain shift can be particularly pronounced due to variations in scene composition, lighting, and object scales, the standard semi-supervised approaches may fall short without careful selection and use of the training images [15, 19].

Parallel to domain adaptation, active machine learning has emerged as a paradigm to optimize the training process by selectively querying the most informative data points [2, 16]. In the context of object detection, this translates into identifying and utilizing the images that would most improve the model's performance if added to the training set. However, active learning typically depends on existing annotations or additional labeling, posing a challenge for its application in real-world scenarios.

In response to these challenges, our research proposes a novel integration of domain adaptation and active learning principles, aimed at enhancing RetinaNet's performance.

[*] Corresponding author

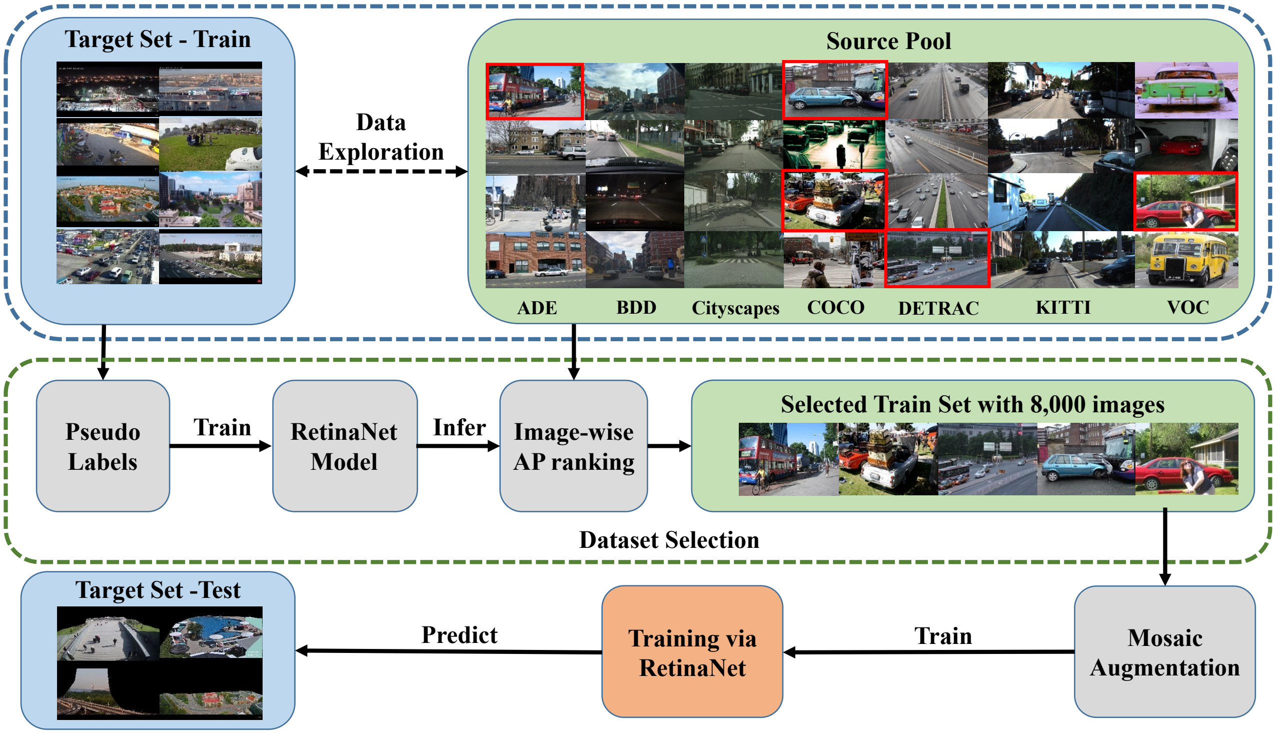


Figure 1. The flowchart of our method for curating a targeted training subset from a diverse source pool to enhance object detection performance using RetinaNet. The methodology commences with data exploration to comprehend the intrinsic characteristics of the target training set, followed by employing the pre-trained RetinaNet model to infer pseudo labels for unsupervised data alignment. Utilizing image-wise Average Precision (**AP**) ranking, the most pertinent subset is selected, which is then augmented through mosaic augmentation techniques to enrich feature representation. The approach culminates in applying the refined model to the target test set, demonstrating the efficacy of strategic dataset curation and augmentation in transfer learning scenarios.

By leveraging unlabelled target data and a strategic image selection based on image-wise average precision (**AP**) rankings, we circumvent the need for additional annotations while effectively addressing the domain shift. This method not only simplifies the adaptation process but also ensures the training focuses on the most relevant and representative samples from the source dataset. Our approach signifies a pragmatic step forward in the application of object detection models, paving the way for more adaptable and efficient solutions in diverse operational environments.

The main contributions of this work include:

- Introduce a novel method that utilizes image-wise AP for selecting the most relevant training data, and experimental results demonstrate its advantage.
- Implement an analysis of bounding box distributions to guide the selection of data sources, thereby ensuring a comprehensive representation of object variations within the training set.
- Employ mosaic augmentation on strategically chosen images to increase the training data's variance, effectively broadening the model's exposure to diverse scenarios and enhancing its generalization capabilities in reality.

2. Related Work

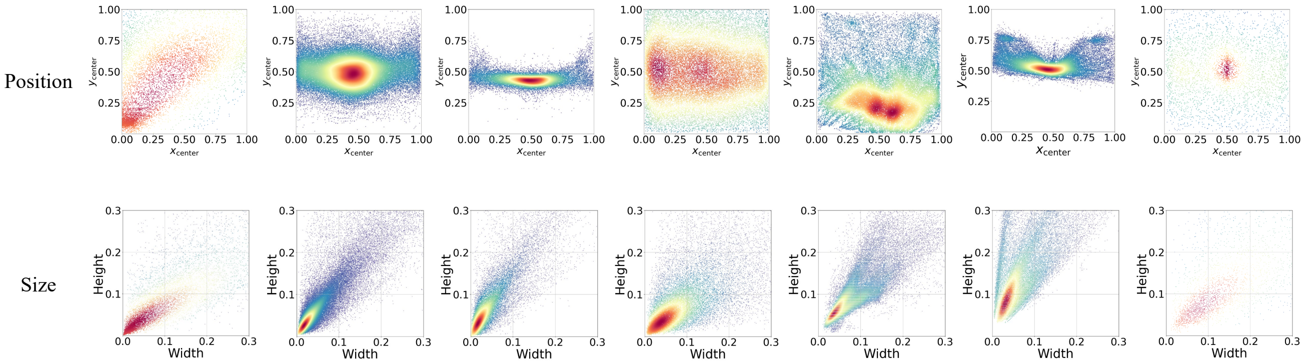
2.1. Clustering Methods

The primary goal of clustering analysis is to reveal the inherent structures and relationships among data points. Within this paper, the partition-based clustering algorithm is selected for detecting data similarity.

K-Means Clustering [7] is a classic partition-based clustering algorithm that operates with a given parameter K: (1) Randomly initialize K cluster centroids; (2) Assign each data point to the nearest centroid, thus forming clusters; (3) Calculate the mean of each cluster and set it as the new centroid. Steps (2) and (3) are iterated until a termination condition is met (e.g., no data points change clusters). In the SnP Framework[21], K-Means is applied to cluster the feature of the source dataset.

2.2. Search and Pruning Framework

The Search and Pruning Framework (SnP) [21] proposed in the baseline is a method for searching training dataset. The classical SnP approach consists of two main steps: Target-specific Subset Search and Budget-



(a) ADE (a) BDD (c) Cityscapes (d) COCO (e) DETRAC (f) KITTI (g) VOC

Figure 2. A composite of several scatter plot matrices from the perspectives of position and size for source datasets. COCO dataset presents significant diversities for both metrics, which is ideal for building a balanced dataset for the object detection compared with the others.

constrained Pruning. These steps collaboratively select a training set from the source pool. The selected set is both budget-compliant and sufficiently similar to the target set.

Target-specific Subset Search initially employs the InceptionV3 network to extract features from all images in the source pool and the target set. Subsequently, K-Means clustering is applied to the images in the source pool, dividing them into K clusters. The Fréchet Inception Distance (FID) [8] from each cluster to the target set is then computed. The clusters are sorted in ascending order by FID and progressively merged to construct S^* until the FID between S^* and the target set no longer decreases.

Budget-constrained Pruning aims to trim the candidate training set S^* to obtain an efficient training set D_S that adheres to budget constraints $b = (n, m)$, where n and m represent the upper limits for the number of identities and images, respectively. All images under n identities are initially selected from S^* to form a subset $\hat{S}$. A sampling method that minimizes the difference between D_S and $\hat{S}$ is then applied to derive the final training set D_S with m images.

2.3. Object Detection

RetinaNet [14] and the YOLO [17] series, specifically YOLOv5 [10] and YOLOv8 [11], have been influential in the domain of object detection. RetinaNet is notable for its innovative use of Focal Loss, which effectively addresses the class imbalance problem in object detection by mainly focusing on hard-to-detect objects, thus significantly improving detection accuracy. The YOLO series, known for its real-time detection capabilities, has evolved from its initial version to YOLOv5 and YOLOv8. These later versions incorporate advanced techniques like Cross Stage Partial networks (CSP) and Path Aggregation Networks (PAN), enhancing both the efficiency and performance of the models. This progression marks a significant shift towards more practical and deployable object detection solutions.

3. Problem Statement and Method

3.1. Problem Statement

The major task of this competition[1, 21] is to search small-scale, yet highly effective training sets from a large-scale data pool such that a competitive target-specific model can be obtained. We need to select a subset D_S from the Source S based on the relationship $R_{S,T}$ we discovered between Source S and Target T. It can be formulated as an optimization issue, with the following structure:

$$\begin{aligned} \underset{D_S, R_{S,T}}{\arg\max} \quad & P_M(D_S, T|R_{S,T}) \\ \text{s.t.} \quad & D_S \in S \\ & |D_S| \leq C, \end{aligned} \tag{1}$$

where P_M is defined as the performance of detection model M which is trained on D_S with a budget C.

3.2. Distance vs. Image-wise Average Precision

In this challenge, the core problem is determining a metric between dataset from source pool and target pool.

Fréchet Inception Distance (FID)[8] is commonly employed to evaluate the distance between two distributions, which can be expressed as:

$$\text{FID} = ||\mu_S - \mu_T||^2 + \text{Tr}(\mathbf{\Sigma}_S + \mathbf{\Sigma}_T - 2(\mathbf{\Sigma}_S\mathbf{\Sigma}_T)^{1/2}) \tag{2}$$

in which μ_S and μ_T are derived from the activation of an Inception-v3 network [18] when applied to two different sets of images, $\mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_T$ are the covariance matrices of

Dataset	AP	AP^{50}	AP^{75}	AP^{S}	AP^{M}	AR^{L}	AR^{S}	AR^{M}	AR^{L}
ADE[25]	0.098	0.241	0.064	0.004	0.097	0.309	0.028	0.177	0.425
BDD[22]	0.078	0.188	0.054	0.001	0.080	0.277	0.029	0.149	0.038
Cityscapes[3]	0.070	0.176	0.044	0.003	0.061	0.254	0.021	0.121	0.359
COCO[13]	**0.119**	**0.295**	0.079	0.006	**0.124**	**0.343**	**0.041**	**0.226**	**0.475**
DETRAC[20]	0.088	0.220	0.056	0.006	0.099	0.244	0.004	0.162	0.398
KITTI[6]	0.060	0.158	0.037	**0.007**	0.065	0.177	0.006	0.109	0.336
VOC[5]	0.064	0.141	0.047	0.005	0.058	0.240	0	0.092	0.421
ALL	0.113	0.267	**0.081**	0.002	0.113	0.339	0.030	0.218	0.470

Table 1. Comparison of experimental results obtained by randomly selecting 1,990 images from each corresponding dataset to use as a training set.

Dataset	Images	Boxes
ADE[25]	3,048	12,543
BDD[22]	68,943	700,703
Cityscapes[3]	2,831	26,929
COCO[13]	12,251	43,867
DETRAC[20]	82,266	503,853
KITTI[6]	7,481	40,570
VOC[5]	1,990	4,008

Table 2. Statistical comparison among datasets from source pool from the perspective of numbers of images and bounding boxes. The total number of images is 178,810.

the activation for the two sets of images, i.e. source and target. The trace operation, Tr(·), sums the diagonal elements of a matrix.

Average Precision (AP) is a widely utilized evaluation metric for the object detection, which involves calculating precision and recall by intersection of union of predicted bounding and ground truth boxes under different thresholds, which is denoted as:

$$AP = \frac{1}{N} \sum_{r \in \mathcal{R}} P_{\text{interp}}(r), \tag{3}$$

in which $P_{\text{interp}}(r)$ denotes the interpolated precision at recall level r, R denotes the set of recall values ranging from 0 to 1 with increments of 0.1, and N represents the total number of recall levels, namely 11 in our case [5].

4. Experiment

4.1. Dataset Description

In this competition [1], source pool includes more than 170,000 images from ADE [25], BDD [22], Cityscapes[3], COCO [13], DETRAC [20], KITTI [6] and VOC datasets [5]. Statictical distribution of source pool is presented from Table 2.

To better understand the data for dataset source selection, we undertake localization distribution comparison by sampling 1,990 images from these datasets. In this experiment, the normalized centers of bounding boxes are taken into consideration, the result of which is demonstrated from the Figure 3.

The target dataset, Region100, which consists of images from 100 static cameras around the world, poses challenges to the participants. To begin with, the angle of view, brightness and image quality varies from different cameras, and inconsistency affects feature extraction, as models must adapt to diverse conditions. Additionally, the dataset features densely vehicles, complicating accurate localization, especially for remote objects. Moreover, manual erasing manipulations on test images can lead to discrepancies between training and testing data, affecting model performance. Addressing these challenges requires innovative training approaches to ensure adaptability across varied imaging conditions.

4.2. Dataset Selection

Our method consists of several stages. First, we select a data source from seven datasets (see Table 2) by analyzing the distribution of bounding boxes. Our experiments indicate that the COCO dataset exhibits an even distribution in terms of localization and size, as depicted in Figure 2.

To compare the efficiency of difference datasets, we randomly select 1,990 images from each source dataset and conduct training based on each dataset respectively.

4.3. Implementation Details

In this scenario, where multiple labeled datasets are available and the target data is unlabeled, one approach for image selection involves extracting features and filtering to retain the top 8000 images from the source dataset that are closest in distance to the target set.

Another approach employs the transductive learning paradigm [27], where the unlabeled data is labeled using

Method	**AP**	$\textbf{AP}^{50}$	$\textbf{AP}^{75}$	$\textbf{AP}^{S}$	$\textbf{AP}^{M}$	$\textbf{AR}^{L}$	$\textbf{AR}^{S}$	$\textbf{AR}^{M}$	$\textbf{AR}^{L}$
Random	0.189	0.427	0.146	0.025	0.212	0.437	0.095	0.316	0.551
Random+Mosaic	0.206	0.463	0.159	0.021	0.226	0.472	0.101	0.315	0.548
SnP	0.209	0.460	0.164	0.029	0.232	0.467	0.111	0.339	0.572
SnP+Mosaic	0.218	0.488	0.168	0.029	0.238	0.487	0.119	0.332	0.564
AP	**0.230**	0.477	**0.191**	**0.035**	**0.252**	**0.520**	0.096	**0.348**	**0.610**
AP+Mosaic	0.226	**0.500**	0.178	0.028	0.248	0.502	**0.126**	0.342	0.579

Table 3. Comparison of inference scores across different methods(trained on COCO[13]) on TestA (maxDets=1000).

Method	**AP**	$\textbf{AP}^{50}$	$\textbf{AP}^{75}$	$\textbf{AP}^{S}$	$\textbf{AP}^{M}$	$\textbf{AR}^{L}$	$\textbf{AR}^{S}$	$\textbf{AR}^{M}$	$\textbf{AR}^{L}$
SnP-baseline	0.176	0.386	0.141	0.016	0.211	0.440	0.087	0.304	0.544
Random-COCO	0.188	0.420	0.147	0.023	0.218	0.430	0.100	0.323	0.541
AP-COCO	0.226	0.466	**0.192**	**0.032**	0.251	**0.525**	0.090	**0.351**	**0.619**
AP-COCO+Mosaic	**0.228**	**0.494**	0.184	0.031	**0.254**	0.509	**0.121**	0.347	0.582

Table 4. Comparison of inference scores across different methods on TestB (maxDets=100).

a pretrained model. As shown in Figure 1, the framework of our work mainly comprises of two parts, dataset selection and mosaic augmentation on the selected images. We introduced YOLOv8x model on the train part of the target set to generate pseudo labels to estimate distribution of the entire target set. We then train a RetinaNet model based on pseudo labels to build a metric to evaluate image-wise similarity from the object detection region. Since ground truth from source set is available, image-wise AP is adopted as the metric for image selection.

We follow the hyper-parameter setting of RetinaNet training. The backbone of the model is ResNet-50, and IoU threshold is set as 0.5. Input images are resized into 1,333×800 with the original ratio. The training protocol initiates with a warmup learning rate for the initial 500 iterations, followed by an adjustment to 1e-3 before the commencement of the 8th epoch. The learning rate is further reduced to 1e-4 at the subsequent two epochs and finally to 1e-5 at the last epoch. For the experiments for dataset selection, batch size is set as 4, whereas, for other experiments, it is reduced to 2.

4.4. Main Results

In evaluating different datasets for the object detection, the COCO dataset stands out, as evidenced by our experimental results, as shown in Table 1. It exhibits superior Average Precision (**AP**) and Average Recall (**AR**), particularly in $\textbf{AP}^{50}$ and large object recall. The COCO dataset's comprehensive annotations and diverse image collection make it an ideal source dataset, facilitating the development of robust and generalized object detection models. Its performance across varied metrics underscores its effectiveness and justifies its selection as the preferred dataset for enhancing object detection methodologies.

Table 3 and 4 show performance evaluation of several methods on TestA and TestB, respectively. "Random" denotes randomly selecting 8,000 images from COCO dataset, and "SnP" takes target-train datasets, namely from target to find out the training image dataset with the nearest distances. Note that our proposed methods, i.e. "**AP**" and "**AP**+Mosaic", outperform others on the majority of of evaluation metrics.

Experimental results underscore the advantages of combining **AP** with Mosaic augmentation: a substantial boost in detection accuracy, especially for larger objects, and a balanced improvement across varying object sizes and detection thresholds. This hybrid approach proves to be highly effective for enhancing the robustness and accuracy of object detection models. Figure 3 demonstrates advantage of our proposed methods when comparing with the baseline approach.

4.5. Ablation Study

Within this comparative analysis, we examine the impact of mosaic augmentation on the spatial distribution of bounding box (bbox) center points within image datasets, visualized through heatmaps. The original dataset's distribution [Figure 4 (a)] indicates a pronounced central concentration of bounding box centers, suggesting a spatial bias that may hinder the performance of object detection models.

Implementing mosaic augmentation, a technique that combines multiple images into a single training sample, results in a markedly altered distribution [Figure 4 (b)]. This

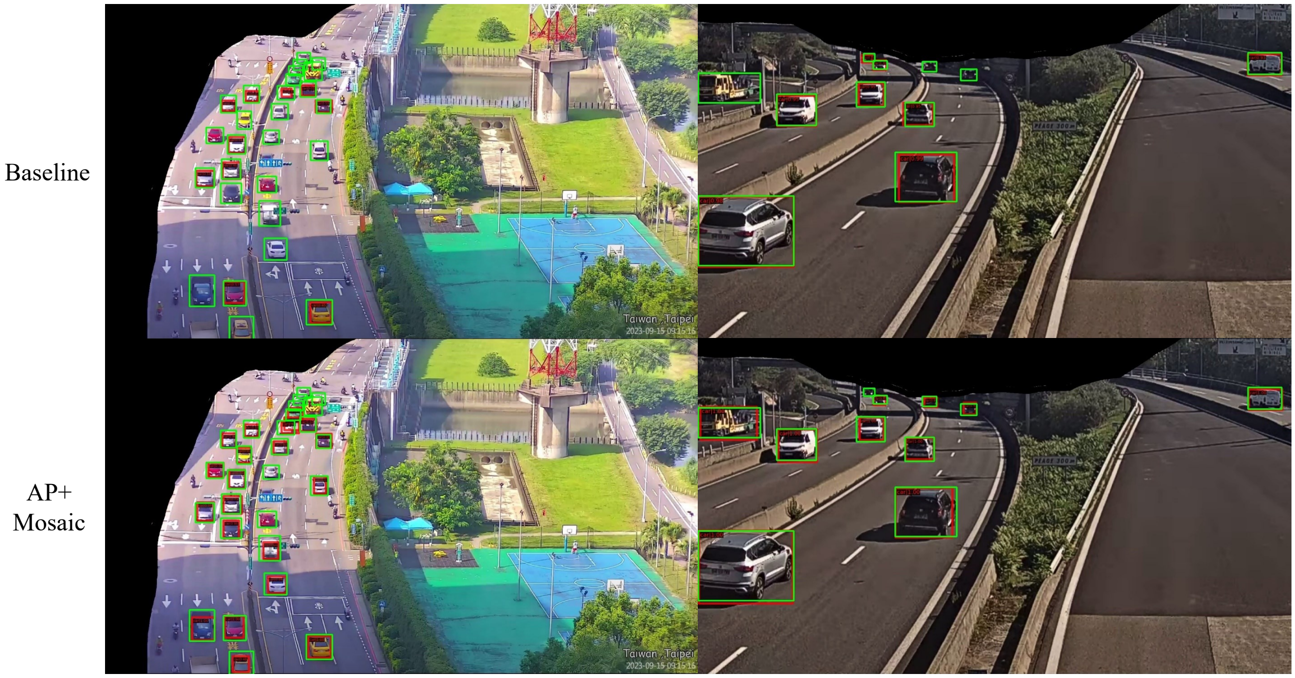


Figure 3. Visualization of predictions on sample images from testA via baseline model and AP+Mosaic model. Green bounding boxes denote groundtruth while red ones are predictions with corrresponding models.

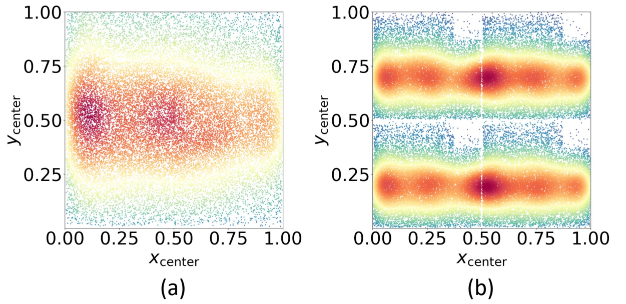


Figure 4. Comparison of bounding box center distributions before and after applying the mosaic augmentation. (a) shows the original dataset distribution, while (b) demonstrates the diversified object locations resulting from the mosaic augmentation.

method not only augments the dataset size but also enhances object location diversity. The post-augmentation heatmap reveals a distinct quadripartite structure, reflecting the mosaic's composition, thereby significantly expanding the diversity of object positions and orientations.

The advantage of mosaic augmentation is twofold: it mitigates the central bias prevalent in many datasets and introduces a wider array of contextual scenarios, thereby enabling models to learn from a more varied set of examples. This approach is particularly beneficial for improving model robustness and generalization, as evidenced by the more uniform distribution across the entire image space, ensuring enhanced object detection across diverse spatial contexts.

5. Conclusion

This work presents the innovative approach for optimizing object detection models by refining the dataset selection process. By leveraging image-wise Average Precision (**AP**) for dataset curation, we ensure the inclusion of the most impactful images, leading to a more relevant and focused training set. Our methodology integrates a detailed analysis of bounding box distributions, facilitating informed decisions that enhance the robustness and diversity of object scenarios within the dataset. The implementation of mosaic augmentation further enhances this effect, expanding the training data's variance and thus, the model's adaptability to diverse real-world situations. Experimental results affirm the advantage of our proposed strategy, particularly in improving detection precision across various object sizes and IoU thresholds. Future work will focus on exploiting fusion of selected images and augmented images.

References

[1] The 2nd datacv challenge @ cvpr 2024. In *https://sites.google.com/view/vdu-cvpr24/competition*. 3, 4

[2] Jiwoong Choi, Ismail Elezi, Hyuk-Jae Lee, Clement Farabet, and Jose M Alvarez. Active learning for deep object detection via probabilistic modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10264–10273, 2021. 1

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 4

[4] Jinhong Deng, Wen Li, Yuhua Chen, and Lixin Duan. Unbiased mean teacher for cross-domain object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4091–4101, 2021. 1

[5] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 4

[6] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 4

[7] Ja Hartingan and Maurice K. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 28(1):100–108, 1979. 2

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 3

[9] Jisoo Jeong, Vikas Verma, Minsung Hyun, Juho Kannala, and Nojun Kwak. Interpolation-based semi-supervised learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11602–11611, 2021. 1

[10] Glenn Jocher. YOLOv5 by Ultralytics, 2020. 3

[11] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLO, 2023. 3

[12] Yu-Jhe Li, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Kan Chen, Bichen Wu, Zijian He, Kris Kitani, and Peter Vajda. Cross-domain adaptive teacher for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7581–7590, 2022. 1

[13] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 4, 5

[14] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 3

[15] Liang Liu, Boshen Zhang, Jiangning Zhang, Wuhao Zhang, Zhenye Gan, Guanzhong Tian, Wenbing Zhu, Yabiao Wang, and Chengjie Wang. Mixteacher: Mining promising labels with mixed scale teacher for semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7370–7379, 2023. 1

[16] Anant Raj and Francis Bach. Convergence of uncertainty sampling for active learning. In *International Conference on Machine Learning*, pages 18310–18331. PMLR, 2022. 1

[17] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2016. 3

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 3

[19] Xinjiang Wang, Xingyi Yang, Shilong Zhang, Yijiang Li, Litong Feng, Shijie Fang, Chengqi Lyu, Kai Chen, and Wayne Zhang. Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3240–3249, 2023. 1

[20] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 193:102907, 2020. 4

[21] Yue Yao, Tom Gedeon, and Liang Zheng. Large-scale training data search for object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15568–15578, 2023. 2, 3

[22] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 4

[23] Jiacheng Zhang, Xiangru Lin, Wei Zhang, Kuo Wang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Semi-detr: Semi-supervised object detection with detection transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23809–23818, 2023. 1

[24] Yangtao Zheng, Di Huang, Songtao Liu, and Yunhong Wang. Cross-domain object detection through coarse-to-fine feature adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13766–13775, 2020. 1

[25] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 4

[26] Qiang Zhou, Chaohui Yu, Zhibin Wang, Qi Qian, and Hao Li. Instant-teaching: An end-to-end semi-supervised object detection framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4081–4090, 2021. 1

[27] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018. 4