Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Advances in Weakly Supervised Object Detection: Leveraging Unlabeled Data for Enhanced Performance

Hao Chen [1], Sicheng Lei [2, *], Zhengliang Lyu [3] and Naitian Zhang [4]

[1] Information Engineering College, Chengdu University of Information Technology, Chengdu, China

[2] School of Integrated Circuits, Huazhong University of Science and Technology, Wuhan, China

[3] School of artificial Intelligence, China University of Mining and Technology-Beijing, Beijing, China

[4] Bangor, Central South University of Forest and Technology, Changsha, China

* Corresponding Author Email: katiekarner1@email.phoenix.edu

Abstract. Weakly supervised object detection represents a burgeoning field within the realm of computer vision, reflecting the growing interest in developing models that can effectively identify and classify objects with minimal labeled data. This paper offers a comprehensive classification of contemporary, state-of-the-art deep learning models tailored for weakly supervised target detection. The classification encompasses four principal categories: Multi-Instance Learning (MIL), Class Activation Mapping (CAM), Deep Weakly Supervised Learning leveraging Attention Mechanisms, and Weakly Supervised Object Detection employing Pseudo-labels. Each category represents a unique approach to the challenge of discerning and localizing objects with limited supervision, emphasizing different aspects of learning from sparse or imprecise annotations. Our analysis delves into the intricate methodologies and theoretical foundations underlying these models, offering insights into their practical applications and performance metrics. Furthermore, we explore the evolutionary trajectory of these techniques, highlighting their advancements and the pivotal role they play in advancing the frontiers of automated object detection in diverse and complex environments. This synthesis not only charts the current landscape of weakly supervised object detection but also paves the way for future research directions in this dynamic and rapidly evolving field.

Keywords: Weak supervised object detection model algorithm.

1. Introduction

In traditional object detection methodologies, the requirement for extensive manual labeling presents a significant challenge, demanding considerable time and labor. However, weakly supervised object detection provides an innovative solution by leveraging image-level labels or partial labeling to train models capable of interpreting pixel-level details. This approach not only reduces the burden of extensive manual annotation but also opens avenues for more scalable and efficient model training. The upcoming paper delves into this paradigm shift in object detection. It offers a comprehensive exploration encompassing the underlying theories that power weakly supervised learning, a detailed review of datasets optimized for this approach, and an analysis of the evaluation metrics that are pivotal in assessing the performance of these models. This critical overview aims to bridge the gap between the labor-intensive traditional methods and the more efficient, weakly supervised techniques, presenting a cohesive understanding of the advancements in the field and their implications for future research and applications.

2. Relevant theories

2.1. WSDDN (Weakly Supervised Deep Detection Network)

In 2016, Bilean et al. introduced an innovative approach by amalgamating multi-instance learning with deep convolutional neural networks, culminating in an end-to-end system [1, 2]. This breakthrough was significant in the field of weakly supervised deep learning. The most notable

advancement of this technique, known as the Weakly Supervised Deep Detection Network (WSDDN), lies in its dual-stream architecture [3, 4]. This dual-stream network elegantly handles the complexities inherent in learning from weak labels, representing a leap forward in object detection methodologies. Then, the spatial pyramid pool layer was used to extract the candidate region feature F_r.After the feature map F_r is processed by two fully connected layers, the output subsequent region feature vector F_k is imported into two parallel data streams.

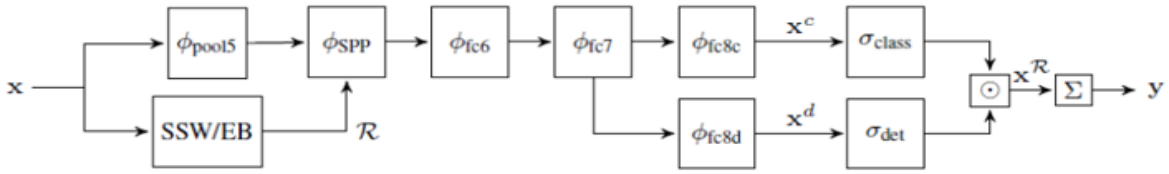

Fig. 1 The network structure of WSDDN (Photo/Picture credit: Original).

As can be seen from the figure 1, WSDDN has two data stream branches, one is classification data stream(fc8c) and the other is detection data stream(fc8d).As the name suggests, the function of the classification data stream is to classify the candidate region, and the output of the data stream is the category probability of each candidate region; The detection data stream is to detect whether the candidate region contains objects, and the output of the data stream is the confidence of the presence of objects in the candidate region. The difference between the two is reflected in that softmax operation is performed at different latitudes to obtain σ_{class} and σ_{det}, and the calculation process is as follows.

$$[\sigma_{class}(X^c)]ij = \frac{X^c_{i,j}}{\sum_{k=1}^{C} X^c_{i,k}}; [\sigma_{class}(X^d)]ij = \frac{X^d_{i,j}}{\sum_{k=1}^{|R|} X^d_{k,j}} \quad (1)$$

The above two formulas are both softmax functions, the difference between them is that for the input of n*c. When calculating σ_{class}, softmax is calculated along the second dimensionc. And in calculating σ_{det}, the softmax value is calculated along the first-dimension n. After the two data are obtained, use Hadamard multiplication (element-by-element multiplication), and then calculate the final data for the entire image along the dimensions of the region box, as shown in the following formula.

$$X^r = \sigma_{class}(X^c) \odot \sigma_{det}(X^d) \quad (2)$$

$$y = \sum_{r=1}^{|R|} X^r \quad (3)$$

Then, the whole WSDDN network is completed. At this time, the category scores of the obtained image are σ class, the detection score is σ_ det, and the whole image is y.

2.2. CAM (Weakly Supervised Deep Detection Network

Class Activation Mapping (CAM), also known as class activation maps, class heatmaps, or significance maps, vividly demonstrates the distribution of informational contribution within an image. Essentially, the more intense the color in these maps, the greater the area's contribution to the network's decision-making process. The operation of Convolutional Neural Networks (CNNs) is akin to a filtering mechanism that extracts salient features from images. Each convolution layer acts as a discerning filter, highlighting areas with pronounced characteristics—larger values in these areas signify more distinct features and garner higher attention from the convolution process. Different layers or channels within the feature maps store varied features, each extracted by its respective convolution. These feature maps hold rich semantic information, the significance of which can be

quantified through meticulous computation, as outlined in [5]. This multifaceted approach allows for a deeper understanding of how neural networks interpret and prioritize different aspects of visual data, offering invaluable insights into their decision-making frameworks. As shown in Figure 2.

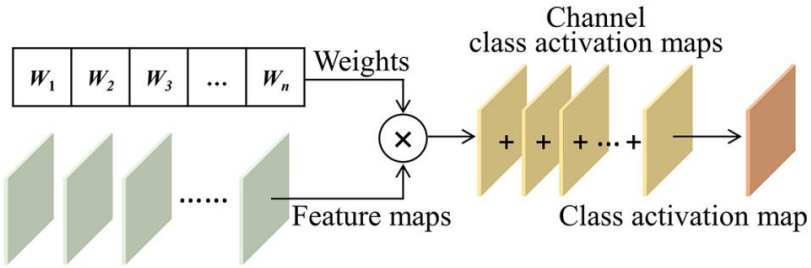

Fig. 2 The process of class activation mapping (Photo/Picture credit: Original).

However, an advanced CAM called Grad-CAM was invented (Selvaraju et al., 2016). Grad-CAM calculates the activation gradient of a specific layer's target category relative to the same layer [6].

2.3. Attention Mechanism in Weakly Supervised Object Detection

The attention mechanism in computational models closely mirrors the human visual system's selective focus, a process where attention is primarily directed towards salient regions of an image, ignoring extraneous background details. This simulation of human focus within algorithms significantly reduces the volume of low-quality data, thereby conserving resources that would otherwise be expended in processing such data. This facet of the attention mechanism finds pivotal application in object detection tasks. In many contemporary object detection algorithms, the inclusion of an attention mechanism module is indispensable. Particularly noteworthy is the development of a monocular three-bit object detection algorithm, which leverages depth information and a multiscale channel attention mechanism. This algorithm employs the channel attention mechanism in a feature extraction network to process monocular image features. These features are then integrated using a convolutional module algorithm, yielding enhanced accuracy in both 2D and 3D feature representation when compared to conventional methods. Further research into feature maps reveals that employing a multi-channel attention mechanism for feature extraction offers significant advantages. This approach incorporates two inputs: a monocular image and counting degree maps, utilizing the overall architecture of the Multi-Scale Segmentation Attention (MSA) module. Notably, inflated convolution replaces the traditional final average pooling layer and the fully connected layer. When benchmarked against the conventional KITTI dataset, the MSA algorithm with the multi-channel attention mechanism demonstrates superior accuracy. We also report on the mean accuracy of object detection in three categories—cars, pedestrians, and bicyclists—using the SPlit dataset, examined under both 2D and 3D perspectives. The enhanced algorithm's performance was further compared with the M3D-RPN algorithm across forty recall positions, exhibiting higher precision at all three difficulty levels. The experimental results, as detailed in Tables 1 and 2, affirm the efficacy of the upgraded algorithm.

Table 1. Comparison of the average precision for multi-class BEV detection.

classification	grade of difficulty	AMC M3DRPN	AMC OURS
pedestrian	simple	5.56	4.35
	medium	4.05	4.17
	difficulty	3.29	3.71
cyclist	simple	1.25	3.22
	medium	0.81	3.15
	difficulty	0.78	3.00
car	simple	20.85	27.53
	medium	15.62	19.40
	difficulty	11.88	14.82

Table 2. Comparison of the average precision for multi-class 3D detection.

classification	grade of difficulty	AMSDRPN	AMC OURS
pedestrian	simple	4.92	3.74
	medium	3.48	3.35
	difficulty	2.92	3.22
cyclist	simple	0.94	2.93
	medium	0.47	2.61
	difficulty	21.18	1.61
car	simple	14.53	22.30
	medium	11.07	15.03
	difficulty	8.65	11.93

The described algorithm leverages the first four convolutional layers of the OSNet architecture as its primary backbone. Positioned between the third and fourth convolutional layers, the algorithm further processes the deep semantic features extracted via the fifth convolutional layer. These features are then channeled into distinct branches designed for the extraction of fine-grained pedestrian characteristics. The model undergoes a supervised training regimen to enhance its performance. A key aspect of the algorithm is the inclusion of an attention module for the input image. This process involves the deformation and transposition of the input, followed by the application of a softmax function to activate and merge the transformed inputs, resulting in a feature map. This map is then refined by multiplying it with the transpose of X and A, thereby restoring the original feature map's attributes. The effectiveness of this enhanced algorithm is evident in its performance metrics, especially when benchmarked against the Market1501 dataset. Here, the algorithm achieves a mean Average Precision (mAP) of 89.2% and a Rank-1 accuracy of 95.3%. Remarkably, when compared to the HA-CNN method utilizing an in-attention mechanism, this algorithm demonstrates a substantial improvement, registering a 13.5% increase in mAP and a 4.1% improvement in Rank-1 accuracy. Further examination and comparative analysis reveal that the incorporation of an attention mechanism module into these distinct target detection algorithms significantly augments their efficacy, as elucidated in Table 3.

Table 3. Comparison between the algorithm and other algorithms.

Method	Publication	Market1501		DukeMTMC-ReID		CUHK03-Labeled		CUHK03-Detected	
		mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
HA-CNN	CVPR'18	75.7	91.2	63.8	80.5	41.0	44.4	38.6	41.7
MGN	ACMMM'18	86.9	95.7	78.4	88.7	67.4	68.0	66.0	66.8
Pyramid	CVPR'19	88.2	95.7	79.0	89.0	76.9	78.9	74.4	78.9
BDB	ICCV'19	86.7	95.7	76.0	89.0	76.7	79.4	73.5	76.4
HOReid	CVPR'20	84.9	94.2	75.6	86.9	-	—	-	
SNR	CVPR'20	84.7	94.4	72.9	84.4			—	
AAformer	Arxiv'21	87.7	95.4	80.0	90.1	77.8	79.9	71.3	74.0
CBDB_Net	TCSVT'21	85.0	94.4	74.3	87.7	76.6	77.8	72.8	75.4
DRL-Net	IEEE'22	86.9	94.7	76.6	88.1	-	—		
DCAL	CVPR'22	87.5	94.7	80.1	89.0				
Ours		**89.2**	**95.3**	**80.5**	**90.1**	**78.9**	**80.4**	**75.4**	**78.1**

2.4. Weakly Supervised Target Detection Algorithm Based on Pseudo Labeling

In weakly supervised learning there is a category of semi-supervised learning. How to make good use of the unlabeled data is becoming a problem. Pseudo-labeling is a good solution to this problem [8]. This method first uses labeled data for training to get a model, then uses this model to predict the

unlabeled data to generate pseudo-labels, and finally packs the labeled data and pseudo-labels together for secondary training to generate the final model. Li reconciles the multimodal input signals in the joint semantic mining (JSM) to maintaining perpixel pseudo-labels with iterative refinements [9]. Zhang mines pseudo-tags from each Image-level example by Pseudo Ground-truth Excavation (PGE) [10]. Lang used previously pre-trained data representation and statistical ranking techniques to purify pseudo labels [11, 12]. As shown in Figure 3.

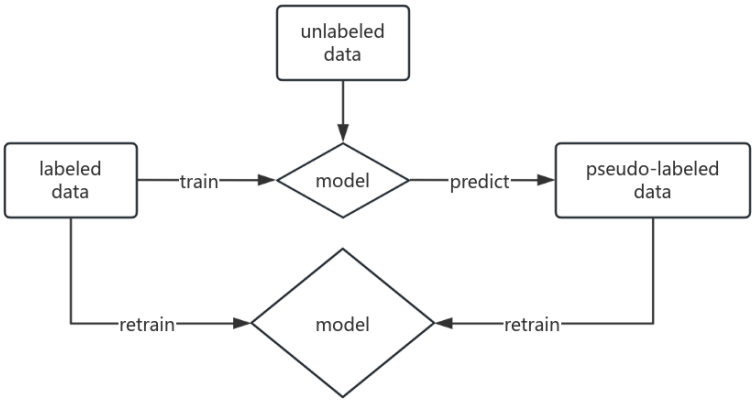

Fig. 3 Illustration of Weakly Supervised Target Detection Algorithm Based on Pseudo Labeling (Photo/Picture credit: Original).

3. Datasets and evaluation metrics

Over the past two decades, the field has seen a surge in the development of weakly supervised target detectors, necessitating a robust framework for their fair evaluation. Central to this evaluation are the datasets and metrics used. Notably, the PASCAL VOC dataset, encompassing 20 diverse object categories such as human faces, birds, dogs, and cars, has been instrumental in this regard. Additionally, other datasets like MSCOCO, ILSVRC, and CUB-200-2011 play a pivotal role in providing a comprehensive range of object classes and real-world scenarios, thus enabling a thorough assessment of these detectors. In terms of evaluation metrics, PASCAL VOC employs both mAP (mean Average Precision) and CorLoc (Correct Localization) to gauge model performance. mAP, an aggregate measure of precision across different recall levels, is particularly apt for assessing the combined accuracy of classification and localization in object detection models. This dual-focus metric aligns well with the intrinsic nature of object detection, where accurate classification is inextricably linked to precise localization.

4. Challenges

Weakly supervised target detection remains a challenging domain, with several key issues yet to be fully resolved. A primary concern is the effective utilization of unlabeled data for model training. This encompasses developing methods that can leverage such data to enhance learning without compromising accuracy. Additionally, there is the persistent problem of the models' insufficient generalization ability. Models often struggle to maintain performance across diverse and unseen datasets, indicating a need for more robust training methodologies. Another significant challenge is improving model adaptability in complex scenarios, such as those involving occlusion, deformation, or varying illumination conditions. Addressing these issues is crucial for advancing the field, as real-world applications frequently present such complexities. Moreover, the integration of more sophisticated attention mechanisms and advanced feature extraction techniques may provide a pathway to more resilient and adaptable models. These developments could potentially lead to breakthroughs in weakly supervised learning paradigms, expanding their applicability in various

fields like remote sensing, medical imaging, and autonomous systems, where precise and reliable target detection is paramount.

5. Conclusion

Our research delves deeply into articles on Weakly Supervised Deep Detection Networks (WSDDN), Class Activation Mapping (CAM), Squeeze-and-Excitation Networks, and Spatial Transformer Networks, which are pivotal in the field of weakly supervised target detection. This approach notably capitalizes on the potential of unlabeled data, thereby significantly diminishing the need for labor-intensive manual annotation while concurrently enhancing target detection capabilities. However, this domain is not without its challenges. One primary issue lies in the optimal utilization of unlabeled data—how can we extract maximal value from these datasets without compromising the quality of the learning process? Furthermore, we face the dilemma of limited generalization. This pertains to the model's ability to apply learned patterns to new, unseen data, a critical factor for the robustness and applicability of these detection systems.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Wang, Z., Zhang, W., & Zhang, M. L. (2023). Transformer-based Multi-Instance Learning for Weakly Supervised Object Detection. preprint, 2303.14999.

[2] Yang, Y., Pan, Z., Hu, Y., et al. (2022). PistonNet: Object separating from background by attention for weakly supervised ship detection. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 15, 5190-5202.

[3] Yao, X., Feng, X., Han, J., et al. (2020). Automatic weakly supervised object detection from high spatial resolution remote sensing images via dynamic curriculum learning. IEEE Transactions on Geoscience and Remote Sensing, 59(1), 675-685.

[4] Shao, F., Chen, L., Shao, J., et al. (2022). Deep learning for weakly-supervised object detection and localization: A survey. Neurocomputing, 496, 192-207.

[5] Eloy, P., M. J. T., Pau, G., et al. (2023). Deep machine learning for meteor monitoring: Advances with transfer learning and gradient-weighted class activation mapping. Planetary and Space Science, 238.

[6] Yueyue, H., Yingyan, H., Hangcheng, D., et al. Continuous gradient fusion class activation mapping: segmentation of laser-induced damage on large-aperture optics.

[7] Liu, Q., Li, W., Yu, S., et al. A monocular 3D target detection algorithm combining depth information guidance and multi-scale channel attention mechanism. Journal of Shandong University (Science Edition).

[8] Zhou, H., Zhan, F., Zhou, C., et al. (2024). Pedestrian re-recognition method based on attention mechanism and multi-branch association. Microelectronics and Computers, (02).

[9] Li, J., Ji, W., Bi, Q., et al. (2021). Joint semantic mining for weakly supervised RGB-D salient object detection. Advances in Neural Information Processing Systems, 34, 11945-11959.

[10] Zhang, Y., Bai, Y., Ding, M., et al. (2018). W2f: A weakly-supervised to fully-supervised framework for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 928-936).

[11] Lang, H., Vijayaraghavan, A., & Sontag, D. (2022). Training subset selection for weak supervision. Advances in Neural Information Processing Systems, 35, 16023-1603.

[12] Xu, M., Zhang, Z., Hu, H., et al. (2021). End-to-end semi-supervised object detection with soft teacher. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3060-3069).