Zero-Shot Object Detection with Partitioned Contrastive Feature Alignment

Haohe Li
NingBo University
Chong Wang (wangchong@nbu.edu.cn)
NingBo University
Shenghao Yu
NingBo University
Zheng Huo
NingBo University
Yujie Zheng
NingBo University
Li Dong
NingBo University
Jiafei Wu
Sensetime reserach

Research Article

Keywords: attribute, contrastive learning, memory bank, zero-shot object detection

Posted Date: July 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1770867/v1

Zero-Shot Object Detection with Partitioned Contrastive Feature Alignment

Haohe Li[1], Chong Wang[1*], Shenghao Yu[1], Zheng Huo[1], Yujie Zheng[1], Li Dong[1] and Jiafei wu[2]

[1*]Faculty of Electrical Engineering and Computer Science, NingBo University, Ningbo, 315000, Zhejiang, China.
[2]Sensetime reserach, Sensetime, Shanghai, 200000, China.

*Corresponding author(s). E-mail(s): wangchong@nbu.edu.cn;
Contributing authors: lhh_1023@hotmail.com; ysh_nbu@163.com; zhenghuo369@163.com; zhengyujie99@foxmail.com; dongli@nbu.edu.cn; wujiafei@sensetime.com;

Abstract

How to properly align the extracted visual features with certain semantic embeddings of unseen objects is crucial to the problem of Zero-Shot Object Detection (ZSD). To give a better guess of those unseen visual features, a partitioned contrast strategy is proposed in this paper to train the visual and attribute feature alignment networks. To be specific, four types of contrast are considered, including the visual-to-visual, visual-to-attribute, attribute-to-visual and attribute-to-attribute contrasts. Combining with two cross-batch memory banks of the visual features and unseen attribute features, it is effective to adjust the alignment rules for unseen visual features. Although the visual features of the unseen classes are missing during training, the common projection rules for visual and attribute features are learned by emphasizing the contrast involving unseen attribute features. Experimental results on the MS-COCO dataset show the superiority of the proposed model. Our code and models are publicly available at: https://github.com/lihh1023/PCFA-ZSD.

Keywords: attribute, contrastive learning, memory bank, zero-shot object detection

1 Introduction

With the extraordinary development of deep learning, object detection has made great progress in recent years [1–5]. However, in order to achieve the best performance, massive and well-labeled image datasets are needed for the generic object detection models. Obviously, it is difficult to collect enough images with bounding box annotations for rare target classes. Therefore, some scholars propose zero-shot learning [6–10] to address the extreme situation when there is no image for training. Naturally, it is extended to zero-shot object detection (ZSD) [11–20] very soon. The characteristic of ZSD lies in how to guide the model to learn visual features and detect the objects without training samples.

Most ZSD models rely on various semantic embedding to build the connection between seen classes and unseen classes, such as word vectors [13–19, 21], textual descriptions [11, 12], attributes [20]. Many works [13, 14, 21],focus on learning the projection from visual features to semantic space. Rahman et al. [21] designed a semantic alignment network with the semantic clustering loss and max-margin loss. Later, they proposed a polarity loss in [13] to address the class-imbalance issue. Mao et al. [20] built an attribute table to connect the seen and unseen classes at the semantic level. Projecting semantic features into visual space is exploited in [11] using the textual descriptions. Similarly, textual descriptions are also used in [12] by projecting visual and semantic features into a common space. In addition, Rahman et al. [16] proposed a self-monitoring mechanism, using pseudo-labeling techniques in a transductive way. Using generative networks, e.g., SU-ZSD [17], GT-Net [18] and DELO [19], to generate unseen class features is becoming very active recently.

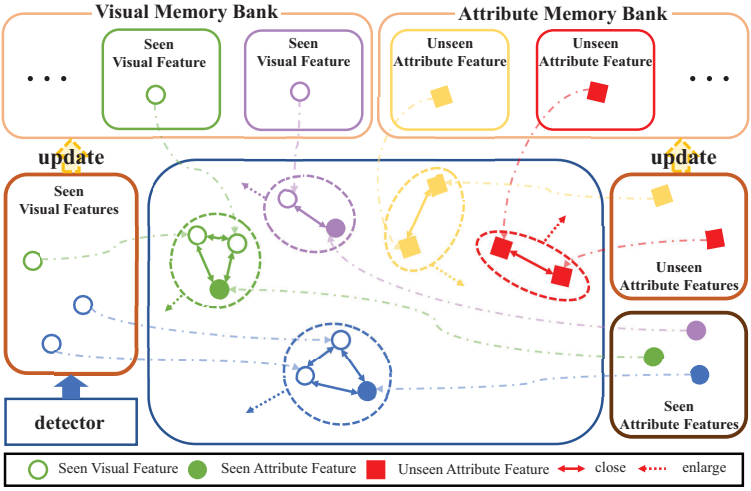

Fig. 1 The illustration of the partitioned contrastive feature alignment.

It is obviously that the key to ZSD is the projection between visual and semantic features. Thus, the goal of this work is to achieve better alignment for both features in a mutual space based on the idea of contrastive learning [9, 22, 23], i.e., simultaneously maximizing consistency between similar instances and encouraging differences between different instances. Unlike previous contrastive learning based work [9, 14, 22, 23], different types of contrasts, including visual-to-visual, visual-to-attribute, attribute-to-visual and attribute-to-attribute for either seen or unseen classes, are considered separately to constrain the feature projection as shown in Fig.1, while the attributes are selected as the semantic descriptions of class. Moreover, to enrich the variety of available features, especially the ones of unseen classes, a cross-batch memory bank mechanism is employed to collect features for the aforementioned contrast. Although the visual features of the unseen classes are missing, the common projection rules for visual and attribute features are learned by emphasizing the contrast involving unseen attribute features.

2 Methodology

As shown in Fig.2, RetinaNet [1] is chosen as the base model for object detection, where ResNet [24] and Feature Pyramid Network (FPN) [25] are used to extract multi-scale visual features for bounding box regression and classification. Inspired by the structure in [6, 16], the projected semantic attribute vectors from ASC-ZSD [20] are chosen as the category centroids, while their distances from the extracted visual features are used for classification. In this section, a new partitioned contrastive feature alignment (PCFA) strategy, including the visual feature sampling, cross-batch memory bank and partitioned contrastive learning, is designed to simultaneously enlarge the inter-class distance and reduce the intra-class distance for both the attribute and visual features.

2.1 General Framework

The structure of the box regression subnet is the same as [13], in the proposed network as shown in Fig.2. However, only the semantic embeddings (word vectors) of seen classes $E \in R^{C_s \times d}$ are used to help box localization with a learnable matrix, where C_s is the number of seen categories and d is the dimension of the semantic embedding.

On the other hand, the semantic embeddings of seen and unseen classes are also introduced in the classification subnet. However, the gap between semantic embeddings (attribute vectors A in this work) and the visual features V of the prediction boxes is huge. Thus, it is natural to project them into a mutual space using a linear mapping δ and convolution operation σ with the parameters θ_A and θ_V,

$$\tilde{A} = \delta(A;\ \theta_A), \tag{1}$$

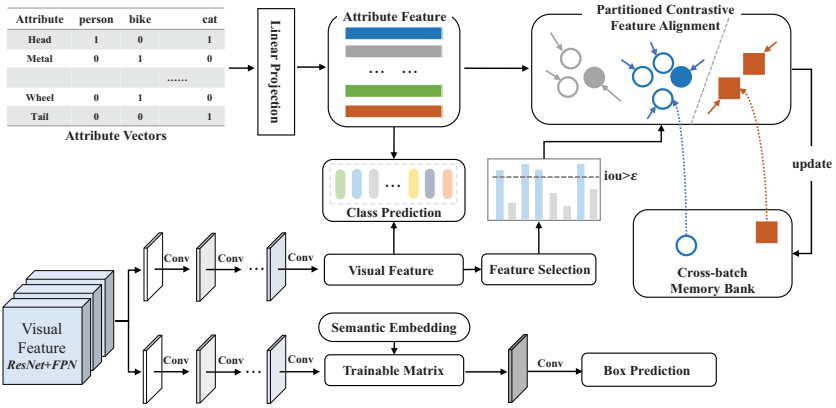

Fig. 2 The overall architecture of our model. The hollow circles, solid circles, and solid squares represent visual features, seen attribute features, and unseen attribute features, respectively.

$$\tilde{V} = \sigma\left(V;\ \theta_V\right). \tag{2}$$

$\tilde{A} = \{\tilde{A}_{C_s},\ \ \tilde{A}_{C_u}\} \in R^{(C_s+C_u)\times m}$ and $\tilde{V} \in R^{B\times m}$ represent the attribute features which include seen and unseen attribute features and visual features in a common space, where C_u, B and m are the number of unseen categories, prediction boxes and dimensions, respectively. Then the cosine similarity between all pairs of $\tilde{A}$ and $\tilde{V}$ are calculated to find the seen or unseen objects with the score threshold.

However, such simple operations in Eq. (1) and Eq.(2) do not perform well in ZSD due to the lack of constraints between the seen and unseen parts in $\tilde{A}$ during the training stage. Hence, it is important to construct a more intuitive relationship between them.

2.2 Partitioned Contrastive Feature Alignment

To find the implicit links between the seen and unseen objects in semantic attributes and visual features, inspired by contrastive learning [9, 22, 23], the contrast between them is utilized to guide the learning of θ_A and θ_V for better feature alignment. To construct such contrast, the visual features need to be sampled first to address the imbalance issue between the foreground and background as well as between the visual features and semantic attribute vectors. Moreover, a cross-batch memory bank mechanism is introduced to enrich the contrast.

2.2.1 Visual Feature Sampling

It should be noted that most of the candidate boxes detected by RetinaNet [1] are easy negatives, i.e., various backgrounds. In order to filter out these background boxes, the Intersection-over-Union (IOU) scores $S = \{s_k\}_{k=1}^{K}$ with

matched ground truth are used as the criteria for visual feature sampling. Only those visual features ($\tilde{v}_k \in \tilde{V}$) whose corresponding s_k are greater than a given threshold ε (0.7 in this paper) are retained for later feature alignment. Then, the new set of visual features $\tilde{V}^{(t)}$ at t-th batch can be defined as,

$$\tilde{V}^{(t)} = \left\{(\tilde{v}_n^{(t)}, i_n^{(t)})\right\}_{n=1}^{N} = \{(\tilde{v}_k, i_k) | s_k \geq \varepsilon\}_{k=1}^{K}, \tag{3}$$

where K is the total boxes detected by RetinaNet [1] and N is the length of $\tilde{V}^{(t)}$.

2.2.2 Cross-batch Memory Bank

Noting that the actual visual features and their corresponding seen categories are limited to one batch during the training stage. Moreover, only one attribute feature is available for either unseen or seen class. Such unbalance on the categories and types of features is not good for the learning process. Inspired by [26, 27], two additional memory banks (MB) are introduced to retain the visual and unseen attribute features in previous batches, which can enrich both the visual features and semantic attribute features of unseen classes. At t-th batch, the visual and attribute memory banks, denoted as $M_v^{(t)}$ and $M_a^{(t)}$, contain the features of current batch and previous $(b-1)$ batches, respectively. They are updated at the beginning of each batch as,

$$M_v^{(t)} = \tilde{V}^{(t)}, \tilde{V}^{(t-1)}, \tilde{V}^{(t-2)}, \ldots, \tilde{V}^{(t-b)}, \tag{4}$$

$$M_a^{(t)} = \tilde{A}^{(t)}, \tilde{A}_u^{(t-1)}, \tilde{A}_u^{(t-2)}, \ldots, \tilde{A}_u^{(t-b)}. \tag{5}$$

In our experiments, the value of b is set as 1, while larger value only brings very slight improvement but huge memory and time cost.

2.2.3 Partitioned Contrastive Feature Alignment

As one of the key contributions of this work, the idea of contrastive learning is adopted in the ZSD framework. Unlike previous works, the partitioned contrast between visual and attribute features is performed in the mutual space to train the operations in Eq.(1) and Eq.(2) for better feature alignment. To be specific, four types of contrast are performed, including the visual-to-visual, visual-to-attribute, attribute-to-visual and attribute-to-attribute contrasts. They can then be utilized to formulate new losses to make sure the projected features, either visual or attribute ones, of the same category will be close to each other. Furthermore, the unseen attribute features should be laid far from seen ones, i.e., having a large contrast with other features.

For each element $\tilde{v}_j'$ in the whole visual memory bank $M_v^{(t)}$, a positive bag $P_j^{v2v} = \left\{\tilde{v}_1^+, \tilde{v}_2^+, \cdots, \tilde{v}_{Z_j}^+\right\}$, $j = 1, 2, \cdots, N$ of size Z_j is constructed to contain all visual features with the same label y_j in $M_v^{(t)}$. Besides that, it also has a corresponded attribute feature $\tilde{a}_{y_j}^+$ in attribute memory bank $M_a^{(t)}$. Then, the

visual-to-visual and visual-to-attribute contrastive loss $L_{v2v}\left(\tilde{v}'_j\right)$ and $L_{v2a}\left(\tilde{v}'_j\right)$ can be defined as,

$$L_{v2v}\left(\tilde{v}'_j\right) = -\frac{1}{Z_j+1}\sum_{\tilde{v}_z^+ \in P_j^{v2v}}^{Z_j} log\frac{d\left(\tilde{v}'_j, \tilde{v}_z^+\right)}{d_{v2all}\left(\tilde{v}'_j\right)}, \tag{6}$$

$$L_{v2a}\left(\tilde{v}'_j\right) = -\frac{1}{Z_j+1} log\frac{d\left(\tilde{v}'_j, \tilde{a}_{y_j}^+\right)}{d_{v2all}\left(\tilde{v}'_j\right)}, \tag{7}$$

where $d(*,*)$ is the cosine similarity between two features, $d_{v2all}\left(\tilde{v}_j^{(t)}\right)$ is the visual-to-visual and visual-to-attribute similarities of all features. They are calculated as,

$$d\left(a,b\right) = \exp(\frac{ab^T}{\|a\| \cdot \|b^T\| \cdot \tau}), \tag{8}$$

$$d_{v2all}\left(\tilde{v}'_j\right) = \sum_{n=1,n\neq i}^{N} d\left(\tilde{v}'_j, \tilde{v}_n\right) + \sum_{c=1}^{C_s+(b+1)C_u} d\left(\tilde{v}'_j, \tilde{a}_c\right), \tag{9}$$

where the hyper-parameter temperature τ is set to 0.2 following [23].

Similarly, for each element of seen attribute feature $\tilde{a}'_l$ in the attribute feature set from $M_a^{(t)}$, all visual features belonging to the l-th class are extracted to form another positive bag $P_l^{a2v} = \left\{\tilde{v}_1^+, \tilde{v}_2^+, \cdots, \tilde{v}_{Z_l}^+\right\}$, $l = 1, 2, \cdots, C_s$. Then, the attribute-to-visual contrastive loss $L_{a2v}\left(\tilde{a}'_l\right)$ can be formulated as,

$$L_{a2v}(\tilde{a}'_l) = -\frac{1}{Z_l}\sum_{\tilde{v}_z^+ \in P_l^{a2v}}^{Z_l} log\frac{d(\tilde{a}'_l, \tilde{v}_z^+)}{d_{a2all}(\tilde{a}'_l)}, \tag{10}$$

$$d_{a2all}\left(\tilde{a}'_l\right) = \sum_{n=1}^{N} d\left(\tilde{a}'_l, \tilde{v}_n\right) + \sum_{c=1,c\neq l}^{C_s+(b+1)C_u} d\left(\tilde{a}'_l, \tilde{a}_c\right). \tag{11}$$

Considering the contrast between seen and unseen attribute features, each unseen attribute feature $\tilde{a}'_r$ from $M_a^{(t)}$ is corresponding to b attribute features with the same label, i.e. $P_r^{a2a} = \left\{\tilde{a}_1^+, \tilde{a}_2^+, \ldots, \tilde{a}_b^+\right\}$. Subsequently, the attribute-to-attribute contrastive loss $L_{a2a}\left(\tilde{a}'_r\right)$ can be formulated as,

$$L_{a2a}(\tilde{a}'_r) = -\frac{1}{b}\sum_{\tilde{a}_e^+ \in P_r^{a2a}}^{b} log\frac{d(\tilde{a}'_r, \tilde{a}_e^+)}{d_{a2all}(\tilde{a}'_r)}, \tag{12}$$

$$d_{a2all}\left(\tilde{a}'_r\right) = \sum_{n=1}^{N} d\left(\tilde{a}'_r, \tilde{v}_n\right) + \sum_{c=1,c\neq r}^{C_s+(b+1)C_u} d\left(\tilde{a}'_r, \tilde{a}_c\right). \tag{13}$$

It is worth noting that the information of unseen categories is introduced by the similarity between the attribute features in Eq.(12). Although the visual

features of unseen classes are not available in ZSD, such contrast on the similarity may give a hint where the unseen visual features will be projected. Finally, the total loss of the proposed PCFA is,

$$\begin{aligned} L_{PCFA} = \alpha \sum_{t=1}^{N} \left\{ L_{v2v}\left(\tilde{v}'_j\right) + L_{v2a}\left(\tilde{v}'_j\right) \right\} \\ +\beta \sum_{l=1}^{C_s} L_{a2v}\left(\tilde{a}'_l\right) + \gamma \sum_{r=1}^{(b+1)C_u} L_{a2a}(\tilde{a}'_r), \end{aligned} \tag{14}$$

where α, β and γ are the weights for different types of contrast, respectively.

3 Experiments

3.1 Dataset and Implementation Details

The proposed PCFA-ZSD model is evaluated in MS-COCO (2014) [28]. There are 80 categories of MS-COCO[28]. Following the protocol in [13, 20], 80 categories are divided into 65 seen classes and 15 unseen classes. There are 62,300 images for training, without examples of unseen classes. At the time of ZSD validation, there are 10,098 images containing 16,388 object detection boxes that are unseen classes. For generalized ZSD (GZSD) verification, the seen images and bounding boxes are included. The dimension m of the mutual space is set to 128 in this experiment. In the inference stage, the predicted bounding boxes with IOUs > 0.5 and similarity scores greater than 0.3 for seen classes and 0.1 for unseen are selected. The weight of L_{PCFA} is set to 0.5, and the α, β and γ are set to $\frac{1}{N+C_s+(b+1)C_u}$ for each batch in our work.

3.2 Experimental Results

Table 1 Comparison results on MS-COCO dataset for ZSD with iou=0.5 and 65/15 data split.The best results are highlighted with **bold**.

Model	mAP(%)	recall@100
FL-ZSD [13]	8.48	20.44
PL-ZSD [13]	12.40	37.72
TL-ZSD [16]	14.57	48.15
ACS-ZSD [20]	15.34	47.83
SU-ZSD [17]	19.00	54.00
ContrastZSD [14]	18.60	**59.50**
Baseline	20.98	50.51
CCFA-ZSD (ours)	**24.62**	55.32

Table 2 Comparison results on MS-COCO dataset for GZSD with iou=0.5 and 65/15 data split.

Model	seen		unseen		HM	
	mAP(%)	recall@100	mAP(%)	recall@100	mAP(%)	recall@100
FL-ZSD [13]	36.96	40.09	8.66	20.45	4.03	27.08
PL-ZSD [13]	34.07	36.38	12.40	37.16	18.18	36.76
TL-ZSD [16]	28.79	54.14	14.05	37.16	18.89	44.07
ACS-ZSD [20]	34.02	37.86	15.34	47.83	21.15	42.26
SU-ZSD [17]	36.90	57.70	19.00	53.90	25.08	55.74
ContrastZSD [14]	**40.20**	**62.90**	16.50	**58.60**	23.40	**60.70**
Baseline	28.75	33.55	20.98	50.20	24.24	40.21
CCFA-ZSD (ours)	33.35	38.64	**24.62**	54.72	**28.31**	45.29

The comparison results for ZSD and GZSD with other state-of-the-art models on MS-COCO dataset are shown in Table1 and Table2. All reported results are the average of five experiments. The mAP is given for both ZSD and GZSD, while the Harmonic Mean (HM) of mAP is provided for GZSD. It can be seen that the proposed PCFA-ZSD achieves the highest mAP for unseen objects in both ZSD and GZSD settings. It is worth noting that mAP of PCFA-ZSD is higher than the baseline by about 3.6–4.6% for not only unseen classes but also seen ones. Here, the baseline refers to the base model without the proposed CFA and memory bank.

Table 3 Average Precision (%) of each unseen class on MS-COCO dataset with 65/15 data split

Model	airplane	train	parking meter	cat	bear	suitcase	frisbee	snowboard	fork	sandwich	hot dog	toilet	mouse	toaster	hair drier	mAP
PL-ZSD[13]	20.0	48.2	0.6	28.3	13.8	12.4	21.8	15.1	8.9	8.5	0.9	5.7	0.0	1.7	0.0	12.4
TL-ZSD[16]	19.6	**63.4**	3.7	43.2	3.7	**13.8**	12.8	24.2	12.6	9.7	6.0	1.5	2.3	**2.0**	0.0	14.6
ACS-ZSD[20]	8.7	25.5	6.6	40.8	54.0	9.6	10.6	26.8	16.4	11.0	5.0	7.8	**6.2**	1.3	0.0	15.4
SU-ZSD[17]	10.1	48.7	1.2	**64.0**	**64.1**	12.2	0.7	**28.0**	16.4	19.4	0.1	18.7	1.2	0.5	**0.2**	19.0
Baseline	13.2	42.4	9.3	48.9	55.0	11.7	**40.3**	11.9	21.0	20.7	21.4	13.6	5.6	0.6	0.0	21.0
CCFA-ZSD	**46.9**	62.2	**11.8**	43.6	55.8	12.3	24.7	17.5	**22.7**	**21.4**	**27.2**	**18.9**	3.7	0.4	0.1	**24.6**

In the ZSD task, our PCFA-ZSD outperforms other RetinaNet [1] based frameworks including PL-ZSD [13], TL-ZSD [16] and ACS-ZSD [20]. Although TL-ZSD [16] uses a transduction learning method to exploit unseen information from the test dataset, its performance is still inferior to ours. Compared with the two-stage detection models (SU-ZSD [17] and ContrastZSD[14]), our method maintains at least 5.62% advantage in mAP, with a second place in Recall@100.

For the GZSD task, our method still surpasses the compared state-of-the-arts models in both Unseen and HM mAP. The HM mAP of the proposed PCFA-ZSD is 3.23% higher than the second place (SU-ZSD [17]) in Table2. Although ContrastZSD [14] also adopted the idea of contrastive learning and achieves the highest mAP on seen objects, its performance on unseen objects

is worse than our method, which matters more in ZSD problem. It shows the effectiveness of the proposed partitioned contrast strategy.

The per-class results on MS-COCO are given in Table 2 to show more details. It can be observed that our work achieves the highest AP in about half (7/15) classes, and the second highest AP in another 4 classes. Since the "mouse", "toaster", and "hair drier" share very limited visually similarity with the seen classes, it is challenging for the ZSD task. Thus, the AP of these classes is low for all methods.

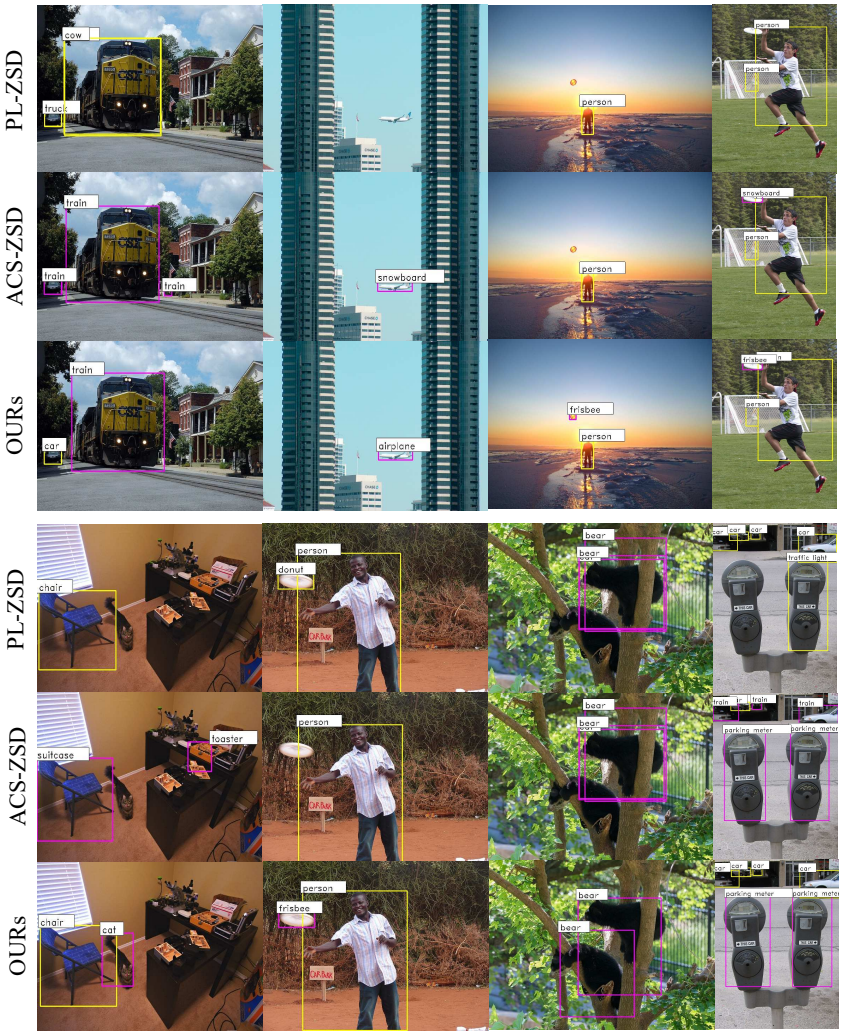

Fig. 3 Qualitative comparison results on MS-COCO. Yellow and purple bounding boxes represent seen and unseen classes, respectively.

A set of qualitative comparison of the GZSD results is presented in Fig.3. It can be noted that our model is more accurate on the detection of "airplane", "frisbee" and "parking meter". The "airplane" is wrongly detected as snowboard by ASC-ZSD [20], while the "frisbee" is mistakenly detected as a donut by PL-ZSD [13]. Moreover, both PL-ZSD [13] and ASC-ZSD [20] miss some unseen objects in the images.

3.3 Ablation Studies

Table 4 ABLATION STUDY ON CFA AND MB MODULES.

Model	ZSD		GZSD(HM)	
	mAP(%)	recall@100	mAP(%)	recall@100
Baseline	20.96	50.30	24.21	40.07
Baseline + CFA	23.30	52.54	26.03	44.73
Baseline + CFA + MB	**24.62**	**55.32**	**28.31**	**45.29**

An ablation study on contrastive feature alignment (CFA) and cross-batch memory bank (MB) are demonstrated in Table3. It is worth mentioning that the mAP and Recall@100 receive significant boost with the CFA for both ZSD and GZSD, which means the partitioned contrast strategy is useful. Furthermore, the additional features introduced by MB can further elevate the detection performance. To a certain extent, the visual features for contrastive alignment are enriched, while previous unseen attribute features are also available for a more robust contrast.

4 CONCLUSION

In this work, a new framework based on partitioned contrastive feature alignment is proposed for zero-shot object detection. Constrained by four types of contrasts, the visual and attribute features can be better aligned in the common space. As a result, the detector gains the enhanced ability to detect unseen classes more accurately for both ZSD and GZSD tasks on the MS-COCO dataset.

Funding The authors gratefully acknowledge funding from the Zhejiang Provincial Natural Science Foundation of China (No.LY20F030005) and the National Natural Science Foundation of China (No.61603202).

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)

[2] Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6154–6162 (2018)

[3] Liu, W., Li, H., Yu, S., Chen, S., Ye, X., Wu, J., et al.: Dynamic relevance learning for few-shot object detection. arXiv preprint arXiv:2108.02235 (2021)

[4] Liu, W., Wang, C., Yu, S., Tao, C., Wang, J., Wu, J.: Novel instance mining with pseudo-margin evaluation for few-shot object detection. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2250–2254 (2022). IEEE

[5] Wang, X., Xiang, X., Zhang, B., Liu, X., Zheng, J., Hu, Q.: Weakly supervised object detection based on active learning. Neural Processing Letters, 1–15 (2022)

[6] Annadani, Y., Biswas, S.: Preserving semantic relations for zero-shot learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7603–7612 (2018)

[7] Xu, W., Xian, Y., Wang, J., Schiele, B., Akata, Z.: Attribute prototype network for zero-shot learning. Advances in Neural Information Processing Systems **33**, 21969–21980 (2020)

[8] Yang, S., Wang, K., Herranz, L., van de Weijer, J.: On implicit attribute localization for generalized zero-shot learning. IEEE Signal Processing Letters **28**, 872–876 (2021)

[9] Jiang, H., Wang, R., Shan, S., Chen, X.: Transferable contrastive network for generalized zero-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 9765–9774 (2019)

[10] Ren, H., Zheng, Z., Lu, H.: Energy-guided feature fusion for zero-shot sketch-based image retrieval. Neural Processing Letters, 1–10 (2022)

[11] Zhang, L., Wang, X., Yao, L., Wu, L., Zheng, F.: Zero-shot object detection via learning an embedding from semantic space to visual space. In: Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20} (2020). International Joint Conferences on

Artificial Intelligence Organization

[12] Li, Z., Yao, L., Zhang, X., Wang, X., Kanhere, S., Zhang, H.: Zero-shot object detection with textual descriptions. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8690–8697 (2019)

[13] Rahman, S., Khan, S., Barnes, N.: Improved visual-semantic alignment for zero-shot object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11932–11939 (2020)

[14] Yan, C., Chang, X., Luo, M., Liu, H., Zhang, X., Zheng, Q.: Semantics-guided contrastive network for zero-shot object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

[15] Li, Q., Zhang, Y., Sun, S., Zhao, X., Li, K., Tan, M.: Rethinking semantic-visual alignment in zero-shot object detection via a softplus margin focal loss. Neurocomputing **449**, 117–135 (2021)

[16] Rahman, S., Khan, S., Barnes, N.: Transductive learning for zero-shot object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6082–6091 (2019)

[17] Hayat, N., Hayat, M., Rahman, S., Khan, S., Zamir, S.W., Khan, F.S.: Synthesizing the unseen for zero-shot object detection. In: Proceedings of the Asian Conference on Computer Vision (2020)

[18] Zhao, S., Gao, C., Shao, Y., Li, L., Yu, C., Ji, Z., Sang, N.: Gtnet: Generative transfer network for zero-shot object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12967–12974 (2020)

[19] Zhu, P., Wang, H., Saligrama, V.: Don't even look once: Synthesizing features for zero-shot detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11693–11702 (2020)

[20] Mao, Q., Wang, C., Yu, S., Zheng, Y., Li, Y.: Zero-shot object detection with attributes-based category similarity. IEEE Transactions on Circuits and Systems II: Express Briefs **67**(5), 921–925 (2020)

[21] Rahman, S., Khan, S., Porikli, F.: Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In: Asian Conference on Computer Vision, pp. 547–563 (2018). Springer

[22] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in Neural Information Processing Systems **33**, 18661–18673

(2020)

[23] Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: Few-shot object detection via contrastive proposal encoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7352–7362 (2021)

[24] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[25] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)

[26] Yu, S., Wang, C., Mao, Q., Li, Y., Wu, J.: Cross-epoch learning for weakly supervised anomaly detection in surveillance videos. IEEE Signal Processing Letters **28**, 2137–2141 (2021)

[27] Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3733–3742 (2018)

[28] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision, pp. 740–755 (2014). Springer