ProposalCLIP: Unsupervised Open-Category Object Proposal Generation via Exploiting CLIP Cues

Hengcan Shi, Munawar Hayat, Yicheng Wu, Jianfei Cai
Department of Data Science and AI, Monash University, Australia
{hengcan.shi, munawar.hayat, yicheng.wu, jianfei.cai}@monash.edu

Abstract

Object proposal generation is an important and fundamental task in computer vision. In this paper, we propose ProposalCLIP, a method towards unsupervised open-category object proposal generation. Unlike previous works which require a large number of bounding box annotations and/or can only generate proposals for limited object categories, our ProposalCLIP is able to predict proposals for a large variety of object categories without annotations, by exploiting CLIP (contrastive language-image pre-training) cues. Firstly, we analyze CLIP for unsupervised open-category proposal generation and design an objectness score based on our empirical analysis on proposal selection. Secondly, a graph-based merging module is proposed to solve the limitations of CLIP cues and merge fragmented proposals. Finally, we present a proposal regression module that extracts pseudo labels based on CLIP cues and trains a lightweight network to further refine proposals. Extensive experiments on PASCAL VOC, COCO and Visual Genome datasets show that our ProposalCLIP can better generate proposals than previous state-of-the-art methods. Our ProposalCLIP also shows benefits for downstream tasks, such as unsupervised object detection.

1. Introduction

Object proposal generation aims to predict a number of category-agnostic bounding box proposals for all objects in an image. It serves as a fundamental and crucial step towards many higher-level tasks, such as object detection [11, 23], object segmentation [12, 34–36] and image captioning [19, 24]. How to effectively generate as few as possible proposals to cover all objects is the key challenge in object proposal generation.

Traditional proposal generation methods [1, 6, 41, 47, 50] often utilizes low-level cues (e.g., color, texture, gradient and/or edge) to select proposals from sliding window boxes. In recent years, deep-learning-based methods [18, 22, 31, 50] take high-level semantics from CNNs or Transformers as cues to select or regress proposals. Although these deep-learning-based methods significantly improve the proposal generation performance, they require a huge number of bounding box annotations for training, which is very labor-intensive, especially for large-scale datasets. Meanwhile, due to the significant annotation effort required, only objects of a limited number of categories can be labeled. Thus, these supervised methods can only generate proposals for limited object categories. However, real-world applications such as object retrieval [2, 13], image captioning [19, 24] and referring grounding [27, 45] usually require object proposals of diverse categories.

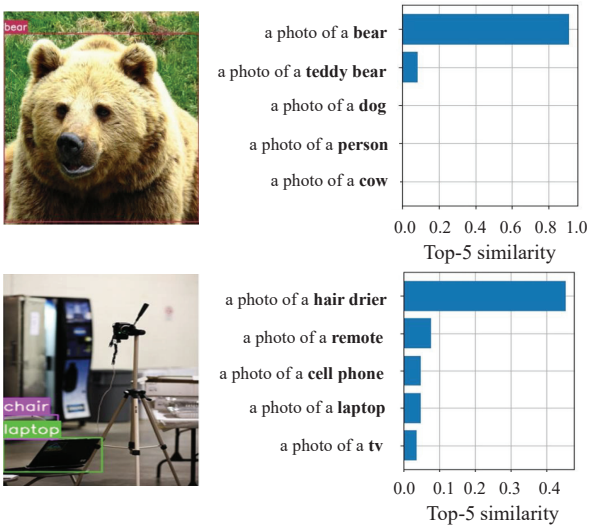


Figure 1. Examples of CLIP [28] image-text matching results. This pre-trained model well recognizes objects of open categories. However, it can only highlight a single object in an image. For example, the "chair" and "laptop" are not recognized in the bottom image.

Some recent efforts [16, 37, 42, 46] aim to address these challenges. ORE [16] and OVR-CNN [46] leverage incremental learning and image caption supervisions to rec-

ognize additional object categories. However, these approaches also need a mass of bounding box and image caption annotations. Without the human-intensive annotations, these methods fail to perform well. LOST [37] and rOSD [42] propose unsupervised deep-learning-based proposal generation, which leverages off-the-shelf knowledge from other tasks to generate proposals. Specifically, they predict proposals based on class activation maps (CAMs) and attention maps from pre-trained classification networks. These unsupervised methods avoid the box annotation, but they are only able to recognize limited object categories. In addition, although CAMs/attention maps activate some salient regions, there are many objects in non-activated regions. As a result, proposals generated by these methods can only cover parts of objects, as shown in Tables 2 and 3.

In this paper, we propose a novel method, called ProposalCLIP, towards unsupervised and open-category object proposal generation. Our method can generate a variety of proposals of different object categories without requiring expensive bounding box annotations, by leveraging the off-the-shelf image-text matching model, CLIP (contrastive language-image pre-training) [28]. We exploit CLIP [28] features because it is trained on millions of image-language pairs from web, and thus have potential to generalize to various object categories, as shown in Fig. 1. Nevertheless, CLIP [28] cannot be directly used in object proposal generation, because it is only trained to recognize single-object images and cannot well handle multi-object images. For example, in the second image in Fig. 1, it only ignores both the "chair" and "laptop" objects. Thus, it is non-trivial to apply CLIP for our task. In our ProposalCLIP, we first analyze CLIP features and build an objectness score based on our analysis for proposal generation. In addition, we design a graph-based proposal merging model, which exploits CLIP features to effectively combine different proposals. We also extract pseudo labels based on CLIP cues to train a box regression model to further improve our proposals. We conduct experiments on three common datasets, PASCAL VOC (20 classes), COCO (80 classes) and Visual Genome (1,600 classes). The experimental results demonstrate the effectiveness of our proposed method.

Our major contributions can be summarized as follows: (1) We propose a novel method that can effectively generate proposals for open categories in real world, without requiring annotations. (2) To the best of our knowledge, this is the first study to analyze and exploit CLIP cues as prior knowledge for object proposal generation. We analyze the CLIP for proposal generation and design a CLIP proposal selection model, a graph-based proposal merging model as well as a proposal regression model to further refine and tailor CLIP cues. (3) Extensive experiments show that our proposed framework obtains significant improvements on three popular datasets and has benefits for downstream tasks.

2. Related Work

Supervised Object Proposal Generation. Fully-supervised object proposal generation methods use bounding box annotations to train models and select proposals from initial bounding boxes. BING [6] used normed image gradients (NG) as the cue and trained a support vector machine (SVM) to select proposals. BING++ [47] further incorporated edges and segments to improve the proposal localization quality. These methods are based on low- and mid-level cues. Many deep-learning-based methods have been developed to explore higher-level cues for the proposal generation task. DeepBox [18] designed a four-layer CNN to re-rank initial proposals generated by Edge Boxes [50]. Faster RCNN [31] built a region proposal network (RPN) including a classifier and a box regressor to select and correct bounding boxes generated by anchors. RFP-Net [14] and Refinedbox [22] modified RPN. RFP-Net [14] used receptive fields (RFs) to generate initial boxes to remove many hyper-parameters of anchor boxes in RPN. Refinedbox [22] replaced the classifier in RPN with a ranking model to re-rank bounding boxes from Edge Boxes [50]. A number of methods [3, 21, 30] also trained deep networks to directly regress object proposals. Although these fully-supervised methods achieve high-quality proposals, they need a huge set of bounding box annotations for training.

To reduce the requirement of human annotations, weakly-supervised methods [5, 15, 33, 39, 40, 48, 49] only use image-level labels rather than bounding box annotations. They usually leverage image-level labels to train a classification network and generate class activation maps (CAMs) from the trained model. Then, they extract proposals from these CAMs. Some of weakly-supervised methods [5,40,49] select high-confidence proposals as pseudo labels to train a proposal generation model in a fully-supervised manner to improve the accuracy. Graph-based technologies are also used in several works [15, 39] to extract seeds and cluster centers for better proposal generation. However, weakly-supervised methods still require costly image classification annotations and human annotators. In addition, both fully- and weakly- supervised method can only generate proposals for a limited number of object categories.

Unsupervised Object Proposal Generation. To avoid these limitations of supervised object proposal generation, unsupervised methods have attracted increasing research interests [41, 50] in recent years, which does not need annotations on the target dataset. Early methods such as Selective Search [41] and Edge Boxes [50] employed color, texture or edge cues to predict proposals. These methods contain no training process and thus avoid human annotations. However, these methods are only able to leverage low-level information. High-level information such as deep learning features are hard to be used in unsupervised settings, because deep learning requires training data and an-

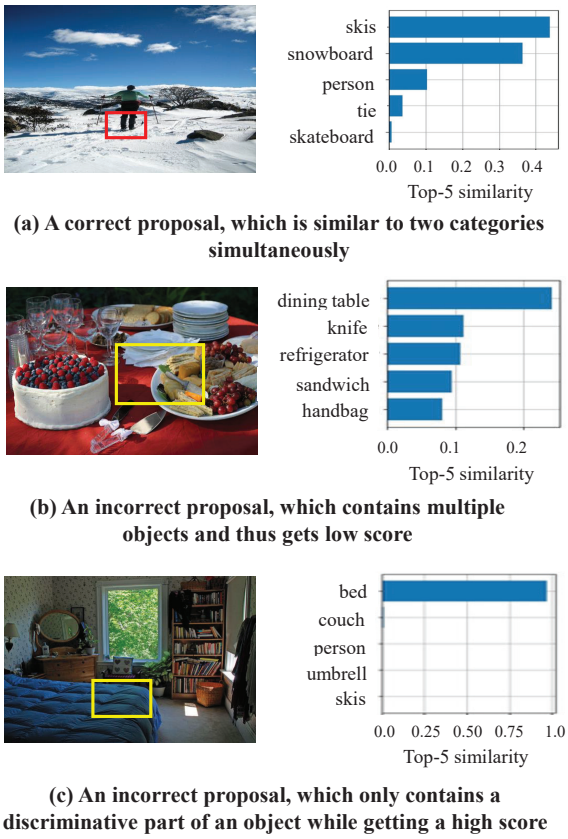
notations. To solve this problem, Detco [44] proposed to use contrastive learning to train the deep learning model in a self-supervised manner. Nevertheless, the model trained by contrastive learning can only extract features. A fully-supervised detector is still needed for bounding boxes prediction. Inspired by weakly-supervised techniques, some recent methods [37,42,43] predict proposals from CAMs or attention maps provided by pre-trained classification models. Wei *et al.* [43], Vo *et al.* [42] and Simeoni *et al.* [37] employed PCA-, saliency- and seed-based methods to generate proposals from CAMs/attention maps, respectively. Moreover, these methods also used image group knowledge provided by human. The image set is divided into multiple groups by human and each group contains a common object. A common pitfall of these approaches is that they are only able to predict proposals for limited object categories, due to the limitation of pre-trained classification models. Moreover, they search object proposals based on CAMs/attention maps, which highlight important regions in images but ignore objects in non-activated regions. Different from these approaches, our method can generate proposal for open categories and cover diverse objects.

CLIP Cue. CLIP [28] is an image-text matching model, which contains a visual feature encoder and a textual feature encoder. It was trained with numerous image-language pairs to embed the visual features and textual features into a unified feature space. Because it shows good ability to recognize diverse language and object categories, its feature space has been used as cues in many applications, such as image generation [26], image-text retrieval [2, 9, 38], image classification [7, 32] and image captioning [4]. Inspired by these works, we exploit CLIP features as cues for unsupervised open-category object proposal generation.

3. Proposed Method

3.1. CLIP Feature Analysis for Proposal Generation

Previous deep-learning-based unsupervised proposal generation methods leverage pre-trained classifiers as prior knowledge to extract salient object regions. However, they are only able to generate proposals for a fixed number of object categories, because their classifiers are trained with a fixed set of classes. In this paper, we exploit prior knowledge from CLIP [28] image-text matching as cues. CLIP consists of a visual feature encoder (ViT [8]) and a textual feature encoder (GPT-2 [29]), and embeds visual and textual features into a same feature space for matching. Unlike classification models, CLIP is trained to match images with their corresponding natural language descriptions, and thus has potential to recognize diverse objects in real world. However, directly extracting objects from CLIP attention maps like previous methods ignores objects in non-salient regions. It is also hard to separate overlapped instances [37].



(a) A correct proposal, which is similar to two categories simultaneously

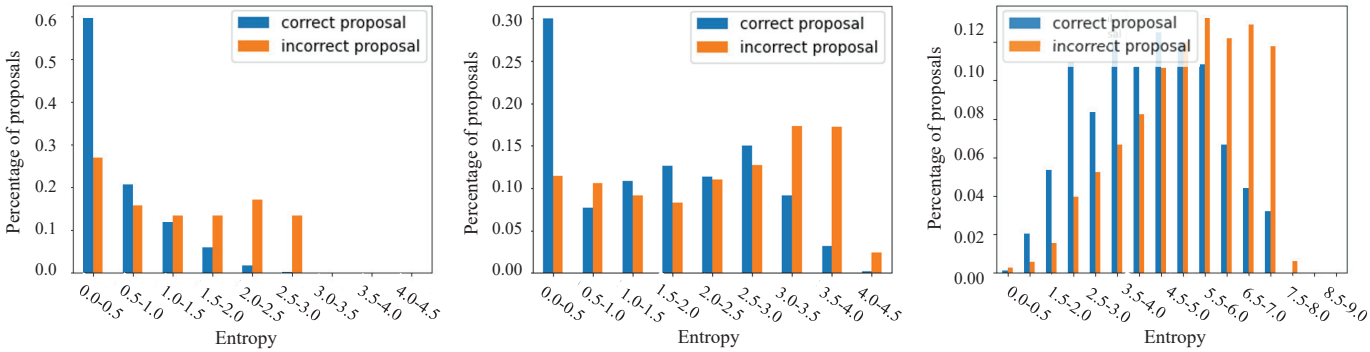
(b) An incorrect proposal, which contains multiple objects and thus gets low score

(c) An incorrect proposal, which only contains a discriminative part of an object while getting a high score

Figure 2. Some examples of CLIP matching results for proposals.

Meanwhile, since CLIP [28] is trained to recognize a single object, it cannot directly encode features for an image containing multiple objects. Therefore, our basic idea is to use an existing proposal method such as Edge Boxes [50] to extract a mass of candidate single-object proposals, while leveraging CLIP prior knowledge to evaluate their objectness to select proposals.

To evaluate the objectness, we first extract visual features of each initial proposal using CLIP image encoder, while using the textual feature encoder to capture textual features of the candidate object categories. In a specific dataset such as Microsoft COCO [20], its object categories (80 categories for COCO) can serve as the candidate object categories. In real world, we can use a large noun dictionary as candidate object categories. After the feature extraction, for each proposal, we calculate its feature similarity with every candidate category and use softmax function to normalize these similarities. A simple way is to use the **maximum similarity** as the objectness score, because one well-extracted proposal usually has an exact category. However, as shown in Fig. 2 (a), we observed that in open-category proposal generation, some well-extracted proposals may be simultaneously assigned to multiple categories,



(a) VOC 2007 (20 Categories) (b) COCO (80 Categories) (c) Visual Genome (1600 Categories)

Figure 3. Distributions of similarity entropies on the PASCAL VOC 2007, COCO and Visual Genome training sets. Correct proposals mean their IoU with ground truths higher than 0.5, while incorrect proposals with IoU lower than 0.5.

Proposals	VOC 2007	COCO	Visual Genome
Correct proposals	0.56	1.29	3.38
Incorrect proposals	1.32	2.41	5.03

Table 1. Average similarity entropies of proposals on PASCAL VOC 2007, COCO and Visual Genome training sets.

because semantically confusing categories are unavoidable in real world. It is also not always accurate to use multiple top similarities to evaluate the proposal objectness. Therefore, rather than using the maximum similarity, we propose a simple yet effective objectness estimation method based on the **similarity entropy**.

Table 1 shows average similarity entropies of different proposals on multiple datasets. Fig. 3 depicts the distributions of similarity entropies. Here, we use a traditional proposal method [50] to generate a large number of proposals and divide them into two types, correct and incorrect proposals. Correct proposals are defined as any proposal whose IoU with ground truths are higher than 0.5, while others are incorrect proposals. From Table 1 and Fig. 3, we can observe the followings: (1) correct proposals show significantly lower average CLIP similarity entropies than incorrect proposals on all datasets; (2) correct proposals are mainly distributed in low-entropy ranges, while incorrect proposals dominate high-entropy ranges; and (3) different datasets show distinctly different entropy ranges.

Based on observations (1) and (2), setting a threshold on CLIP similarity entropy can directly filter out about 40% incorrect proposals while keeping most of the correct proposals. However, as pointed out in the observation (3), different datasets require to carefully set different thresholds, which substantially reduces the generalization ability in real-world applications. Hence, we propose to filter out proposals through the percentage instead of a threshold. For example, we can select 60% low-similarity-entropy initial proposals to remove a mass of incorrect proposals while keeping correct proposals. Meanwhile, we also propose to use the CLIP similarity entropy to re-score proposals for better generation, as described in Sec. 3.2.2.

We also observe **limitations** of the CLIP similarity entropy. As shown in Fig. 2 (c), for incorrect proposals only containing a discriminative part of an object, CLIP shows low entropies and cannot remove them based on the similarity entropy. To solve this problem, we propose a graph-based proposal merging model in Sec. 3.2.3, which is able to merge such fragmented proposals by leveraging CLIP features. We next introduce our ProposalCLIP based on these observations and analysis.

3.2. ProposalCLIP

Our ProposalCLIP contains four modules, as illustrated in Fig. 4: (a) an initial proposal generation model to get candidate proposals from the input image, (b) a CLIP proposal selection model that leverages CLIP feature space to refine proposals from the candidate proposals, (c) a graph-based merging model that merges fragmented proposals based on both spatial and CLIP cues, and (d) a proposal regression model that further refines proposals. Below we introduce the details of each module.

3.2.1 Initial Proposal Generation

Given the input image I, we first obtain initial proposals $\{O_m\}_{m=1}^{M}$, where M is the number of initial proposals and $O_m \in R^4$ represents the coordinates of the m-th proposal. Our method can use any existing proposal generation model to predict initial proposals. Here, we use Edge Boxes [50] as an example, which generates proposals based on low-level image information. Besides proposal coordinates, the existing proposal generation model also predicts an object-

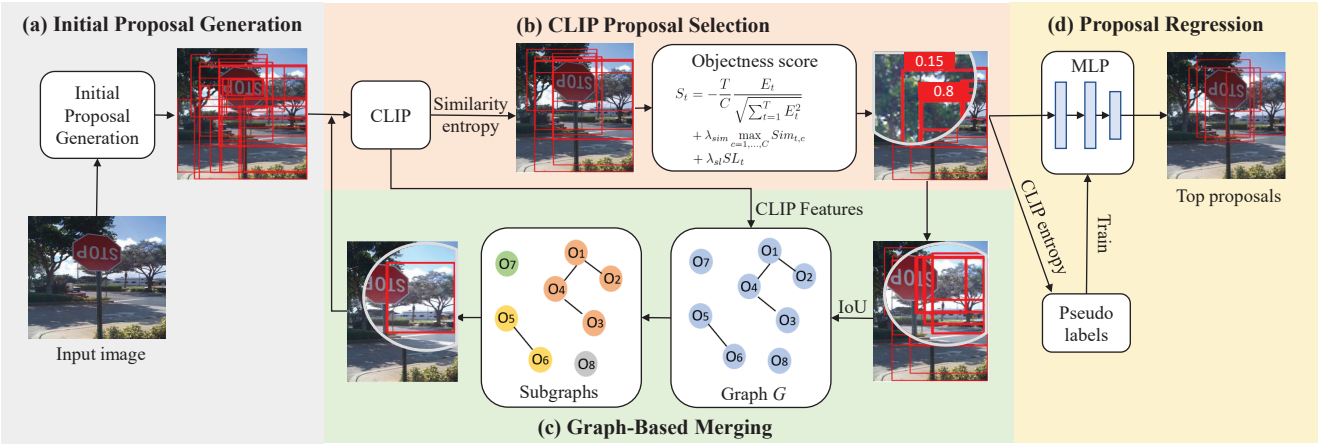


Figure 4. Illustration of our ProposalCLIP. (a) The initial proposal generation model extracts initial proposals. (b) The CLIP proposal selection model selects and re-scores proposals based on CLIP cues. (c) The graph-based proposal merging model corrects fragmented proposals based on CLIP features. (d) The proposal regression model refines proposals.

ness score SL_m for each proposal O_m. This score can also be used in the following selection.

3.2.2 CLIP Proposal Selection

Based on the analysis in Sec. 3.1, we propose to leverage CLIP similarity entropy to estimate objectness of initial proposals and select proposals. Specifically, for each initial proposal O_m, we first use CLIP vision encoder to extract its features V_m. Then, we encode textual features $\{T_c\}_{c=1}^{C}$ of every candidate object category, where C is the number of candidate categories. Next, we calculate the cosine similarity between the visual features of the proposal and the textual features of each category:

$$\widetilde{Sim}_{m,c} = \frac{V_m \cdot T_c}{\|V_m\|\|T_c\|}. \tag{1}$$

The cosine similarity $\widetilde{Sim}_{m,c}$ is then normalized by a softmax function over all classes, and the normalized similarity is denoted as $Sim_{m,c}$.

After than, the CLIP similarity entropy for the proposal O_m can be obtained as follows:

$$E_m = -\sum_{c=1}^{C} Sim_{m,c} \times log(Sim_{m,c}). \tag{2}$$

We then filter out high-entropy proposals to remove a large number of incorrect proposals but retain most of correct proposals. The T retained proposals are represented by $\{O_t\}_{t=1}^{T}$. In our experiments, we remove 40% high-entropy proposals.

However, as shown in Fig. 3, there are still many incorrect proposals in the retained proposals. Therefore, we propose a CLIP-based objectness score to re-rank retained proposals for further selection as:

$$S_t = -\frac{T}{C}\frac{E_t}{\sqrt{\sum_{t=1}^{T} E_t^2}} + \lambda_{sim} \max_{c=1,\dots,C} Sim_{t,c} + \lambda_{sl} SL_t \tag{3}$$

where the first item is the negative entropy score, the second and third items are the maximum similarity and the initial score, respectively. We use them as references, because proposals are also probably correct when they have extremely high maximum similarities and the initial scores. Similarity entropies for different proposals are normalized by L2 normalization. We use $\frac{T}{C}$ to automatically weight the first item to be in an appropriate range. λ_{sim} and λ_{sl} are coefficients to control the proportion of each item. We then select out correct proposals based on this objectness score.

3.2.3 Graph-Based Proposal Merging

Although our CLIP proposal selection model can filter out a large number of incorrect proposals, some fragmented proposals which only contain a discriminative part of an object cannot be easily filtered. Thus, we propose a graph-based proposal merging model to solve this limitation.

Specifically, we first build an undirected graph $\mathcal{G}$: $\mathcal{G} =< \mathcal{N}, \mathcal{E} >$. Nodes $\mathcal{N}$ in the graph are proposals $\{O_t\}_{t=1}^{T}$ selected by our CLIP proposal selection model. Edges $\mathcal{E}$ are computed by spatial and semantic similarities between these proposals. We use the IoU (intersection over union) between two proposals to evaluate their spatial similarity as follows:

$$IoU_{i,j} = \frac{O_i \cap O_j}{O_i \cup O_j} \tag{4}$$

where $IoU_{i,j}$ is the IoU between proposals O_i and O_j. Their semantic similarity $PSim_{i,j}$ is estimated by the cosine similarity between their CLIP visual features:

$$PSim_{i,j} = \frac{V_i \cdot V_j}{\|V_i\|\|V_j\|}. \tag{5}$$

After capturing both the spatial and semantic similarities, edges in the graph are calculated as:

$$\begin{aligned} e_{i,j} &= U(IoU_{i,j} - Thr_{IoU}) \\ &\times U(PSim_{i,j} - Thr_{PSim}) \end{aligned} \tag{6}$$

where $e_{i,j} \in \mathcal{E}$ represents the edge between nodes O_i and O_j. $Thr_{IoU} = 0.5$ and $Thr_{PSim} = 0.9$ are thresholds for IoU and visual feature similarity, respectively. $U(\cdot)$ represents a unit step function. Here, we use a strict criterion to generate edges. There is an edge ($e_{i,j} = 1$) between two proposals only when they are well overlapped and have very similar features.

Once the graph $\mathcal{G}$ is built, we then determine all maximal connected subgraphs $\{\mathcal{H}_k\}_{k'=1}^{K'}$. We delete subgraphs which only contain one node (i.e., one proposal) and merge proposals in each of the remaining subgraph $\mathcal{H}_{k'}$. Finally, K merged proposals are generated and represented by $\{\tilde{O}_k\}_{k=1}^{K}$.

We use the CLIP proposal selection model to evaluate the objectness of the merged proposals, by generating their similarity entropies $\{\tilde{E}_k\}_{k=1}^{K}$ and objectness scores $\{\tilde{S}_k\}_{k=1}^{K}$. Some merged proposals are removed, if their entropies are higher than the maximum entropy in selected proposals (i.e., $\tilde{E}_k > \max_{t=1,\dots,T} E_t$). The rest of the merged proposals are added to the set of selected proposals.

3.2.4 Proposal Regression

We further propose a proposal regression model as an optional part to refine proposals. The first three parts in our method do not need any training. If there is an images set without annotations, our proposal regression model can extract pseudo labels from this set and train a regression model for proposal refinement. To extract pseudo labels, we leverage the first three parts in our method to generate proposals. We then select the intersection between top 1% low-entropy proposals and top 5% high-initial-score proposals as pseudo labels $\{Y_n\}_{n=1}^{N}$, where N is the number of pseudo labels.

A lightweight MLP (Multilayer Perceptron) is built to regress proposals. Inspired by RPN [31], we take the proposal visual features as the input. We also input the visual features of the entire image and the normalized coordinates of the proposal as references. Our MLP consists of three fully-connected layers with batch normalization and ReLU activation. The first layer is used to fuse input features, the second layer is to transform the fused features and the final layer outputs the normalized coordinates of the refined proposal $\hat{O}$.

The model is trained by Smooth L1 loss as follows:

$$Loss = \begin{cases} 0.5(\hat{O}_n - Y_n)^2, & \|\hat{O}_n - Y_n\| < 1 \\ \|\hat{O}_n - Y_n\| - 0.5, & \|\hat{O}_n - Y_n\| > 1 \end{cases} \tag{7}$$

where $\hat{O}_n$ is the regressed proposals on the training set and Y_n is the pseudo label.

The objectness of refined proposals are also estimated by the CLIP proposal selection model. If the CLIP entropy of the refined proposal is lower than the original one and their IoU is higher than 0.75, we replace the original proposal with the refined one. If not, we keep the original proposal.

4. Experiments

4.1. Datasets and Metrics

We verify our method on three object proposal generation datasets, PASCAL VOC 2007 [10], COCO 2017 [20] and Visual Genome [17]. PASCAL VOC 2007 [10] consists of 9,963 images and 20 object categories. It is split into training, validation and test sets containing 2,501, 2,510 and 4,952 images, respectively. Similar to previous works [22, 31], we verify our method on the test set, while combining images in training and validation sets to train the proposal regression model. COCO 2017 [20] contains 123,287 images and 80 object categories, which is divided into training set (118,287 images) and validation set (5,000 images). We use the validation set to test, while employing the training set to extract pseudo labels. Visual Genome [17] includes 107,228 images. Since it has over 1,600 object categories, we choose it to evaluate the open-category ability of our method. We randomly select 5,000 images which contain about 50,000 bounding boxes for testing and 2,000 images for training. We call it as "Visual Genome mini".

We adopt common proposal generation metrics, Recall and AR (Average Recall), to evaluate the performance. Recall@X is the ratio of well found ground truth objects whose IoU with a proposal is higher than the threshold X. AR is the average recall at IoU thresholds from 0.5 to 0.95. We use AP (Average Precision) to evaluate the performance of unsupervised object detection.

4.2. Implementation Details

Our method can use any existing proposal generation technique as our initial proposal generation module. In our experiments, we use Edge Boxes [50] as an example and generate 300 initial proposals per image. In the CLIP proposal selection model , we select 60% low-entropy proposals and set λ_{sim} to 0.06 and λ_{sl} to 1. In the graph-based proposal merging model, we set Thr_{IoU} and Thr_{PSim} to

Method	VOC 2007										COCO									
	Recall@0.5 (%)					AR(%)					Recall@0.5 (%)					AR(%)				
	1	10	30	50	100	1	10	30	50	100	1	10	30	50	100	1	10	30	50	100
Fully-supervised																				
DeepBox [18]	-	58.1	71.8	77.2	84.5	-	33.9	44.5	49.2	54.9	-	21.9	32.3	38.4	47.5	-	12.5	18.9	22.5	27.8
RPN [31]	-	60.1	73.8	80.7	89.0	-	28.4	38.1	42.7	48.9	-	30.6	46.2	55.1	65.0	-	16.1	25.0	30.2	36.1
RefinedBox [22]	-	79.5	88.6	90.8	92.4	-	49.8	56.1	57.7	59.0	-	44.6	57.3	62.4	68.1	-	30.4	38.2	41.1	44.3
Unsupervised																				
Selective search [41]	11.3	35.7	52.3	59.8	69.1	4.9	16.5	27.7	33.9	42.0	3.3	11.1	19.6	24.2	31.0	1.5	4.1	8.5	11.2	15.7
Edge boxes [50]	15.2	42.5	58.3	64.7	72.5	7.6	24.2	35.1	39.9	46.3	5.5	17.1	25.9	30.5	36.5	3.0	10.9	16.2	18.7	23.5
rOSD [42]	16.6	33.2	42.9	45.3	49.8	6.8	15.3	21.5	22.1	25.5	4.7	13.5	22.7	25.4	27.1	1.5	4.6	9.8	12.2	13.9
LOST [37]	18.8	23.7	25.4	26.3	27.7	7.2	10.2	11.7	12.6	13.7	5.0	6.6	7.4	7.8	8.5	1.6	2.4	2.8	3.1	3.6
Ours	**22.1**	**52.1**	**65.8**	**71.7**	**78.0**	**10.6**	**29.6**	**39.3**	**43.5**	**48.3**	**11.2**	**27.1**	**33.5**	**35.7**	**38.3**	**4.8**	**14.3**	**20.2**	**23.9**	**26.8**

Table 2. Proposal generation results on the PASCAL VOC 2007 test set and COCO validation set.

Method	Recall@0.5 (%)				
	1	10	30	50	100
Cross-domain					
Faster RCNN [31] (trained on COCO)	**10.3**	22.5	29.0	31.4	32.1
Unsupervised					
Selective search [41]	4.0	11.3	17.8	28.1	38.7
Edge boxes [50]	4.9	14.5	23.9	33.8	45.8
rOSD [42]	5.6	15.7	25.5	28.3	33.5
LOST [37]	5.8	12.1	13.2	13.6	13.9
Ours	8.5	**24.1**	**33.0**	**38.7**	**47.1**

Table 3. Proposal generation results on the Visual Genome mini dataset.

0.5 and 0.9, respectively. In the proposal regression model, we train the MLP for 30 epochs and the learning rate is 1e-5. All experiments are conducted on the Pytorch deep learning platform [25] on one Nvidia RTX 3090 GPU.

4.3. Comparison with State of the Art

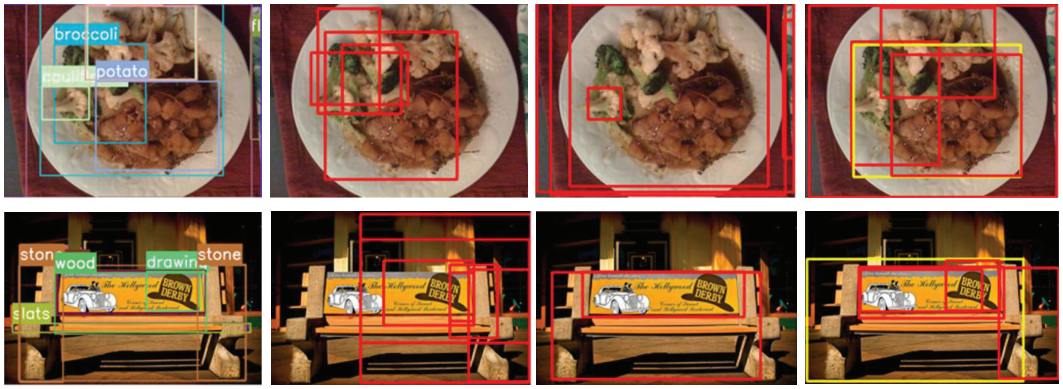
We first compare the unsupervised object proposal generation performance on the VOC 2007 and COCO datasets. The results are shown in Table 2. It can be seen that compared with Edge boxes [50], our method achieves improvements by a large margin, especially when extracting a small number of proposals (e.g., 1, 10 and 30). When extracting 10 proposals, our method outperforms Edge boxes [50] by about 10% on VOC 2007 and COCO, in the term of Recall@0.5. When extracting a large number of proposals, our method also shows significantly improvements. Compared with the CNN-based method rOSD [42] and the Transformer-based method LOST [37], our method achieves gains of 3.3% on VOC 2007 and 6.2% on COCO in Recall@0.5, when extracting only one proposal. Moreover, rOSD [42] and LOST [37] are both based on CAMs/attention maps. They do not well generate multiple proposals, while our method can simultaenously predict multiple proposals for an image. We also report the results from some fully-supervised methods, as reported in [22]. It can be observed that our method outperforms the fully-supervised method DeepBox [18] on COCO, when extracting a small number of proposals. These superior results demonstrate the effectiveness of our method.

We then compare the open-world proposal generation capability on Visual Genome mini [10]. As shown in Table 3, our methods outperforms all previous unsupervised methods. Faster RCNN [31] trained on COCO achieves good performance when extracting a few objects (1 and 10 objects), because it is well trained to capture 80-category objects. However, it cannot well extract more objects, due to no annotation for other object categories. Our method shows better recalls when extracting more objects. Some qualitative results are shown in Fig. 5. It is can be seen that Faster RCNN [31] trained on COCO fails to capture some objects, such as the "potato" and "drawing". This is because Faster RCNN [31] requires annotations for these objects. Our method can recognize open-category objects without annotations, and thus successfully extracts these objects. It is also seen that the top 100 proposals generated by our ProposalCLIP are better than those from Edge boxes [50].

4.4. Ablation Study

The contribution of each component. Table 4 shows the effects of different components in our method. We find that our CLIP proposal selection model achieves the most improvements while only slightly reducing the speed. The graph-based proposal merging also significantly improves the performance. We visualized proposals generated by our graph-based merging model in Fig. 5. It can be found that our model successfully captures some objects by the graph-based merging. The proposal regression model decreases the recall when generating a few proposals. This could be caused by unavoidable noises in pseudo labels. Nevertheless, the proposal regression model increases the performance in most cases. These results demonstrate the effectiveness of our CLIP proposal selection, graph-based proposal merging and proposal regression models.

Objectness scores. We report the effects of different objectness scores in Table 5. It can been seen that our sim-



Ground Truth EdgeBox Faster RCNN Ours

Figure 5. Qualitative results on the Visual Genome mini dataset. Top to bottom: ground truth, correct proposals from Edge boxes [50], Faster RCNN [31] trained on COCO and our ProposalCLIP with 100 proposals. Yellow boxes in our results are generated by our graph-based merging model.

Method	Recall@0.5 (%) 1	10	30	50	Time (s)
Initial proposals (Edge boxes [50])	15.2	42.5	58.3	64.7	0.9
+ CLIP Proposal Selection	21.0	51.8	62.1	68.5	1.2
+ Graph-Based Proposal Merging	21.6	**52.4**	64.6	70.8	1.5
+ Proposal Regression	**22.1**	52.1	**65.8**	**71.7**	1.9

Table 4. The effects of different components on the PASCAL VOC 2007 test set. Time means the running time per image.

Objectness score	Recall@0.5 (%) 1	10	30	50	100
Edge boxes [50]	15.2	42.5	58.3	64.7	72.5
Initial score	19.1	46.5	61.2	65.7	74.4
Maximum similarity	17.6	50.7	60.0	65.3	74.1
CLIP similarity entropy	20.2	**52.5**	60.4	66.6	75.0
Our final objectness score	**21.0**	51.8	**62.1**	**68.5**	**76.3**

Table 5. The effects of different objectness scores on the PASCAL VOC 2007 test set.

ilarity entropy outperforms the initial score and maximum similarity. The initial score gains better performance than the original Edge boxes [50], thanks to our CLIP entropy selection. Our final objectness score achieves the best performance.

ProposalCLIP for downstream tasks. We further conduct an experiment to demonstrate the usefulness of our proposals for downstream tasks. Table 6 presents the performance on unsupervised object detection on COCO. We use CLIP [28] as a classifier to classify each proposal and leverage NMS (Non-Maximum Suppression) to generate the final results. It can be observed that our Proposal CLIP achieves an 8.5% AP in an unsupervised manner.

Limitations. Our method cannot well deal with small objects, such as the "cauliflower" object in the first image

Method	AP@0.5 (%)
Edge boxes [50] + CLIP [28] + NMS	6.3
LOST [37] + CLIP [28] + NMS	5.2
Ours + CLIP [28] + NMS	**8.5**

Table 6. Unsuperivsed object detection results on the COCO validation set.

in Fig. 5. Because small objects are usually in low resolution, it is hard to recognize them using CLIP. Meanwhile, the initial proposal generation model is also hard to capture small objects. We leave the exploration of super-resolution techniques to solve this problem for future work.

5. Conclusion

In this paper, we have presented ProposalCLIP, an effective approach for unsupervised open-category object proposal generation. In our approach, a proposal selection model is first introduced to recognize open-category objects and select proposals for them by CLIP-based objectness scores. Secondly, a graph-based merging model unifies fragmented proposals based on CLIP feature cues. Thirdly, we introduce a regressor module that leverages CLIP cues to refine proposals. Experimental results have demonstrated that our proposed method is capable to find open-category proposals in an unsupervised manner. Our method also outperforms existing state-of-the-art methods by a large margin on three popular datasets, and shows the benefits for downstream tasks. **Social impacts**: We test our method on three popular datasets, which might contain ethic biases. It would be better to collect more unbiased data in real world to reduce this impact.

Acknowledgement: This research is supported in part by Monash FIT Start-up Grant.

References

[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2189–2202, 2012. 1

[2] Shuai Bai, Zhedong Zheng, Xiaohan Wang, Junyang Lin, Zhu Zhang, Chang Zhou, Hongxia Yang, and Yi Yang. Connecting language and vision for natural language-based vehicle retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4034–4043, 2021. 1, 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 2

[4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 3

[5] Gong Cheng, Junyu Yang, Decheng Gao, Lei Guo, and Junwei Han. High-quality proposals for weakly supervised object detection. *IEEE Transactions on Image Processing*, 29:5794–5804, 2020. 2

[6] Ming-Ming Cheng, Yun Liu, Wen-Yan Lin, Ziming Zhang, Paul L Rosin, and Philip HS Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. *Computational Visual Media*, 5(1):3–20, 2019. 1, 2

[7] Ruizhe Cheng, Bichen Wu, Peizhao Zhang, Peter Vajda, and Joseph E Gonzalez. Data-efficient language-supervised zero-shot learning with self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3119–3124, 2021. 3

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021. 3

[9] Maksim Dzabraev, Maksim Kalashnikov, Stepan Komkov, and Aleksandr Petiushko. Mdmmt: Multidomain multimodal transformer for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3363, 2021. 3

[10] 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(2):303–338, 2010. 6, 7

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 1

[12] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 42(02):386–397, 2020. 1

[13] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016. 1

[14] Lin Jiao, Shengyu Zhang, Shifeng Dong, and Hongqiang Wang. Rfp-net: Receptive field-based proposal generation network for object detection. *Neurocomputing*, 405:138–148, 2020. 2

[15] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1377–1385, 2017. 2

[16] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5840, 2021. 1

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2016. 6

[18] Weicheng Kuo, Bharath Hariharan, and Jitendra Malik. Deepbox: Learning objectness with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2479–2487, 2015. 1, 2, 7

[19] Xiangyang Li and Shuqiang Jiang. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia*, PP(99):1–1, 2019. 1

[20] 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 *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014. 3, 6

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 2

[22] Yun Liu, Shijie Li, and Ming-Ming Cheng. Refinedbox: Refining for fewer and high-quality object proposals. *Neurocomputing*, 406:106–116, 2020. 1, 2, 6, 7

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, 2021. 1

[24] Shweta Mahajan and Stefan Roth. Diverse image captioning with context-object split latent spaces. *Advances in Neural Information Processing Systems*, 2020. 1

[25] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 7

[26] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[27] Heqian Qiu, Hongliang Li, Qingbo Wu, Fanman Meng, Hengcan Shi, Taijin Zhao, and King Ngi Ngan. Language-aware fine-grained object representation for referring expression comprehension. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4171–4180, 2020. 1

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 3, 8

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 3

[30] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788, 2016. 2

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(06):1137–1149, 2017. 1, 2, 6, 7, 8

[32] Mert Bulent Sariyildiz, Yannis Kalantidis, Diane Larlus, and Karteek Alahari. Concept generalization in visual representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9629–9639, 2021. 3

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[34] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 38–54, 2018. 1

[35] Hengcan Shi, Hongliang Li, Fanman Meng, Qingbo Wu, Linfeng Xu, and King Ngi Ngan. Hierarchical parsing net: Semantic scene parsing from global scene to objects. *IEEE Transactions on Multimedia*, 20(10):2670–2682, 2018. 1

[36] Hengcan Shi, Hongliang Li, Qingbo Wu, and Zichen Song. Scene parsing via integrated classification model and variance-based regularization. In *IEEE conference on computer vision and pattern recognition*, 2019. 1

[37] Oriane Siméoni, Gilles Puy, Huy V Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *BMVC*, 2021. 1, 2, 3, 7, 8

[38] Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 982–997, 2021. 3

[39] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan Yuille. Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE transactions on pattern analysis and machine intelligence*, 42(1):176–191, 2018. 2

[40] Peng Tang, Xinggang Wang, Angtian Wang, Yongluan Yan, Wenyu Liu, Junzhou Huang, and Alan Yuille. Weakly supervised region proposal network and object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 352–368, 2018. 2

[41] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 1, 2, 7

[42] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *European Conference on Computer Vision*, pages 779–795. Springer, 2020. 1, 2, 3, 7

[43] Xiu-Shen Wei, Chen-Lin Zhang, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. *Pattern Recognition*, 88:113–126, 2019. 3

[44] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8392–8401, 2021. 3

[45] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 1

[46] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 1

[47] Ziming Zhang, Yun Liu, Xi Chen, Yanjun Zhu, Ming-Ming Cheng, Venkatesh Saligrama, and Philip HS Torr. Sequential optimization for efficient high-quality object proposal generation. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1209–1223, 2017. 1, 2

[48] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2

[49] Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1841–1850, 2017. 2

[50] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014. 1, 2, 3, 4, 6, 7, 8