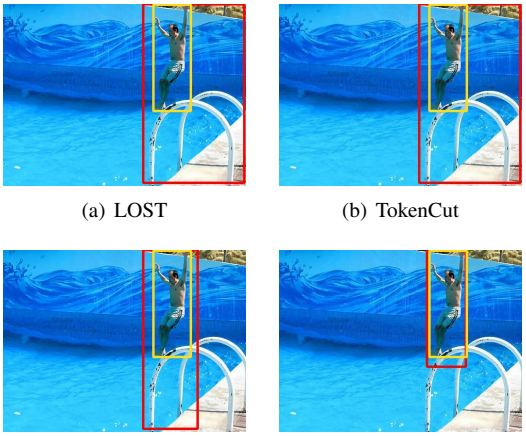
Foreground Guidance and Multi-Layer Feature Fusion for Unsupervised Object Discovery with Transformers

Zhiwei Lin* Zengyu Yang*† Yongtao Wang✉
Wangxuan Institute of Computer Technology, Peking University
zwlin@pku.edu.cn, yzyysj@gmail.com, wyt@pku.edu.cn

Abstract

*Unsupervised object discovery (UOD) has recently shown encouraging progress with the adoption of pre-trained Transformer features. However, current methods based on Transformers mainly focus on designing the localization head (*e.g.*, seed selection-expansion and normalized cut) and overlook the importance of improving Transformer features. In this work, we handle UOD task from the perspective of feature enhancement and propose **FOR**eground guidance and **MU**lti-**LA**yer feature fusion for unsupervised object discovery, dubbed FORMULA. Firstly, we present a foreground guidance strategy with an off-the-shelf UOD detector to highlight the foreground regions on the feature maps and then refine object locations in an iterative fashion. Moreover, to solve the scale variation issues in object detection, we design a multi-layer feature fusion module that aggregates features responding to objects at different scales. The experiments on VOC07, VOC12, and COCO_20k show that the proposed FORMULA achieves new state-of-the-art results on unsupervised object discovery. The code will be released at https://github.com/VDIGPKU/FORMULA.*



(a) LOST (b) TokenCut (c) FORMULA-L (ours) (d) FORMULA-TC (ours)

Figure 1. **Example results of UOD on VOC12.** In (a) and (b), we show results obtained by LOST [40] and TokenCut [52] respectively, which are two state-of-the-art UOD methods. The results of our method are presented in (c) and (d). Predictions are in red, and ground-truth boxes are in yellow. We can find that the proposed FORMULA localizes object more accurately. Best viewed in color.

1. Introduction

Object detection is one of the fundamental problems in computer vision, which serves a wide range of applications such as face recognition [42], pose estimation [57], and autonomous driving [44]. In recent years, significant success has been achieved in the field [36, 35] thanks to the increasing amount of annotated training data. However, the labeling of large-scale datasets [31, 38] is rather costly. Although multiple techniques, including semi-supervised learning [2], weakly-supervised learning [17], and self-supervised learning [19] have been proposed to alleviate this issue, manual labeling is still required.

Here we concentrate on a fully unsupervised task for object detection, named unsupervised object discovery. Previous CNN-based methods [13, 47, 48, 54] leverage region proposal and localize objects by comparing the proposed bounding boxes of each image across a whole image collection. However, these approaches are difficult to scale to large datasets due to the quadratic complexity brought by the comparison process [49]. Recently, DINO [6] has found that the attention maps of a Vision Transformer (ViT) [14] pre-trained with self-supervised learning reveal salient foreground regions. Motivated by DINO, LOST [40] and TokenCut [52] are proposed to discover objects by leveraging the high-quality ViT features. Both methods first construct an undirected graph using the similarity of patch-wise features from the last layer of the ViT. Then, a two-step seed-selection-expansion strategy and Normalized Cut [39] are adopted respectively to segment the foreground objects. While both approaches have achieved superior results over previous state-of-the-art [48, 49], we have found that they

*Equal contribution. †As an intern at PKU. ✉Corresponding author.

mainly focus on the construction of the localization head and overlook the potential of improving the ViT features.

In this paper, we propose a simple but effective feature enhancement method for existing ViT-based UOD frameworks, named FORMULA. Our method consists of two parts, *i.e.*, the foreground guidance module and the multi-layer feature fusion module. For the foreground guidance module, we utilize the object mask predicted by an off-the-shelf UOD detector to highlight the foreground object region and then refine the object location through an iterative process. Specifically, we first generate an object mask from the original ViT feature map using an existing UOD detector (*e.g.*, LOST or TokenCut). Then, we construct a probability map with 2D Gaussian distribution from the mask, which roughly localizes the foreground objects. After that, we highlight the foreground area by applying the probability map to the original ViT feature map. Finally, the updated feature map is used for the UOD detector to obtain a refined object mask. The whole process can be iterated. In this way, we enhance the ViT feature map by introducing the foreground object information and suppressing background interference. Our method can localize objects much more accurately with only a few iterations.

Besides, we note that LOST and TokenCut only use the feature map from the last layer of ViT. However, the scale of objects in non-object-centric images, like those in COCO [31], can vary greatly. The feature from the last layer of a pre-trained ViT mainly captures the key areas for classification, which is usually at a larger scale. Thus, the performance on smaller objects is hurt. To address this issue, we propose the multi-layer feature fusion module. In detail, we simply merge the features from the last several layers through a weighted summation to aggregate multi-scale information for unsupervised object discovery.

Our main contributions can be summarized as follows:

- We introduce foreground guidance from the object mask predicted by an existing UOD detector to the original ViT feature map and propose an iterative process to refine the object location.
- We further design a multi-layer feature fusion module to address the scale variation issues in object detection, releasing the potential of ViT feature representation for object discovery.
- The proposed method can be easily incorporated into any existing UOD methods based on ViT and achieves new state-of-the-art results on the popular PASCAL VOC and COCO benchmarks.

2. Related Work

Self-supervised learning. Learning powerful feature representations in a self-supervised manner that dispenses with human annotation has made great progress recently. This is performed by defining a pretext task that provides surrogate supervision for feature learning [34, 59, 19]. Despite no labels, models trained with self-supervision have outperformed their supervised counterparts on several downstream tasks, such as image classification [19, 55, 8, 23, 9, 5, 20, 10, 6, 50, 11, 22] and object detection [53, 7]. While [55, 8, 23, 5, 20, 10] have adopted CNN as pre-training backbones, recent works [6, 50, 11, 22, 29, 28] have explored Transformers [46] for self-supervised visual learning, demonstrating their superiority over traditional CNN. Our work utilizes the strong localization capability of self-supervised ViT for unsupervised object discovery.

Unsupervised object discovery. The goal of unsupervised object discovery is to localize objects without any supervision. Early works generally rely on image features encoded by a CNN [24, 48, 54, 49, 12]. These methods need to compare the extracted features of each image across those of every other image in the dataset, leading to quadratic computational overhead [49]. Besides, the dependence on the inter-image similarity results in these methods being unable to run on a single image. Recently, LOST [40] and TokenCut [52] have been proposed to address these issues by leveraging the high-quality feature representation generated from a self-supervised ViT. Their motivation is that the attention map extracted from the last layer of a pre-trained ViT contains explicit information about the foreground objects. Specifically, both methods first propose the construction of an intra-image similarity graph using features extracted from the ViT backbone [6, 14]. A heuristic seed-selection-expansion strategy and Normalized Cut [39] are then adopted respectively by the two methods to segment a single foreground object in an image.

Although achieving excellent performance, these methods mainly concentrate on localization design and fail to further improve ViT features for unsupervised object discovery. Instead, our work starts from the perspective of feature enhancement. Concretely, we introduce the foreground guidance to highlight the foreground regions on ViT features and propose a multi-layer feature fusion module to aggregate multi-scale features.

Multi-layer Feature Representations. One of the main challenges in object detection is to effectively represent features on different scales. Extensive works have been proposed over the years to deal with the multi-scale problem with multi-layer features. These methods leverage the pyramidal features of CNN to compute a multi-scale feature representation. [33, 21, 26, 1, 37, 18, 30] combine low-resolution and high-resolution features with upsampling or lateral connections to aggregate semantic information from all levels. [32, 4, 30] make predictions at different scales from different layers and use post-processing to filter the

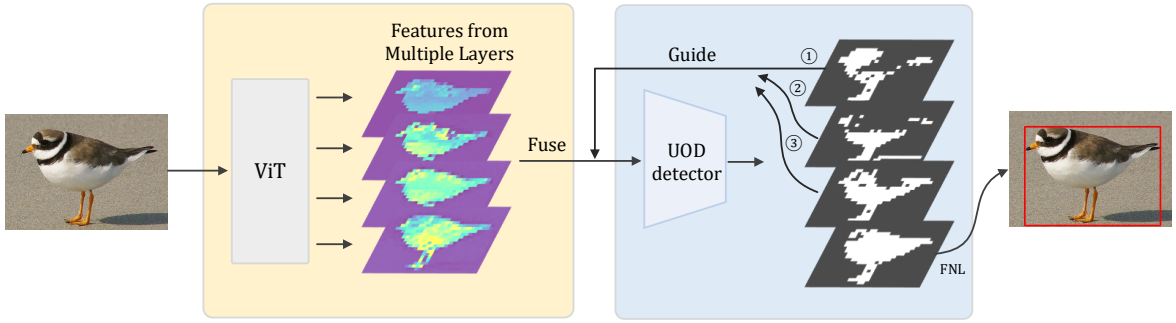


Figure 2. **The pipeline of FORMULA.** To enhance ViT features for unsupervised object discovery, we extract multi-layer features to aggregate information from different scales and introduce foreground guidance from the predicted segmentation to the input of the UOD detector.

final predictions. In addition to CNN, several works have recently exploited the multi-layer representation for Transformer networks. [51] aggregates the class tokens from each Transformer layer to gather the local, low-level, and middle-level information that is crucial for fine-grained visual categorization. In [56], the authors divide the multi-layer representations of the Transformer's encoder and decoder into different groups and then fuse these group features to fully leverage the low-level and high-level features for neural machine translation.

These works inspire us to explore the multi-layer features of ViT for better object localization. Instead of designing complicated fusion modules, we propose a simple and efficient fusion method that sums the feature from each layer of the Transformer with different weights.

3. Approach

In this section, we introduce our method for unsupervised object discovery, *i.e.*, FORMULA. The overall pipeline of FORMULA is presented in Fig. 2. Firstly, we briefly review Vision Transformers and their previous applications in UOD as preliminary knowledge. Then, we describe the two modules of FORMULA, namely foreground guidance and multi-layer feature fusion.

3.1. Preliminary

Vision Transformers [14] receive a sequence of image patches and use stacked multi-head self-attention blocks to extract feature maps from images. It divides an input image of $H \times W$ into a sequence of $N = HW/P^2$ patches of fixed resolution $P \times P$. Patch embeddings are then formed by mapping the flattened patches to a D-dimensional latent space with a trainable linear projection. An extra learnable `[CLS]` token is attached to the patch embeddings and position embeddings are added to form the standard transformer input in $\mathbb{R}^{(N+1)\times D}$.

DINO [6] has shown that the attention map extracted from the last layer of a self-supervised ViT indicates prominent foreground areas. Following this observation, LOST and TokenCut propose to localize objects using the key features $k \in \mathbb{R}^{N\times D}$ from the last layer in two steps. First, an intermediate feature map $F_{int} \in \mathbb{R}^{h\times w}$ is constructed from the inter-patch similarity graph, where $h = H/P$ and $w = W/P$. Specifically, for LOST, it is the map of inverse degrees; for TokenCut, it is the second smallest eigenvector of the graph. Second, a object mask $m \in \{0,1\}^{h\times w}$ is generated from F_{int} to segment the foreground object.

3.2. Foreground Guidance with Self-iteration

In the foreground guidance module, the predicted object mask is treated as the foreground guidance to highlight the foreground region and guide the segmentation process. Specifically, given an existing unsupervised object detector $\mathcal{D}$ and an intermediate feature map F_{int} extracted from the pre-trained ViT, the binary foreground object mask $m \in \{0,1\}^{h\times w}$ of the object can be generated as follows:

$$m = \mathcal{D}(F_{int}). \tag{1}$$

Here, $\mathcal{D}$ could be any ViT-based object discovery methods, *e.g.*, LOST and TokenCut. Moreover, the value of $m(\boldsymbol{x}_i)$ equals to 1 if the corresponding patch i with coordinates $\boldsymbol{x}_i$ is predicted to belong to the foreground object. With the foreground mask m, the approximate coordinates of the object center $\mathcal{O}$ can be calculated by

$$\boldsymbol{x}_{\mathcal{O}} = \frac{1}{\sum_{i=1}^{h\times w} m(\boldsymbol{x}_i)} \sum_{i=1}^{h\times w} m(\boldsymbol{x}_i)\boldsymbol{x}_i. \tag{2}$$

Then, we construct a probability map $P \in \mathbb{R}^{h\times w}$ using the 2D Gaussian distribution function g:

$$P(i) = g(i|\boldsymbol{x}_{\mathcal{O}}, \sigma^2) = \frac{1}{2\pi\sigma^2} e^{-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_{\mathcal{O}}\|^2}{2\sigma^2}}, \tag{3}$$

where σ is a hyper-parameter. Intuitively, the value of P indicates the regions in an image that are likely to belong

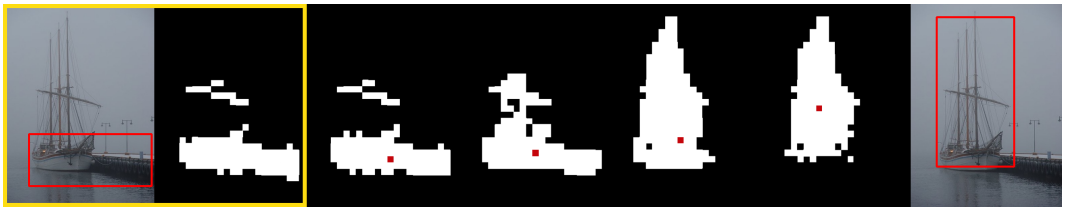


Figure 3. **Illustration of the iteration process.** The initial prediction made by LOST [40] (framed in yellow) and the final prediction after four iterations (right). The red dot in each iteration is the object center calculated from the previous segmentation.

to an object. The probability map P can be viewed as the foreground guidance for object localization, guiding the detector $\mathcal{D}$ to refine the final prediction in each iteration step. Specifically, we achieve this by applying the Hadamard product to P and F_{int}:

$$\widetilde{F_{int}} = F_{int} \circ P. \tag{4}$$

The new feature map $\widetilde{F_{int}}$ can be interpreted as a re-weighting of F_{int} and the foreground part of F_{int} is emphasized. Thus, during unsupervised object discovery, the detector will focus more on the foreground part instead of the whole image.

Finally, we can generate a new object mask through Eq. 1 from $\widetilde{F_{int}}$ and iterate the whole procedure until convergence, *i.e.*, the distance between the centers in two consecutive iterations is smaller than τ. The overall process is presented in Algorithm 1. An example of the iteration process is illustrated in Fig. 3. It is worth noting that the efficiency bottleneck of ViT-based UOD methods lies in the process of feature extraction rather than localization. Thus, our iteration module only brings marginal extra computational overhead. More details can be found in 4.3.

Algorithm 1 Foreground Guidance with Self-iteration

Input: Unsupervised Object Detector $\mathcal{D}$, Intermediate Feature Map F_{int}, Standard Deviation σ of Gaussian Distribution g.
Initialize: $m = \mathcal{D}(F_{int})$, $\boldsymbol{x}_{\mathcal{O}} = 0$, $\widehat{\boldsymbol{x}_{\mathcal{O}}} = \infty$.
while $\|\widehat{\boldsymbol{x}_{\mathcal{O}}} - \boldsymbol{x}_{\mathcal{O}}\|^2 \geqslant \tau$ **do**
  $\widehat{\boldsymbol{x}_{\mathcal{O}}} = \boldsymbol{x}_{\mathcal{O}}$
  $\boldsymbol{x}_{\mathcal{O}} = Eq.\ 2(m)$
  $P = g(\boldsymbol{x}_{\mathcal{O}}, \sigma^2)$
  $\widetilde{F_{int}} = F_{int} \circ P$
  $m = \mathcal{D}(\widetilde{F_{int}})$
end while
Output: $m_{ref} = m$.

3.3. Multi-layer Feature Fusion

Different layers of Transformers encode features at different scales. Deeper layers tend to gather global and semantic information and focus on the discriminative parts of objects. Consequently, the activated areas in their feature maps are smaller (Fig. 7(b), 7(c)). In contrast, shallower layers focus on local information and thus the activated areas in their feature maps are broader (Fig. 7(d), 7(e)). However, the scale of the objects in non-object-centric images can vary greatly. Only using the feature from one layer is insufficient to deal with the scale variation problem. To address this issue, we propose the multi-layer feature fusion module to aggregate information from various scales.

In detail, from each multi-head attention layer l of a ViT, we can extract the key feature $k_l \in \mathbb{R}^{(N+1)\times D}$. Then, we drop the `[CLS]` token to be consistent with LOST and TokenCut. The aggregated feature for unsupervised object discovery is obtained by a weighted summation of the key features from all layers:

$$f = \sum_{l=1}^{L} \alpha_l k_l, \tag{5}$$

where α_l is the weight of layer l. The features from different layers of ViT contain information of objects at various scales. The aggregated feature incorporates more comprehensive object information at different scales to better localize objects.

4. Experiments

In this section, we conduct extensive experiments on various datasets to demonstrate the effectiveness of our method. FORMULA achieves new state-of-the-art results in unsupervised object discovery tasks. In addition, we conduct ablation studies to discuss the effect of foreground guidance and multi-layer feature fusion modules.

4.1. Experimental Settings

Datasets. Following LOST and TokenCut, we evaluate the proposed approach on three widely-adopted benchmarks for object discovery: VOC07 [15], VOC12 [16], and COCO [31]. Specifically, for VOC07 and VOC12, we use `trainval` set to evaluate our method. For COCO, we only use a subset of the COCO2014 `trainval` dataset, named COCO_20k [48].
Evaluation metric. Same to [48, 54, 49, 40, 52], we use the Correct Localization (CorLoc) metric for evaluation. In this

Method	Backbone	VOC07(↑)	VOC12(↑)	COCO_20k(↑)
Selective Search [45]	-	18.8	20.9	16.0
EdgeBoxes [60]	-	31.1	31.6	28.8
Kim *et al*. [25]	-	43.9	46.4	35.1
Zhange *et al*. [58]	VGG16 [41]	46.2	50.5	34.8
DDT+ [54]	VGG19 [41]	50.2	53.1	38.2
rOSD [48]	VGG16&19 [41]	54.5	55.3	48.5
LOD [49]	VGG16 [41]	53.6	55.1	48.5
DINO-seg [6, 40]	ViT-S/16 [6, 14]	45.8	46.2	42.1
LOST [40]	ViT-S/16 [6, 14]	61.9	64.0	50.7
TokenCut [52]	ViT-B/16 [6, 14]	68.8	72.4	59.0
FORMULA-L	ViT-S/16 [6, 14]	64.4	67.7	54.0
FORMULA-TC	ViT-B/16 [6, 14]	**69.4**	**73.2**	**59.7**
LOD + CAD [40]	VGG16 [41]	56.3	61.6	52.7
rOSD + CAD [40]	VGG16&19 [41]	58.3	62.3	53.0
LOST + CAD [40]	ViT-S/16 [6, 14]	65.7	70.4	57.5
TokenCut + CAD* [40]	ViT-B/16 [6, 14]	71.4	75.5	62.7
FORMULA-L + CAD [40]	ViT-S/16 [6, 14]	66.8	71.5	59.0
FORMULA-TC + CAD [40]	ViT-B/16 [6, 14]	**72.2**	**76.2**	**63.7**

Table 1. **Unsupervised single-object discovery.** We compare FORMULA with current state-of-the-art unsupervised object discovery methods. We use DINO pre-trained ViT [6] as our backbone. * is our implementation. '-L/TC' means to adopt LOST/TokenCut as the unsupervised object detector $\mathcal{D}$. '+CAD' means to train a second-stage object detector using pseudo-labels generated through each method.

Method	Supervis.	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean(↑)
WSDDN [3]	weak	39.4	50.1	31.5	16.3	12.6	64.5	42.8	42.6	10.1	35.7	24.9	38.2	34.4	55.6	9.4	14.7	30.2	40.7	54.7	46.9	34.8
PCL [43]	weak	54.4	9.0	39.3	19.2	15.7	62.9	64.4	30.0	25.1	52.5	44.4	19.6	39.3	67.7	17.8	22.9	46.6	57.5	58.6	63.0	43.5
rOSD [48] + OD [40]	-	38.8	44.7	25.2	15.8	0.0	52.9	45.4	38.9	0.0	16.6	24.4	43.3	57.2	51.6	8.2	0.7	0.0	9.1	65.8	9.4	27.4
LOST [40] + OD [40]	-	57.4	0.0	40.0	19.3	0.0	53.4	41.2	72.2	0.2	24.0	28.1	55.0	57.2	25.0	8.3	1.1	0.9	21.0	61.4	5.6	28.6
FORMULA-L + OD [40]	-	62.8	3.2	45.6	23.9	0.0	53.3	41.2	74.3	0.1	18.9	32.7	60.7	59.8	27.2	11.4	0.0	0.0	38.5	39.8	2.3	29.8
FORMULA-TC + OD [40]	-	61.4	19.2	49.4	26.1	0.0	60.5	46.6	78.7	0.4	21.2	31.8	73.7	55.2	15.8	12.2	0.0	0.0	41.5	51.5	8.9	32.7

Table 2. **Unsupervised object detection (OD).** We evaluate FORMULA on VOC07 `test` using AP@0.5. All the '+OD' methods are trained on VOC07 `trainval`.

metric, a predicted bounding box is considered correct if its intersection over union (IoU) score with one of the ground truth boxes in an image is greater than 0.5.

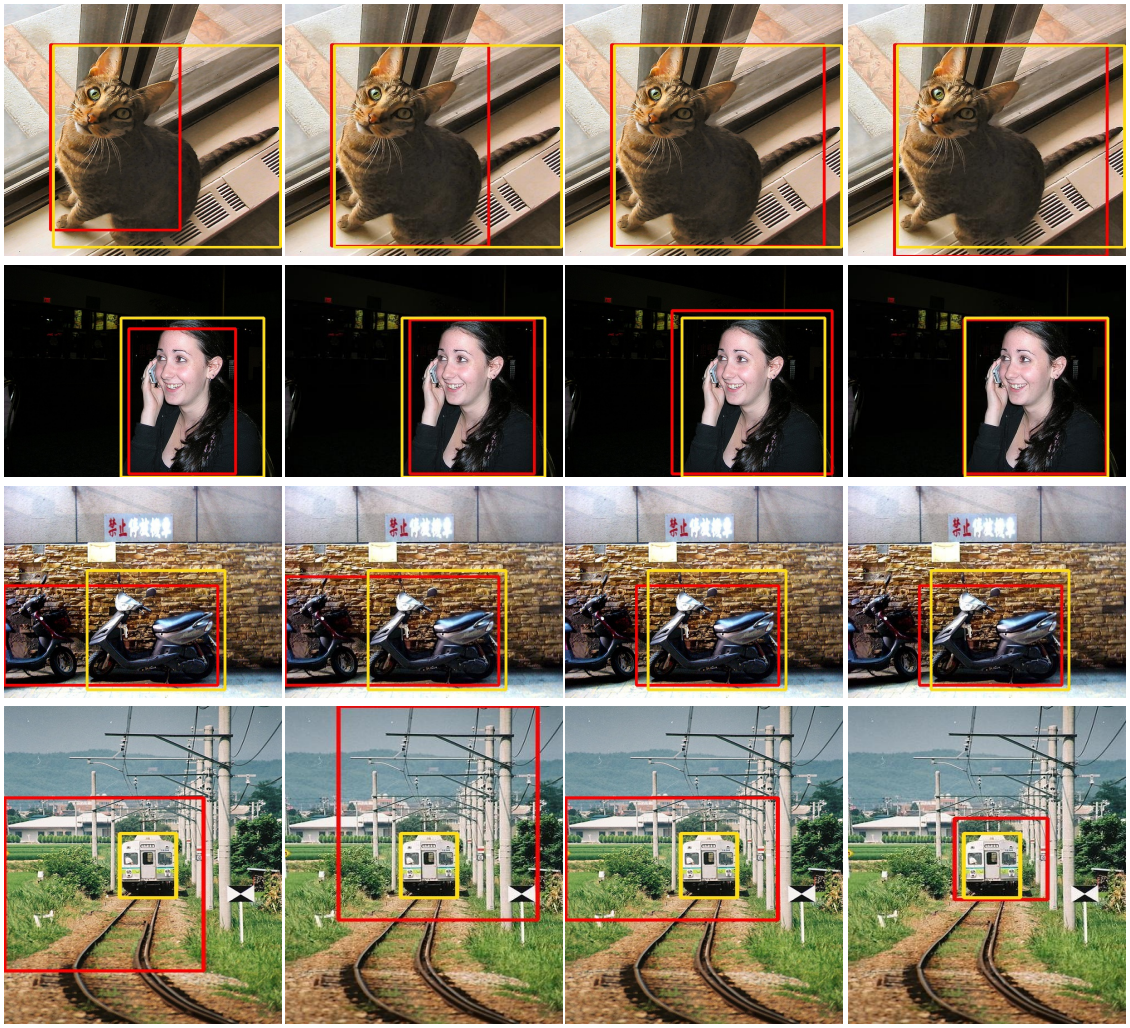
Implementation details. We use the weights from the publicly available DINO pre-trained ViT model [6]. We adopt LOST/TokenCut as our basic unsupervised object detector $\mathcal{D}$, named FORMULA-L/FORMULA-TC. For backbone architecture, we adopt ViT-S/16 and ViT-B/16 [14]. We set σ=0.1 for FORMULA-L and σ=1 for FORMULA-TC. τ is set to $\sqrt{2}$ in our experiments. For multi-layer feature fusion, we only fuse the features from the last four layers.

4.2. Main Results

4.2.1 Unsupervised Single-object Discovery

In Table 1, we present unsupervised single-object discovery performance of FORMULA on three popular datasets. As shown in the table, our method outperforms the previous state-of-the-art, LOST and TokenCut, on all three datasets. Specifically, on VOC07, our method surpasses LOST and TokenCut by 2.5% and 0.6%, respectively. On VOC12, FORMULA obtains 67.7% with ViT-S and 73.2% with ViT-B, improving the performance of LOST and TokenCut by 3.7% and 0.8%, respectively. On the COCO_20k dataset, FORMULA achieves 54.0% and 59.7%, which significantly surpasses LOST and TokenCut by 3.3% and 0.7%, respectively. In Fig. 4, we provide some visual results obtained by our method and the two baselines, *i.e.*, LOST and TokenCut. It can be seen that our method boosts detection performance on various scales.

Additionally, following [40], we also train a second-stage class-agnostic object detector (CAD) for unsupervised single-object discovery. Concretely, we train a Faster R-CNN [36] using the bounding boxes generated by FORMULA as pseudo-labels. It can be seen from Table 1 that our method outperforms the state-of-the-art by an average



(a) LOST (b) TokenCut (c) FORMULA-L (ours) (d) FORMULA-TC (ours)

Figure 4. **Example results of unsupervised single-object discovery on VOC07 and VOC12.** In (a) and (b), we show the results obtained by LOST [40] and TokenCut [52]. Our results are illustrated in (c) and (d). Red and yellow bounding boxes indicate the predicted bounding boxes and the ground-truth respectively. From top to bottom, the scale of the objects becomes smaller.

of 0.8% on the three datasets. The consistent gain over baseline methods shows the effectiveness of FORMULA for unsupervised object discovery.

4.2.2 Unsupervised Object Detection

Following LOST [40], we also evaluate our method on unsupervised object detection. Similar to training a CAD, a Faster R-CNN detector is trained using pseudo-labels generated by our method. Concretely, to generate pseudo-labels for the class-aware detectors, we cluster the boxes produced by FORMULA via K-means algorithm and then map them to ground-truth classes via Hungarian matching [27] for evaluation. It is worth noting that no human supervision is involved during the training process. We use the *Average Precision* metric with the threshold of IoU being 0.5, a common setting in PASCAL VOC detection. The results are shown in Table 2. As we can see, we improve the state-of-the-art results of unsupervised object detection on VOC07 `test` by a significant 4.1%.

4.3. Analysis and Discussion

In this section, we investigate the effectiveness of foreground guidance and multi-layer feature fusion modules. We conduct the following experiments with FORMULA-L and use ViT-S/16 as the backbone.

4.3.1 Ablation Study

We perform a set of ablation experiments to show the effectiveness of each component of our FORMULA. Results of the ablations are shown in Table 3. We can see that the fore-

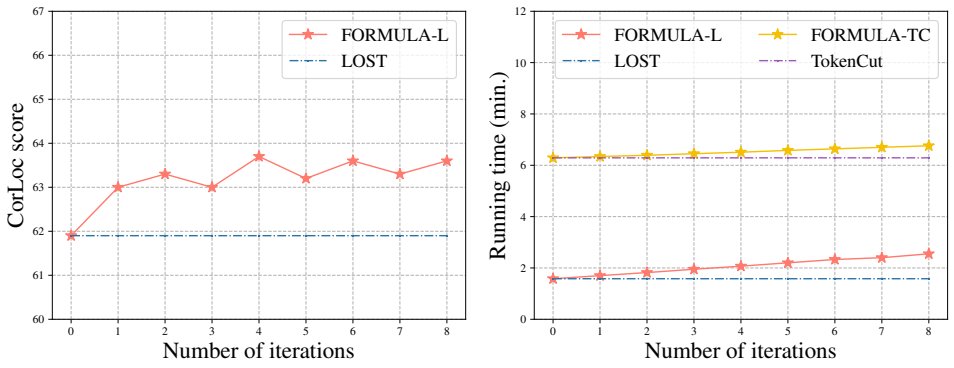


Figure 5. **Study of the iteration process.** (a) result variations with the number of iterations; (b) the running time of our method with different iteration numbers. Both experiments are conducted on VOC07 `trainval`.

Foreground guidance	*Multi-layer*	*CorLoc*(↑)
		61.9
✓		63.7
	✓	63.3
✓	✓	**64.4**

Table 3. **Ablation experiments.** Results on VOC07 `trainval`. "*Foreground guidance*" and "*Multi-layer*" represent foreground guidance module and multi-layer feature fusion module, respectively.

ground guidance module can bring an improvement of 1.8% and the multi-layer feature fusion module improves model performance by 1.4%. When both modules are applied, we have a total 2.5% improvement. The results demonstrate the effectiveness of the two modules.

4.3.2 Main Properties

We discussed some of the properties of FORMULA here.

Foreground guidance. To better understand the effect of foreground guidance, we manually change the $\boldsymbol{x}_{\mathcal{O}}$ of probability map P in Eq. 3 to different locations in the image. The effects are shown in Fig. 6. When $\boldsymbol{x}_{\mathcal{O}}$ falls into the background, the model would miss the most area of the object. By contrast, if it is located inside the foreground object and closer to the center of the object, the model could better capture the semantic layout of the object and improve its performance. The result suggests that our guidance mechanism can help the model focus more on the region of interest.

Backbone architecture. In Table 4, we show the results of FORMULAwith different Transformer backbones. We compare the results between ViT [14] small (ViT-S) and base (ViT-B) with a patch size of 8 or 16. It can be seen that FORMULA consistently outperforms LOST across different backbone architectures, demonstrating the effectiveness and architecture generalizability of FORMULA.

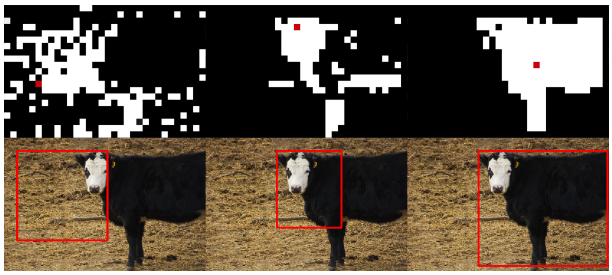


Figure 6. **Study of foreground guidance.** We manually place the center $\boldsymbol{x}_{\mathcal{O}}$ (red points in the first row) in Eq. 3 at different locations. The center $\boldsymbol{x}_{\mathcal{O}}$ is placed outside the foreground object (left), at a corner part of the object (middle), and near the center of the object (right). The first and second rows are the resulting object masks and bounding boxes.

Iteration number. Fig. 5(a) shows the influence of the number of iterations from 1 to 8. We find surprisingly that just one iteration can improve performance significantly. An optimal value for our method is 4, which achieves the best performance of 63.7%. More iterations may slightly reduce performance, which we assign to the random perturbations after convergence.

Running Time. We present the running time of our method for different numbers of iterations in Fig. 5(b). Results of LOST and TokenCut are shown as well. We measure the inference time on all images of VOC07 `trainval` with a single GTX TITAN X GPU. It can be seen that our method only brings marginal extra computational overhead even with 8 iterations.

Analysis of σ. We report the results of using different value of σ in Table 5. It can be observed that the performance slightly decreases as σ increases above 0.1, which we attribute to the fact that, with a higher value of σ, the broader 2D Gaussian distribution would make the model focus on

Method	Backbone	VOC07(↑)	VOC12(↑)	COCO_20k(↑)
LOST [40]	ViT-S/8 [6, 14]	55.5	57.0	49.5
FORMULA-L	ViT-S/8 [6, 14]	**57.9**	**61.3**	**49.6**
LOST [40]	ViT-S/16 [6, 14]	61.9	64.0	50.7
FORMULA-L	ViT-S/16 [6, 14]	**64.4**	**67.7**	**54.0**
LOST [40]	ViT-B/16 [6, 14]	60.1	63.3	50.0
FORMULA-L	ViT-B/16 [6, 14]	**62.8**	**66.5**	**53.4**

Table 4. **Analysis of the backbone.** CorLoc score comparison across three different backbones.

σ	VOC07(↑)	VOC12(↑)	COCO_20k(↑)
0.01	**63.7**	**66.6**	**52.8**
0.05	**63.7**	**66.6**	**52.8**
0.1	**63.7**	**66.6**	**52.8**
0.3	63.1	66.3	52.6
0.5	62.4	66.0	52.3

Table 5. **Ablation of σ.** CorLoc scores for different σ on three datasets. The results are obtained using only the foreground guidance module.

Datasets	FORMULA-L	FORMULA-TC
VOC07	[2,1,1,6]	[3,5,1,1]
VOC12	[1,1,2,6]	[1,6,1,2]
COCO_20k	[0,2,3,5]	[2,7,0,1]

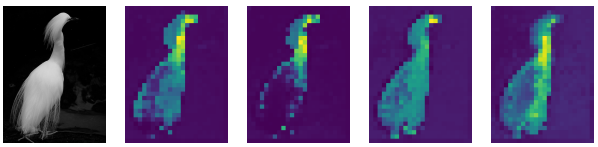
Table 6. **Layer weights.** The layer weights are the relative weight of each layer, with the sum being one. The weight of the last layer is in front and the others are in the following order.

$layer^{-1}$	$layer^{-2}$	$layer^{-3}$	$layer^{-4}$	*CorLoc*(↑)
✓				61.9
	✓			61.5
		✓		62.9
			✓	**63.3**

Table 7. **Ablation of the contribution of each layer.** We use features of each of the last four layers as input. '$layer^{-1}$' represents the last layer. The results are acquired on VOC07 `trainval`.

larger areas and thus hurt detection on small objects. Additionally, our method performs consistently with a σ between 0.01 and 0.1. Thus, we simply adopt a typical value of 0.1 in our experiments.

Multi-layer Fusion weights. The weights of multi-layer fusion weights for our results in Table 1 are presented in Table 6. FORMULA with LOST is better at capturing local information of the foreground, thus requiring more global knowledge from low layers to perform better. Different



(a) Input (b) $layer^{-1}$ (c) $layer^{-2}$ (d) $layer^{-3}$ (e) $layer^{-4}$

Figure 7. **The activated areas of attention maps from the last four layers of ViT.** (a) Image sampled from VOC07 [15]; (b)-(e) are the visualized attentions F_{int}. The deeper layers gather global information and focus on the discriminative parts of the object.

from FORMULA-L, the information from the last two layers plays a more important role for FORMULA-TC.

Besides, to better understand how each layer contributes to the overall performance, we conduct experiments using each of the four layers. The results are presented in Table 7. We can see that low layers, such as the third and fourth ones, play a more important role in gathering features at different scales, which aligns with the results in Table 6. We also visualize the attention maps of the last four layers in Fig. 7. The activated areas of the attention map vary at different scales for different layers. These results together indicate the ViT features from different layers can help detect objects at various scales.

5. Conclusion

In this work, we propose FORMULA, a simple and effective feature enhancement method for unsupervised object discovery. We utilize the foreground guidance acquired by an existing UOD detector to highlight the foreground regions on the feature maps and iteratively refine the segmentation predictions. In addition, by fusing the multi-layer features from a self-supervised ViT, we further aggregate multi-scale information for the feature representation. Our approach can work with any existing ViT-based unsupervised object discovery methods. Moreover, FORMULA achieves new state-of-the-art results on three datasets for unsupervised object discovery task. We hope our work could inspire more future research on enhancing ViT features for unsupervised visual learning.

References

[1] Sean Bell, C. Lawrence Zitnick, Kavita Bala, and Ross B. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] David Berthelot, Nicholas Carlini, Ian J. Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Neural Information Processing Systems (NeurIPS)*, 2019.

[3] Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] Zhaowei Cai, Quanfu Fan, Rogério Schmidt Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European Conference on Computer Vision (ECCV)*, 2016.

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[7] Kai Chen, Lanqing Hong, Hang Xu, Zhenguo Li, and Dit-Yan Yeung. Multisiam: Self-supervised multi-instance siamese representation learning for autonomous driving. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[9] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[11] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[12] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Show, match and segment: Joint learning of semantic matching and object co-segmentation. *IEEE Transactions on Pattern Recognition and Machine Intelligence (PAMI)*, 2020.

[13] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[15] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007.

[16] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012.

[17] Xiaoxu Feng, Xiwen Yao, Gong Cheng, and Junwei Han. Weakly supervised rotation-invariant aerial object detection network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[18] Golnaz Ghiasi and Charless C. Fowlkes. Laplacian pyramid reconstruction and refinement for semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 2016.

[19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018.

[20] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Neural Information Processing Systems (NeurIPS)*, 2020.

[21] Bharath Hariharan, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv preprint arXiv:2111.06377*, 2021.

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Co-attention cnns for unsupervised object co-segmentation. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.

[25] Gunhee Kim and Antonio Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *Neural Information Processing Systems (NeurIPS)*, 2009.

[26] Tao Kong, Anbang Yao, Yurong Chen, and Fuchun Sun. Hypernet: Towards accurate region proposal generation and joint object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955.

[28] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient

self-supervised vision transformers for representation learning. *International Conference on Learning Representations (ICLR)*, 2022.

[29] Zhaowen Li, Zhiyang Chen, Fan Yang, Wei Li, Yousong Zhu, Chaoyang Zhao, Rui Deng, Liwei Wu, Rui Zhao, Ming Tang, and Jinqiao Wang. MST: masked self-supervised transformer for visual representation. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[30] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, 2014.

[32] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott E. Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: single shot multibox detector. In *European Conference on Computer Vision (ECCV)*, 2016.

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[34] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, 2016.

[35] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[36] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *Neural Information Processing Systems (NeurIPS)*, 2015.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal on Computer Vision (IJCV)*, 2015.

[39] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Recognition and Machine Intelligence (PAMI)*, 2000.

[40] 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. In *British Machine Vision Conference (BMVC)*, 2021.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[42] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[43] Peng Tang, Xinggang Wang, Song Bai, Wei Shen, Xiang Bai, Wenyu Liu, and Alan L. Yuille. PCL: proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Recognition and Machine Intelligence (PAMI)*, 2020.

[44] Marvin Teichmann, Michael Weber, J. Marius Zöllner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *IEEE Intelligent Vehicles Symposium (IV)*, 2018.

[45] Jasper R. R. Uijlings, Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders. Selective search for object recognition. *International Journal on Computer Vision (IJCV)*, 2013.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems (NeurIPS)*, 2017.

[47] Huy V. Vo, Francis R. Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[48] Huy V. Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *European Conference on Computer Vision (ECCV)*, 2020.

[49] Huy V. Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[50] Feng Wang, Tao Kong, Rufeng Zhang, Huaping Liu, and Hang Li. Self-supervised learning by estimating twin class distributions. *arXiv preprint arXiv:2110.07402*, 2021.

[51] Jun Wang, Xiaohan Yu, and Yongsheng Gao. Feature fusion vision transformer for fine-grained visual categorization. In *British Machine Vision Conference (BMVC)*, 2021.

[52] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[53] Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. In *Neural Information Processing Systems (NeurIPS)*, 2021.

[54] 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 (PR)*, 2019.

[55] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[56] Jian Yang, Yuwei Yin, Shuming Ma, Haoyang Huang, Dongdong Zhang, Furu Wei, and Zhoujun Li. Gtrans: Grouping

and fusing transformer layers for neural machine translation. *arXiv preprint arXiv:2207.14467*, 2022.

[57] Sen Yang, Zhibin Quan, Mu Nie, and Wankou Yang. Transpose: Keypoint localization via transformer. In *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[58] Runsheng Zhang, Yaping Huang, Mengyang Pu, Qingji Guan, Jian Zhang, and Qi Zou. Mining objects: Fully unsupervised object discovery and localization from a single image. *arXiv preprint arXiv:1902.09968*, 2019.

[59] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, 2016.

[60] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision (ECCV)*, 2014.