# ECAD: EXPANDING CLASS-AGNOSTIC DETECTION BEYOND THING-CENTRIC OBJECTNESS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Object detection is a fundamental task in visual perception, providing structured region representations for recognition, grounding, reasoning, and interaction. However, existing detection paradigms largely inherit a thing-centric notion of objectness, where detectors are mainly trained to localize discrete and countable object instances. Consequently, many semantically meaningful visual elements, such as sky, road, grassland, water, and sports courts, are often absorbed into the background despite their importance for scene understanding and spatial reasoning. In this paper, we formulate **Expanded Class-Agnostic Detection (ECAD)**, a new setting that aims to discover category-agnostic visual candidates beyond conventional thing-centric objects. To support this setting, we construct **BTCO-Bench**, a Beyond Thing-Centric Objectness benchmark with category-agnostic box annotations covering both real-world and cross-domain scenarios. We further propose **ECADet**, a lightweight DETR-based detector built upon a frozen DINOv3 encoder, and introduce **Geometry-Aware Expert Regression (GAER)** and **Prototype-Guided Query Modulation (PGQM)** to improve localization and objectness estimation for diverse visual elements, respectively. Extensive experiments show that ECADet consistently outperforms representative class-agnostic and proposal-based detectors on BTCO-Bench, demonstrating the effectiveness of expanded objectness discovery. *Code and benchmark will be released.*

## 1 INTRODUCTION

Object detection is fundamental to visual perception, providing structured region representations for downstream recognition, reasoning, and interaction. Yet, what constitutes an object is often implicitly assumed. Mainstream datasets, such as COCO (Lin et al., 2014), LVIS (Gupta et al., 2019), and PASCAL VOC (Everingham et al., 2010), mainly annotate discrete, countable thing instances, while other semantically meaningful visual elements are treated as background. Although this thing-centric notion of objectness has driven substantial progress, it limits detector coverage in complex scenes.

As shown in Fig. 1, real-world scenes contain many visual elements that fall outside the conventional definition of object instances yet carry clear semantic, functional, or interactive value. Roads guide navigation, walls and floors define spatial layouts, while sky, water, grassland, and courts provide important scene context. Although often non-countable and less clearly bounded than common objects, these elements are not mere background; they serve as semantic anchors for scene understanding, spatial reasoning, embodied

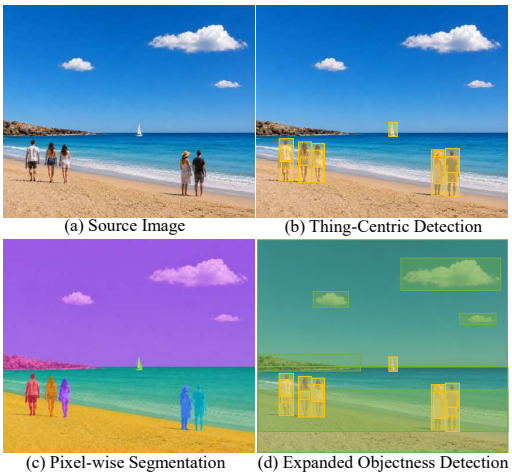


Figure 1: Expanded Class-Agnostic Detection. Given a source image in (a), existing detectors in (b) are constrained by thing-centric objectness, while segmentation in (c) provides dense pixel-wise labels rather than sparse detection candidates. In contrast, ECAD in (d) expands the object discovery space by discovering both thing objects and semantically meaningful visual elements as category-agnostic boxes.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

navigation, and multimodal grounding. Expressions such as people on the 'grassland', buildings under the 'sky', players at the center of the 'court', and boats on the 'sea' all rely on such anchors. If the visual front end detects only conventional thing objects, these references cannot be effectively grounded or used for reasoning. Thus, the limitation of existing detectors lies not only in which categories they can recognize, but also in which visual content they can discover as objects.

Recent advances in class-agnostic, open-vocabulary, and open-world detection have relaxed category constraints from different perspectives. Class-agnostic detection (CAD) (Jaiswal et al., 2021; Maaz et al., 2022; Syuen Lim et al., 2024; Tang et al., 2026) learns general objectness without category-specific classification, open-vocabulary detection (OVD) (Gu et al., 2021; Fu et al., 2025; Zhang et al., 2026) recognizes categories unseen during training, and open-world detection (OWD) (Joseph et al., 2021; Ma et al., 2025) addresses unknown discovery and incremental learning. However, these paradigms still inherit thing-centric supervision from existing datasets. They expand the category space or improve category generalization, but rarely reconsider what visual content should be discovered as objects. Consequently, many semantically or functionally meaningful visual elements beyond conventional object instances remain outside the object discovery space.

Motivated by these observations, we revisit the scope of objectness in CAD. Rather than asking which category a visual element belongs to, we ask **which visual content should be discovered as category-agnostic object candidates**. Conventional class-agnostic detectors mainly target thing instances, whereas our goal is to cover both common objects, such as humans, vehicles, and animals, and semantically meaningful elements often treated as background, including sky, sea, grassland, roads, and courts. *Although semantic and panoptic segmentation also model such stuff-like regions, they perform dense pixel-wise labeling rather than sparse candidate discovery. Their flat output spaces are also less suited to representing overlapping or hierarchical candidates, such as a person and its parts.* Importantly, our goal is neither to enlarge the detection vocabulary nor to convert segmentation labels into detection categories. The model still predicts category-agnostic candidate boxes without assigning class names. We therefore introduce **Expanded Class-Agnostic Detection (ECAD)**, which broadens objectness from thing-centric instances to a wider range of semantically meaningful visual elements while preserving the category-agnostic box prediction formulation.

A dedicated benchmark is needed to evaluate this setting. Existing detection benchmarks largely inherit thing-centric annotations, leaving many semantically meaningful non-thing elements unlabeled and treated as background. They therefore cannot reliably assess a detector's ability to discover visual content beyond conventional object instances. To address this gap, we construct **BTCO-Bench**, a category-agnostic benchmark for Beyond Thing-Centric Objectness, covering both real-world and cross-domain scenarios. Further details are provided in the **BTCO-Bench** section.

We further propose **ECADet**, an Expanded Class-Agnostic Detector for discovering semantically meaningful visual elements beyond thing-centric objectness. Built on RF-DETR, ECADet freezes the large-scale pretrained DINOv3 encoder to preserve its transferable visual representations and reduce overfitting to limited training data, while training only a lightweight projector, Transformer decoder, and two objectness-oriented modules. **Geometry-Aware Expert Regression (GAER)** improves localization for elements with diverse shapes and boundary ambiguity, while **Prototype-Guided Query Modulation (PGQM)** injects structural foreground priors into object queries. This lightweight design broadens the object discovery space without retraining the visual foundation model, providing richer category-agnostic proposals for downstream recognition, grounding, segmentation, and multimodal reasoning.

Our main contributions are summarized as follows:

- We revisit the implicit assumption of thing-centric objectness in existing detection paradigms and introduce **ECAD**, which extends category-agnostic object discovery from conventional thing instances to broader semantically meaningful visual elements.
- We construct **BTCO-Bench**, a Beyond Thing-Centric Objectness benchmark with category-agnostic bounding-box annotations across real-world and cross-domain scenarios. We further propose **ECADet**, a lightweight DETR-based baseline equipped with **GAER** and **PGQM**.
- Extensive experiments show that ECADet consistently outperforms representative class-agnostic and proposal-based detectors on BTCO-Bench, demonstrating its effectiveness in discovering visual content beyond thing-centric objectness.

## 2 Related Work

**Class-agnostic detection (CAD)** localizes object candidates without assigning category labels, improving generalization to unseen categories and domains. Early work decoupled localization from category prediction (Jaiswal et al., 2021), while later methods leveraged vision-language supervision and large-scale grounding data to improve proposal quality. MAVL (Maaz et al., 2022) uses box-level supervision from vision-language datasets, DiPEx (Syuen Lim et al., 2024) expands textual queries to better cover the object concept space, and PF-RPN (Tang et al., 2026) progressively refines a learnable visual embedding for prompt-free proposal generation across unseen domains. However, these methods still learn objectness from thing-centric annotations of discrete, countable, and nameable entities, leaving their discovery space constrained by existing detection datasets. In contrast, we extend category-agnostic discovery from conventional thing instances to semantically meaningful non-thing elements.

**Open-vocabulary detection (OVD) and open-world detection (OWD)** relax the category constraints of closed-set detection from complementary perspectives. OVD methods, such as ViLD (Gu et al., 2021), LLMDet (Fu et al., 2025), and NoOVD (Zhang et al., 2026), associate visual regions with open-ended textual concepts to recognize categories unseen during training. OWD methods, such as OWD (Joseph et al., 2021) and SKDF (Ma et al., 2025), further address unknown object discovery and incremental category learning during deployment. However, both paradigms largely inherit the object definition of standard detection datasets, where detection targets remain thing-centric instances. Consequently, semantically or functionally meaningful elements outside conventional instance annotations are still treated as background. In contrast, our goal is not to recognize more category names, but to broaden the visual content included in category-agnostic object discovery.

**Semantic, panoptic, and open-vocabulary segmentation** provide dense scene parsing through pixel-level prediction. Representative works include FCN (Long et al., 2015) for semantic segmentation, the panoptic segmentation formulation (Kirillov et al., 2019), and LSeg (Li et al., 2022) and OpenSeg (Ghiasi et al., 2022) for open-vocabulary segmentation. Benchmarks such as ADE20K (Zhou et al., 2017) and COCO-Stuff (Caesar et al., 2018) naturally cover stuff categories including sky, road, ground, and wall. However, these tasks differ fundamentally from our setting. Segmentation produces dense pixel-level outputs, typically assigning each pixel a predefined or text-specified semantic label, whereas our task discovers a sparse set of category-agnostic candidate boxes with objectness scores. Thus, our work is complementary to segmentation: rather than determining the semantic label of each pixel, we identify which semantically meaningful visual elements should be explicitly proposed as object candidates.

## 3 BTCO-Bench: Beyond Thing-Centric Objectness Benchmark

### 3.1 Motivation

Existing detection benchmarks are not designed for ECAD. Datasets such as COCO and LVIS largely annotate countable thing instances, leaving many semantically meaningful, especially stuff-like, elements unlabeled and treated as background. Consequently, they cannot assess a detector's ability to discover visual content beyond conventional object instances. Moreover, this expanded notion of objectness should generalize from standard natural images to artistic and degraded domains. Yet existing cross-domain benchmarks likewise lack category-agnostic box annotations for such elements. To address this gap, we construct **BTCO-Bench**, a Beyond Thing-Centric Objectness benchmark for evaluating category-agnostic discovery of both conventional thing instances and semantically meaningful visual elements beyond the thing-centric annotation space across real-world and cross-domain scenarios.

### 3.2 Benchmark Construction

BTCO-Bench consists of two complementary splits: BTCO-Real and BTCO-XDomain. **BTCO-Real** is built upon LVIS, which contains diverse real-world scenes and everyday objects. We augment the LVIS minival split with bounding-box annotations for semantically meaningful elements beyond its original annotation space. These include stuff-like regions, such as sky, roads, grassland, water, mountains, and courts, as well as previously unannotated object-like elements, such as police

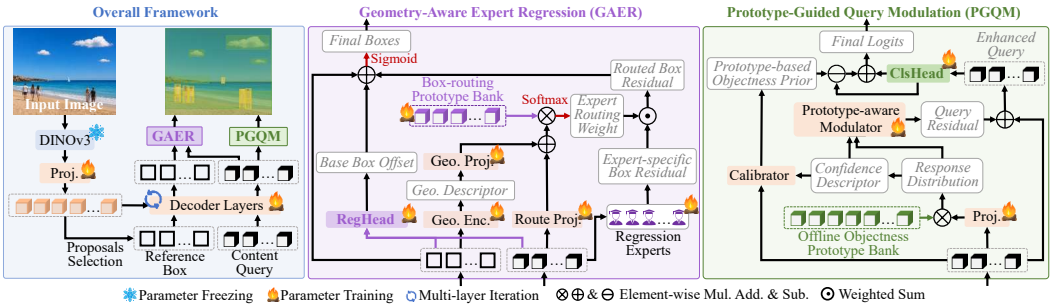


Figure 2: Overview of the proposed ECADet framework. A frozen DINOv3 extracts dense visual features, followed by a lightweight DETR decoder. Geometry-Aware Expert Regression (GAER) routes queries to specialized regression experts for diverse geometric patterns, while Prototype-Guided Query Modulation (PGQM) enhances query representations with offline objectness prototypes for more reliable objectness prediction.

lights, rocks, flowers, trees, windows, billboards, and signs. We further sample 1,000 LVIS training images and annotate them under the same protocol, forming a compact training split for expanded objectness learning. All regions are represented as category-agnostic boxes with a unified foreground label, *i.e.*, *object*. BTCO-Real thus evaluates the discovery of conventional thing instances together with previously unannotated object-like and semantically meaningful stuff-like elements. **BTCO-XDomain** evaluates the cross-domain generalization of expanded objectness under domain shifts. It comprises three subsets: artistic images from MSOSB (Zhang et al., 2024), curated by removing style-transferred and implausible AI-generated samples and retaining watercolor, cartoon, and oil-painting styles; monochrome and colored pencil drawings collected from the Internet; and adverse-condition images from SDGOD (Wu & Deng, 2022), excluding style-transferred samples and covering `Daytime foggy`, `Dusk rainy`, and `Night clear`. Compared with BTCO-Real, this split covers **seven** conditions involving artistic abstraction, sketch-like contours, adverse weather, and illumination changes, enabling evaluation of expanded objectness in unseen domains.

Five trained annotators participated in image collection, data cleaning, bounding-box annotation, and annotation standardization. All annotations follow the COCO format and assign each region the unified foreground label *object*. Category names are not used for training or evaluation, ensuring strict adherence to the class-agnostic detection setting. The statistics of BTCO-Bench are summarized in Table 1. The table reports the number of images and annotated boxes in each split, together with their scale distributions, highlighting the diverse object sizes covered by the benchmark.

Table 1: Statistics of BTCO-Bench.

| Split | Source | Domain | Images | Boxes | Box Size | | |
|---|---|---|---|---|---|---|---|
| | | | | | Small | Medium | Large |
| BTCO-Real (Train) | LVIS train | Real | 1,000 | 28,242 | 8,827 | 10,832 | 8,583 |
| BTCO-Real (Test) | LVIS minival | Real | 5,000 | 145,631 | 48,939 | 53,296 | 43,396 |
| BTCO-XDomain | MSOSB/Web/SDGOD | Cross-domain | 18,639 | 134,143 | 10,105 | 38,144 | 85,894 |

### 3.3 Design Principles and Characteristics

BTCO-Bench follows four key design principles. 1) It broadens objectness coverage from conventional thing instances to semantically meaningful non-thing elements. 2) All annotations are category-agnostic, focusing evaluation on objectness estimation and localization rather than category recognition. 3) The benchmark covers elements with diverse scales, aspect ratios, shapes, and boundary ambiguity; non-thing elements are often less compact and less clearly bounded than regular objects, posing greater localization challenges. 4) BTCO-XDomain introduces substantial appearance shifts through artistic images, pencil drawings, and adverse-condition scenes, enabling evaluation of objectness generalization beyond standard natural-image domains. ***More details are provided in the Appendix A.***

## 4 The Proposed Method

### 4.1 Problem Formulation

We consider the **Expanded Class-Agnostic Detection (ECAD)** problem. Given an input image $x$, the ground-truth set is $\mathcal{Y} = \{\mathbf{b}_i\}_{i=1}^{N}$, where all annotated regions share a unified foreground label

regardless of their semantic categories. The model predicts a set of category-agnostic candidates:

$$\hat{\mathcal{Y}} = \{(\hat{s}_j, \hat{\mathbf{b}}_j)\}_{j=1}^{Q}, \tag{1}$$

where $\hat{s}_j$ denotes the objectness score and $\hat{\mathbf{b}}_j \in [0,1]^4$ is the normalized bounding box of the $j$-th candidate. Unlike closed-set and open-vocabulary detection, ECAD does not assign category labels to predicted regions. Instead, it treats both conventional thing instances and semantically meaningful non-thing elements as foreground candidates. The goal is to learn an expanded objectness function that discovers both conventional thing instances and semantically meaningful visual elements beyond the thing-centric annotation space.

### 4.2 Overview of the Method

Our framework builds on RF-DETR, as shown in Fig. 2. A frozen DINOv3 encoder extracts visual features, followed by a lightweight projector and Transformer decoder that produce query features $\mathbf{Q}^l = \{\mathbf{q}_i^l\}_{i=1}^N$ and reference boxes $\mathbf{R}^l = \{\mathbf{r}_i^l\}_{i=1}^N$ at each decoder layer. Although all foreground regions share a unified label in ECAD, they remain highly diverse in appearance and geometry. We therefore introduce two complementary modules. Geometry-Aware Expert Regression (GAER) routes queries to specialized regression experts based on their features and reference-box geometry, while Prototype-Guided Query Modulation (PGQM) modulates query representations with offline DINOv3-derived objectness prototypes to improve objectness estimation. *Both modules retain the original RF-DETR heads as the primary prediction paths and introduce lightweight residual refinements for ECAD.*

### 4.3 Geometry-Aware Expert Regression (GAER)

In ECAD, all foreground regions share a unified label despite substantial geometric diversity in scale, aspect ratio, shape, and localization difficulty. A shared regression head must model these heterogeneous patterns with a single mapping, limiting region-specific box refinement. We therefore introduce Geometry-Aware Expert Regression (GAER), which routes each query to specialized regression experts based on its representation and reference-box geometry.

Specifically, given the query features $\mathbf{Q}^l = \{\mathbf{q}_i^l\}_{i=1}^N$ and the corresponding reference boxes $\mathbf{R}^l = \{\mathbf{r}_i^l\}_{i=1}^N$ at the $l$-th decoder layer, we retain the original regression pathway of RF-DETR. For the $i$-th query, the **original regression head (RegHead)** predicts a ***Base Box Offset***:

$$\Delta\mathbf{b}_{\text{base},i}^l = \text{RegHead}(\mathbf{q}_i^l), \tag{2}$$

where $\mathbf{q}_i^l \in \mathbb{R}^d$ denotes the $i$-th query feature, $d$ is the hidden dimension, and $\Delta\mathbf{b}_{\text{base},i}^l \in \mathbb{R}^4$ is the base box refinement predicted by the original regression head.

Meanwhile, we encode the reference-box geometry for routing. Let $\tilde{\mathbf{r}}_i^l = (c_x, c_y, w, h)$ denote the normalized reference box obtained from the decoder reference state. We construct a 6-dimensional ***Geometry Descriptor***:

$$\mathbf{g}_i^l = \psi(\tilde{\mathbf{r}}_i^l) = [c_x, c_y, \log w, \log h, \log(wh), \log(w/h)], \tag{3}$$

where $\psi(\cdot)$ denotes the box **Geometry Encoder**, capturing center, scale, area, and aspect ratio.

To determine which **Regression Experts** are suitable for the current query, we jointly use the query feature and its reference-box geometry. The query feature and ***Geometry Descriptor*** are projected into the same routing space and then fused by addition:

$$\mathbf{t}_i^l = f_{\text{route}}(\mathbf{q}_i^l) + f_{\text{geo}}(\mathbf{g}_i^l), \tag{4}$$

where $\mathbf{t}_i^l \in \mathbb{R}^d$ is the routing feature. Here, $f_{\text{route}}(\cdot)$ is **Route Projecter**, a two-layer MLP that maps the query feature to the routing space, and $f_{\text{geo}}(\cdot)$ is **Geometry Projecter**, a two-layer MLP that maps the 6-dimensional ***Geometry Descriptor*** to the same space.

We maintain a learnable **Box-routing Prototype Bank** $\mathbf{P}_{\text{box}} = \{\mathbf{p}_k\}_{k=1}^K$, where $K$ is the number of **Regression Experts** and each prototype $\mathbf{p}_k \in \mathbb{R}^d$ represents a latent localization pattern. The routing score between the $i$-th query and the $k$-th prototype is computed by cosine similarity:

$$a_{i,k}^l = \text{sim}(\mathbf{t}_i^l, \mathbf{p}_k), \tag{5}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. The ***Expert Routing Weight*** is obtained by a temperature-scaled softmax:

$$w_{i,k}^{l} = \frac{\exp(a_{i,k}^{l}/\tau_{\mathrm{box}})}{\sum_{j=1}^{K}\exp(a_{i,j}^{l}/\tau_{\mathrm{box}})}, \tag{6}$$

where $w_{i,k}^{l}$ is the weight assigned to the $k$-th **Regression Expert**, and $\tau_{\mathrm{box}}$ is the temperature.

The $k$-th **Regression Expert** $E_k(\cdot)$, comprising an adapter and a linear box regressor, predicts the ***Expert-specific Box Residual*** as $\Delta\mathbf{b}_{\mathrm{expert},i,k}^{l} = E_k(\mathbf{q}_i^{l})$.

The final ***Routed Box Residual*** is obtained by weighted aggregation over all experts:

$$\Delta\mathbf{b}_{\mathrm{route},i}^{l} = \eta\sum_{k=1}^{K} w_{i,k}^{l}\Delta\mathbf{b}_{\mathrm{expert},i,k}^{l}, \tag{7}$$

where $\Delta\mathbf{b}_{\mathrm{route},i}^{l}$ denotes the ***Routed Box Residual***, with scale controlled by $\eta$.

Finally, the ***Routed Box Residual*** complements rather than replaces the ***Base Box Offset***, yielding the final prediction $\mathbf{b}_i^{l} = \sigma(\mathbf{r}_i^{l} + \Delta\mathbf{b}_{\mathrm{base},i}^{l} + \Delta\mathbf{b}_{\mathrm{route},i}^{l})$, where $\sigma(\cdot)$ denotes the sigmoid function and $\mathbf{b}_i^{l}$ is the final predicted box.

Overall, GAER preserves the original regression head's stable localization while augmenting it with query-adaptive expert residuals, enabling shared localization knowledge and selective modeling of heterogeneous geometric patterns across class-agnostic foreground regions.

### 4.4 Prototype-Guided Query Modulation (PGQM)

In parallel with the regression branch's modeling of geometric diversity, the classification branch handles the appearance and contextual diversity of class-agnostic foreground regions. Since ECAD reduces classification to objectness estimation without category supervision, a single classifier may struggle to capture diverse foreground modes. We therefore propose Prototype-Guided Query Modulation (PGQM), which uses objectness prototypes as structural foreground priors to modulate query features before classification.

We first construct an **Offline Objectness Prototype Bank** from the training set. A frozen DINOv3 encoder extracts dense features, from which foreground regions are cropped using ground-truth boxes and pooled into object-level embeddings for clustering into objectness prototypes. Derived from self-supervised representations and foreground annotations, these prototypes provide diverse category-agnostic priors for foreground appearance.

Given the query features $\mathbf{Q}^{l} = \{\mathbf{q}_i^{l}\}_{i=1}^{N}$, we project each query into the prototype-matching space as $\mathbf{z}_i^{l} = f_{\mathrm{proj}}(\mathbf{q}_i^{l})$, where $f_{\mathrm{proj}}(\cdot)$ is a lightweight **Projector**. We then compute its cosine similarity to the offline objectness prototype bank $\mathbf{P}_{\mathrm{obj}} = \{\mathbf{p}_m\}_{m=1}^{M}$ as $s_{i,m}^{l} = \mathrm{sim}(\mathbf{z}_i^{l}, \mathbf{p}_m)$.

The prototype ***Response Distribution*** is obtained by temperature-scaled softmax:

$$\pi_{i,m}^{l} = \frac{\exp(s_{i,m}^{l}/\tau_{\mathrm{obj}})}{\sum_{n=1}^{M}\exp(s_{i,n}^{l}/\tau_{\mathrm{obj}})}, \tag{8}$$

where $\tau_{\mathrm{obj}}$ is a temperature parameter.

Instead of using prototypes as an independent classifier, we derive a ***Confidence Descriptor*** as $\mathbf{c}_i^{l} = \Phi(\mathbf{s}_i^{l}, \boldsymbol{\pi}_i^{l}, \mathbf{q}_i^{l})$, where $\mathbf{s}_i^{l}$ and $\boldsymbol{\pi}_i^{l}$ denote the raw ***Similarity Scores*** and normalized ***Response Distribution***, respectively. The descriptor summarizes normalized entropy, the top-1/top-2 probability margin, pooled prototype response, maximum prototype similarity, query feature norm, and a derived confidence score, indicating whether a query matches the foreground prototypes confidently and consistently or responds ambiguously across multiple prototypes.

The **Prototype-aware Modulator** uses the ***Response Distribution*** and ***Confidence Descriptor*** to generate a ***Query Residual***. Specifically, we construct an entropy-aware response distribution as $\bar{\boldsymbol{\pi}}_i^{l} = (1 - H_i^{l})\boldsymbol{\pi}_i^{l} + H_i^{l}\mathbf{u}$, where $H_i^{l}$ is the normalized entropy of $\boldsymbol{\pi}_i^{l}$ and $\mathbf{u}$ is the uniform prototype distribution. This operation suppresses unreliable prototype assignments under uncertain responses.

The modulator maps $\bar{\boldsymbol{\pi}}_i^{l}$ through a two-layer MLP and the normalized query direction through a linear layer. Their sum is normalized and filtered by a confidence-aware gate to produce $\Delta\mathbf{q}_i^{l} =$

$\mathcal{M}(\mathbf{q}_i^l, \bar{\boldsymbol{\pi}}_i^l, \mathbf{c}_i^l)$, where $\mathcal{M}(\cdot)$ denotes the **Prototype-aware Modulator**. The gate is predicted by another two-layer MLP using the entropy-aware response distribution and confidence cues in $\mathbf{c}_i^l$, including the response margin, confidence score, entropy, and query norm.

Using a learnable modulation scale $\lambda$, the ***Enhanced Query*** is formed as $\hat{\mathbf{q}}_i^l = \mathbf{q}_i^l + \lambda \Delta \mathbf{q}_i^l$ and then fed into the **original classification head (ClsHead)** to obtain the base objectness logit $o_{\text{base},i}^l = \text{ClsHead}(\hat{\mathbf{q}}_i^l)$.

For conservative objectness calibration, we estimate a ***Prototype-based Objectness Prior*** as $o_{\text{prior},i}^l = G_{\text{obj}}(\mathbf{q}_i^l, \mathbf{c}_i^l)$, where $G_{\text{obj}}(\cdot)$ is a low-capacity **Calibrator** that combines query norm, maximum prototype similarity, pooled prototype similarity, and prototype entropy using positive learnable weights and a learnable bias.

The final objectness logit is obtained through bounded residual calibration as $o_i^l = o_{\text{base},i}^l + \beta \gamma_i^l \tanh(o_{\text{prior},i}^l - o_{\text{base},i}^l)$, where $\gamma_i^l$ is a dynamic confidence gate predicted by a two-layer MLP using the objectness prior, query norm, confidence score, response margin, and pooled similarity. The calibration scale $\beta$ is learnable, while $\tanh(\cdot)$ bounds the correction magnitude.

Overall, PGQM retains the original classification head as the main prediction path, using offline DINOv3-driven objectness prototypes only for confidence-aware query modulation and conservative logit calibration to improve objectness estimation.

### 4.5 Training Objective and Implementation Details

**Training objective.** ECADet follows RF-DETR's set prediction objective. After Hungarian matching, it is optimized with classification, $\ell_1$ regression, and GIoU losses. GAER and PGQM are trained end-to-end under this objective without additional supervision.

**Implementation details.** We adopt RF-DETR with a frozen DINOv3 encoder as the baseline. GAER uses $K = 6$ regression experts and a learnable routing prototype bank, while PGQM employs a fixed objectness prototype bank with $M = 4$ prototypes. We set $\tau_{\text{box}} = \tau_{\text{obj}} = 0.07$ in Eqs. (6) and (8), and $\eta = 1.1$ in Eq. (7). All experiments use distributed data-parallel training on two NVIDIA RTX 3090 GPUs. We train for 48 epochs with a batch size of 4 per GPU and gradient accumulation over 4 steps. AdamW is used with an initial learning rate and weight decay of $1 \times 10^{-4}$, together with gradient clipping at a maximum norm of 0.1.

## 5 Experimental

### 5.1 Setup

**Settings.** Under ECAD, all methods ignore category labels and treat every annotated visual element as foreground. They predict category-agnostic boxes ranked by objectness confidence without class-specific post-processing. ECADet builds on RF-DETR with a frozen DINOv3 encoder and is trained on BTCO-Real (Train) and evaluated on BTCO-Real (Test) and BTCO-XDomain.

**Comparison methods.** We compare ECADet with RPN (Ren et al., 2016), CAOD (Jaiswal et al., 2021), MAVL (Maaz et al., 2022), DiPEx (Syuen Lim et al., 2024), PF-RPN (Tang et al., 2026), SAM3 (Carion et al., 2026), and two RF-DETR baselines using either a trainable DINOv2 encoder (original RF-DETR) or a frozen DINOv3 encoder. These methods cover region proposal, CAD, vision-language-assisted discovery, prompt-free proposal generation, and class-agnostic segmentation. RPN, CAOD, DiPEx, both RF-DETR baselines, and ECADet are trained on BTCO-Real (Train), whereas MAVL, PF-RPN, and SAM3 are evaluated using their publicly released models pretrained on larger-scale datasets. For SAM3, we use the generic text prompt "objects" to generate category-agnostic masks, take the tightest axis-aligned rectangle enclosing each mask as its predicted box, and retain the mask confidence as the objectness score. For methods with class-specific predictions, category labels are discarded and confidence scores are treated as objectness scores under the unified class-agnostic evaluation protocol.

**Evaluation protocol and metrics.** All methods are evaluated as single-foreground-class detectors. We report COCO-style AP averaged over IoU thresholds from 0.50 to 0.95 as the primary metric, together with $AP_{50}$, $AP_{75}$, $AP_s$, $AP_m$, and $AP_l$. We also report total parameters (trainable parameters in parentheses) and FPS, measured on a single NVIDIA RTX 3090 GPU with a batch size of 1, excluding data loading and post-processing. For proposal-generation methods, we additionally report $AR_K$ using the top-$K$ predictions ranked by objectness. This fixed budget prevents inflated recall from excessive proposals and ensures fair comparison. Unless otherwise specified, we set $K = 100$ and report $AR_{100}$.

### 5.2 Main Results

**Results on BTCO-Real (Test).** As shown in Table 2, ECADet outperforms all comparison methods. The segmentation-to-box baseline SAM3 achieves 9.6 AP, indicating that generic segmentation outputs do not directly provide competitive sparse and ranked candidates for ECAD. With DINOv3-L, ECADet achieves 41.0 AP, exceeding the strongest competitor, DiPEx, by 14.4 points. It also attains the best $AP_{50}$, $AP_{75}$, and $AR_{100}$, surpassing the best competing results by 16.2, 13.7, and 8.8 points, respectively. Consistent gains across visual-element scales demonstrate its ability to capture diverse expanded objectness. ECADet also offers favorable accuracy–efficiency trade-offs across backbones. The DINOv3-S variant reaches 30.1 AP at 44.6 FPS with only 32.2M parameters, outperforming all competitors. Compared with the RF-DETR (DINOv3-L) baseline, ECADet improves AP from 24.5 to 41.0 while adding only 1.0M parameters and incurring modest inference overhead, validating the effectiveness of GAER and PGQM.

Table 2: Quantitative results on the BTCO-Real (Test) (%).

| Methods | $AP_s$ | $AP_m$ | $AP_l$ | AP | $AP_{50}$ | $AP_{75}$ | $AR_{100}$ | Params. (M) | FPS↑ |
|---|---|---|---|---|---|---|---|---|---|
| SAM3 | 7.0 | 12.8 | 10.4 | 9.6 | 15.7 | 10.0 | 15.6 | 840.5 (840.5) | 5.6 |
| RPN | 10.5 | 14.5 | 7.0 | 10.8 | 26.2 | 7.5 | 44.2 | 27.5 (27.2) | 49.5 |
| CAOD | 17.1 | 25.6 | 28.7 | 23.4 | 49.2 | 19.7 | 38.1 | 41.3 (41.1) | 5.8 |
| MAVL | 4.4 | 18.4 | 41.6 | 20.8 | 31.7 | 22.2 | 39.3 | 59.0 (58.8) | 19.7 |
| DiPEx | 16.3 | 29.5 | 35.6 | 26.6 | 36.1 | 28.6 | 39.7 | 172.8 (*) | 8.0 |
| PF-RPN | 13.5 | 32.0 | 38.0 | 26.4 | 34.9 | 29.0 | 48.5 | 135.6 (78.3) | 9.4 |
| RF-DETR (DINOv2-S) | 13.5 | 23.5 | 26.8 | 20.4 | 39.6 | 18.8 | 53.1 | 135.6 (135.6) | 50.7 |
| RF-DETR (DINOv3-L) | 14.2 | 23.1 | 22.3 | 24.5 | 44.5 | 20.8 | 53.0 | 315.1 (11.9) | 18.7 |
| **ECADet** (DINOv3-S) | 13.4 | 33.1 | 47.5 | 30.1 | 54.4 | 29.1 | 52.3 | 32.2 (10.6) | 44.6 |
| **ECADet** (DINOv3-B) | 19.6 | 39.2 | 51.7 | 36.0 | 60.9 | 36.1 | 58.0 | 98.2 (12.6) | 42.3 |
| **ECADet** (DINOv3-L) | 25.5 | 44.7 | 55.0 | 41.0 | 65.4 | 42.7 | 61.9 | 316.1 (12.9) | 17.3 |

*Note:* The '*' denotes 256 trainable prompt parameters.

**Results on BTCO-XDomain.** As shown in Table 3, all ECADet variants achieve the best AP, demonstrating strong cross-domain robustness and favorable backbone scaling. The segmentation-to-box baseline SAM3 obtains 12.4 AP, suggesting that directly converting generic segmentation masks into boxes remains insufficient for reliable expanded object discovery under domain shifts. With DINOv3-S, ECADet reaches 45.8 AP, outperforming PF-RPN by 2.9 points; DINOv3-L further raises AP to 65.1, yielding a 22.2-point gain. It also achieves the best $AP_{50}$, $AP_{75}$, and $AR_{100}$, exceeding the strongest competing results by 21.5, 19.3, and 9.3 points, respectively. Consistent gains across scales indicate effective expanded objectness modeling under appearance shifts. These results show that GAER and PGQM generalize well to sketches, paintings, cartoons, and adverse-condition scenes rather than overfitting domain-specific appearance cues.

Table 3: Quantitative results on the BTCO-XDomain (%).

| Methods | $AP_s$ | $AP_m$ | $AP_l$ | AP | $AP_{50}$ | $AP_{75}$ | $AR_{100}$ |
|---|---|---|---|---|---|---|---|
| SAM3 | 6.2 | 13.0 | 14.8 | 12.4 | 18.4 | 13.8 | 18.6 |
| RPN | 5.5 | 13.6 | 7.3 | 8.8 | 20.4 | 6.5 | 57.5 |
| CAOD | 11.7 | 25.7 | 30.0 | 27.4 | 54.0 | 25.0 | 50.7 |
| MAVL | 5.9 | 20.7 | 46.3 | 36.1 | 50.2 | 39.3 | 69.6 |
| DiPEx | 2.9 | 9.0 | 29.8 | 22.2 | 33.6 | 22.0 | 40.3 |
| PF-RPN | 17.1 | 36.7 | 49.8 | 42.9 | 54.1 | 47.0 | 82.4 |
| RF-DETR (DINOv2-S) | 14.8 | 24.9 | 25.2 | 23.9 | 37.1 | 25.3 | 77.0 |
| RF-DETR (DINOv3-L) | 24.1 | 30.7 | 24.9 | 26.0 | 37.6 | 28.6 | 80.4 |
| **ECADet** (DINOv3-S) | 21.5 | 42.4 | 50.6 | 45.8 | 65.3 | 49.3 | 79.9 |
| **ECADet** (DINOv3-B) | 28.8 | 49.4 | 53.5 | 50.0 | 68.7 | 53.8 | 83.4 |
| **ECADet** (DINOv3-L) | 57.5 | 71.4 | 64.6 | 65.1 | 75.6 | 66.3 | 91.7 |

### 5.3 Component Ablation

Table 4 reports component ablations of GAER and PGQM on BTCO-Real (Test). Both modules improve RF-DETR, with distinct effects. GAER raises AP from 24.5 to 27.7, with larger gains in $AP_{50}$ and $AR_{100}$, indicating that geometry-aware expert routing mainly improves localization quality and proposal coverage across diverse visual elements. PGQM achieves a slightly higher AP of 28.0, showing that prototype-guided query modulation enhances objectness representation and ranking reliability. Combining both modules yields

Table 4: Ablation study of GAER and PGQM (%).

| Method | $AP_s$ | $AP_m$ | $AP_l$ | AP | $AP_{50}$ | $AP_{75}$ | $AR_{100}$ |
|---|---|---|---|---|---|---|---|
| RF-DETR (DINOv3-L) | 14.2 | 23.1 | 22.3 | 24.5 | 44.5 | 20.8 | 53.0 |
| + GAER | 19.3 | 30.5 | 31.7 | 27.7 | 51.2 | 26.5 | 59.5 |
| + PGQM | 18.8 | 29.7 | 31.7 | 28.0 | 49.0 | 25.8 | 58.5 |
| + GAER + PGQM (ECADet (DINOv3-L)) | 25.5 | 44.7 | 55.0 | **41.0** | 65.4 | 42.7 | 61.9 |

41.0 AP, a 16.5-point gain over the baseline and a substantial improvement over either module alone. It also reaches 65.4 $AP_{50}$ and 42.7 $AP_{75}$, with pronounced gains on medium and large visual elements. These results confirm the complementary roles and strong synergy of GAER and PGQM.

### 5.4 Structural Hyperparameter Analysis

We analyze the configurations of GAER and PGQM under single-module settings. The number of regression experts is studied with RF-DETR+GAER at $\eta = 0.5$, while the offline prototype count is examined with RF-DETR+PGQM. Based on the results, we set the expert and prototype counts to 6 and 4, respectively, and evaluate sensitivity to $\eta$ using the full ECADet.

**Effect of the number of regression experts.** Table 5 investigates the effect of the number of regression experts in GAER on BTCO-Real (Test). AP first improves and then declines as the expert count increases. Using 6 experts achieves the best AP of 27.7, outperforming other configurations. This suggests that a moderate expert count provides sufficient capacity to model diverse geometric patterns while maintaining effective specialization. Too many experts may reduce the training samples assigned to each expert, leading to under-trained experts and less reliable routing. Therefore, we adopt 6 regression experts.

Table 5: Effect of regression expert count in GAER (%).

| #Reg. Exp. | $AP_s$ | $AP_m$ | $AP_l$ | AP | $AP_{50}$ | $AP_{75}$ | $AR_{100}$ |
|---|---|---|---|---|---|---|---|
| 5 | 19.5 | 29.7 | 30.6 | 26.0 | 45.3 | 25.9 | 59.5 |
| <u>6</u> | 19.3 | 30.5 | 31.7 | **27.7** | 51.2 | 26.5 | 59.5 |
| 7 | 18.3 | 28.1 | 29.3 | 24.5 | 43.7 | 24.0 | 58.6 |
| 8 | 20.1 | 30.9 | 30.5 | 23.5 | 43.3 | 25.3 | 59.9 |
| 9 | 17.7 | 27.1 | 27.8 | 23.5 | 43.2 | 22.7 | 57.3 |

**Effect of the number of offline objectness prototypes.** Table 6 investigates the number of offline objectness prototypes in PGQM on BTCO-Real (Test). Using 4 prototypes achieves the best AP of 28.0 and the highest $AP_{50}$ of 49.0. Although 3 prototypes yield slightly higher scale-specific APs, $AP_{75}$, and $AR_{100}$, the 4-prototype setting offers the best overall accuracy, suggesting a better balance between prototype diversity and matching stability. Increasing the count beyond 4 consistently degrades performance, likely because excessive partitioning produces redundant or weakly supported prototypes, reducing matching reliability. Therefore, we adopt 4 offline objectness prototypes in PGQM.

Table 6: Effect of objectness prototype count in PGQM (%).

| #Obj. Proto. | $AP_s$ | $AP_m$ | $AP_l$ | AP | $AP_{50}$ | $AP_{75}$ | $AR_{100}$ |
|---|---|---|---|---|---|---|---|
| 3 | 20.1 | 31.5 | 33.1 | 27.6 | 48.0 | 27.6 | 59.4 |
| <u>4</u> | 18.8 | 29.7 | 31.7 | **28.0** | 49.0 | 25.8 | 58.5 |
| 5 | 18.7 | 28.4 | 30.7 | 25.2 | 44.9 | 24.8 | 58.3 |
| 6 | 17.9 | 27.8 | 28.9 | 24.1 | 43.4 | 23.6 | 58.0 |

**Sensitivity to the routed residual scale $\eta$.** Table 7 analyzes the sensitivity to $\eta$ using the full ECADet with 6 regression experts and 4 offline objectness prototypes. As $\eta$ increases from 0.1 to 1.1, AP generally improves from 36.0 to 41.0, indicating that an appropriate residual scale allows the expert-routed branch to complement the base box prediction. Performance peaks at $\eta = 1.1$, with 41.0 AP, 65.4 $AP_{50}$, and 42.7 $AP_{75}$. Increasing $\eta$ to 1.3 causes a drop, suggesting that excessive residual refinement may perturb the base prediction and reduce localization stability. Meanwhile, the stable performance within $\eta \in [0.7, 1.1]$ shows limited sensitivity to moderate variations. Therefore, we set $\eta = 1.1$ in the final model. ***See Appendix B for qualitative results.***

Table 7: Sensitivity to the routed residual scale $\eta$ (%).

| $\eta$ | $AP_s$ | $AP_m$ | $AP_l$ | AP | $AP_{50}$ | $AP_{75}$ | $AR_{100}$ |
|---|---|---|---|---|---|---|---|
| 0.1 | 22.7 | 41.4 | 47.4 | 36.0 | 62.0 | 36.7 | 59.2 |
| 0.3 | 23.8 | 42.0 | 51.4 | 38.4 | 63.0 | 39.8 | 59.2 |
| 0.5 | 25.5 | 44.0 | 52.2 | 39.7 | 64.2 | 41.2 | 62.0 |
| 0.7 | 25.4 | 43.8 | 53.4 | 40.0 | 64.6 | 41.4 | 62.1 |
| 0.9 | 25.5 | 44.6 | 54.7 | 40.8 | 65.3 | 42.4 | 61.8 |
| <u>1.1</u> | 25.5 | 44.7 | 55.0 | **41.0** | 65.4 | 42.7 | 61.9 |
| 1.3 | 24.4 | 41.5 | 49.0 | 37.4 | 61.0 | 38.9 | 61.6 |

## 6 Conclusion

In this work, we introduce Expanded Class-Agnostic Detection (ECAD), extending thing-centric objectness to semantically meaningful visual elements. We construct BTCO-Bench for real-world and cross-domain evaluation and propose ECADet, a lightweight detector built on a frozen VLM. ECADet employs GAER and PGQM to model geometric diversity and improve objectness estimation. Extensive experiments demonstrate consistent gains over existing class-agnostic detection and proposal-generation methods, together with strong cross-domain generalization and favorable efficiency. We hope this work inspires broader exploration of objectness beyond conventional instances and advances visual perception in open and diverse environments.

### AI USE STATEMENT

In this work, generative AI tools were used only to assist with translation and language polishing during manuscript preparation. They were not used for problem formulation, benchmark design, method development, data annotation or processing, implementation, experimental design, or interpretation of results. All AI-assisted text was carefully reviewed and revised by the authors, who take full responsibility for the final content of the paper.

### ETHICS STATEMENT

This work does not involve the collection of new images, human-subject data, or personally identifiable information. BTCO-Bench is constructed by providing category-agnostic bounding-box annotations and evaluation splits for images from publicly available datasets. The proposed annotations focus exclusively on semantically meaningful visual regions and do not introduce identity-related labels. We will release the annotations and split definitions in accordance with the licenses and terms of use of the original datasets.

### REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the ECAD setting, the construction and annotation protocol of BTCO-Bench, the ECADet architecture, training settings, evaluation protocols, and ablation configurations. To facilitate reproducibility, we will release the source code, category-agnostic annotations, dataset split definitions, and evaluation tools.

## REFERENCES

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1209–1218, 2018.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

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

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14987–14997, 2025.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pp. 540–557. Springer, 2022.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5356–5364, 2019.

Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Premkumar Natarajan. Class-agnostic object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 919–928, 2021.

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*, pp. 5830–5840, 2021.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9404–9413, 2019.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, 2014.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Shuailei Ma, Yuefeng Wang, Ying Wei, Enming Zhang, Jiaqi Fan, Xinyu Sun, and Peihao Chen. Skdf: a simple knowledge distillation framework for distilling open-vocabulary knowledge to open-world object detector. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *European conference on computer vision*, pp. 512–531. Springer, 2022.

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 and machine intelligence*, 39(6):1137–1149, 2016.

Jia Syuen Lim, Zhuoxiao Chen, Mahsa Baktashmotlagh, Zhi Chen, Xin Yu, Zi Huang, and Yadan Luo. Dipex: Dispersing prompt expansion for class-agnostic object detection. *arXiv e-prints*, pp. arXiv–2406, 2024.

Qihong Tang, Changhan Liu, Shaofeng Zhang, Wenbin Li, Qi Fan, and Yang Gao. Prompt-free universal region proposal network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13080–13090, 2026.

Aming Wu and Cheng Deng. Single-domain generalized object detection in urban scene via cyclic-disentangled self-distillation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 847–856, 2022.

Yupeng Zhang, Shuqi Zheng, Ruize Han, Yuzhong Feng, Junhui Hou, Linqi Song, Wei Feng, and Liang Wan. Rethinking the one-shot object detection: Cross-domain object search. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9573–9581, 2024.

Yupeng Zhang, Ruize Han, Zhiwei Chen, Wei Feng, and Liang Wan. Noovd: Novel category discovery and embedding for open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6304–6313, 2026.

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

# APPENDIX

## A VISUALIZATION OF BTCO-BENCH

Fig. 3 and Fig. 4 present representative samples from BTCO-Real and BTCO-XDomain, respectively. BTCO-Bench extends conventional objectness beyond category-defined thing instances by including visually meaningful regions that provide semantic, functional, or contextual information. All annotations are represented as sparse bounding boxes without category labels, following the expanded class-agnostic detection setting.

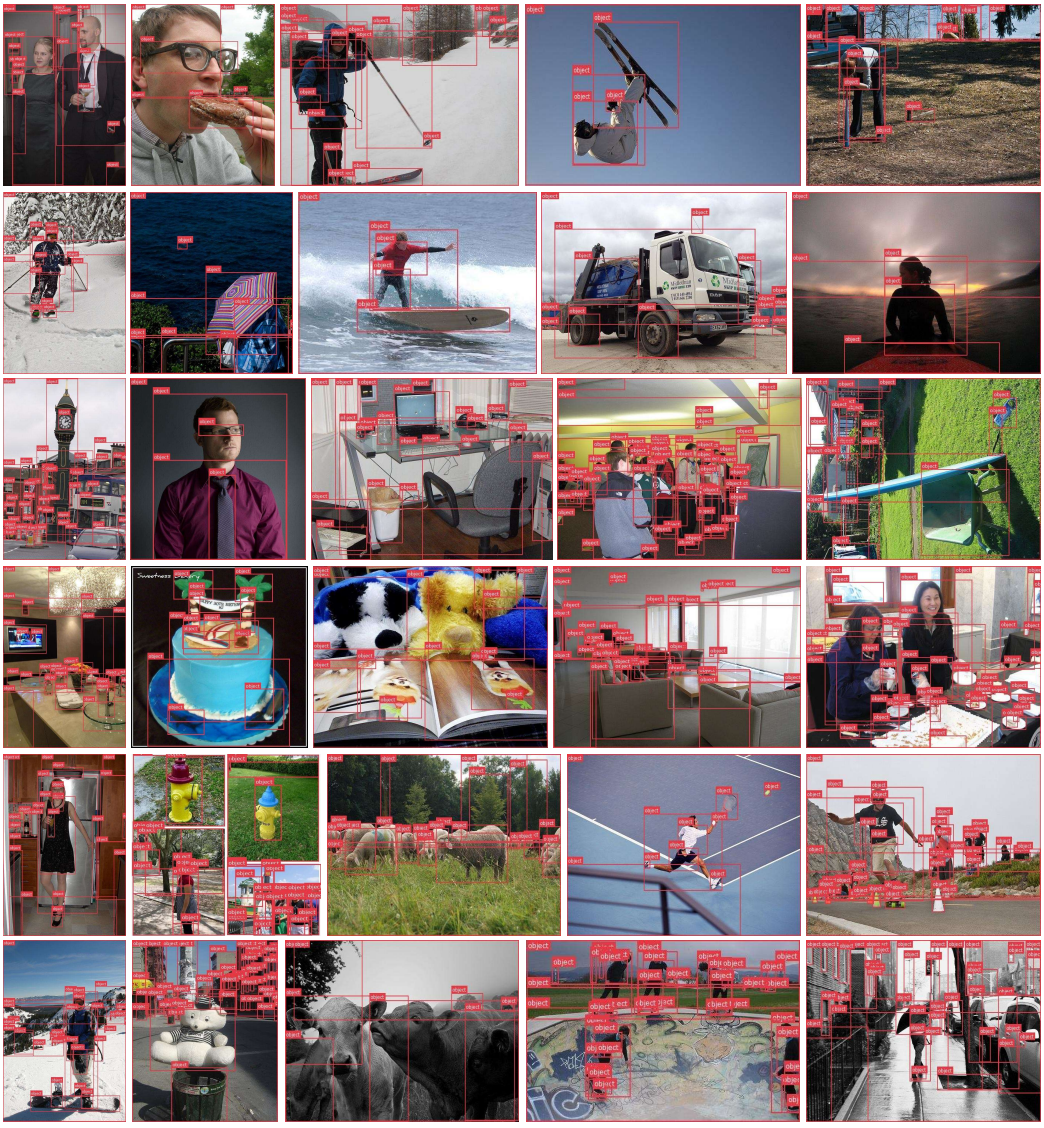

Figure 3: Representative samples from BTCO-Real with expanded objectness annotations.

BTCO-Real contains diverse real-world scenes with annotations covering both conventional object instances and expanded visual elements. Besides typical things such as people, vehicles, and animals, the annotations also include stuff-like regions and contextual elements, such as sky, water, roads, grasslands, floors, and countertops, which are often treated as background in existing detection benchmarks. These annotations aim to evaluate whether detectors can discover meaningful foreground regions beyond predefined object categories.

BTCO-XDomain further evaluates the generalization ability of expanded objectness under substantial appearance variations. It contains artistic transformations, including pencil sketches, cartoons,



Figure 4: Representative samples from BTCO-XDomain under diverse appearance shifts.

oil paintings, and watercolor images, as well as adverse environmental conditions such as fog, rain, and nighttime scenes. The examples demonstrate that the benchmark challenges detectors to identify semantically meaningful regions even when low-level appearance cues undergo significant changes.

## B Qualitative Results

To further understand the behavior of different methods beyond quantitative comparisons, we provide qualitative evaluations on both in-domain and cross-domain scenarios.

Fig. 5 presents qualitative comparisons on the BTCO-Real test set. MAVL and DiPEx mainly focus on conventional thing instances and fail to capture many semantically meaningful stuff-like regions, such as sky, sea, floors, and countertops. Although MAVL and PF-RPN are trained on large-scale datasets, their supervision remains dominated by conventional thing-centric annotations, limiting their ability to discover expanded visual elements beyond predefined object categories. DiPEx, despite being trained on the same BTCO-Real (Train) split as ECADet, still struggles to localize these regions, suggesting that the challenge lies not only in data coverage but also in the modeling capability required for expanded objectness. In contrast, PF-RPN generates dense and highly redundant proposals, achieving broad coverage at the cost of overlapping predictions and limited semantic selectivity. ECADet consistently localizes both conventional thing instances and semantically meaningful visual regions while maintaining a better balance between coverage and selectivity.

Fig. 6 further evaluates the cross-domain generalization of ECADet on BTCO-XDomain, which covers diverse appearance shifts including pencil sketches, cartoons, oil paintings, watercolor images, and adverse weather conditions. Despite substantial visual appearance variations, ECADet




Figure 5: Qualitative comparison on the BTCO-Real test set. ECADet provides more complete expanded objectness predictions by localizing both conventional thing instances and semantically meaningful stuff-like regions, while avoiding excessive duplicate detections.

maintains reliable localization of both thing instances and semantically meaningful stuff-like regions. By contrast, existing methods suffer from missing detections or noisy predictions under severe domain shifts, as they remain dependent on conventional appearance-driven object cues. In particular, MAVL and DiPEx overlook many meaningful regions, while PF-RPN generates excessive noisy proposals with limited semantic selectivity. These qualitative results further validate that the proposed expanded objectness formulation enables robust region discovery beyond conventional object cues under both in-domain and cross-domain scenarios.

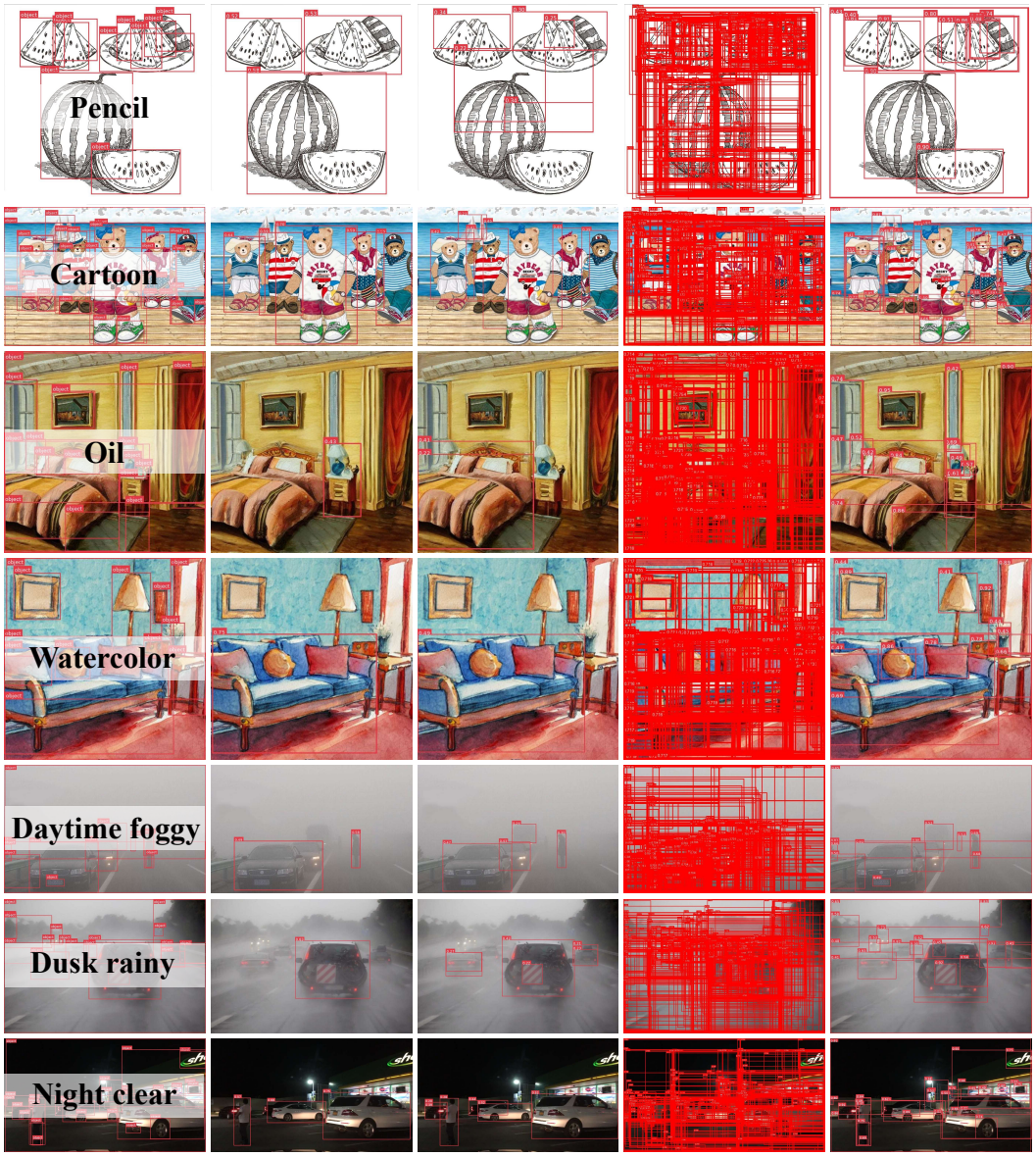


Figure 6: Qualitative comparison on the BTCO-XDomain under diverse appearance shifts, including pencil sketches, cartoons, oil paintings, watercolor images, and adverse weather scenes. ECADet preserves reliable localization of both thing instances and semantically meaningful stuff-like regions despite substantial domain variations.