CLIFF: Continual Latent Diffusion for Open-Vocabulary Object Detection

Wuyang Li[1], Xinyu Liu[1], Jiayi Ma[2], and Yixuan Yuan[1](✉)

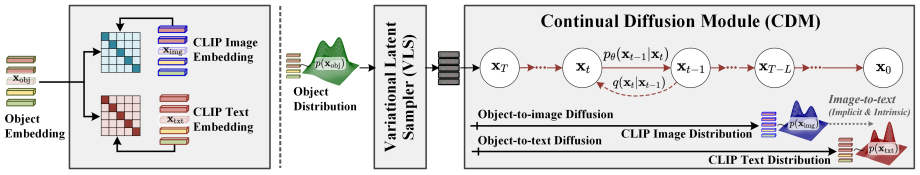
[1] The Chinese University of Hong Kong
[2] Wuhan University
wymanbest@outlook.com, yxyuan@ee.cuhk.edu.hk

Abstract. Open-vocabulary object detection (OVD) utilizes image-level cues to expand the linguistic space of region proposals, thereby facilitating the detection of diverse novel classes. Recent works adapt CLIP embedding by minimizing the object-image and object-text discrepancy combinatorially in a discriminative paradigm. However, they ignore the underlying distribution and the disagreement between the image and text objective, leading to the misaligned distribution between the vision and language sub-space. To address the deficiency, we explore the advanced generative paradigm with distribution perception and propose a novel framework based on the diffusion model, coined Continual Latent Diffusion (CLIFF), which formulates a continual distribution transfer among the object, image, and text latent space probabilistically. CLIFF consists of a Variational Latent Sampler (VLS) enabling the probabilistic modeling and a Continual Diffusion Module (CDM) for the distribution transfer. Specifically, in VLS, we first establish a probabilistic object space with region proposals by estimating distribution parameters. Then, the object-centric noise is sampled from the estimated distribution to generate text embedding for OVD. To achieve this generation process, CDM conducts a short-distance object-to-image diffusion from the sampled noise to generate image embedding as the medium, which guides the long-distance diffusion to generate text embedding. Extensive experiments verify that CLIFF can significantly surpass state-of-the-art methods on benchmarks. The code is available at https://github.com/CUHK-AIM-Group/CLIFF.

Keywords: Open-vocabulary object detection · Diffusion model · Continual distribution transfer

1 Introduction

Object detection [20, 37, 69] has made significant strides in recent years, as demonstrated by the impressive results on benchmarks [17, 47]. Nonetheless, due to the unsatisfactory closed-set constraint [77, 82], object detectors are limited to discovering the novel-class objects [38, 39] unseen during training, impeding a reliable scene understanding in the real world. Further, it hinders the journey toward achieving true artificial intelligence since it requires the human-like perceptual ability to recognize unexpected objects appearing everywhere.



(a) Existing OVD works (b) The proposed CLIFF framework modeling a continual distribution transfer

Fig. 1: Illustration of (a) existing discriminative paradigm [15, 68, 75] with a combinatorial objective (object→image and object→text in parallel), and (b) the proposed method conducting a continual diffusion (object→image→text) with generative nature.

To break through this barrier, open-vocabulary object detection (OVD) has been studied that leverages auxiliary linguistic cues to generalize the detection of novel classes [82]. Recent advances adapt CLIP [67], a vision-language (VL) foundation model with strong zero-shot capacity, to boost the novel-class generalization of the object detector. Some works [13, 14] utilize the CLIP prediction (image-level) to generate pseudo labels at the object level, enabling the self-training of the object detector. To fully explore the hidden knowledge, some other works strive to align the object-level representation with CLIP embeddings in the hidden space. During training the object detector, they minimize the object-image [15, 68, 75] and object-text [46, 68, 76] discrepancy to *align the sub-space among the object, image, and text modalities*, thereby unleashing the novel-class generalization capacity from the pre-trained CLIP model.

Despite the significant success, the existing OVD paradigm, as depicted in Fig. 1(a), has two limitations regarding the sub-space alignment. First, existing works utilize text embedding as the classifier weight to explicitly optimize the *discriminative* class separation, which is agnostic to the underlying distribution and may lead to class bias [18]. This distribution-agnostic property results in the distribution misalignment [40, 42, 46] and severely hinders the novel-class generalization [78]. Besides, due to the uncalibrated latent space in discriminative models [45], *e.g.*, numerical high-variance, the object distribution is prone to overfit to the labeled classes, worsening the misalignment with the unbiased CLIP distribution. With the rise of *generative* paradigm [2], this barrier has been gradually broken via diverse probabilistic modeling [18, 45, 59]. Hence, we aim to incorporate probabilistic advantages into discriminative detection pipelines, which can facilitate the distribution transfer across the object and text sub-spaces, thereby achieving an unbiased novel-class perception [45, 59].

Second, some existing works minimize the object-image and object-text discrepancy simultaneously in a *combinatorial* manner, ignoring the disagreement between the image and text two objectives[3]. With this combinatorial nature, optimizing object-image alignment as the same role as the object-text counterpart is suboptimal due to the potential conflict between the two objectives [25].

[3] As the first attempt, we discover the ever-overlooked disagreement issue and justify it in Appendix via experiments.

Intuitively, the image and object share visual representation, while the image and text are coupled with CLIP pre-training, which motivates us to introduce the image as an interpolated sub-space [61] between the object and text counterpart, yielding a *continual* distribution transfer among three modalities.

To address the above issues, we propose Continual Latent Diffusion (CLIFF) for OVD, as shown in Fig. 1(b). CLIFF breaks through the discriminative bottleneck via a probabilistic Variational Latent Sampler (VLS) and then formulates a diffusion-based distribution transfer [21] in a Continual Diffusion Module (CDM) to avoid the combinatorial limitation. Specifically, VLS leverages variational approximation [24] to establish a probabilistic object space with learned distribution parameters, and then samples object-centric noise to boost the diffusion process. CDM formulates the continual diffusion process to transfer the object distribution to the image and text counterpart, avoiding disagreement in the existing combinatorial objective. Besides, different from existing diffusion-based works [27, 78] with unsatisfactory efficiency, CLIFF achieves comparable and even better model efficiency compared with state-of-the-art OVD works, which is verified in Appendix. To be summarized, the contributions are four-fold:

- We propose Continual Latent Diffusion (CLIFF) for OVD, which achieves a continual distribution transfer with DDPM-based generation. To the best of our knowledge, this work represents the first attempt to study latent diffusion in VL discriminative tasks and push forward benchmark results.
- Variational Latent Sampler is proposed to generate a probabilistic object space with variational modeling, in which the object-centric noise can be sampled with reparameterization to boost the generation.
- We design an efficient Continual Diffusion Module to facilitate the distribution transfer among three cross-modality sub-spaces, namely the object, image, and text space, using a new diffusion module and unified DDPM formulation.
- CLIFF is extensively validated through different benchmarks, which verify its state-of-the-art role and show great adaptability to various detectors.

2 Related Works

2.1 Open-Vocabulary Object Detection

Open-vocabulary object detection (OVD) leverages auxiliary image-level labels with linguistic cues, *e.g.*, classification [68] and caption datasets [80], to enlarge the semantic space of the region proposals [5, 69], thereby facilitating the novel-class generalization for the real-world applications, such as automatic driving [41, 43, 50, 52], medical imaging analysis [31, 35, 36, 49] and diagnostic decision-making [1, 7, 44, 53, 79]. Zareian *et al.* [82] formulate the standard OVD problem setting and propose OVR-CNN to project the visual space to the linguistic counterpart on Faster-RCNN [69]. Some other works [22, 23, 60] delve into OVD with Vision Transformer (ViT) by designing large-scale region-aware pretraining [23], feature masking [22], which fully unleash the potential of ViT in fine-grained detection tasks. Recent works adapt CLIP [67] to objects detectors for the novel-class

generalization. The other works adapt the CLIP-encoded latent embedding and aim to align three critical sub-spaces: the object space, the CLIP image space, and the CLIP text space. They bridge the gap between the object and text sub-space via region-text contrastive pre-training [85], region-word bipartite matching [46], CLIP-based conditional matching [81], and learnable detection-friendly prompts [9, 11, 12]. PromptDet [12] and DetPro [9] use learnable text prompts to learn with noisy uncurated web images, thereby enhancing the object-text alignment. Additionally, recent advances [15, 68, 75] adapt the image sub-space with the embedding distillation [15], weight transfer [68] and bags of regions [75], to fully use CLIP embedding, serving as a combinatorial objective with the object-text alignment. Existing works follow a similar discriminative paradigm for OVD, which cannot model the underlying distribution with a sub-optimal distribution alignment. Differently, we formulate a continual transfer in a unified denoising diffusion module, enabling the distribution-level perception and adaptation among the object, image, and text sub-spaces.

2.2 Denoising Diffusion Probabilistic Model

As the advanced generative paradigm, Denoising Diffusion Probabilistic Model (DDPM) [21] ("diffusion model" for short) generates samples by formulating a gradual denoising process. This process can be treated as a distribution transfer [51] from the normal distribution to the data in a flow-like manner [51, 54, 84]. Instead of diffusing in the data space, LDM [70] and VDM [66] deploy the denoising diffusion in the latent space to enable the high-resolution synthesis. Recent works can be generally categorized into faster sampling [55, 73], noise corruption designs [3, 62], and diffusion model architecture [33, 34, 65], etc. The diffusion model has achieved remarkable generative performance in the image, text, video, audio, graph, and motion generation [21, 28, 30, 32, 74] with different diffusion conditions for downstream applications.

In addition to generative tasks, recent works [6,16,27,56,78,83] have discovered that the diffusion feature has adequate semantic discriminability with richer distribution-level perception and data-efficient property [4, 64] than conventional discriminative models [67]. Without fine-tuning, an off-the-shelled diffusion model can achieve satisfactory zero-shot classification [29]. Along this trend, some bottlenecks in discriminative models are gradually broken by involving the generative paradigm [78]. CARD [19] uses the diffusion model to estimate the uncertainty to improve image classification. Unlike existing works considering the noise→data diffusion, we formulate a continual distribution transfer among three modalities, yielding the object→image→text transfer in the latent space.

3 Preliminaries

Problem Formulation. Assuming a set of base classes $\mathcal{C}_B$ and novel classes $\mathcal{C}_N$, we have a detection dataset $\mathcal{D}_{det} = \{I_i, O_i\}_{i=1}^{n}$ with object-level labels $\{O_i\}_{i=1}^{n}$ in $\mathcal{C}_B$ and another auxiliary dataset $\mathcal{D}_{aux} = \{I_i, Y_i\}_{i=1}^{n'}$ with image-level

Algorithm 1 Reverse Diffusion Process (Denoising): RevDif($\mathbf{x}_T, \mathbf{c}, L$)

Require: $\mathbf{x}_T$: initial noise; $\mathbf{c}$: diffusion condition; L: the number reverse steps.
Ensure: $\mathbf{x}_{T-L}$: generated samples with L steps.
1: **for** $t = T, ..., T-L+1$ **do**
2: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
3: $\quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta \left(\mathbf{x}_t, \mathbf{c}, t \right) \right) + \sigma_t \mathbf{z}$
4: **end for**
5: **return** $\mathbf{x}_{T-L}$

labels, *e.g.,* the image caption [21] and classification dataset [68], which has the weakly supervised labels in $\mathcal{C}_B \cup \mathcal{C}_N$. In $\mathcal{D}_{det}$, each image I_i is associated with a set of objects $O_i = \{(x, y, w, h), \hat{y}\}$, including the bounding-box coordinates $\{(x, y, w, h)\}$ and class labels $\{\hat{y}\} \subset \mathcal{C}_B$. OVD aims to use $\mathcal{D}_{det}$ and $\mathcal{D}_{aux}$ for training and detect base and novel objects in $\mathcal{C}_B \cup \mathcal{C}_N$ during the test stage.

Diffusion Model. Formulating a generative paradigm with the Markov chain [21, 63], DDPM is able to generate data $\mathbf{x}_0$ from the noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ via a T-step denoising process. Concretely, DDPM begins with the clean data $\mathbf{x}_0$ and gradually adds Gaussian noise using the fixed forward process $q(\mathbf{x}_t|\mathbf{x}_{t-1})$, where $t = 1, 2, ..., T$. This process can be denoted as follows,

$$q\left(\mathbf{x}_t|\mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, \left(1-\bar{\alpha}_t\right)\mathbf{I}\right), \tag{1}$$

where $\bar{\alpha}_t$ is noise schedule [62]. To recover data $\mathbf{x}_0$ from noise $\mathbf{x}_T$, the reverse process parameterized with θ is learned to denoise inputs with condition $\mathbf{c}$:

$$p_\theta(\mathbf{x}_{0:T}|\mathbf{c}) = p(\mathbf{x}_T) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}), \tag{2}$$

where $p(\mathbf{x}_T)$ is fixed to Gaussian. In this reverse step, the diffusion model typically learns a neural network $\boldsymbol{\epsilon}_\theta\left(\cdot\right)$ to generate the values for computing the mean of $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c})$ given by the noisy input $\mathbf{x}_t$, which is optimized to minimize the variational lower bound (ELBO). As proved in [21, 29], this training objective can be re-written as:

$$-\mathbb{E}_{t, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})} \left[\left\| \epsilon - \boldsymbol{\epsilon}_\theta \left(\mathbf{x}_t, \mathbf{c}, t \right) \right\|^2 \right], \tag{3}$$

where the diffusion model $\boldsymbol{\epsilon}_\theta(\cdot)$ can estimate the added noise of each noisy input $\mathbf{x}_t$. After optimizing with Eq. 3, the diffusion model can generate a data sample $\mathbf{x}_{T-L}$ at arbitrary time-step $T-L$ from a random noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ with L steps of the reverse diffusion, optionally guided by the condition $\mathbf{c}$. This process [21] is represented in Algorithm 1. In the following sections, we use RevDif$(\cdot)$ to represent the reverse diffusion (noise→data) for simplicity.

4 Methodology

The overall workflow is illustrated in Fig. 2. We feed input images $\{I_i\}_{i=1}^{B}$ to the backbone and RPN [69] to extract N region proposals, which are sent to the

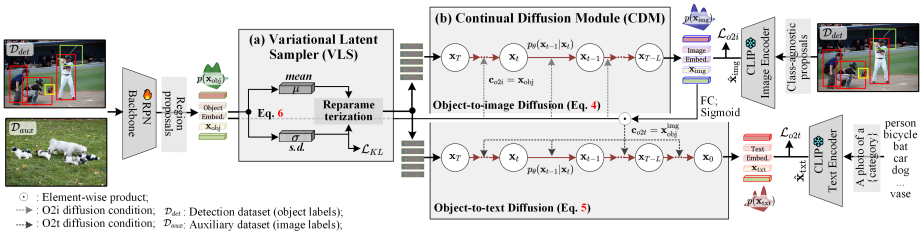


Fig. 2: Illustration of the proposed CLIFF. VLS establishes a probabilistic embedding space with region proposals, where we sampled object-centric noise $\mathbf{x}_T$. Then, CDM achieves a continual distribution transfer from the object to the image and text.

feature extractor to generate object embedding $\mathbf{x}_{\text{obj}}$. Then, VLS, parameterized with ϕ, establishes and estimates the mean and variance of a probabilistic object space $p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}})$, where the reparameterization is conducted to sample object-centric noise $\mathbf{x}_T$, as shown in Fig. 2(a). After that, the sampled noise $\mathbf{x}_T$ is sent to CDM, denoted as $\boldsymbol{\epsilon}_\theta(\cdot)$, to generate text embedding $\mathbf{x}_{\text{txt}}$ with diffusion, as depicted in Fig. 2(b). Conditioned on $\mathbf{c}_{o2i}$, we first conduct object-to-image diffusion to generate the image embedding $\mathbf{x}_{\text{img}}$ with L steps (top sub-figure), which is used to generate refined diffusion condition $\mathbf{c}_{o2t}$. Conditioned on $\mathbf{c}_{o2t}$, we further perform a T-step object-to-text diffusion to generate text embedding $\mathbf{x}_{\text{txt}}$ (bottom sub-figure), whose similarity with the CLIP text embedding $\hat{\mathbf{x}}_{\text{txt}}$ are measured to serve as the final class prediction.

Formally, by extending Eq. 2, CLIFF consists of two diffusion processes among the object, image, and text embedding sub-space. Guided by the condition $\mathbf{c}_{o2i}$, the **object-to-image** diffusion, written as $p_{\theta,\phi}(\mathbf{x}_{\text{img}}|\mathbf{c}_{o2i}, \mathbf{x}_{\text{obj}})$, aims to generate the associated CLIP image embedding from the object $\mathbf{x}_{\text{obj}}$ with L steps:

$$p_{\theta,\phi}(\mathbf{x}_{\text{img}}|\mathbf{c}_{o2i}, \mathbf{x}_{\text{obj}}) = p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}}) \prod_{t=T-L+1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}_{o2i}). \tag{4}$$

Similarly, with condition $\mathbf{c}_{o2t}$, the **object-to-text** diffusion is designed to generate text embedding with T steps for the final prediction, denoted as follows,

$$p_{\theta,\phi}(\mathbf{x}_{\text{txt}}|\mathbf{c}_{o2t}, \mathbf{x}_{\text{obj}}) = p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}}) \prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}_{o2t}). \tag{5}$$

The proposed CLIFF offers two significant improvements compared to the standard diffusion process (Eq. 2). First, instead of beginning from a random noise space $p(\mathbf{x}_T)$, we start from a differential probabilistic space $p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}})$ with objects to generate the text embedding. Second, going beyond a single diffusion between two sub-spaces, we formulate the effective diffusion process among the object, image, and text in different modalities.

4.1 Variational Latent Sampler

With input images $\{I_i\}_{i=1}^{B}$, the backbone, RPN[4], and the feature extractor are adopted to obtain N object embeddings $\mathbf{x}_{\text{obj}} \in \mathbb{R}^{N\times D}$. Then, we send $\mathbf{x}_{\text{obj}}$ to VLS to establish a probabilistic embedding space $p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}})$ satisfying the DDPM w.r.t the normally distributed assumption for the diffusion-based transfer. This probabilistic space $p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}})$ can break through the discriminative limitation of [14,15,46,75] without class bias [18]. Besides, it introduces object-level semantics to enable object-to-text transfer with a differential nature, different from the non-optimizable random space $p(\mathbf{x}_T)$ of noise-to-data transfer [21].

Formally, inspired by VAE [24], we establish the probabilistic space with centered isotropic multivariate Gaussian $p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}}) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$, whose distribution parameters are computed from the observed object embedding $\mathbf{x}_{\text{obj}}$ with non-linear layers [18]. To this end, two fully connected layers are introduced on the shared region feature extractor of the basic object detector, denoted as f_{obj} *e.g.*, the ResNet blocks in Faster RCNN [69] and the MLP in Mask RCNN [20], to learn to estimate the distribution parameters. Specifically, given the object embedding $\mathbf{x}_{\text{obj}}$, the variational posterior $p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}})$ is written as $\log p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}}^{(i)}) = \log \mathcal{N}(\mathbf{x}_T; \boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{2(i)}\mathbf{I})$. Without losing the differential property [24], the *mean* and *s.d.* of the probabilistic object space can be estimated,

$$\begin{aligned} \boldsymbol{\mu}^{(i)} &= \mathbf{W}_\mu f_{\text{obj}}(\mathbf{x}_{\text{obj}}^{(i)}) + \mathbf{b}_\mu, \\ \boldsymbol{\sigma}^{(i)} &= \mathbf{W}_\sigma f_{\text{obj}}(\mathbf{x}_{\text{obj}}^{(i)}) + \mathbf{b}_\sigma, \end{aligned} \tag{6}$$

where $\mathbf{W}_{(\cdot)}$ and $\mathbf{b}_{(\cdot)}$ are learned weight and bias. Then, with the probabilistic modeling, we sample object-centric noise from the latent object space $\mathbf{x}_T \sim p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}})$ with reparameterization [24]: $\mathbf{x}_T^{(i)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \boldsymbol{\epsilon}$, where $\odot$ is the element-wise product and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

To optimize the probabilistic space, we deploy Kullback–Leibler (KL) divergence $D_{\text{KL}}\left(p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}})||p(\mathbf{x}_T)\right)$ with $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ that encourages the Gaussian distributed [24], which has **two advantages**. First, the probabilistic nature of normal Gaussian removes the class bias within the object space [18] and promotes the generalization capacity towards discovering novel classes [45, 59]. Second, it naturally satisfies the DDPM constraint (Eq. 1) with the Gaussian formulation [21] for the following diffusion process. The implemented KL optimization objective, as proved in the Appendix, can be re-written in the following format,

$$\mathcal{L}_{KL} = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{D}\left(1 + \log(\boldsymbol{\sigma}_j^{2(i)}) - \boldsymbol{\mu}_j^{2(i)} - \boldsymbol{\sigma}_j^{2(i)}\right). \tag{7}$$

VLS can establish a probabilistic space with differential property and sample object-centric noise $\mathbf{x}_T$ for the diffusion process, serving to generate probabilistic object-text pairs in the diffusion-based distribution transfer [48, 51].

[4] Besides N_{rpn} RPN proposals, we follow [68, 81] to add additional proposals for training, including N_{cs} class-specific proposals (from $\mathcal{D}_{aux}$) and N_{ca} class-agnostic proposals (from $\mathcal{D}_{det}$) from [58]. The number of proposals is $N = N_{rpn} + N_{cs} + N_{ca}$.

4.2 Continual Diffusion Module

With object-centric noise $\mathbf{x}_T$, we further deploy the object-to-image (o2i) and object-to-text (o2t) diffusion to achieve a continual distribution transfer for OVD. To achieve this, an efficient diffusion module $\boldsymbol{\epsilon}_\theta(\cdot)$ with MLP architecture is designed, the details of which are presented in the Appendix.

Object-to-image Diffusion. Given class-agnostic region proposals, we send them to the object detector to generate the object embedding $\mathbf{x}_{\text{obj}}$ and feed them into frozen CLIP image encoder [67] to obtain static image embedding $\hat{\mathbf{x}}_{\text{img}}$. Then, with a randomly sampled object-centric noise $\mathbf{x}_T \sim p_\phi(\mathbf{x}_T|\mathbf{x}_{\text{obj}})$ from VLS, we encourage the diffusion module to generate the image embedding $\mathbf{x}_{\text{img}}$ with object-level condition $\mathbf{c}_{o2i} = \mathbf{x}_{\text{obj}}$ using a L-step reverse diffusion:

$$\mathbf{x}_{\text{img}} = \text{RevDif}(\mathbf{x}_T, \mathbf{c}_{o2i}, L), \tag{8}$$

where RevDif$(\cdot)$ (Algorithm 1) recovers image embedding from object-centric noise $\mathbf{x}_T$. To this end, the object-to-image diffusion, denoted as $p_{\theta,\phi}(\mathbf{x}_{\text{img}}|\mathbf{c}_{o2i}, \mathbf{x}_{\text{obj}})$, uses object embedding as the diffusion condition $\mathbf{c}_{o2i} = \mathbf{x}_{\text{obj}}$, and is optimized by minimizing the distance between the generated embedding $\mathbf{x}_{\text{img}}$ and CLIP image embedding $\hat{\mathbf{x}}_{\text{img}}$, which is achieved via the following loss function,

$$\mathcal{L}_{o2i} = \frac{1}{N_{ca}} \sum_{i=1}^{N_{ca}} ||\mathbf{x}_{\text{img}}^{(i)} - \hat{\mathbf{x}}_{\text{img}}^{(i)}||_1, \tag{9}$$

where N_{ca} is the number of class-agnostic proposals. With the optimized diffusion process, the object distribution could be transferred to the CLIP image counterpart, facilitating the adaptation to the CLIP latent space.

Object-to-text Diffusion. With diffusion-based formulation, we further encourage the model to reconstruct the associated text embedding $\mathbf{x}_{\text{txt}}$ for region proposals $\mathbf{x}_{\text{obj}}$, which conducts a generative procedure in latent space. Note that this generative objective is different from existing OVD works that directly optimize the discriminative class separation. Specifically, we design the object-to-text diffusion with a T-step diffusion $p_{\theta,\phi}(\mathbf{x}_{\text{txt}}|\mathbf{c}_{o2t}, \mathbf{x}_{\text{obj}})$ guided by condition $\mathbf{c}_{o2t}$. Considering the dominant role of condition [21] in the generated content, we propose a **simple yet effective** attentive mechanism to obtain the refined condition $\mathbf{c}_{o2t} = \mathbf{x}_{\text{obj}}^{\text{img}}$ with the generated image image-level context:

$$\mathbf{x}_{\text{obj}}^{\text{img}} = \text{Sigmoid}\left(\text{FC}(\mathbf{x}_{\text{img}})\right) \odot \mathbf{x}_{\text{obj}}, \tag{10}$$

where $\mathbf{x}_{\text{img}}$ is generated via Eq. 8, $\odot$ is the element-wise product, and FC is a liner projection. Then, we use the refined condition $\mathbf{c}_{o2t} = \mathbf{x}_{\text{obj}}^{\text{img}}$ to guide the text embedding generation for $\mathbf{x}_{\text{txt}}$ with a T-step reverse diffusion:

$$\mathbf{x}_{\text{txt}} = \text{RevDif}(\mathbf{x}_T, \mathbf{c}_{o2t}, T). \tag{11}$$

Considering the mutual benefits of generative and discriminative objectives for model learning as explored in [45], the loss function for the object-to-text

diffusion process is designed with these two distinct components. To this end, the optimization objective can be expressed as follows,

$$\mathcal{L}_{o2t} = \underbrace{\frac{1}{N_l}\sum_{i}^{N_l} ||\boldsymbol{\epsilon}_\theta(\mathbf{x}_t^{(i)}, \mathbf{c}_{o2t}^{(i)}, t) - \boldsymbol{\epsilon}||_1}_{\text{Generative}} + \underbrace{\mathcal{L}_{CE}\left(\text{Sigmoid}\Big(\frac{1}{\tau} cos(\mathbf{x}_{\text{txt}}, \hat{\mathbf{x}}_{\text{txt}})\Big), \hat{y}\right)}_{\text{Discriminative}}, \tag{12}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{x}_t$ is the noisy CLIP text embedding[5] at a randomly sampled time-step $t \sim \mathcal{U}[1, T]$, $\tau = 0.02$ is a scaling factor as [15, 68], $\hat{y}$ is the class label, $\mathcal{L}_{CE}$ is the Cross-Entropy loss, and $N_l = N_{cs} + N_{rpn}$ counts the labeled proposals. The first term optimizes the diffusion module $\boldsymbol{\epsilon}_\theta(\cdot)$ to reconstruct the text embedding in a generative manner [21], serving for the objective of Eq. 3. The second term can give an interpretable score for NMS [69] in detectors and ensure the generated embedding with sufficient discriminability.

Remark. We use a shared diffusion module with a T-step diffusion formulation to optimize two processes, where the first L steps are assigned for object-to-image, and the whole T steps are for object-to-text diffusion. Note that even though we do not explicitly optimize image-to-text diffusion, this transition can be intrinsically achieved between the L-th and T-th steps, as illustrated by the gray-colored part in Fig. 1(b). The reason is based on the continual property of the Markov chain in DDPM, *i.e.*, formulating a random process that undergoes transitions from one state to another within a finite number of possible states [21, 63].

4.3 Training and Inference

Training Stage. We implement the overall optimization objective on the basic object detector, Faster RCNN [69], to train the whole framework:

$$\mathcal{L}_{CLIFF} = \mathcal{L}_{rpn} + \mathcal{L}_{reg} + \alpha\mathcal{L}_{KL} + \beta_1\mathcal{L}_{o2i} + \beta_2\mathcal{L}_{o2t}, \tag{13}$$

where $\mathcal{L}_{rpn}$ and $\mathcal{L}_{reg}$ are RPN loss and the bounding box regression loss in the RCNN head [69] borrowed from object detectors. α, β_1, β_2 are the loss weight.

Inference Stage. We first extract the region proposals $\mathbf{x}_{\text{obj}}$ and then conduct reverse diffusion to generate the associated text embedding $\mathbf{x}_{\text{txt}}$ for region proposals, whose similarity with $\hat{\mathbf{x}}_{\text{txt}}$ is measured for class prediction $\tilde{y}$:

$$\tilde{y} = \text{Sigmoid}\left(\frac{1}{\tau} cos(\mathbf{x}_{\text{txt}}, \hat{\mathbf{x}}_{\text{txt}})\right), \tag{14}$$

where $\tilde{y}$ is the per-class confidence, $\mathbf{x}_{\text{txt}}$ is the generated text embedding via object-to-text diffusion (Eq. 11), and $\hat{\mathbf{x}}_{\text{txt}}$ is CLIP text embedding for all classes.

[5] Following Eq. 1, $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_{\text{txt}} + (1 - \bar{\alpha}_t)\,\boldsymbol{\epsilon}$ as [21]

Table 1: Experimental results (%) on COCO benchmark. † indicates the reproduced results with the consistent baseline setting as [68], including ImageNet pre-training and CLIP prompting. ‡ indicates using the longer training schedule [68].

Method	Object Detector	Sub-space Alignment	mAP_n	mAP_b	mAP	Reference
OVR-CNN [82]	Faster RCNN C4	Obj → Txt	22.8	46.0	39.9	CVPR-2021
RegionCLIP [85]	Faster RCNN C4	Obj → Txt & Img	26.8	54.8	47.5	CVPR-2022
ViLD [15]	Faster RCNN C4	Obj → Txt & Img	27.6	59.5	51.3	ICLR-2022
OV-DETR [81]	Deformable DETR	Obj → Txt & Img	29.4	61.0	52.7	ECCV-2022
Detic [86]	Faster RCNN C4	Obj → Txt	27.8	47.1	45.0	ECCV-2022
Detic‡ [86]	Faster RCNN C4	Obj → Txt	28.4	53.8	47.2	ECCV-2022
PB-OVD [14]	Faster RCNN C4	Obj → Txt	30.8	46.1	42.1	ECCV-2022
MAVL-OVD [58]	Faster RCNN C4	Obj → Txt	30.4	52.6	46.8	ECCV-2022
OCD‡ [68]	Faster RCNN C4	Obj → Txt & Img	36.6	54.0	49.4	NeurIPS-2022
F-VLM [26]	Faster RCNN FPN	Obj → Txt & Img	28.0	-	39.6	ICLR-2023
VLDet [46]	Faster RCNN C4	Obj → Txt	32.0	50.6	45.8	ICLR-2023
BARON†,‡ [75]	Faster RCNN C4	Obj → Txt & Img	41.0	53.3	50.1	CVPR-2023
EdaDet [71]	Faster RCNN C4	Obj → Txt	37.8	57.7	52.5	ICCV-2023
GOAT [72]	Faster RCNN C4	Obj → Txt & Img	36.4	53.0	48.6	ICCV-2023
CLIFF (ours)	Faster RCNN C4	Obj → Image → Txt	**41.3**	54.1	50.6	-
CLIFF‡ (ours)	Faster RCNN C4	Obj → Image → Txt	**43.2**	54.0	51.2	-

5 Experiments

5.1 Datasets and Evaluation Metrics

We follow existing works [68, 75] to compare with state-of-the-art methods. **COCO Setting.** The COCO-2017 dataset [47] is used for training and evaluation, containing 80 object-level classes. As for the auxiliary dataset $\mathcal{D}_{aux}$, we follow [82] to use COCO caption dataset in $\mathcal{C}_B \cup \mathcal{C}_N$. The core evaluation metric is the novel-class Average Precision (AP) with an IoU threshold of 0.5, denoted as mAP_n, for the novel-class generalization. The results are shown in Table 1.
LVIS Setting. LVIS v1.0 [17] contains 1203 classes and is split into frequent, common, and rare types. For the auxiliary datasets $\mathcal{D}_{aux}$, we follow [68] to use the subset of ImageNet-21K. The core evaluation metric is mask AP of rare classes, denoted as AP_r^m. We report the comparison results in Table 2.
Pascal VOC Setting. We follow [72] to evaluate the generalization to unseen datasets with Pascal VOC [10] and LVIS [17] using COCO-trained model. We use AP_{50} to evlaute the performance as [72]. The results are given in Table 7.

5.2 Implementation Details

Object Detector. Considering that previous works use inconsistent detectors, we strictly follow the setting of [68,82] in different benchmarks. Concretely, for the COCO setting, we use Faster-RCNN C4 [69] with the ResNet-50 backbone. The backbone is pre-trained on ImageNet [8] without bells and whistles. For the LVIS setting, we follow the two main streams of existing works by using Mask-RCNN-FPN [20] and CenterNet v2 [87] as the basic object detectors for a thorough comparison, which is based on ReseNet-50 backbone. For the experiments of all benchmarks, we use the SGD optimizer with a weight decay of $1e^{-4}$ and a

Table 2: Experimental results (%) on LVIS benchmark with the consistent object detector setting. Ens indicates using the ensemble trick [11].

Method	Object Detector	AP_r^m	AP_c^m	AP_f^m	AP	Reference
ViLD-Ens [15]	Mask RCNN	16.6	24.6	30.3	25.5	ICLR-2022
MVAL-OVD [58]	Mask RCNN	17.0	21.2	26.1	22.4	ECCV-2022
OCD [68]	Mask RCNN	21.1	25.0	29.1	25.9	NeurIPS-2022
DetPro [9]	Mask RCNN	19.8	25.6	28.9	25.9	CVPR-2022
RegionCLIP [85]	Mask RCNN	17.1	27.4	34.0	28.2	CVPR-2022
F-VLM [26]	Mask RCNN	18.6	24.0	26.9	24.2	ICLR-2023
BARON [75]	Mask RCNN	18.0	24.4	28.9	25.1	CVPR-2023
BARON-Ens [75]	Mask RCNN	19.2	26.8	29.4	26.5	CVPR-2023
CLIFF (ours)	Mask RCNN	**22.4**	26.9	29.8	26.8	-
VLDet [46]	CenterNet v2	21.7	29.8	34.3	30.1	ICLR-2023
CoDet [57]	CenterNet v2	23.4	30.0	34.6	30.7	NeirIPS-2023
GOAT [72]	CenterNet v2	23.3	29.7	34.3	30.4	ICCV-2023
EdaDet [71]	CenterNet v2	23.7	27.5	29.1	27.5	ICCV-2023
CLIFF (ours)	CenterNet v2	**24.5**	27.7	29.8	28.2	-

momentum of 0.9 following [68]. We use 4 NVIDIA V100 for training with a batch size of 16. We pre-train the base object detector with a 1× schedule as [68], and then train the whole model with a 2× schedule. We also explore the longer 8× training schedule to make fair comparisons with specific methods.

CLIP Model. As the preious work [68], we use the single prompt, '`a photo of a {category}`', to compute CLIP text embedding $\hat{\mathbf{x}}_{\text{txt}}$, and use 224×224 input-scale for image embedding $\hat{\mathbf{x}}_{\text{img}}$ via ViT-B/32 model [67].

CLIFF. The number of the time-step in object-to-text (T) and object-to-image (L) diffusion are set $T = 10, L = 3$, which are justified in Table 4. α, β_1, and β_2 are empirically set to 2.0, 15.0, and 1.0, respectively, which are verified by the experiments in Appendix. L_2 normalization is deployed on CLIP embedding as [15, 68]. Class-specific and RPN proposals are used to optimize object-to-text diffusion [68, 75], while class-agnostic ones are for object-to-image diffusion.

5.3 Comparison with State-of-the-Art Methods

Results on COCO Benchmark. As shown in Table 1, we compare the proposed method with state-of-the-art counterparts on the COCO benchmark. CLIFF achieves 41.3% mAP_n, which outperforms OCD [68] (36.6% mAP_n), VLDet [46] (32.0% mAP_n) and BARON [75] (41.0% mAP_n) with 4.7%, 9.3% and 0.3% mAP_n, comprehensively verifying the effectiveness of the proposed CLIFF. The proposed method also surpasses EdaDet [71] and GOAT [72] by a noticeable margin. Besides, after using fair multi-scale training [68, 75], our method (43.2% mAP_n) outperforms existing works by a large margin. Unlike existing works aligning object-text and object-image separately, we formulate a continual diffusion process among the three sub-spaces in a different learning paradigm.

Results on LVIS Benchmark. We present a thorough comparison with corresponding state-of-the-art methods using Mask RCNN and CenterNet v2 detectors, which is shown in Table 2. For Mask RCNN, it can be observed that CLIFF achieves the best 22.4% AP_r^m, which surpasses state-of-the-art counterparts

Table 3: Ablation study results on COCO and LVIS benchmarks with 1× schedule.

	CDM		VLS	COCO			LVIS-Box				LVIS-Mask			
	o2t	o2i		mAP_n	mAP_b	mAP	AP^b_r	AP^b_c	AP^b_f	AP^b	AP^m_r	AP^m_c	AP^m_f	AP^m
1	(baseline)			32.2	49.9	45.3	15.8	20.1	26.2	21.4	15.2	19.7	25.0	20.5
2	✓			36.6	49.5	46.1	16.2	20.9	27.4	22.6	15.8	20.1	25.7	21.6
3	✓	✓		39.5	52.4	49.2	16.7	21.7	27.7	23.1	16.5	20.9	25.9	22.1
4	✓		✓	38.8	**53.3**	49.3	17.3	21.9	27.8	23.3	16.7	21.1	26.1	22.2
5	✓	✓	✓	**40.0**	52.9	**49.5**	**18.9**	**22.1**	**28.0**	**23.8**	**18.1**	**21.4**	**26.2**	**22.7**

Table 4: Analysis on the diffusion step. o2i and o2t indicates the object-to-image and object-to-text diffusion process respectively. * indicates sharing the same diffusion procedure and optimizing o2i and o2t in a combinatorial style [15, 68].

	o2i	o2t	mAP_n	mAP_b	mAP
1	1	10	40.2	53.4	50.0
2	3	10	41.3	**54.1**	**50.6**
3	5	10	41.3	53.7	50.3
4	7	10	41.2	53.9	50.5
5	9	10	**41.4**	53.5	50.4
6	3	5	40.9	53.8	50.3
7	3	10	41.3	**54.1**	**50.6**
8	3	20	39.7	53.4	49.8
9	10	10	**41.4**	53.6	50.4
10	*10	*10	39.9	52.8	49.5

Table 5: Analysis of the discriminative and generative terms in Eq. 12 by replacing the random noise (Rand) with object-centric noise (OC) in different time steps (2× schedule). *rows* 7 and 8 indicate removing individual terms.

	Discriminative		Generative		mAP_n
	T	T-1: 0	T	T-1: 0	
1	Rand	Rand	Rand	Rand	39.2
2	OC	Rand	OC	Rand	41.0
3	Rand	OC	Rand	OC	40.4
4	OC	OC	OC	Rand	41.1
5	OC	Rand	OC	OC	**41.3**
6	OC	OC	OC	OC	**41.3**
7	w/o Discriminative		OC	OC	38.2
8	OC	OC	w/o Generative		40.1

BARON [75] (19.2% AP^m_r), OCD [68] (21.3% AP^m_r) with 3.1% AP^m_r and 1.3% AP^m_r, respectively. Besides, our method surpasses the baseline model MAVL-OVD [58] with a significant 5.4% AP^m_r, verifying the effectiveness of our method. For the comparison with CenterNet v2, we can find that our method gives 24.5% $AP^m_r$and also surpasses the latest published works noticeably, outperforming CoDet [57] and GOAT [72] with 1.1% and 1.2% gains. This can reveal that CLIFF is a generic method in terms of multi-modal alignment and can be deployed in different object detectors with great transferability and adaptability.

5.4 Quantitative Analysis

Ablation Study. The ablative results are shown in Table 3 with a 1× training schedule. The first line indicates the model trained with extra MAVL proposals [58, 68], which is a fair baseline model as [68, 75]. Compared with the baseline model (32.2% mAP_n, 15.8% AP^b_r, 15.2% AP^m_r), introducing the object-to-text (*row* 2) diffusion gives satisfactory improvements with 36.6% mAP_n, 16.2% AP^b_r, and 15.8% AP^m_r. After integrating object-to-image diffusion (*row* 3), *i.e.*, the full CDM, we can observe a further improvement with 7.3% mAP_n, 0.9% AP^b_r, and 1.3% AP^m_r gains, compared with the baseline. This verifies the superior effectiveness in modeling a continual distribution transfer than the existing discriminative paradigm. Besides, after replacing the randomly sampled noise by

Table 6: Analysis of the refined condition $\mathbf{c}_{o2t}$ in object-to-image diffusion (Eq. 5) with 2× training schedule.

	Setting of $\mathbf{c}_{o2t}$	mAP_n	mAP_b	mAP
1	Object Embed: $\mathbf{x}_\text{obj}$	40.4	53.4	49.9
2	Image Embed: $\mathbf{x}_\text{img}$	39.6	52.9	49.4
3	Additive: $\mathbf{x}_\text{obj} + \mathbf{x}_\text{img}$	40.8	53.5	50.2
4	Concated: $[\mathbf{x}_\text{obj} : \mathbf{x}_\text{img}]$	41.0	53.6	50.4
5	Attentive: Eq. 10	**41.3**	**54.1**	**50.6**

Table 7: Cross-dataset evaluation with AP_{50} (%) transferring COCO-trained models to the PASCAL VOC and LVIS.

Method	PASCAL	LVIS
OVR-CNN [82]	52.9	5.2
PB-OVD [14]	59.2	8.0
VLDet [46]	61.7	10.0
GOAT [72]	63.6	14.0
CLIFF (Ours)	**66.1**	**16.7**

object-centric noise with VLS (*row* 5), it can be observed that the CDM (*row* 3) can be improved with 40.0% mAP_n, 18.9% $\text{AP}_\text{r}^\text{b}$, and 18.1% $\text{AP}_\text{r}^\text{m}$, verifying the effectiveness of each component in the proposed CLIFF.

The Number of Diffusion Steps. We study different diffusion steps for o2i and o2t in Table 4, and have the following three observations. First, in *rows* 1-5, we observe that the results are stable when $L > 3$ (around 41.3% mAP_n), but give notable performance drops when the step is insufficient $L = 1$. Considering more computation requirements when enlarging the diffusion step, we select $L = 3$ with the best trade-off. Second, in *rows* 6-8, setting too small and large o2t steps both give significant performance drops, which may be caused by unsatisfactory diffusion optimization. Third, we set consistent diffusion steps and compare the proposed continual diffusion (*row* 9) with existing combinatorial style, *i.e.*, sharing the diffusion process and generated embedding between the o2i and o2t (*row* 10). With a 1.5% performance gain over the combinatorial objective, the effectiveness of our continual transfer can be justified.

Exploring Object-centric Noise. We further give a detailed analysis of the object-centric (OC) noise by replacing the random noise for the discriminative and generative terms in object-to-text objective (the core process for OVD) in different time steps, as shown in Table 5. Compared with the random-noise baseline with 39.2% mAP_n, controlling the initial stage (*row* 2) and middle stages (*row* 3) gives 41.0% ane 40.4% mAP_n, respectively. Introducing more OC noise in the diffusion (*rows* 4-6) gives stable performance improvements, indicating the effectiveness of the VLS. Further, we conduct an ablative study on o2t diffusion (*rows* 7-8), *i.e.*, the first and second terms in Eq. 12, verifying their critical role in optimizing the diffusion process with both aspects.

Analyzing the Condition Design. The strategy of refining the object-to-text (Eq. 5) condition $\mathbf{c}_{o2t} = \mathbf{x}_\text{obj}^\text{img}$ is analyzed in Table 6. Compared with individually using the object and image (*rows* 1-2) condition with 40.4% and 39.6% mAP_n, combining the two aspects for the diffusion condition generally works better (*rows* 3-5), *e.g.*, addition with 40.8% mAP_n, concatenation with 41.0%. Besides, we observe that the proposed attentive refinement on the condition (Eq. 10) achieves optimal performance and gives the best 41.3% mAP_n, due to the adaptive fusion of image-level context. This can verify the effectiveness of our design.

Generalization to Unseen Datasets. We further conduct the cross-dataset comparison following [72] by considering two different settings. Compared with

Fig. 3: Qualitative comparison among (a) the baseline [68], (b) our CLIFF, and (c) ground-truth on the COCO benchmark. We only visualize the novel-class predictions for a clearer evaluation of open-vocabulary generalization.

state-of-the-art counterpart GOAT [72], the proposed CLIFF gives 66.1% and 16.7% AP_{50} on the PASCAL VOC and LVIS datasets using the COCO-trained model, which surpasses GOAT with significant 2.5% and 2.7% gains, showing the superior generalization capacity of our method.

5.5 Qualitative Analysis

As shown in Fig. 3, we visualize the novel-class prediction and conduct a qualitative comparison among (a) the baseline [68, 75] trained with MAVL proposals [68], (b) the proposed CLIFF, and (c) ground-truth. It can be observed that the proposed method gives more accurate novel-class predictions, especially for occlusion situations and complex scenarios. Our method reduces the confusion among novel classes, *e.g.*, correcting the wrong prediction (couch) given by the baseline in *col.* 1. Besides, the proposed CLIFF can relieve some missing errors, as illustrated in *col.* 2-5, revealing its superior novel-class generalization capacity. Hence, the proposed CLIFF is able to relieve the bias on the base classes and encourage the model to discover the novel-class objects with probabilistic modeling.

6 Conclusion

This paper proposes a novel diffusion-based framework for OVD, coined CLIFF, which breaks through the bottlenecks of the combinatorial and discriminative limitation in existing OVD pipelines. The proposed CLIFF consists of a Variational Latent Sampler (VLS) to sample object-centric noise in the probabilistic object space, and leverages a Continual diffusion Module (CDM) to formulate a continual distribution transfer among three sub-spaces (object, image, and text) with cross-modal nature. Extensive experiments validate that the proposed method significantly outperforms existing approaches on large-scale benchmarks.

Acknowledgement. This work was supported by Innovation and Technology Commission- Innovation and Technology Fund ITS/229/22 and Hong Kong Research Grants Council (RGC) General Research Fund 11211221.

References

1. Ali, S., Ghatwary, N., Jha, D., Isik-Polat, E., Polat, G., Yang, C., Li, W., Galdran, A., Ballester, M.Á.G., Thambawita, V., et al.: Assessing generalisability of deep learning-based polyp detection and segmentation methods through a computer vision challenge. Sci. Rep. (2024)
2. Amit, T., Nachmani, E., Shaharbany, T., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. ArXiv:2112.00390 (2021)
3. Bansal, A., Borgnia, E., Chu, H.M., Li, J.S., Kazemi, H., Huang, F., Goldblum, M., Geiping, J., Goldstein, T.: Cold diffusion: Inverting arbitrary image transforms without noise. ArXiv:2208.09392 (2022)
4. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: ICLR (2022)
5. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020)
6. Chen, S., Sun, P., Song, Y., Luo, P.: Diffusiondet: Diffusion model for object detection. ArXiv:2211.09788 (2022)
7. Chen, Z., Li, W., Xing, X., Yuan, Y.: Medical federated learning with joint graph purification for noisy label learning. MedIA (2023)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
9. Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., Li, G.: Learning to prompt for open-vocabulary object detection with vision-language model. ArXiv:2203.14940 (2022)
10. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV **88**, 303–338 (2010)
11. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Expand your detector vocabulary with uncurated images. ArXiv:2203.16513 (2022)
12. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images. In: ECCV. pp. 701–717. Springer (2022)
13. Gao, M., Xing, C., Niebles, J.C., Li, J., Xu, R., Liu, W., Xiong, C.: Towards open vocabulary object detection without human-provided bounding boxes. ArXiv:2111.09452 (2021)
14. Gao, M., Xing, C., Niebles, J.C., Li, J., Xu, R., Liu, W., Xiong, C.: Open vocabulary object detection with pseudo bounding-box labels. In: ECCV (2022)
15. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. In: ICLR (2022)
16. Gu, Z., Chen, H., Xu, Z., Lan, J., Meng, C., Wang, W.: Diffusioninst: Diffusion model for instance segmentation. ArXiv:2212.02773 (2022)
17. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: CVPR (2019)
18. Han, J., Ren, Y., Ding, J., Yan, K., Xia, G.S.: Few-shot object detection via variational feature aggregation. ArXiv:2301.13411 (2023)

19. Han, X., Zheng, H., Zhou, M.: Card: Classification and regression diffusion models. In: NeurIPS (2022)
20. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: CVPR (2017)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
22. Kim, D., Angelova, A., Kuo, W.: Contrastive feature masking open-vocabulary vision transformer. In: ICCV. pp. 15602–15612 (2023)
23. Kim, D., Angelova, A., Kuo, W.: Region-aware pretraining for open-vocabulary object detection with vision transformers. In: CVPR. pp. 11144–11154 (2023)
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. ArXiv:1312.6114 (2013)
25. Korte, B.H., Vygen, J., Korte, B., Vygen, J.: Combinatorial optimization, vol. 1. Springer (2011)
26. Kuo, W., Cui, Y., Gu, X., Piergiovanni, A., Angelova, A.: Open-vocabulary object detection upon frozen vision and language models. In: ICLR (2023)
27. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. In: ICLR (2023)
28. Leng, Y., Chen, Z., Guo, J., Liu, H., Chen, J., Tan, X., Mandic, D., He, L., Li, X.Y., Qin, T., et al.: Binauralgrad: A two-stage conditional diffusion probabilistic model for binaural audio synthesis. ArXiv:2205.14807 (2022)
29. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. ArXiv:2303.16203 (2023)
30. Li, C., Feng, B.Y., Fan, Z., Pan, P., Wang, Z.: Steganerf: Embedding invisible information within neural radiance fields. In: ICCV. pp. 441–453 (2023)
31. Li, C., Lin, X., Mao, Y., Lin, W., Qi, Q., Ding, X., Huang, Y., Liang, D., Yu, Y.: Domain generalization on medical imaging classification using episodic training with task augmentation. Computers in biology and medicine **141**, 105144 (2022)
32. Li, C., Liu, H., Fan, Z., Li, W., Liu, Y., Pan, P., Yuan, Y.: Gaussianstego: A generalizable stenography pipeline for generative 3d gaussians splatting. ArXiv:2407.01301 (2024)
33. Li, C., Liu, H., Liu, Y., Feng, B.Y., Li, W., Liu, X., Chen, Z., Shao, J., Yuan, Y.: Endora: Video generation models as endoscopy simulators. ArXiv:2403.11050 (2024)
34. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Yuan, Y.: U-kan makes strong backbone for medical image segmentation and generation. ArXiv:2406.02918 (2024)
35. Li, C., Zhang, Y., Li, J., Huang, Y., Ding, X.: Unsupervised anomaly segmentation using image-semantic cycle translation. ArXiv:2103.09094 (2021)
36. Li, C., Zhang, Y., Liang, Z., Ma, W., Huang, Y., Ding, X.: Consistent posterior distributions under vessel-mixing: a regularization for cross-domain retinal artery/vein classification. In: ICIP. pp. 61–65 (2021)
37. Li, W., Chen, Z., Li, B., Zhang, D., Yuan, Y.: Htd: Heterogeneous task decoupling for two-stage object detection. TIP (2021)
38. Li, W., Guo, X., Yuan, Y.: Novel scenes & classes: Towards adaptive open-set object detection. In: ICCV. pp. 15780–15790 (2023)
39. Li, W., Liu, J., Han, B., Yuan, Y.: Adjustment and alignment for unbiased open set domain adaptation. In: CVPR. pp. 24110–24119 (2023)
40. Li, W., Liu, X., Yao, X., Yuan, Y.: Scan: Cross domain object detection with semantic conditioned adaptation. In: AAAI. pp. 1421–1428 (2022)
41. Li, W., Liu, X., Yuan, Y.: Scan++: Enhanced semantic conditioned adaptation for domain adaptive object detection. TMM (2022)

42. Li, W., Liu, X., Yuan, Y.: Sigma: Semantic-complete graph matching for domain adaptive object detection. In: CVPR (2022)
43. Li, W., Liu, X., Yuan, Y.: Sigma++: Improved semantic-complete graph matching for domain adaptive object detection. TPAMI (2023)
44. Li, W., Yang, C., Liu, J., Liu, X., Guo, X., Yuan, Y.: Joint polyp detection and segmentation with heterogeneous endoscopic data. In: ISBI Workshop: EndoCV 2021. pp. 69–79. CEUR-WS Team (2021)
45. Liang, C., Wang, W., Miao, J., Yang, Y.: Gmmseg: Gaussian mixture based generative semantic segmentation models. ArXiv:2210.02025 (2022)
46. Lin, C., Sun, P., Jiang, Y., Luo, P., Qu, L., Haffari, G., Yuan, Z., Cai, J.: Learning object-language alignments for open-vocabulary object detection. In: ICLR (2023)
47. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
48. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
49. Liu, H., Liu, Y., Li, C., Li, W., Yuan, Y.: Lgs: A light-weight 4d gaussian splatting for efficient surgical scene reconstruction. ArXiv:2406.16073 (2024)
50. Liu, X., Li, W., Yamaguchi, T., Geng, Z., Tanaka, T., Tsai, D.P., Chen, M.K.: Stereo vision meta-lens-assisted driving vision. ACS Photonics (2024)
51. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. ArXiv:2209.03003 (2022)
52. Liu, X., Li, W., Yang, Q., Li, B., Yuan, Y.: Towards robust adaptive object detection under noisy annotations. In: CVPR. pp. 14207–14216 (2022)
53. Liu, X., Li, W., Yuan, Y.: Decoupled unbiased teacher for source-free domain adaptive medical object detection. TNNLS (2023)
54. Liu, Y., Li, C., Yang, C., Yuan, Y.: Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. ArXiv:2401.12561 (2024)
55. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. ArXiv:2211.01095 (2022)
56. Luo, J., Li, Y., Pan, Y., Yao, T., Feng, J., Chao, H., Mei, T.: Semantic-conditional diffusion networks for image captioning. ArXiv:2212.03099 (2022)
57. Ma, C., Jiang, Y., Wen, X., Yuan, Z., Qi, X.: Codet: Co-occurrence guided region-word alignment for open-vocabulary object detection. NeurIPS **36** (2024)
58. Maaz, M., Rasheed, H., Khan, S., Khan, F.S., Anwer, R.M., Yang, M.H.: Class-agnostic object detection with multi-modal transformer. In: ECCV (2022)
59. Miller, D., Sünderhauf, N., Milford, M., Dayoub, F.: Uncertainty for identifying open-set errors in visual object detection. Robot. Autom. Lett. **7**(1), 215–222 (2021)
60. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: ECCV. pp. 728–755 (2022)
61. Ni, J., Qiu, Q., Chellappa, R.: Subspace interpolation via dictionary learning for unsupervised domain adaptation. In: CVPR (June 2013)
62. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021)
63. Norris, J.R.: Markov chains. No. 2, Cambridge university press (1998)
64. Pan, P., Fan, Z., Feng, B.Y., Wang, P., Li, C., Wang, Z.: Learning to estimate 6dof pose from limited data: A few-shot, generalizable approach using rgb images. In: 2024 International Conference on 3D Vision (3DV). pp. 1059–1071. IEEE (2024)
65. Peebles, W., Xie, S.: Scalable diffusion models with transformers. ArXiv:2212.09748 (2022)

66. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: CVPR. pp. 10619–10629 (2022)
67. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
68. Rasheed, H.A., Maaz, M., Khattak, M.U., Khan, S., Khan, F.: Bridging the gap between object and image-level representations for open-vocabulary detection. In: NeurIPS (2022)
69. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In: NeurIPS (2015)
70. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
71. Shi, C., Yang, S.: Edadet: Open-vocabulary object detection using early dense alignment. In: ICCV. pp. 15724–15734 (2023)
72. Wang, J., Zhang, H., Hong, H., Jin, X., He, Y., Xue, H., Zhao, Z.: Open-vocabulary object detection with an open corpus. In: ICCV. pp. 6759–6769 (2023)
73. Wizadwongsa, S., Suwajanakorn, S.: Accelerating guided diffusion sampling with splitting numerical methods. In: ICLR (2023)
74. Wu, L., Gong, C., Liu, X., Ye, M., Liu, Q.: Diffusion-based molecule generation with informative prior bridges. ArXiv:2209.00865 (2022)
75. Wu, S., Zhang, W., Jin, S., Liu, W., Loy, C.C.: Aligning bag of regions for open-vocabulary object detection. ArXiv:2302.13996 (2023)
76. Wu, X., Zhu, F., Zhao, R., Li, H.: Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. ArXiv:2303.13076 (2023)
77. Xu, H., Zhang, Y., Sun, L., Li, C., Huang, Y., Ding, X.: Afsc: Adaptive fourier space compression for anomaly detection. ArXiv:2204.07963 (2022)
78. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. ArXiv:2303.04803 (2023)
79. Yang, Q., Li, W., Li, B., Yuan, Y.: Mrm: Masked relation modeling for medical image pre-training with genetics. In: ICCV. pp. 21452–21462 (2023)
80. Ye, K., Zhang, M., Kovashka, A., Li, W., Qin, D., Berent, J.: Cap2det: Learning to amplify weak caption supervision for object detection. In: ICCV (2019)
81. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary detr with conditional matching. In: ECCV. pp. 106–122 (2022)
82. Zareian, A., Rosa, K.D., Hu, D.H., Chang, S.F.: Open-vocabulary object detection using captions. In: CVPR (2021)
83. Zbinden, L., Doorenbos, L., Pissas, T., Huber, A.T., Sznitman, R., Márquez-Neila, P.: Stochastic segmentation with conditional categorical diffusion models. In: ICCV (2023)
84. Zhang, Y., Li, C., Lin, X., Sun, L., Zhuang, Y., Huang, Y., Ding, X., Liu, X., Yu, Y.: Generator versus segmentor: Pseudo-healthy synthesis. In: MICCAI. pp. 150–160 (2021)
85. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: CVPR. pp. 16793–16803 (2022)
86. Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: ECCV. pp. 350–368 (2022)
87. Zhou, X., Koltun, V., Krähenbühl, P.: Probabilistic two-stage detection. In: ArXiv:2103.07461 (2021)