

Unknown-Aware Object Detection: Learning What You Don't Know from Videos in the Wild

Xuefeng Du¹, Xin Wang², Gabriel Gozum¹, and Yixuan Li¹

¹University of Wisconsin-Madison, ²Microsoft Research

{xfdu, sharonli}@cs.wisc.edu, wanxin@microsoft.com, ggozum@wisc.edu

Abstract

Building reliable object detectors that can detect out-of-distribution (OOD) objects is critical yet underexplored. One of the key challenges is that models lack supervision signals from unknown data, producing overconfident predictions on OOD objects. We propose a new unknown-aware object detection framework through Spatial-Temporal Unknown Distillation (STUD), which distills unknown objects from videos in the wild and meaningfully regularizes the model's decision boundary. STUD first identifies the unknown candidate object proposals in the spatial dimension, and then aggregates the candidates across multiple video frames to form a diverse set of unknown objects near the decision boundary. Alongside, we employ an energy-based uncertainty regularization loss, which contrastively shapes the uncertainty space between the in-distribution and distilled unknown objects. STUD establishes the state-of-the-art performance on OOD detection tasks for object detection, reducing the FPR95 score by over 10% compared to the previous best method. Code is available at <https://github.com/deeplearning-wisc/stud>.

1. Introduction

Object detection models have achieved remarkable success in known contexts for which they are trained. Yet, they often struggle with out-of-distribution (OOD) data—samples from unknown classes that the network has not been exposed to during training, and therefore should not be predicted by the model in testing. Teaching the object detectors to be aware of unknown objects is critical for building a reliable vision system, especially in safety-critical applications like autonomous driving [8] and medical analysis [2].

While much research progress is made in OOD detection for classification models [17, 20, 29, 31, 33, 36, 59], the problem remains underexplored in the context of object detection. Unlike image-level OOD detection, detecting un-

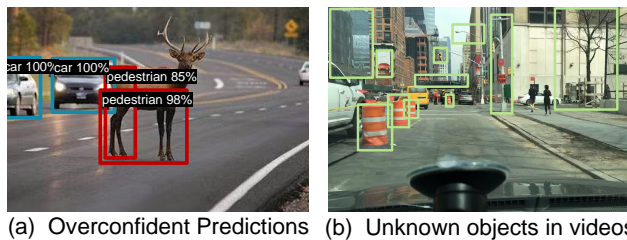


Figure 1. (a) Vanilla object detectors can predict OOD objects (e.g., deer) as an ID class (e.g., pedestrian) with high confidence. (b) Unknown objects (in bounding boxes) naturally exist in the video datasets, such as billboards, traffic cones, overbridges, street lights, etc. Image is taken from the BDD100K dataset [67].

knowns for object detection requires a finer-grained understanding of the complex scenes. In practice, an image can be OOD in specific regions while being in-distribution (ID) elsewhere. Taking autonomous driving as an example, we observe that an object detection model trained to recognize ID objects (e.g., cars, pedestrians) can produce a high-confidence prediction for an unseen object such as a deer; see Figure 1(a). This happens when our object detector minimizes its training error without explicitly accounting for the uncertainty that could appear outside the training categories. Unfortunately, the plethora of ways that unknown objects can emerge are innumerable in an open world. It is arguably expensive to annotate a large number of OOD objects in complex scenes—in addition to the already costly process of ID data collection.

In this paper, we propose a new *unknown-aware object detection* framework through Spatial-Temporal Unknown Distillation (STUD), which distills unknown objects from videos in the wild and meaningfully regularizes the model's decision boundary. Video data naturally captures the open-world environment that the model operates in, and encapsulates a mixture of both known and unknown objects; see Figure 1(b). For example, buildings and trees (OOD) may appear in the driving video, though they are not labeled explicitly for training an object detector for cars and pedestrians (ID). Our approach draws an analogy to the concept of distillation in chemistry, which refers to the “process of

separating the substances from a mixture” [46]. While classic object detection models primarily use the labeled known objects for training, we attempt to capitalize on the unknown ones for model regularization by jointly optimizing object detection and OOD detection performance.

Concretely, our framework consists of two components, tackling challenges of (1) distilling diverse unknown objects from videos, and (2) regularizing object detector with the distilled unknown objects. To address the first problem, we introduce a new *spatial-temporal unknown distillation* approach, which automatically constructs *diverse* unknown objects (Section 3.1). In the spatial dimension, for each ID object in a frame, we identify the unknown object candidates in the reference frames based on an OOD measurement. We then distill the unknown object by linearly combining the selected objects in the feature space, weighted by the dissimilarity measurement. The distilled unknown object therefore captures a more diverse distribution over multiple objects than using single ones. In the temporal dimension, we propose aggregating unknown objects from multiple video frames, which captures additional diversity of unknowns in the temporal dimension.

Leveraging the distilled unknown objects, we further employ an unknown-aware training objective (Section 3.2). Unlike vanilla object detection, we train the object detector with an uncertainty regularization branch. Our regularization facilitates learning a more conservative decision boundary between ID and OOD objects, which helps flag unseen OOD objects during inference. To achieve this, the regularization contrastively shapes the uncertainty surface, which produces larger probabilistic scores for ID objects and vice versa, enabling effective OOD detection in testing. Our key contributions are summarized as follows:

- We propose a new framework *STUD*, addressing a challenging yet underexplored problem of unknown-aware object detection. To the best of our knowledge, we are the first to exploit the rich information from videos to enable OOD identification for the object detection models.
- *STUD* effectively regularizes object detectors by distilling diverse unknown objects in both spatial and temporal dimensions without costly human annotations of OOD objects. Moreover, we show that *STUD* is more advantageous than synthesizing unknowns in the high-dimensional pixel space (e.g., using GAN [30]) or using negative proposals as unknowns [23].
- We extensively evaluate the proposed *STUD* on large-scale BDD100K [67] and Youtube-VIS datasets [66]. *STUD* obtains state-of-the-art results, outperforming the best baseline by a large margin (10.88% in FPR95 on BDD100K) while preserving the accuracy of object detection on ID data.

2. Problem Setup

We start by formulating the OOD detection problem for the object detection task. Most previous formulations of OOD detection treat entire images as anomalies, which can lead to ambiguity shown in Figure 1(a). In particular, natural images are not monolithic entities but instead are composed of numerous objects and components. Knowing which regions of an image are anomalous allows for the safe handling of unfamiliar objects. Compared to image-level OOD detection, object-level OOD detection is more relevant in realistic perception systems, yet also more challenging as it requires reasoning OOD uncertainty at the fine-grained object level. We design reliable object detectors that are aware of unknown OOD objects in testing. That is, an object detector trained on the ID categories (e.g., cars, trucks) can identify test-time objects (e.g., deer) that do not belong to the training categories and refrain from making a confident prediction on them.

Setup. We denote the input and label space by $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{1, 2, \dots, K\}$, respectively. Let $\mathbf{x} \in \mathcal{X}$ be the input image, $\mathbf{b} \in \mathbb{R}^4$ be the bounding box coordinates associated with objects in the image, and $y \in \mathcal{Y}$ be the semantic label of the object. An object detection model is trained on ID data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{b}_i, y_i)\}_{i=1}^M$ drawn from an unknown joint distribution \mathcal{P} . We use neural networks with parameters θ to model the bounding box regression $p_\theta(\mathbf{b}|\mathbf{x})$ and the classification $p_\theta(y|\mathbf{x}, \mathbf{b})$.

OOD detection for object detection. The OOD detection can be formulated as a binary classification problem, distinguishing between the in vs. out-of-distribution objects. Let $P_{\mathcal{X}}$ denote the marginal probability distribution on \mathcal{X} . Given a test input $\mathbf{x}^* \sim P_{\mathcal{X}}$, as well as an object \mathbf{b}^* predicted by the object detector, the goal is to predict $p_\theta(g|\mathbf{x}^*, \mathbf{b}^*)$. We use $g = 1$ to indicate a detected object being ID, and $g = 0$ being OOD, with semantics outside the support of \mathcal{Y} .

3. Unknown-Aware Object Detection

Our unknown-aware object detection framework trains an object detector in tandem with the OOD uncertainty regularization branch. Both share the feature extractor and the prediction head and are jointly trained from scratch (see Figure 2). Our framework encompasses two novel components, which address: (1) how to distill diverse unknown objects in the spatial and temporal dimensions (Section 3.1), and (2) how to leverage the unknown objects for effective model regularization (Section 3.2).

3.1. Spatial-Temporal Unknown Distillation

Our approach *STUD* distills unknown objects guided by the rich spatial-temporal information in videos, without ex-

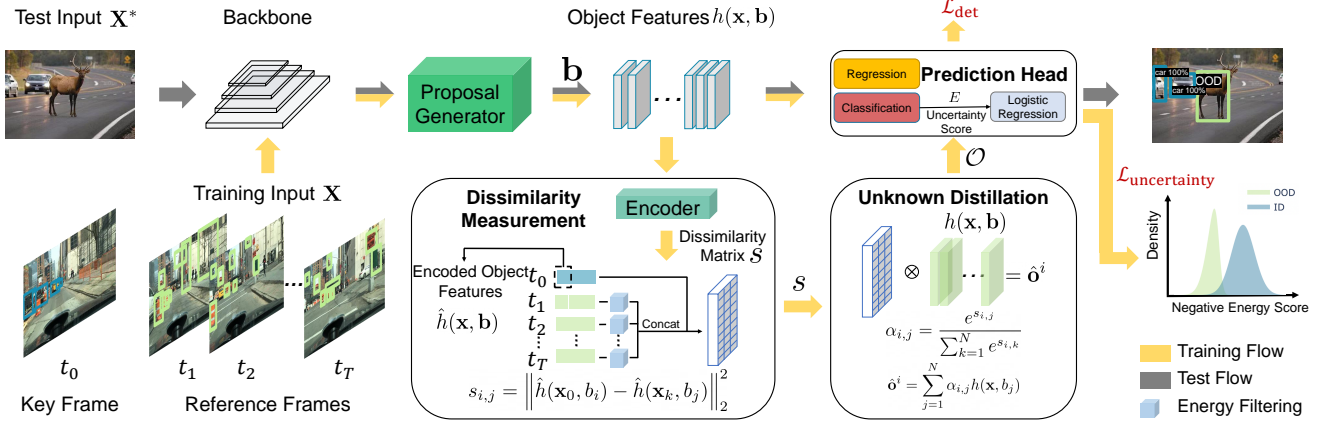


Figure 2. **Overview of the proposed unknown-aware object detection framework STUD.** For an ID object from the key frame encoded as $\hat{h}(\mathbf{x}_0, \mathbf{b}_i)$, we perform energy filtering to identify the unknown object candidates in the reference frames. We then distill the unknown object \hat{o}^i by linearly combining the unknown objects in the feature space, weighted by the dissimilarity score $s_{i,j}$. The distilled unknowns, along with the ID objects, are used to train the uncertainty regularization branch ($\mathcal{L}_{uncertainty}$). $\mathcal{L}_{uncertainty}$ contrastively shapes the uncertainty surface, which produces a larger score for ID objects and vice versa. During testing, we use the output of the logistic regression for OOD detection. \otimes denotes the operation in Equation (3) and $1 \leq k \leq T$ is the index of the reference frames.

PLICIT supervision signals of unknown objects. Video data naturally encapsulates a mixture of both known and unknown objects. While classic object detection models primarily use the labeled known objects for training, we attempt to capitalize on the unknown ones for model regularization. For this reason, we term our approach *unknown distillation*—extracting unknown objects *w.r.t* the known objects. Notably, our distillation process for object detection is performed at the object level, in contrast to constructing the image-level outliers [18]. That is, for every ID object in a given frame, we construct a corresponding OOD counterpart. The distilled unknowns will be used for model regularization (Section 3.2).

While intuition is straightforward, challenges arise in constructing unknown objects in an unsupervised manner. The plethora of ways that unknown objects can emerge are innumerable in high-dimensional space. Taking the ID object *car* as an example (*c.f.* Figure 3), the objects such as billboards, trees, buildings, etc. can all be considered as unknowns *w.r.t* the car. This undesirably increases the sample complexity and demands a diverse collection of unknown objects to be observed. We tackle the challenge through distilling diverse unknown objects by leveraging the rich information in the spatial and temporal dimensions of videos.

Spatial unknown distillation. In the *spatial* dimension, for each ID object in a given frame, we create the unknown counterpart through a linear combination of the object features from the reference frames, weighted by the dissimilarity measurement. Utilizing multiple objects captures a more diverse distribution of unknowns than using single ones. STUD operates on the feature outputs from the proposal generator to calculate dissimilarity. Specifically, we consider a pair of frames $\mathbf{x}_0, \mathbf{x}_1$ at timestamps t_0



Figure 3. **The dissimilarity measurement.** For each ID object at timestamp t_0 (in blue), we discover the objects in the reference frame that are dissimilar to it (in green), which are more likely to contain OOD objects for model regularization. The red numbers show the dissimilarity after normalization (Equation (2)).

and t_1 , designated key frame and reference frame, respectively. For an object (\mathbf{x}, \mathbf{b}) , we denote its feature representation as $h(\mathbf{x}, \mathbf{b}) \in \mathbb{R}^m$, where m is the feature dimension. We collect a set of object features $\{h(\mathbf{x}_0, \mathbf{b}_i)\}_{i=1}^{N_0}$ and $\{h(\mathbf{x}_1, \mathbf{b}_j)\}_{j=1}^{N_1}$ with the objectiveness score above a threshold. We adopt a dissimilarity measurement using the L_2 distance between two features:

$$s_{i,j} = \left\| \hat{h}(\mathbf{x}_0, \mathbf{b}_i) - \hat{h}(\mathbf{x}_1, \mathbf{b}_j) \right\|_2^2, \quad (1)$$

where $\hat{h}(\mathbf{x}_0, \mathbf{b}_i)$ and $\hat{h}(\mathbf{x}_1, \mathbf{b}_j)$ are encoded feature vectors obtained by a small network using the object features $h(\mathbf{x}, \mathbf{b})$ as input. In our experiments, the encoder consists of two convolutional layers with kernel size of 3×3 and an average pooling layer. The larger $s_{i,j}$ is, the more dissimilar the object features are. The dissimilarity measurement results are illustrated in Figure 3. The OOD objects in the reference frame, such as street lights and billboards, have a more significant dissimilarity.

Lastly, we perform a weighted average of the object features from frame \mathbf{x}_1 . Using multiple objects captures a diverse distribution of unknowns. The weights α are defined as the normalized exponential of the dissimilarity scores:

$$\hat{\mathbf{o}}^i = \sum_{j=1}^{N_1} \alpha_{i,j} h(\mathbf{x}_1, \mathbf{b}_j), \quad \alpha_{i,j} = \frac{e^{s_{i,j}}}{\sum_{k=1}^{N_1} e^{s_{i,k}}}, \quad (2)$$

where $\hat{\mathbf{o}}^i$ is the distilled unknown object (in the feature space), corresponding to the i -th object at frame \mathbf{x}_0 .

Temporal unknown distillation. Our spatial unknown distillation mechanism operates on a single reference frame, which can be extended to multiple video frames to capture additional diversity of unknowns in the *temporal* dimension. For example, consider a video of a car driving on the highway, the more frames we observe, the more unknown objects can be observed, such as trees, buildings, and rocks.

Given a frame \mathbf{x}_0 at timestamp t_0 , we propose distilling the unknown objects from multiple frames $\mathbf{x}_1, \dots, \mathbf{x}_T$. We randomly sample T frames within a range $[t_0 - R, t_0 + R]$. As a special case, $T = 1$ reduces to the previous pair-frame setting. To distill spatial-temporal unknown objects, we concatenate the object feature vectors from T frames, and then measure their dissimilarity *w.r.t* the objects in frame \mathbf{x}_0 by Equation (1). For the i -th object in frame \mathbf{x}_0 , the unknown counterpart is defined as follows:

$$\hat{\mathbf{o}}^i = \sum_{j=1}^N \alpha_{i,j} h(\mathbf{x}, \mathbf{b}_j), \quad \mathbf{x} \in \{\mathbf{x}_1, \dots, \mathbf{x}_T\}, \quad (3)$$

where $\alpha_{i,j}$ denotes the normalized dissimilarity scores defined in Equation 2. $N = \sum_{k=1}^T N_k$ is the total number of objects across T reference frames. The temporal aggregation mechanism allows searching through multiple frames for meaningful and diverse unknown discovery.

Later in Section 4.3, we provide comprehensive ablation studies on the frame sampling range R and the number of selected frames T , and show the benefits of temporal aggregation for improved OOD detection.

Unknown candidate object selection. A critical step in unknown distillation is to filter unknowns in the reference frame \mathbf{x}_1 that may be ID objects or simple background. Without selection, the model may be confused to separate the distilled unknown objects from the ID objects or quickly memorize the simple OOD pattern during training. To prevent this, we pre-filter the proposals based on the energy score, and then use the selected ones for the spatial-temporal unknown distillation. It is shown that the energy score is an effective indicator of OOD data in image classification [36]. To calculate the energy score for object detection network, we feed the object features $\{h(\mathbf{x}_1, \mathbf{b}_j)\}_{j=1}^{N_1}$ to the prediction head and follow the definition:

$$E(\mathbf{x}_1, \mathbf{b}_j) = -\log \sum_{k=1}^K \exp^{f_k(h(\mathbf{x}_1, \mathbf{b}_j); \mathbf{w}_{\text{pred}})}, \quad (4)$$

where $f_k(h(\mathbf{x}_1, \mathbf{b}_j); \mathbf{w}_{\text{pred}})$ is the logit output of the k -way classification branch. A higher energy indicates more OOD-ness and vice versa. Then, we select objects with

mild energy scores, *i.e.*, those in a specific percentile $p\% \leq \text{Rank}(E(\mathbf{x}_1, \mathbf{b}_j))/N_1 \leq q\%$ among all objects. In case of multiple frames $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$, the object selection is performed on each individual frame before temporal aggregation. Ablation study on the effect of the energy filtering and the selection percentile are provided in Section 4.3.

3.2. Unknown-Aware Training Objective

Leveraging the distilled unknown objects from Section 3.1, we now introduce our training objective for unknown-aware object detection. Our key idea is to perform object detection task while regularizing the model to produce a low uncertainty score for ID objects, and a high uncertainty score for the unknown ones. The overall objective function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \beta \cdot \mathcal{L}_{\text{uncertainty}}, \quad (5)$$

where β is the scaling weight when combining the detection loss \mathcal{L}_{det} and the uncertainty regularization loss $\mathcal{L}_{\text{uncertainty}}$. Next we describe the details of $\mathcal{L}_{\text{uncertainty}}$.

Uncertainty regularization. Following Du *et al.* [8], we employ a loss function that contrastively shapes the uncertainty surface, amplifying the separability between known ID objects and unknown OOD objects. To measure the uncertainty, we use the energy score in Equation (4), which is derived from the output of the classification branch. Here we calculate the energy score $E(\mathbf{x}, \mathbf{b})$ for the ID objects and the distilled unknown object features $E(\hat{\mathbf{o}})$. The uncertainty score is then passed into a logistic regression classifier with weight coefficient θ_u , which predicts high probability for ID object (\mathbf{x}, \mathbf{b}) and low probability for the unknown ones $\hat{\mathbf{o}}$. The regularization loss is calculated as:

$$\mathcal{L}_{\text{uncertainty}} = \mathbb{E}_{\hat{\mathbf{o}} \sim \mathcal{O}} \left[-\log \frac{1}{1 + \exp^{-\theta_u \cdot E(\hat{\mathbf{o}})}} \right] + \mathbb{E}_{(\mathbf{x}, \mathbf{b}) \sim \mathcal{D}} \left[-\log \frac{\exp^{-\theta_u \cdot E(\mathbf{x}, \mathbf{b})}}{1 + \exp^{-\theta_u \cdot E(\mathbf{x}, \mathbf{b})}} \right], \quad (6)$$

where \mathcal{O} contains all the unknown object features (*c.f.* Section 3.1). In Figure 4(a), we show the uncertainty regularization loss $\mathcal{L}_{\text{uncertainty}}$ over the course of training on Youtube-VIS dataset [66]. Upon convergence, Figure 4(b) shows the energy score distribution for both the ID and distilled unknown objects. This demonstrates that STUD converges properly and is able to separate the distilled unknown objects and the ID objects.

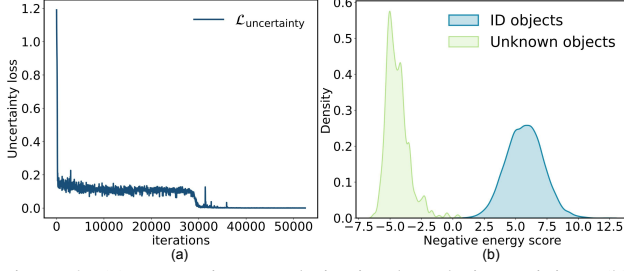


Figure 4. (a) Uncertainty regularization loss during training. (b) The negative energy score distribution for both the ID and the distilled unknown objects after training.

Algorithm 1 STUD: Spatial-Temporal Unknown Distillation for OOD detection

Input: ID data $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{b}_i, y_i)\}_{i=1}^M$, randomly initialized object detector with parameter θ , energy filtering percentile $[p\%, q\%]$, sampling range R , the number of reference frames T , and weight for uncertainty regularization β .

Output: Object detector with parameter θ^* , and OOD detector G .

while *train* **do**

1. Select unknown objects in the reference frames with mild energy scores as defined by Equation (4).
2. Calculate the dissimilarity (using Equation (2)) between an object in the key frame *w.r.t* selected objects in the reference frames.
3. Distill the unknown objects by Equation (3).
4. Calculate the uncertainty regularization loss by Equation (6), update the parameter θ based on the total loss in Equation (5).

end

while *eval* **do**

1. Calculate the uncertainty score by Equation (7).
2. Perform thresholding comparison by Equation (8).

end

Compared to \mathcal{L}_{det} for the vanilla object detector, our loss intends to facilitate learning a more conservative decision boundary between ID and OOD objects, which helps flag unseen OOD objects in testing. We proceed by describing the test-time OOD detection procedure.

Test-time OOD detection. During inference, we use the output of the logistic regression uncertainty branch for OOD detection. In particular, given a test input \mathbf{x}^* , the object detector produces a box prediction \mathbf{b}^* . The uncertainty score for the predicted object $(\mathbf{x}^*, \mathbf{b}^*)$ is given by:

$$p_{\theta}(g \mid \mathbf{x}^*, \mathbf{b}^*) = \frac{\exp^{-\theta_u \cdot E(\mathbf{x}^*, \mathbf{b}^*)}}{1 + \exp^{-\theta_u \cdot E(\mathbf{x}^*, \mathbf{b}^*)}}. \quad (7)$$

For OOD detection, we use the common thresholding mechanism to distinguish between ID and OOD objects:

$$G(\mathbf{x}^*, \mathbf{b}^*) = \begin{cases} 1 & \text{if } p_{\theta}(g \mid \mathbf{x}^*, \mathbf{b}^*) \geq \gamma, \\ 0 & \text{if } p_{\theta}(g \mid \mathbf{x}^*, \mathbf{b}^*) < \gamma. \end{cases} \quad (8)$$

The threshold γ is typically chosen so that a high fraction of ID data (e.g., 95%) is correctly classified. For objects that are classified as ID, one can obtain the bounding box and class prediction using the prediction head as usual. Our approach STUD is summarized in Algorithm 1.

Synergy between unknown distillation and contrastive regularization. The two key components of STUD—unknown distillation (Section 3.1) and contrastive regularization (Section 3.2) work collaboratively. First, a set of well distilled unknown objects may improve the energy-based contrastive regularization and help learn a more accurate decision boundary between known and unknown objects. Second, as the contrastive uncertainty loss amplifies an energy gap between known and unknown objects, the unknown distillation module can benefit from more accurate unknown object selection (via energy-based filtering). The entire training process converges when the two components perform satisfactorily. Our experiments in Section 4 further justify the efficacy of our framework.

4. Experiments

In this section, we provide empirical evidence to show the effectiveness of STUD on two large-scale video datasets (Section 4.1). We show that STUD outperforms other commonly used OOD detection baselines on detecting OOD data in Section 4.2. Ablation studies of STUD and qualitative analysis are presented in Sections 4.3 and 4.4.

4.1. Benchmark Construction

Datasets. We use two large-scale video datasets as ID data: BDD100K [67] and Youtube-Video Instance Segmentation (Youtube-VIS) 2021 [66]. For both tasks, we evaluate on two OOD datasets containing diverse visual categories: MS-COCO [34] and nuImages [1]. We perform careful deduplication to ensure there is no semantic overlap between the ID and OOD data. Extensive details on the datasets are described in the appendix.

Implementation details. We adopt Faster R-CNN [53] as the base object detector. We use Detectron2 library [11] and train with the backbone of ResNet-50 [15] and the default hyperparameters. We set the weight β for $\mathcal{L}_{\text{uncertainty}}$ to be 0.05 for BDD100K and 0.02 for Youtube-VIS dataset. For both datasets, we use $T = 3$ frames and set the sampling range $R = 9$. We set the energy filtering percentile to be 40% – 60% among all proposals. Ablation studies on different hyperparameters are detailed in Section 4.3.

Metrics. For evaluating the OOD detection performance, we report: (1) the false positive rate (FPR95) of OOD samples when the true positive rate of ID samples is at 95%; (2) the area under the receiver operating characteristic curve (AUROC). For evaluating the object detection performance on the ID task, we report the common metric of mAP.

In-distribution \mathcal{D}	Method	FPR95 ↓	AUROC ↑	mAP (ID) ↑	Cost (h)
OOD: MS-COCO / nuImages					
BDD100K	MSP [17]	90.11 / 93.98	66.32 / 59.21	31.0	9.1
	ODIN [33]	80.32 / 87.75	68.49 / 66.51	31.0	9.1
	Mahalanobis [31]	63.06 / 79.02	79.95 / 68.94	31.0	9.1
	Gram matrices [54]	68.78 / 82.60	66.13 / 71.56	31.0	9.1
	Energy score [36]	78.36 / 86.02	73.75 / 67.08	31.0	9.1
	Generalized ODIN [20]	75.99 / 92.15	78.63 / 67.23	30.9	10.5
	CSI [59]	69.38 / 80.06	80.85 / 72.59	29.8	15.3
	GAN-synthesis [30]	67.95 / 88.53	78.33 / 66.50	30.1	14.6
	STUD (ours)	52.18±2.2 / 77.57±3.0	85.67±0.6 / 75.67±0.7	30.5±0.2	10.1
Youtube-VIS	MSP [17]	90.17 / 94.52	70.26 / 54.59	24.8	9.2
	ODIN [33]	87.17 / 97.69	71.46 / 57.46	24.8	9.2
	Mahalanobis [31]	85.60 / 95.65	72.16 / 62.02	24.8	9.2
	Gram matrices [54]	88.68 / 93.20	61.96 / 60.04	24.8	9.2
	Energy score [36]	91.77 / 91.78	70.58 / 59.05	24.8	9.2
	Generalized ODIN [20]	83.90 / 93.18	71.33 / 62.16	24.3	10.5
	CSI [59]	80.21 / 84.85	73.89 / 68.84	23.3	15.7
	GAN-synthesis [30]	84.57 / 94.59	71.59 / 64.43	24.4	15.0
	STUD (ours)	79.82±0.2 / 76.93±0.4	75.55±0.3 / 71.48±0.6	24.5±0.3	10.2

Table 1. **Main results.** Comparison with competitive out-of-distribution detection methods. All baseline methods are based on a model trained on **ID data only** using ResNet-50 as the backbone. ↑ indicates larger values are better, and ↓ indicates smaller values are better. All values are percentages. **Bold** numbers are superior results. We report standard deviations estimated across three runs. The training time is reported in the “cost” column on four NVIDIA GeForce RTX 2080Ti GPUs.

4.2. Comparison with Baselines

STUD establishes SOTA performance. In Table 1, we compare STUD with competitive OOD detection methods in literature, where STUD significantly outperforms baselines on both datasets. For a fair comparison, all the methods use the same ID training data, trained with the same number of epochs. Our comprehensive baselines include Maximum Softmax Probability [17], ODIN [33], Mahalanobis distance [31], Generalized ODIN [20], energy score [36], Gram matrices [54], and a latest method CSI [59]. These baselines rely on the classification output or backbone feature, and therefore can be seamlessly evaluated on the object detection model.

The results show that STUD can outperform these baselines by a considerable margin because the majority of baselines rely on object detection models trained on ID data only, without being regularized by unknown objects. Such a training scheme is prone to produce overconfident predictions on OOD data (Figure 1) while STUD incorporates unknown objects to regularize the model more effectively.

We also compare with GAN-based approach for synthesizing outliers in the pixel space [30], where STUD effectively improves the OOD detection performance (FPR95) by **15.77%** on BDD100K (COCO as OOD) and **17.66%** on Youtube-VIS (nuImages as OOD). Moreover, we show in Table 1 that STUD achieves stronger OOD detection performance while preserving a high object detection accuracy on ID data (measured by mAP). This is in contrast with CSI, which displays significant degradation, with mAP decreasing by 1.2% on Youtube-VIS. *Details of reproducing baselines are in the Appendix Section D.*

Method	AUROC ↑	mAP ↑
COCO / nuImages as OOD		
◊Farthest object	83.04 / 71.38	30.2
◊Random object	79.61 / 70.42	30.3
◊Object with mild energy	83.60 / 71.24	30.3
◊Negative proposal [23]	80.94 / 72.92	30.0
♣GAN [30]	78.33 / 66.50	30.1
♣Mixup [70]	81.76 / 70.17	27.6
♢Gaussian noise	83.64 / 71.50	30.3
STUD (ours)	85.67 / 75.67	30.5

Table 2. Ablation on different unknown distillation approaches (on backbone of ResNet-50, COCO / nuImages are the OOD data).

4.3. Ablation Studies

This section provides comprehensive ablation studies to understand the efficacy of STUD. For consistency, all ablations are conducted on the BDD100K dataset, using ResNet-50 as the backbone. We refer readers to Appendix Section E for more ablations on using a different backbone architecture.

Ablation on different unknown distillation approaches.

We compare STUD with three types of unknown distillation approaches, *i.e.*, (I◊) using independent objects without spatial-temporal aggregation, (II♣) synthesizing unknowns in the pixel space, and (III♢) using noise as unknowns.

- For **type I**, we utilize objects from the reference frame without aggregating multiple objects across spatial and temporal dimensions—a key difference from STUD. The unknown objects can be constructed by: using the object in the reference frame that has the largest dissimilarity, using random objects, using the negative object as

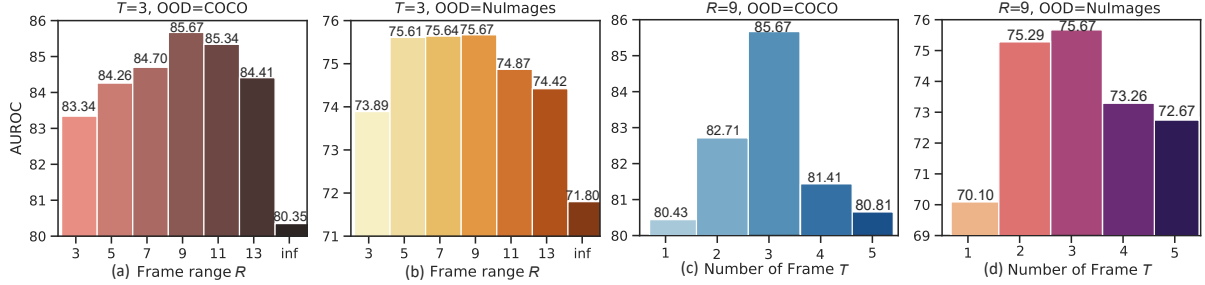


Figure 5. (a)-(b) Ablation study on the sampling range R . We vary the range from 3 to infinity. Metrics are AUROC. We set $T = 3$. (c)-(d) Ablation study on the number of reference frames T during unknown distillation. We fix the sampling range as $R = 9$.

Variants	FPR95 ↓	AUROC ↑	mAP ↑
	COCO / nuImages as OOD		
w/o unknown filtering	62.23 / 83.54	82.87 / 72.29	30.6
w/ ratio 0%-20%	61.41 / 82.33	83.66 / 74.86	30.2
w/ ratio 20%-40%	57.73 / 82.13	85.43 / 74.09	30.3
w/ ratio 40%-60%	52.18 / 77.57	85.67 / 75.67	30.5
w/ ratio 60%-80%	62.29 / 85.12	83.47 / 73.44	30.2
w/ ratio 80%-100%	65.86 / 88.47	82.46 / 72.50	30.3

Table 3. Ablation study on the energy filtering module. Here we set $T = 3$ and $R = 9$.

in [23], and using objects with mild energy scores (percentile 40% – 60%) in the reference frame.

- For **type II**, we consider GAN-based [30] and mixup-based [70] methods. For [30], the classification outputs of the objects in the synthesized images are forced to be closer to a uniform distribution. For mixup, we use a beta distribution of Beta(1), and interpolate ID objects in the pixel space for the reference frames.
- For **type III**, we add fixed Gaussian noise to the ID objects to create unknown object features.

The results are summarized in Table 2, where STUD outperforms alternative approaches. Exploiting objects without spatial-temporal distillation (\diamond) is less effective than STUD, because the generated unknowns either lack diversity (e.g., using object with the biggest dissimilarity or mild energy) or are too simple to effectively regularize the decision boundary between ID and OOD (e.g., using negative or random objects). Synthesizing unknowns in the pixel space (\clubsuit) is either unstable (GAN) or harmful for the object detection performance (mixup). Lastly, Gaussian noise as unknowns is relatively simple and does not outperform STUD.

Ablation on candidate object selection. Table 3 investigates the importance of filtering unknown objects based on the energy score. We contrast performance by either removing the filtering, or using different filtering percentile (c.f. Section 3.1). Using the objects with a mild energy score in the reference frames performs the best. This strategy distills unknown objects with a proper difficulty level, which is effective during contrastive uncertainty regularization.

Ablation on the frame sampling range R . Recall our spatial-temporal unknown distillation requires concatena-

β	FPR95 ↓	AUROC ↑	mAP ↑
	COCO / nuImages as OOD		
0.03	63.52 / 86.18	83.49 / 70.70	30.4
0.04	59.52 / 84.01	84.03 / 72.09	30.3
0.05	52.18 / 77.57	85.67 / 75.67	30.5
0.06	57.37 / 85.53	84.59 / 72.60	30.2
0.07	55.03 / 84.43	84.18 / 71.21	30.2

Table 4. Ablation study on the weight β for the uncertainty regularization loss. In this case, we set $T = 3$ and $R = 9$.

tion of objects from T reference frames. We ablate the effect of randomly selecting T frames within different temporal horizons w.r.t the key frame, modulated by the sampling range R . The results with varying R are shown in Figure 5 (a)-(b) with $T = 3$. We observe that OOD detection benefits from using the reference frames that are mildly close to the key frame. The trend is consistent for both COCO and nuImages OOD datasets. A larger sampling range translates into more dissimilar scenes, resulting in relatively easier unknowns to be distilled. When R becomes infinity, STUD randomly samples frames from the entire video, where the distilled unknowns are much less effective with AUROC significantly degrades (from 85.67% to 80.35% on COCO).

Ablation on the number of reference frames T . We contrast performance under different number of reference frames T and report the OOD detection results in Figure 5 (c)-(d). This ablation shows that STUD indeed benefits from aggregating objects from multiple frames across the temporal dimension. For example, the model trained on BDD100K with $T = 3$ achieves an AUROC improvement of 5.24% (COCO as OOD) compared to $T = 1$. This highlights the importance of temporal distillation with multiple frames. However, a larger T hurts the OOD detection performance. We hypothesize this is because many redundant object features are used during unknown distillation.

Ablation on the uncertainty regularization weight β . Table 4 reports the OOD detection results as we vary the weight β for $\mathcal{L}_{\text{uncertainty}}$. The model is evaluated on both COCO and nuImages datasets as OOD. The results suggest that a mild weight is desirable. In most cases, STUD outperforms the baseline OOD detection methods in Table 1 in terms of AUROC.

auxiliary information [50], such as class attributes, to perform object detection on unseen data—both differing from our focus of OOD detection. Wang *et al.* [62] adopted dissimilarity measurement in the cycle forward step, but their focus is OOD generalization (label space remains the same) rather than OOD detection. Additionally, it did not consider aggregating temporal information from multiple frames.

Video anomaly detection (VAD) aims to identify anomalous events on both the object level [7, 22, 68] and frame level [35, 39, 51] by techniques such as skeleton trajectory modeling [43], weakly supervised learning [69], attention [47], temporal pose graph [38], self-supervised learning [10] and autoencoders [3]. Compared with STUD, the anomalies in VAD do not necessarily have different semantics from the ID training data. Moreover, none of the approaches considered synthesizing unknowns with the help of videos or energy-based model regularization.

6. Conclusion

In this paper, we propose STUD, an unknown-aware object detection framework for OOD detection. STUD distills diverse unknown objects during training by exploiting the rich spatial-temporal information from videos. The distilled unknowns meaningfully improve the decision boundary between the ID and OOD data, resulting in state-of-the-art OOD detection performance while preserving the performance of the ID task. We hope our work will inspire future research towards unknown-aware deep learning in real-world settings.

7. Acknowledgement

Research is supported by Wisconsin Alumni Research Foundation (WARF), Facebook Research Award, and funding from Google Research.

References

- [1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nusenes: A multi-modal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 11618–11628, 2020. 5
- [2] Tianshi Cao, Chinwei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *CoRR*, abs/2007.04250, 2020. 1
- [3] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. Clustering driven deep autoencoder for video anomaly detection. In *European Conference on Computer Vision, ECCV 2020*, pages 329–345, 2020. 9
- [4] Kumari Deepshikha, Sai Harsha Yelleni, P. K. Srijith, and C. Krishna Mohan. Monte carlo dropblock for modelling uncertainty in object detection. *CoRR*, abs/2108.03614, 2021. 8
- [5] Akshay Raj Dhamija, Manuel Günther, and Terrance E. Boult. Reducing network agnostophobia. In *Advances in Neural Information Processing Systems 31, NeurIPS 2018*, pages 9175–9186, 2018. 8
- [6] Akshay Raj Dhamija, Manuel Günther, Jonathan Ventura, and Terrance E. Boult. The overlooked elephant of object detection: Open set. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, pages 1010–1019, 2020. 8
- [7] Keval Doshi and Yasin Yilmaz. Any-shot sequential anomaly detection in surveillance videos. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020*, pages 4037–4042, 2020. 9
- [8] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don’t know by virtual outlier synthesis. *Proceedings of the International Conference on Learning Representations*, 2022. 1, 4, 8
- [9] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016. 8
- [10] Mariana-Iuliana Georgescu, Antonio Barbalau, Radu Tudor Ionescu, Fahad Shahbaz Khan, Marius Popescu, and Mubarak Shah. Anomaly detection in video via self-supervised and multi-task learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, pages 12742–12752, 2021. 9
- [11] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. <https://github.com/facebookresearch/detectron>, 2018. 5
- [12] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Zero-shot detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 8
- [13] David Hall, Feras Dayoub, John Skinner, Haoyang Zhang, Dimity Miller, Peter Corke, Gustavo Carneiro, Anelia Angelova, and Niko Sünderhauf. Probabilistic object detection: Definition and evaluation. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, pages 1020–1029, 2020. 8
- [14] Ali Harakeh and Steven L. Waslander. Estimating and evaluating regression predictive uncertainty in deep object detectors. In *International Conference on Learning Representations*, 2021. 8
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pages 770–778, 2016. 5
- [16] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019. 8
- [17] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations, ICLR 2017*, 2017. 1, 6, 8, 12, 13

- [18] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. 3, 8
- [19] Tobias Hinz, Stefan Heinrich, and Stefan Wermter. Generating multiple objects at spatially distinct locations. In *International Conference on Learning Representations, ICLR 2019*, 2019. 8
- [20] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020. 1, 6, 8, 12, 13
- [21] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Advances in Neural Information Processing Systems*, 2021. 8
- [22] Radu Tudor Ionescu, Fahad Shahbaz Khan, Mariana-Iuliana Georgescu, and Ling Shao. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 7842–7851, 2019. 9
- [23] K. J. Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N. Balasubramanian. Towards open world object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2021*, 2021. 2, 6, 7, 8
- [24] K. J. Joseph, Jathushan Rajasegaran, Salman H. Khan, Fahad Shahbaz Khan, Vineeth Balasubramanian, and Ling Shao. Incremental object detection via meta-learning. *arXiv preprint arXiv:2003.08798*, 2020. 8
- [25] Dahun Kim, Tsung-Yi Lin, Anelia Angelova, In So Kweon, and Weicheng Kuo. Learning open-world object proposals without learning to classify. *CoRR*, abs/2108.06753, 2021. 8
- [26] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015*, 2015. 12
- [27] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems 2018, NeurIPS 2018*, pages 10236–10245, 2018. 8
- [28] Polina Kirichenko, Pavel Izmailov, and Andrew G Wilson. Why normalizing flows fail to detect out-of-distribution data. *Advances in Neural Information Processing Systems*, 33, 2020. 8
- [29] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017. 1
- [30] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018. 2, 6, 7, 8, 12, 13
- [31] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018. 1, 6, 8, 12, 13
- [32] Yi Li and Nuno Vasconcelos. Background data resampling for outlier-aware classification. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 13215–13224, 2020. 8
- [33] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations, ICLR 2018*, 2018. 1, 6, 8, 12, 13
- [34] 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 *European Conference on Computer Vision*, pages 740–755, 2014. 5
- [35] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection - A new baseline. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 6536–6545, 2018. 9
- [36] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020. 1, 4, 6, 8, 12, 13
- [37] Xialei Liu, Hao Yang, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Multi-task incremental learning for object detection. *CoRR*, abs/2002.05347, 2020. 8
- [38] Amir Markovitz, Gilad Sharir, Itamar Friedman, Lihi Zelnik-Manor, and Shai Avidan. Graph embedded pose clustering for anomaly detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 10536–10544, 2020. 9
- [39] Ramin Mehran, Alexis Oyama, and Mubarak Shah. Abnormal crowd behavior detection using social force model. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pages 935–942, 2009. 9
- [40] Dimity Miller, Feras Dayoub, Michael Milford, and Niko Sünderhauf. Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *International Conference on Robotics and Automation, ICRA 2019*, pages 2348–2354, 2019. 8
- [41] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *2018 IEEE International Conference on Robotics and Automation, ICRA 2018*, pages 1–7, 2018. 8
- [42] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5216–5223, 2020. 8
- [43] Romero Moraes, Vuong Le, Truyen Tran, Budhaditya Saha, Moussa Reda Mansour, and Svetha Venkatesh. Learning regularity in skeleton trajectories for anomaly detection in videos. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 11996–12004, 2019. 9
- [44] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022. 8
- [45] Eric T. Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Görür, and Balaji Lakshminarayanan. Do deep generative

- models know what they don't know? In *International Conference on Learning Representations, ICLR 2019*, 2019. 8
- [46] Donald F. Othmer. Symposium on distillation separation of water from acetic acid by azeotropic distillation. *Industrial & Engineering Chemistry*, 27(3):250–255, 1935. 2
- [47] Hyunjong Park, Jongyoun Noh, and Bumsub Ham. Learning memory-guided normality for anomaly detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 14360–14369, 2020. 9
- [48] Juan-Manuel Pérez-Rúa, Xiatian Zhu, Timothy M. Hospedales, and Tao Xiang. Incremental few-shot object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 13843–13852, 2020. 8
- [49] Ilija Radosavovic, Raj Prateek Kosaraju, Ross B. Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 10425–10433, 2020. 12, 13
- [50] Shafin Rahman, Salman H. Khan, and Fatih Porikli. Zero-shot object detection: Joint recognition and localization of novel concepts. *International Journal of Computer Vision*, 128(12):2979–2999, 2020. 8, 9
- [51] Mahdyar Ravanbakhsh, Moin Nabi, Hossein Mousavi, Enver Sangineto, and Nicu Sebe. Plug-and-play CNN for crowd motion analysis: An application in abnormal event detection. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018*, pages 1689–1698, 2018. 9
- [52] Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems 2019, NeurIPS 2019*, pages 14680–14691, 2019. 8
- [53] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015. 5
- [54] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, pages 8491–8501, 2020. 6, 8, 12, 13
- [55] Robin Schirrmeister, Yuxuan Zhou, Tonio Ball, and Dan Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. In *Advances in Neural Information Processing Systems 33, NeurIPS 2020*, 2020. 8
- [56] Vikash Sehwal, Mung Chiang, and Prateek Mittal. SSD: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021. 8
- [57] Joan Serra, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. In *International Conference on Learning Representations, ICLR 2020*, 2020. 8
- [58] Yiyu Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, 2021. 8
- [59] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *Advances in Neural Information Processing Systems*, 2020. 1, 6, 8, 12, 13
- [60] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*, pages 4790–4798, 2016. 8
- [61] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don't know? *Advances in Neural Information Processing Systems*, 2021. 8
- [62] Xin Wang, Thomas E. Huang, Benlin Liu, Fisher Yu, Xiaolong Wang, Joseph E. Gonzalez, and Trevor Darrell. Robust object detection via instance-level temporal cycle confusion. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV*, 2021. 9
- [63] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in Neural Information Processing Systems*, 33, 2020. 8
- [64] Jingkan Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. *CoRR*, abs/2108.11941, 2021. 8
- [65] Jingkan Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 8
- [66] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019*, pages 5187–5196, 2019. 2, 4, 5
- [67] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*, pages 2633–2642, 2020. 1, 2, 5
- [68] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *28th ACM International Conference on Multimedia*, pages 583–591, 2020. 9
- [69] Muhammad Zaigham Zaheer, Arif Mahmood, Marcella Astrid, and Seung-Ik Lee. CLAWS: clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In *16th European Conference on Computer Vision, ECCV 2020*, pages 358–376, 2020. 9
- [70] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018*, 2018. 6, 7
- [71] Jingyang Zhang, Nathan Inkawhich, Yiran Chen, and Hai Li. Fine-grained out-of-distribution detection with mixup outlier exposure. *CoRR*, abs/2106.03917, 2021. 8

Supplementary Material

A. Experimental details

We summarize the OOD detection evaluation task in Table 6. The OOD test dataset is selected from MS-COCO and nuImages dataset, which contains disjoint labels from the respective ID dataset. For the Youtube-VIS dataset, we use the dataset released in year 2021. Since there are no ground truth labels available for the validation images, we select the last 597 videos in the training set as the in-distribution evaluation dataset. The remaining 2,388 videos are used for training. The BDD100K and Youtube-VIS model are both trained for a total of 52,500 iterations. See detailed ablations on the hyperparameters in Section 4.3 of the main paper.

	Task 1	Task 2
ID train dataset	BDD100K train	Youtube-VIS train
ID val dataset	BDD100K val	Youtube-VIS val
OOD dataset	COCO / nuImages	COCO / nuImages
#ID train images	273,406	67,861
#ID val images	39,973	21,889
#OOD images from COCO	1,914	28,922
#OOD images from nuImages	2,100	2,100

Table 6. OOD detection evaluation tasks.

B. In-distribution classes

We provide a detailed description of the in-distribution classes for the two video datasets as follows.

BDD100K dataset contains 8 classes, which are *pedestrian, rider, car, truck, bus, train, motorcycle, bicycle*.

The Youtube-VIS dataset contains 40 classes, which are *airplane, bear, bird, boat, car, cat, cow, deer, dog, duck, earless_seal, elephant, fish, flying_disc, fox, frog, giant_panda, giraffe, horse, leopard, lizard, monkey, motorbike, mouse, parrot, person, rabbit, shark, skateboard, snake, snowboard, squirrel, surfboard, tennis_racket, tiger, train, truck, turtle, whale, zebra*.

C. Software and hardware

We run all experiments with Python 3.8.5 and PyTorch 1.7.0, using NVIDIA GeForce RTX 2080Ti GPUs.

D. Baselines

To evaluate the baselines, we follow the original methods in MSP [17], ODIN [33], Generalized ODIN [20], Mahalanobis distance [31], CSI [59], energy score [36] and gram matrices [54] and apply them accordingly on the classification branch of the object detectors. For ODIN [33], the temperature is set to be $T = 1000$ following the original work. For both ODIN and Mahalanobis distance [31], the noise magnitude is set to 0 because the region-based object detector is not end-to-end differentiable given the existence of region cropping and ROIAlign. For GAN [30], we follow the original paper and use a GAN to generate OOD images. The prediction of the OOD images/objects is regularized to be close to a uniform distribution, through a KL divergence loss with a weight of 0.05. We set the shape of the generated images to be 100×100 and resize them to have the same shape as the real images. We optimize the generator and discriminator using the Adam optimizer [26], with a learning rate of 0.001. For CSI [59], we use the rotations (0° , 90° , 180° , 270°) as the self-supervision task. We set the temperature in the contrastive loss to 0.5. We use the features right before the classification branch (with the dimension to be 1024) to perform contrastive learning. The weights of the losses that are used for classifying shifted instances and instance discrimination are both set to 0.1 to prevent training collapse. For Generalized ODIN [20], we replace and train the classification head of the object detector by the most effective Deconf-C head shown in the original paper.

E. Ablation study on a different backbone architecture

In this section, we evaluate the proposed STUD using a different backbone architecture of the Faster-RCNN, which is RegNetX-4.0GF [49]. Similarly, we compare with the same set of OOD detection baselines as stated in the main paper. The

In-distribution \mathcal{D}	Method	FPR95 ↓	AUROC ↑	mAP (ID) ↑
OOD: MS-COCO / nuImages				
BDD100K	MSP [17]	80.09 / 93.05	74.19 / 63.14	32.0
	ODIN [33]	64.74 / 82.08	77.65 / 67.09	32.0
	Mahalanobis [31]	54.02 / 79.85	82.38 / 75.48	32.0
	Gram matrices [54]	63.96 / 63.61	67.56 / 67.47	32.0
	Energy score [36]	64.79 / 81.62	78.78 / 69.43	32.0
	Generalized ODIN [20]	60.76 / 82.00	80.14 / 70.74	32.5
	CSI [59]	52.98 / 80.00	83.57 / 74.91	31.8
	GAN-synthesis [30]	58.35 / 83.65	81.43 / 70.39	31.5
	STUD (ours)	52.51 / 79.75	84.03 / 76.55	32.3
Youtube-VIS	MSP [17]	89.86 / 97.42	67.04 / 54.02	26.7
	ODIN [33]	89.28 / 96.30	67.54 / 60.82	26.7
	Mahalanobis [31]	90.00 / 94.44	70.47 / 54.83	26.7
	Gram matrices [54]	87.64 / 91.25	69.76 / 61.43	26.7
	Energy score [36]	88.54 / 90.21	67.83 / 58.02	26.7
	Generalized ODIN [20]	85.15 / 98.00	71.57 / 64.23	27.3
	CSI [59]	82.43 / 88.61	71.81 / 54.00	24.2
	GAN-synthesis [30]	85.75 / 93.75	72.95 / 56.94	25.5
	STUD (ours)	81.14 / 80.77	74.82 / 69.52	27.2

Table 7. Comparison with competitive out-of-distribution detection methods. All baseline methods are based on a model trained on ID data only using RegNetX-4.0GF as the backbone. ↑ indicates larger values are better, and ↓ indicates smaller values are better. All values are percentages. **Bold** numbers are superior results.

results are shown in Table 7.

From Table 7, we demonstrate that STUD is effective on alternative neural network architectures. In particular, using RegNet [49] as backbone yields better OOD detection performance compared with the baselines. Moreover, we show that STUD achieves stronger OOD detection performance while preserving or even slightly increasing the object detection accuracy on ID data (measured by mAP). This is in contrast with CSI, which displays significant degradation, with mAP decreasing by 3% on Youtube-VIS.

F. Additional visualization examples

We provide additional visualization of the detected objects on different OOD datasets with models trained on different in-distribution datasets. The results are shown in Figures 7-10.



Figure 7. Additional visualization of detected objects on the OOD images (from MS-COCO) by a vanilla Faster-RCNN (*top*) and STUD (*bottom*). The in-distribution is BDD100K dataset. **Blue**: Objects detected and classified as one of the ID classes. **Green**: OOD objects detected by STUD, which reduce false positives among detected objects.



Figure 8. Additional visualization of detected objects on the OOD images (from nuImages) by a vanilla Faster-RCNN (*top*) and STUD (*bottom*). The in-distribution is BDD100K dataset. **Blue**: Objects detected and classified as one of the ID classes. **Green**: OOD objects detected by STUD, which reduce false positives among detected objects.



Figure 9. Additional visualization of detected objects on the OOD images (from MS-COCO) by a vanilla Faster-RCNN (*top*) and STUD (*bottom*). The in-distribution is Youtube-VIS dataset. **Blue**: Objects detected and classified as one of the ID classes. **Green**: OOD objects detected by STUD, which reduce false positives among detected objects.



Figure 10. Additional visualization of detected objects on the OOD images (from nuImages) by a vanilla Faster-RCNN (*top*) and STUD (*bottom*). The in-distribution is Youtube-VIS dataset. **Blue**: Objects detected and classified as one of the ID classes. **Green**: OOD objects detected by STUD, which reduce false positives among detected objects.