Supplementary Material

A. Glossary

$\mathcal{X}$	The training dataset with $\mathcal{X} := (x_1, x_2, ..., x_N)$ containing N images or videos
$\mathcal{Y}$	The ground-truth labels with $\mathcal{Y} := \{1, 2, ..., K\}$ for $K + 1$ inlier object classes
$\mathcal{B}$	The ground-truth coordinates of the bounding boxes with $\boldsymbol{b} \in \mathcal{B}$
$l(x, \hat{\boldsymbol{b}})$	The fixed-size box features of both inlier and background patches given the predicted bounding boxes $\hat{\boldsymbol{b}}$
$l_{ID}(x, \hat{\boldsymbol{b}})$	The fixed-size box features of only inlier patches given the predicted bounding boxes $\hat{\boldsymbol{b}}$
$\mathcal{L}_{det}$	The standard object detection loss consisting of object classification and bounding-box regression losses
$\mathcal{L}_{nll}$	The negative log-likelihood loss to train the normalizing flow model on inlier features
$\mathcal{L}_{reg}$	The regularization loss for discriminative training using inlier and synthetic outlier features
ξ	The energy-based threshold, when 95% of inlier objects in the validation set are correctly detected
h	The classification head in the ROI head module of the Faster R-CNN architecture
f	The normalizing flow network to maximize the likelihood of inlier features
$\mathcal{Z}$	The latent space of the flow network f defined as a multivariate Gaussian with zero mean and unit variance
Φ	The binary classifier to differentiate the inlier from the synthesized outlier features via discriminative training
θ	The learnable parameters of the standard object detector, i.e. Faster R-CNN
γ	The learnable parameters of the normalizing flow f
ψ	The learnable parameters of the binary classifier Φ
g_k	The generated synthetic features after randomly sampling k samples from flow's latent space
o_s	The selected synthetic outlier features from the generated features g_k such that $o_s \subset g_k$
τ	The step size of the gradient descent for the projection sampling based outlier synthesis
δ	The log-likelihood threshold to determine the synthesized outlier features based on projection sampling
$E(h(.))$	The energy-score calculated from the output of the classification head h
T	The temperature coefficient to compute the energy score $E(h(.))$
α	The weightage of the regularization loss $\mathcal{L}_{reg}$ in the overall training objective
β	The weightage of the negative log-likelihood loss $\mathcal{L}_{nll}$ in the overall training objective

B. Visualization of inlier and synthesized outlier features

We provide the visualization of inlier features (in color) of PASCAL-VOC along with the synthesized outliers (in black) after reducing the number of feature embeddings using Principal Component Analysis (PCA). We compare the results from VOS [10] and our FFS approach in Figure 6. It is noticeable that VOS synthesizes outliers separately for each inlier class. In contrast, our approach synthesizes outliers after estimating the accurate data distribution of all inlier classes using the normalizing flow model, thereby leading to more effective regularization.

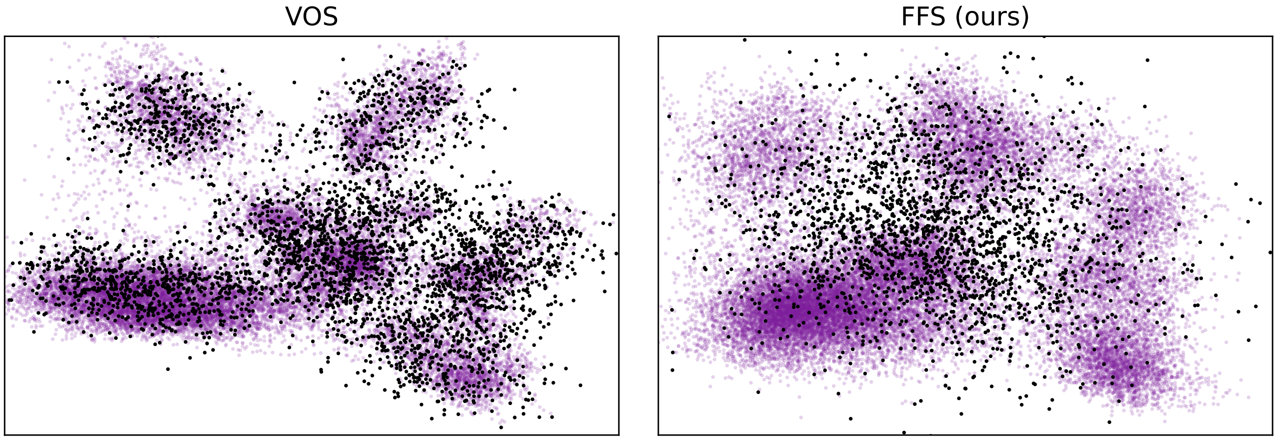

Figure 6. PCA visualization of synthesized outliers (in black) and inlier features (in color) of all 20 PASCAL-VOC object classes.

C. Step size τ for projection sampling

In Table 7, we show the OD detection results of FFS approach for varying step size τ when projection sampling was used to synthesize outliers precisely at the decision boundary. For this experiment, we fixed PASCAL-VOC as the inlier and MS-COCO as the outlier dataset, respectively. We analyzed the number s of synthetic outliers, o_s, and τ to evaluate the performance. It can be seen from the results that, irrespective of the step size τ, the OD detection performance improves as we increase the number s of synthesized outliers o_s. However, the performance gets worse when the s is increased further.

# samples, s	Step size, τ		
	τ = 1	τ = 0.5	τ = 0.1
	FPR95 ↓ / AUROC ↑		
1	55.56 / 85.90	51.13 / 86.73	52.34 / 86.21
2	50.56 / 87.39	49.48 / 87.95	49.94 / 86.92
4	**48.93** / 88.83	**47.23** / 89.44	**47.49** / **88.72**
8	49.79 / **88.96**	48.28 / **89.61**	48.95 / 88.84
16	59.02 / 81.47	59.60 / 82.09	59.32 / 83.54
32	60.44 / 82.38	57.92 / 81.99	58.54 / 82.65
64	57.72 / 82.43	58.67 / 82.22	59.91 / 83.43

Table 7. OD detection results after varying the step size τ of our projection sampling based strategy to synthesize outliers.

D. Datasets

Our experimental setup consists of three inlier and outlier datasets, where we train FFS on both image and video-based inlier datasets. The image-based inlier dataset, i.e., PASCAL-VOC, does not contain the sequence of image frames in terms of time. In contrast, the video-based datasets, BDD100K and Youtube-VIS, are a sequence of image frames dependent on time. We follow the experimental setup of VOS [10] and STUD [9] and utilize the datasets provided for training and inference of FFS. In addition, we show the object classes for each of the inlier datasets for a better interpretation of the subsequent visual results:

- ***PASCAL-VOC:*** There are 16,551 training and 4,952 validation images in the dataset. The object classes are *person, bird, cat, cow, dog, horse, sheep, airplane, bicycle, boat, bus, car, motorcycle, train, bottle, chair, dining table, potted plant, couch and tv.*
- ***BDD100K:*** There are 273,406 training and 39,973 validation images in the dataset. The object classes are *pedestrian, rider, car, truck, bus, train, motorcycle and bicycle.*
- ***Youtube-VIS:*** There are 67,861 training and 21,889 validation images in the dataset. The object classes 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 and zebra.*

We evaluate our trained FFS framework on three different outlier datasets, namely MS-COCO, OpenImages, and nuImages. There are 930 images in MS-COCO and 1,761 images in OpenImages datasets, so objects in these images do not fall into any of the inlier classes of PASCAL-VOC. Similarly, there are 2,100 images of the nuImages dataset for evaluating FFS trained on BDD100K and 28,922 images of MS-COCO for evaluating FFS trained on Youtube-VIS. In all outlier datasets, the objects present in the inlier dataset are mutually independent of objects in the outlier dataset.

E. Visualization of OD detection results for video datasets

In Figure 7 and Figure 8, we show the OD detection results when FFS was trained on video datasets. The results showcase that we obtain significantly better results in detecting outliers and reducing the number of incorrect bounding box predictions compared to STUD [9]. Additionally, for some image pairs, we reduce the model's confidence when the object was wrongly detected as an inlier by our approach and STUD [9].

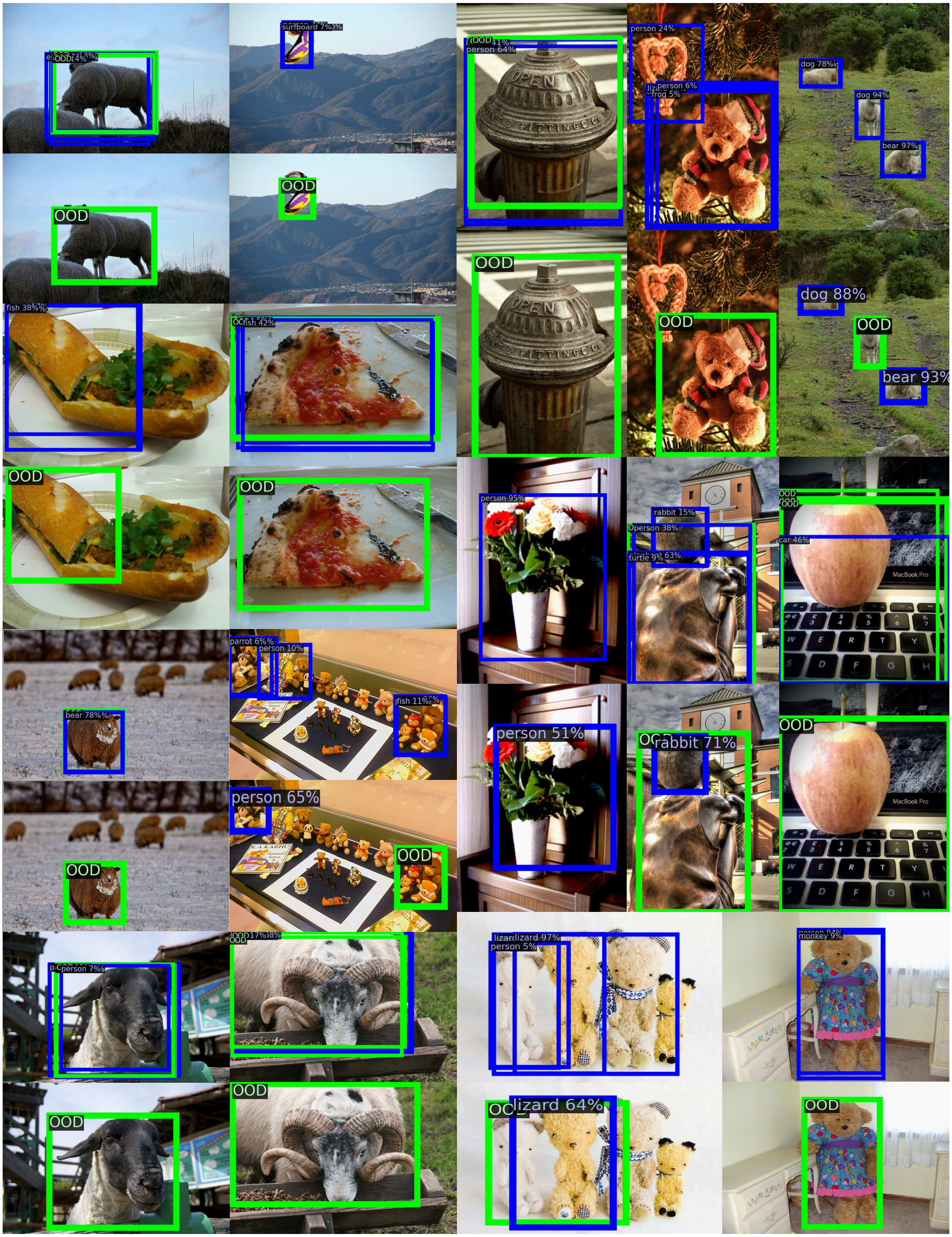

Figure 7. OD detection results on MS-COCO images when FFS was trained on Youtube-VIS dataset. In each image pair, the top image is the results from STUD [9], and the bottom image is the results from FFS.

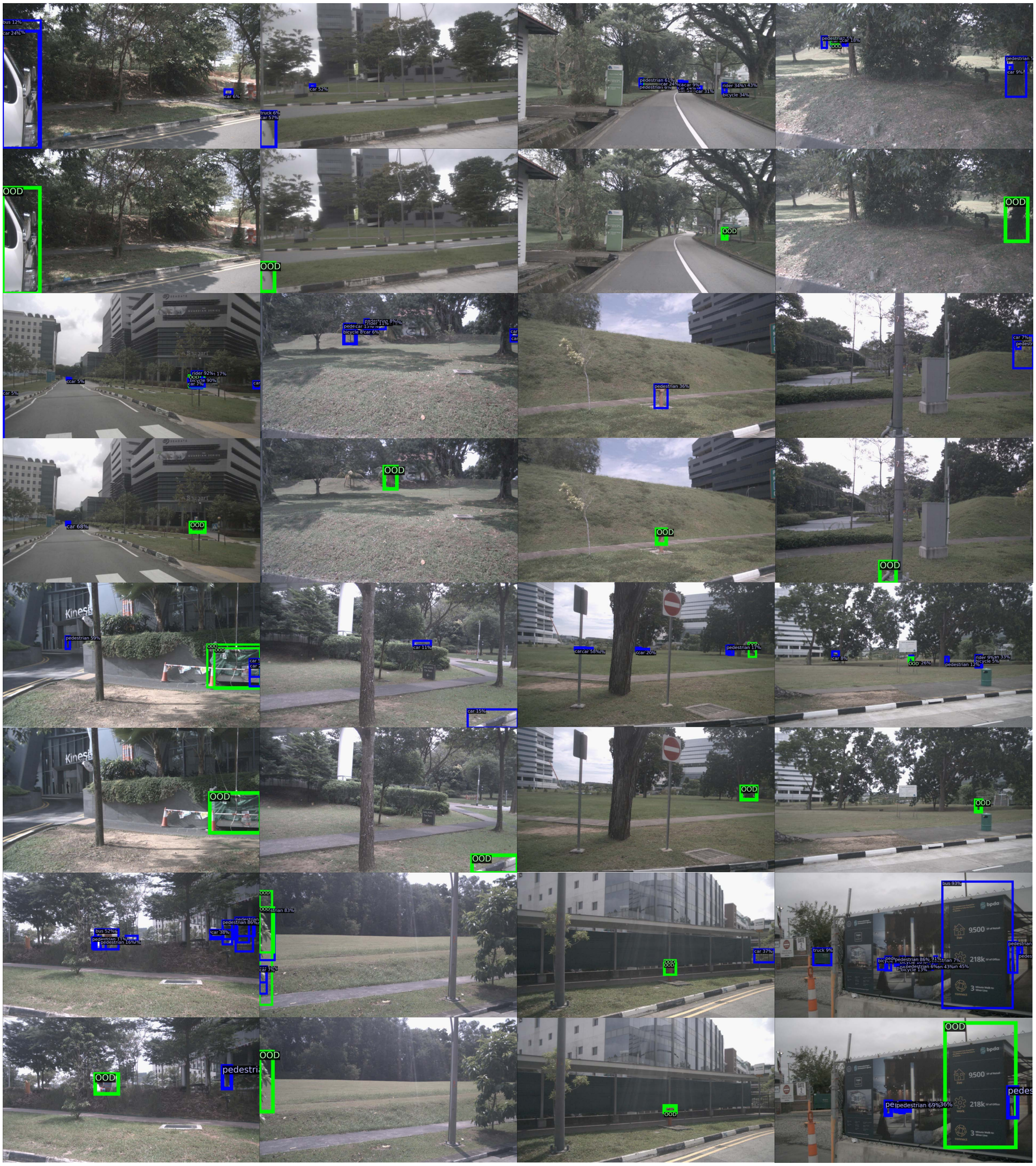

Figure 8. OD detection results on nuImages when FFS was trained on BDD100K dataset. In each image pair, the top image is the results from STUD [9], and the bottom image is the results from FFS.