Extremely Simple Multimodal Outlier Synthesis for Out-of-Distribution Detection and Segmentation

Moru Liu[1*] **Hao Dong**[2*] **Jessica Kelly**[3] **Olga Fink**[4] **Mario Trapp**[1,3]
[1]Technical University of Munich [2]ETH Zürich [3]Fraunhofer IKS [4]EPFL

Abstract

Out-of-distribution (OOD) detection and segmentation are crucial for deploying machine learning models in safety-critical applications such as autonomous driving and robot-assisted surgery. While prior research has primarily focused on unimodal image data, real-world applications are inherently multimodal, requiring the integration of multiple modalities for improved OOD detection. A key challenge is the lack of supervision signals from unknown data, leading to overconfident predictions on OOD samples. To address this challenge, we propose **Feature Mixing**, an extremely simple and fast method for multimodal outlier synthesis with theoretical support, which can be further optimized to help the model better distinguish between in-distribution (ID) and OOD data. Feature Mixing is modality-agnostic and applicable to various modality combinations. Additionally, we introduce CARLA-OOD, a novel multimodal dataset for OOD segmentation, featuring synthetic OOD objects across diverse scenes and weather conditions. Extensive experiments on SemanticKITTI, nuScenes, CARLA-OOD datasets, and the MultiOOD benchmark demonstrate that Feature Mixing achieves state-of-the-art performance with a $\mathbf{10}\times$ to $\mathbf{370}\times$ speedup. Our source code and dataset will be available at https://github.com/mona4399/FeatureMixing.

1 Introduction

Classification and segmentation are fundamental computer vision tasks that have seen significant advancements with deep neural networks [22, 40]. However, most models operate under a closed-set assumption, expecting identical class distributions in training and testing. In real-world applications, this assumption often fails, as out-of-distribution (OOD) objects frequently appear. Ignoring OOD instances poses critical safety risks in domains like autonomous driving and robot-assisted surgery, motivating research on OOD detection [36] and segmentation [7] to identify *unknown* objects that are unseen during training.

Most existing OOD detection and segmentation methods focus on unimodal inputs, such as images [36] or point clouds [7], despite the inherently multimodal nature of real-world applications. Leveraging multiple modalities can provide complementary information to improve performance [15, 14]. Recent work by Dong et al. [17] introduced the first multimodal OOD detection benchmark and framework and also extended the framework to the multimodal OOD segmentation task. A key challenge in OOD detection and segmentation is the tendency of neural networks to assign high confidence scores to OOD inputs [44] due to the lack of explicit supervision for unknowns during training. While real outlier datasets [25] can help mitigate this, they are often costly and impractical to obtain. Alternatively, synthetic outliers [19, 48, 43] have been proven effective for regularization, but existing methods are designed for unimodal scenarios and struggle in multimodal settings [17]. Dong

[*]Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

et al. [17] proposed a multimodal outlier synthesis technique using nearest-neighbor information, but its computational cost remains prohibitive for segmentation tasks.

To address this, we propose ***Feature Mixing***, an extremely simple and efficient multimodal outlier synthesis method with theoretical support. Given in-distribution (ID) features from two modalities, Feature Mixing randomly swaps a subset of N feature dimensions between them to generate new multimodal outliers. By maximizing the entropy of these outliers during training, our method effectively reduces overconfidence and enhances the model's ability to distinguish OOD from ID samples. Feature Mixing is modality-agnostic and applicable to various modality combinations, such as images and point clouds or video and optical flow. Moreover, its lightweight design enables a $10\times$ speedup for multimodal OOD detection and a $370\times$ speedup for segmentation compared to [17] (Fig. 1).

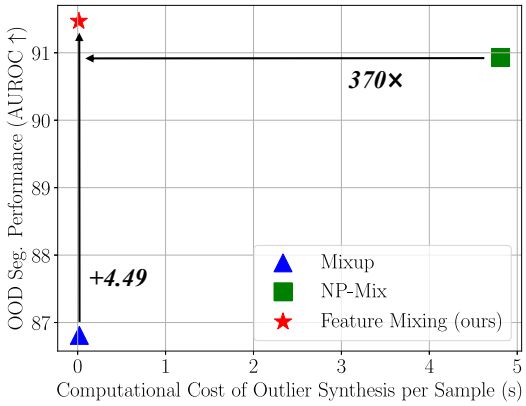


Figure 1: Mixup [52] is efficient for outlier synthesis but performs poorly in OOD segmentation. In contrast, NP-Mix [17] achieves strong OOD segmentation but is computationally expensive. Our Feature Mixing combines both speed and performance, benefiting from its simple yet effective design. Results are on SemanticKITTI dataset.

We conduct extensive evaluations across *eight datasets* and *four modalities* to validate the effectiveness of Feature Mixing. For multimodal OOD detection, we use five datasets from the MultiOOD benchmark [17] with video and optical flow modalities. For multimodal OOD segmentation, we evaluate on large-scale real-world datasets, including SemanticKITTI [3] and nuScenes [6], with image and point cloud modalities. To address the lack of multimodal OOD segmentation datasets, we introduce ***CARLA-OOD***, a synthetic dataset generated using CARLA simulator [18], featuring diverse OOD objects in various challenging scenes and weather conditions (Fig. 6). Our experiments on both synthetic and real-world datasets demonstrate that Feature Mixing outperforms existing outlier synthesis methods in most cases with a significant speedup. In summary, the main contributions of this paper are:

1. We introduce Feature Mixing, an extremely simple and fast method for multimodal outlier synthesis, applicable to diverse modality combinations.
2. We provide theoretical insights in support of the efficacy of Feature Mixing.
3. We present the challenging CARLA-OOD dataset with diverse scenes and weather conditions, addressing the scarcity of multimodal OOD segmentation datasets.
4. We conduct extensive experiments across eight datasets and four modalities to demonstrate the effectiveness of our proposed approach.

2 Related Work

2.1 Out-of-Distribution Detection

OOD detection aims to detect test samples with semantic shift without losing the ID classification accuracy. Numerous OOD detection algorithms have been developed. Post hoc methods [24, 23, 36] aim to design OOD scores based on the classification output of neural networks, offering the advantage of being easy to use without modifying the training procedure and objective. Methods like Mahalanobis [32] and k-nearest neighbor [47] use distance metrics in feature space for OOD detection, while virtual-logit matching [50] integrates information from both feature and logit spaces to define the OOD score. Additionally, some approaches propose to synthesize outliers [19, 48] or normalize logits [51] to address prediction overconfidence by training-time regularization. However, all these approaches are designed for unimodal scenarios without accounting for the complementary nature of multiple modalities.

2.2 Out-of-Distribution Segmentation

OOD segmentation focuses on pixel- or point-level segmentation of OOD objects and has been widely studied in medical images [1], industrial inspection [46], and autonomous driving [5] in recent years. Approaches to pixel-level segmentation are generally categorized into uncertainty-based [27, 49], outlier exposure [38, 31], and reconstruction-based methods [2, 42]. Point-level segmentation has gained attention in recent years due to its practical applications in real-world environments. For example, Cen et al. [7] address OOD segmentation on LiDAR point cloud and train redundancy classifiers to segment unknown object points by simulating outliers through the random resizing of known classes. Li et al. [33] separate ID and OOD features using a prototype-based clustering approach and employing a generative adversarial network [21] to synthesize outlier features. Similarly, these techniques are exclusively focused on unimodal contexts, neglecting the inherent complementarity among different modalities.

2.3 Multimodal OOD Detection and Segmentation

Multimodal OOD detection and segmentation are emerging research areas with limited prior work. Dong et al. [17] introduced the first multimodal OOD detection benchmark, identifying modality prediction discrepancy as a key indicator of OOD performance. They proposed the agree-to-disagree algorithm to amplify this discrepancy during training and developed a multimodal outlier synthesis method that expands the feature space using nearest-neighbor class information. Their approach was later extended to multimodal OOD segmentation on SemanticKITTI [3]. More recently, Li et al. [34] introduced dynamic prototype updating, which adjusts class centers to account for intra-class variability in multimodal OOD detection. In this work, we propose a novel multimodal outlier synthesis method applicable to both OOD detection and segmentation tasks.

3 Methodology

3.1 Problem Setup

In this work, we focus on multimodal OOD detection and segmentation, where multiple modalities are involved to help the model better identify unknown objects. We define the problem setups for each task below.

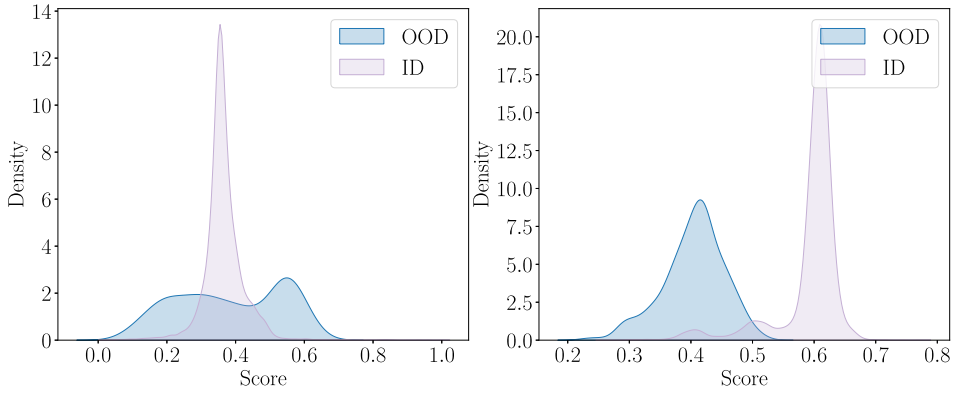
Multimodal OOD Segmentation aims to accurately segment both ID and OOD objects in a point cloud using LiDAR and image data. Given a training set with classes $\mathcal{Y} = \{1, 2, ..., C\}$, unlike traditional closed-set segmentation where test classes match training classes, OOD segmentation introduces unknown classes $\mathcal{U} = \{C + 1\}$ in the test set. A paired LiDAR point cloud and RGB image can be represented as $\mathcal{D} = \{\mathbf{P}, \mathbf{X}, \mathbf{y}\}$, where $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_M\}$ denotes the LiDAR point cloud consisting of M points, with each point $\mathbf{p}$ represented by three coordinates and intensity $\mathbf{p} = (x, y, z, i)$. Let $\mathbf{X} \in \mathbb{R}^{3\times H\times W}$ represent the RGB image, where H and W denote height and width. The label $\mathbf{y} = \{y_1, y_2, ..., y_M\}$ provides semantic labels for each point, where $y \in \mathcal{Y}$ for the training data and $y \in \mathcal{Y} \cup \mathcal{U}$ for the test data.

Given a model $\mathcal{M}$ trained under the closed-set assumption, with its outputs $\mathbf{O} = \mathcal{M}(\mathbf{P}, \mathbf{X}) \in \mathbb{R}^{M\times C}$ within the domain of $\mathcal{Y}$. During deployment, $\mathcal{M}$ should accurately classify known samples in $\mathcal{Y}$ as ID and identify *unknown* samples in $\mathcal{U}$ as OOD. A separate score function $S(\mathbf{p})$ is typically used as an OOD module to decide whether a sample point $\mathbf{p} \in \mathbf{P}$ is from ID or OOD:

$$G_\eta(\mathbf{p}) = \begin{cases} \text{ID} & S(\mathbf{p}) \geq \eta \\ \text{OOD} & S(\mathbf{p}) < \eta \end{cases}, \tag{1}$$

where samples with higher scores $S(\mathbf{p})$ are classified as ID and vice versa, and η is the threshold.

Multimodal OOD Detection aims to identify samples with semantic shifts in the test set using video and optical flow, where unknown classes are introduced. The setup is similar to segmentation but differs in input and output types. The input consists of a paired video $\mathbf{V}$ and optical flow $\mathbf{F}$, represented as $\mathcal{D} = \{\mathbf{V}, \mathbf{F}, y\}$, where y is the sample-level label rather than point-level label in segmentation. Similarly, the model produces sample-level outputs $\mathbf{O} = \mathcal{M}(\mathbf{V}, \mathbf{F}) \in \mathbb{R}^C$ instead of point-level. The remaining setup follows that of segmentation, and we refer the reader to [17] for a detailed definition.



(a) Without Outlier Optimization (b) With Outlier Optimization

Figure 2: (a) Uncertainty-based OOD methods face overconfidence issues, resulting in significant overlap between the score distributions of ID and OOD samples. (b) After training with outlier optimization, the confidence scores for ID and OOD samples become more distinct, enabling the model to better differentiate them. Results are on CARLA-OOD dataset.

3.2 Motivation for Outlier Synthesis

Uncertainty-based OOD detection and segmentation methods [24, 36, 37] are computationally efficient but suffer from overconfidence issues, as illustrated in Fig. 2 (a). Outlier exposure-based methods [25, 38, 31] mitigate this issue by training models using auxiliary OOD datasets to calibrate the confidences of both ID and OOD samples. However, such datasets are often unavailable, especially in multimodal settings. To address this challenge, we introduce Feature Mixing, an extremely simple and efficient multimodal outlier synthesis method that operates in the feature space with negligible computational overhead (Sec. 3.3). These synthesized outliers are further optimized via entropy maximization, enhancing the model's ability to distinguish ID from OOD data (Sec. 3.4). As shown in Fig. 2 (b), training with outlier optimization results in well-separated confidence scores, leading to improved OOD detection and segmentation.

3.3 Feature Mixing for Multimodal Outlier Synthesis

Existing Methods. Some prior works [49, 8] generate outliers in the pixel space by extracting OOD objects from external datasets and pasting them into inlier images. However, such methods are impractical for multimodal scenarios, where we need to generate outliers for paired multimodal data. Instead, generating outliers in the feature space is more effective and scalable. Mixup [52] interpolates features of randomly selected samples to generate outliers but inadvertently introduces noise samples within the ID distribution (Fig. 4 (a)). VOS [19] samples outliers from low-likelihood regions of the class-conditional feature distribution but is designed for unimodal settings and struggles with multimodal data. Moreover, it generates outliers too close to ID samples (Fig. 4 (b)) and is slow for high-dimensional features. NP-Mix [17] explores broader embedding spaces using nearest-neighbor class information but remains computationally expensive for segmentation tasks and introduces unwanted noise (Fig. 4 (c)).

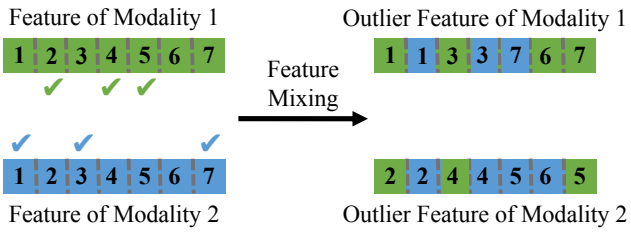
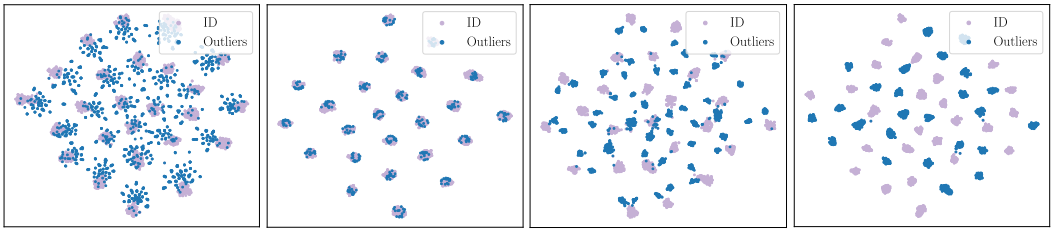


Figure 3: Illustration of Feature Mixing.

Our Solution. To overcome these limitations, we propose Feature Mixing, an extremely simple yet effective approach that generates multimodal outliers directly in the feature space. Our method ensures that the synthesized features remain distinct from ID features (Theorem 1) while preserving semantic consistency (Theorem 2). Given ID features $\mathbf{F} = [\mathbf{F}_c; \mathbf{F}_l]$, where $\mathbf{F}_c$ is from modality 1 and $\mathbf{F}_l$ is from modality 2, Feature Mixing randomly selects a subset of N feature dimensions from each modality and swaps them to obtain new features $\widetilde{\mathbf{F}}_c$ and $\widetilde{\mathbf{F}}_l$, which are then concatenated to form the



(a) Mixup (b) VOS (c) NP-Mix (d) Feature Mixing (ours)

Figure 4: t-SNE Visualization of multimodal outlier synthesis results on the HMDB51 dataset. Our Feature Mixing excels at generating outlier samples by spanning wider embedding spaces without injecting noise at an extremely fast speed.

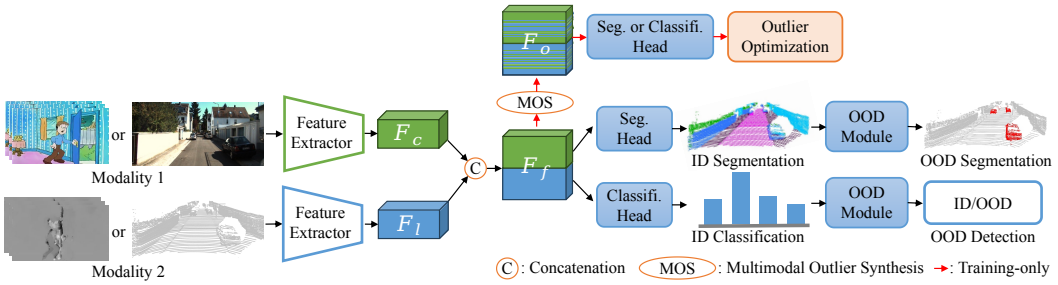


Figure 5: Overview of the proposed framework that integrates Feature Mixing for multimodal OOD detection and segmentation.

multimodal outlier features $\mathbf{F}_o = [\widetilde{\mathbf{F}}_c; \widetilde{\mathbf{F}}_l]$. Fig. 3 and Algorithm 1 provide detailed illustrations of the outlier synthesis process.

As shown in Fig. 4 (d), Feature Mixing excels at generating multimodal outliers by covering a broader embedding space without introducing noisy samples. The generated outliers exhibit ***two key properties***: (1) These outliers share the same embedding space with the ID features but lie in low-likelihood regions (Theorem 1). (2) Their deviation from ID features is bounded, preventing excessive shifts while maintaining diversity (Theorem 2). These properties ensure that the outliers align with real OOD characteristics and can be supported by the following theorems. Due to space limits, the proofs are provided in the Appendix.

Algorithm 1 Feature Mixing

Input: ID feature $\mathbf{F} = [\mathbf{F}_c; \mathbf{F}_l]$, where $\mathbf{F}_c$ is from modality 1 with N_c channels, $\mathbf{F}_l$ is from modality 2 with N_l channels; number of selected feature dimensions for mixing N.

Python-like Code:

$select_c = random.sample(range(N_c), N)$
$select_l = random.sample(range(N_l), N)$
$\widetilde{\mathbf{F}}_c = \mathbf{F}_c.clone()$
$\widetilde{\mathbf{F}}_l = \mathbf{F}_l.clone()$
$\widetilde{\mathbf{F}}_c[select_c, :, :] = \mathbf{F}_l[select_l, :, :]$
$\widetilde{\mathbf{F}}_l[select_l, :, :] = \mathbf{F}_c[select_c, :, :]$
$\mathbf{F}_o = torch.cat([\widetilde{\mathbf{F}}_c, \widetilde{\mathbf{F}}_l], dim = 0)$

Output: Multimodal outlier feature $\mathbf{F}_o$.

Theorem 1. *Outliers $\mathbf{F}_o$ synthesized by Feature Mixing lie in low-likelihood regions of the distribution of the ID features $\mathbf{F}$, complying with the criterion for real outliers.*

Theorem 2. *Outliers $\mathbf{F}_o$ are bounded in their deviation from $\mathbf{F}$, such that $|\mathbf{F}_o - \mathbf{F}|_2 \leq \sqrt{2N} \cdot \delta$, where $\delta = \max_{i,j} \left| \mathbf{F}_c^{(i)} - \mathbf{F}_l^{(j)} \right|$.*

Feature Mixing enables the online generation of multimodal outlier features and can be seamlessly integrated into existing training pipelines (Sec. 3.4). Besides, its simplicity makes it *modality-agnostic*, allowing application to diverse multimodal setups, such as images and point clouds or video and optical flow.

3.4 Framework for Outlier Optimization

Multimodal outlier features generated by Feature Mixing can be optimized using entropy maximization to help the model better distinguish between ID and OOD data, similar to outlier exposure-based methods [25, 38, 31]. Fig. 5 illustrates our framework, which integrates Feature Mixing into multi-

modal OOD detection and segmentation, comprising two key components: *Basic Multimodal Fusion* and *Multimodal Outlier Synthesis and Optimization*.

Basic Multimodal Fusion. Our framework employs a dual-stream network to extract features from different modalities using separate backbones. For example, for multimodal OOD segmentation, features extracted from the image and point cloud backbones, denoted as $\mathbf{F}_c \in \mathbb{R}^{N_c \times H \times W}$ and $\mathbf{F}_l \in \mathbb{R}^{N_l \times H \times W}$ respectively, are concatenated to form the fused representation F_f. F_f contains both 2D and 3D scene information, which is then passed to a segmentation head for ID class segmentation. To enable efficient OOD segmentation at inference, we append an uncertainty-based OOD detection module to compute confidence scores for each prediction. This module supports various post-hoc OOD scoring methods [27, 24, 36], offering flexibility in design choices. Furthermore, the simple late-fusion design facilitates the integration of advanced cross-modal training strategies [26, 17] and generalizes easily to other modalities and tasks.

Multimodal Outlier Synthesis and Optimization. To mitigate overconfidence in uncertainty-based OOD detection, we incorporate outlier samples during training. These can be generated using existing methods such as Mixup [52], VOS [19], NP-Mix [17], or our proposed Feature Mixing. We then apply entropy-based optimization in Eq. (5) to maximize the entropy of outlier features. In this way, we can better separate the confidence scores between ID and OOD samples (Fig. 2), thereby improving the model's ability to distinguish OOD samples.

3.5 Training Strategy

Our training objective is to enhance OOD detection and segmentation while maintaining strong ID classification and segmentation performance.

Multimodal OOD Segmentation. Since accurate ID segmentation is crucial for the effectiveness of post-hoc OOD detection methods, optimizing ID segmentation is a priority. We employ focal loss [35] and Lovász-softmax loss [4], which are widely used in existing segmentation work [10, 53]. The focal loss $\mathcal{L}_{foc}$ addresses class imbalance by focusing on hard examples and is defined as:

$$\mathcal{L}_{foc} = \frac{1}{M} \sum_{m=1}^{M} \sum_{c=1}^{C} \alpha_c \mathbb{1}\{y_m = c\} FL(\mathbf{O}_{m,c}), \tag{2}$$

where $FL(p) = -(1-p)^{\lambda} \log(p)$ denotes the focal loss function and α_c is the weight w.r.t the c-th class. $\mathbb{1}\{\cdot\}$ is the indicator function. The Lovász-Softmax loss $\mathcal{L}_{lov}$ directly optimizes the mean IoU and is expressed as:

$$\mathcal{L}_{lov} = \frac{1}{C} \sum_{c=1}^{C} \overline{\Delta_{J_c}}(\mathbf{m}(c)), \tag{3}$$

where

$$\mathbf{m}_m(c) = \begin{cases} 1 - \mathbf{O}_{m,c} & \text{if } c = y_m, \\ \mathbf{O}_{m,c} & \text{otherwise.} \end{cases} \tag{4}$$

$\overline{\Delta_{J_c}}$ indicates the Lovász extension of the Jaccard index for class c. $\mathbf{m}(c) \in [0,1]^M$ indicates the vector of errors. For the generated multimodal outlier feature $\mathbf{F}_o$, we obtain a prediction output $\widetilde{\mathbf{O}} \in \mathbb{R}^{M \times C}$ using the segmentation head and aim to maximize the prediction entropy of the outlier features:

$$\mathcal{L}_{ent} = \frac{1}{M} \sum_{m=1}^{M} \sum_{c=1}^{C} \widetilde{\mathbf{O}}_{m,c} \log \widetilde{\mathbf{O}}_{m,c}. \tag{5}$$

The final loss is defined as:

$$\mathcal{L} = \mathcal{L}_{foc} + \mathcal{L}_{lov} + \gamma_1 \mathcal{L}_{ent}, \tag{6}$$

where γ_1 is a weighting factor that balances the contributions of each loss term.

Multimodal OOD Detection. The OOD detection loss combines classification loss with entropy regularization, which is defined as:

$$\mathcal{L} = \mathcal{L}_{cls} + \gamma_1 \mathcal{L}_{ent}, \tag{7}$$

where the cross-entropy loss is used for $\mathcal{L}_{cls}$.

4 Experiments

We evaluate Feature Mixing across eight datasets and four modalities to demonstrate its versatility. Specifically, we use nuScenes, SemanticKITTI, and our CARLA-OOD dataset for **Multimodal OOD Segmentation** using image and point cloud data. Additionally, we utilize five action recognition datasets from MultiOOD benchmark [17] for **Multimodal OOD Detection**, employing video and optical flow modalities.

4.1 Experimental Setup

Datasets and Settings. For multimodal OOD segmentation, we follow [17] to treat all vehicle classes as OOD on the SemanticKITTI [3] and nuScenes [6] datasets. During training, the labels of OOD classes are set to void and ignored. During inference, we aim to segment ID classes with high Intersection over Union (IoU) while detecting OOD classes as *unknown*. We also introduce the ***CARLA-OOD*** dataset, created using the CARLA simulator [18], which includes RGB images, LiDAR point clouds, and 3D semantic segmentation ground truth, comprising a total of 245 samples. We select 34 anomalous objects as OOD, which are randomly positioned in front of the ego-vehicle across varied scenes and weather conditions, as shown in Fig. 6. Further details on CARLA-OOD are provided in the Appendix. The model is trained on the KITTI-CARLA [12] dataset with the same sensor setup and evaluated on CARLA-OOD with OOD objects. For multimodal OOD detection, we use HMDB51 [30], UCF101 [45], Kinetics-600 [28], HAC [16], and EPIC-Kitchens [11] datasets from the MultiOOD [17] benchmark. We evaluate using video and optical flow, where we train the model on one ID dataset and treat other datasets as OOD during testing.

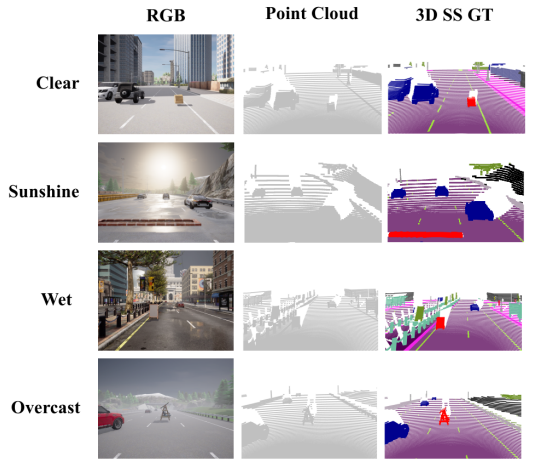


Figure 6: The proposed CARLA-OOD dataset for multimodal OOD segmentation. Points with red color are OOD objects.

Implementation Details. For multimodal OOD segmentation, our implementation follows [17] to build upon the fusion framework proposed in PMF [53], utilizing ResNet-34 [22] as the camera backbone and SalsaNext [10] as the LiDAR backbone. After feature extraction and fusion, we employ two 2D convolution layers as the segmentation head for ID segmentation. The OOD detection module uses MaxLogit [23] as the default scoring function. For multimodal OOD detection, we adopt the framework proposed in MultiOOD [17] and replace the multimodal outlier generation method with our Feature Mixing. Additional implementation details are provided in the Appendix.

Evaluation Metrics. For OOD segmentation, we evaluate both closed-set and OOD segmentation performance at the point level. For closed-set evaluation, we use the mean Intersection over Union for known classes (mIoU_c). For OOD performance, we report the area under the receiver operating characteristic curve (AUROC), the area under the precision-recall curve (AUPR), and the false positive rate at 95% true positive rate (FPR@95). For multimodal OOD detection, we report average accuracy (ACC) instead of mIoU_c for closed-set evaluation, as well as AUROC and FPR@95 for OOD performance.

4.2 Main Results

4.2.1 Evaluation on Multimodal OOD Segmentation

We first evaluate our method on multimodal OOD segmentation. For baselines without outlier optimization, we consider basic Late Fusion, A2D [17] and xMUDA [26], with A2D being the state-of-the-art method. For baselines incorporating outlier optimization, we integrate Mixup [52], NP-Mix [17], and our Feature Mixing into the framework. Due to its inefficiency, VOS [19] is

Method	SemanticKITTI				nuScenes				CARLA-OOD			
	FPR@95↓	AUROC↑	AUPR↑	mIoU_c↑	FPR@95↓	AUROC↑	AUPR↑	mIoU_c↑	FPR@95↓	AUROC↑	AUPR↑	mIoU_c↑
w/o Outlier Optimization												
Late Fusion	53.43	86.98	46.02	61.43	47.55	82.60	26.42	76.79	98.83	57.24	20.56	61.84
xMUDA [26]	55.37	89.67	51.41	60.61	44.32	83.47	20.20	78.79	97.00	57.86	10.35	65.15
A2D [17]	49.02	91.12	55.44	61.98	44.27	83.43	23.55	77.69	97.98	64.21	22.45	63.79
w/ Outlier Optimization												
Mixup [52]	52.04	86.81	48.05	61.36	42.94	83.82	27.89	75.67	99.23	57.94	9.02	62.07
NP-Mix [17]	48.57	90.93	56.85	60.37	41.69	84.88	28.54	76.16	41.81	88.45	29.68	62.56
Feature Mixing (ours)	38.10	91.47	58.74	61.18	40.48	86.83	**38.80**	77.61	**25.85**	92.98	33.37	63.38
xMUDA + FM (ours)	36.63	91.54	53.89	60.43	39.49	85.29	28.74	77.69	30.35	92.45	33.44	65.92
A2D + FM (ours)	**31.76**	**92.83**	**61.99**	60.41	**32.92**	**87.55**	29.39	76.47	25.95	**93.37**	**37.28**	66.41

Table 1: Evaluation results on multimodal OOD segmentation datasets. FM: Feature Mixing.

Methods	OOD Datasets										ID ACC ↑
	Kinetics-600		UCF101		EPIC-Kitchens		HAC		Average		
	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	
Baseline	32.95	92.48	44.93	87.95	8.10	97.70	32.95	92.28	29.73	92.60	87.23
Mixup [52]	25.31	94.10	36.37	90.49	14.37	96.40	22.57	94.85	24.67	93.96	86.89
VOS [19]	31.70	93.22	38.77	89.93	15.39	96.82	31.58	93.03	29.36	93.25	87.34
NPOS [48]	25.31	93.94	37.17	89.71	13.00	96.50	24.17	93.94	24.91	93.52	87.12
NP-Mix [17]	24.52	93.96	36.49	89.67	6.96	97.53	22.92	94.41	22.72	93.89	86.89
Feature Mixing (ours)	19.61	94.72	34.32	90.06	10.15	96.34	15.96	95.54	**20.01**	**94.17**	87.00

Table 2: Multimodal OOD Detection using video and optical flow, with **HMDB51** as ID. Energy is used as the OOD score.

excluded from the segmentation task. Additionally, we combine A2D and xMUDA with Feature Mixing to demonstrate its versatility.

As shown in Tab. 1, Late Fusion without outlier optimization suffers from overconfidence, leading to high FPR@95 values, indicating poor ID-OOD separation. While A2D improves performance in most cases, it remains suboptimal. Integrating outlier optimization yields significant improvements for both NP-Mix and Feature Mixing, underscoring the importance of outlier synthesis. On SemanticKITTI, Feature Mixing improves Late Fusion by 15.33% on FPR@95, 4.49% on AUROC, and 12.72% on AUPR. On nuScenes, Feature Mixing improves the Late Fusion baseline by 7.07% on FPR@95, 4.23% on AUROC, and 12.38% on AUPR. At the same time, Feature Mixing introduces a negligible negative impact on mIoU_c value.

Notably, all baselines without outlier optimization perform poorly on CARLA-OOD, with FPR@95 exceeding 97%, highlighting the dataset's difficulty and the overconfidence issue in uncertainty-based OOD methods. Feature Mixing significantly enhances Late Fusion on CARLA-OOD, reducing FPR@95 by 72.98%, improving AUROC by 35.74%, and increasing AUPR by 12.81%. Furthermore, A2D + Feature Mixing achieves the best results in most cases, demonstrating our framework's adaptability to advanced cross-modal training strategies.

4.2.2 Evaluation on Multimodal OOD Detection

To assess the generalizability of Feature Mixing across tasks and modalities, we evaluate it on MultiOOD for multimodal OOD detection in action recognition, where video and optical flow serve as distinct modalities. We replace the outlier generation method in MultiOOD framework with Feature Mixing and compare it against Mixup [52], VOS [19], NPOS [48], and NP-Mix [17]. Models are trained on HMDB51 [30] or Kinetics-600 [39], and other datasets are treated as OOD during testing. As shown in Tab. 2, our Feature Mixing outperforms other outlier generation methods in most cases, achieving the lowest FPR@95 of 20.01% and the highest AUROC of 94.17% on average when using HMDB51 as ID. Due to space limits, we put the results on Kinetics-600 in the Appendix. These results highlight the effectiveness of Feature Mixing in improving OOD detection across various tasks and modalities. Similarly, Feature Mixing introduces a negligible impact on ID ACC.

4.3 Ablation Studies

Computational Cost. Tab. 3 compares the computational cost of different outlier synthesis methods. For OOD detection, the reported time corresponds to generating 2048 multimodal outlier samples of shape 4352. For OOD segmentation, it represents the time to synthesize 256×352 samples of shape 48.

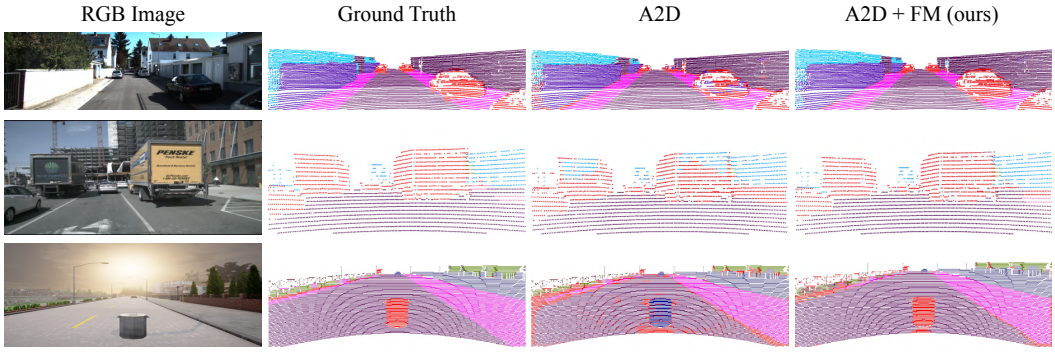


Figure 7: Visualization results on different datasets. From top to down: SemanticKITTI, nuScenes, and CARLA-OOD. Points with red color are OOD objects. Our method can segment OOD objects accurately. More visualization results are in the Appendix.

While Mixup [52] is efficient, it performs poorly in OOD detection. In contrast, NP-Mix [17] achieves strong performance but is computationally expensive. Feature Mixing, benefiting from its simple design, is both highly efficient and effective. Compared to NP-Mix, Feature Mixing provides a $\mathbf{10}\times$ speedup for multimodal OOD detection and a $\mathbf{370}\times$ speedup for segmentation, making it well-suited for real-world applications.

	OOD Detection	OOD Segmentation
Mixup [52]	0.038	0.019
VOS [19]	152.05	61.49
NP-Mix [17]	0.545	4.81
Feature Mixing (ours)	0.058	0.013

Table 3: Computational cost of outlier synthesis methods (s).

Visualization. Fig. 7 presents the visualization of OOD segmentation results across different datasets, showcasing the RGB image, 3D semantic ground truth, and predictions from A2D and our best-performing A2D+FM model. The baseline method A2D struggles to identify OOD objects, whereas our method accurately segments OOD with minimal noise, demonstrating the effectiveness of the proposed framework. Additional visualizations can be found in the Appendix.

Hyperparameter Sensitivity. We evaluate the sensitivity of Feature Mixing to the hyperparameter N, using HMDB51 as ID dataset and Kinetics-600 as OOD dataset. Our findings, as illustrated in Fig. 8, demonstrate that Feature Mixing is robust and consistently outperforms the baseline across all parameter settings.

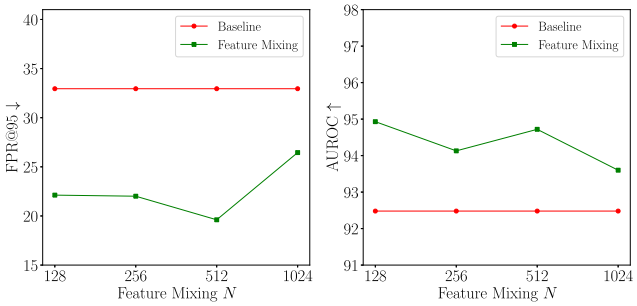


Figure 8: Ablation on the number N in Feature Mixing.

Feature Mixing in Tri- and Uni-modal Settings. To further demonstrate the generality of our method, we conduct additional experiments using Feature Mixing (FM) in both tri-modal and unimodal settings.

Tri-Modal Setting (Video + Optical Flow + Audio). We extend FM to a tri-modal setting on the EPIC-Kitchens dataset, where modalities include video, optical flow, and audio. FM is applied by randomly selecting a subset of N feature dimensions from each modality and performing a cyclic swap: video $\rightarrow$ audio, audio $\rightarrow$ optical flow, and optical flow $\rightarrow$ video. This procedure synthesizes outlier features for all three modalities. As shown in Tab. 4a, FM consistently outperforms NP-Mix across all evaluation metrics, demonstrating robustness and versatility in complex multimodal scenarios.

Unimodal Setting (Video Only). FM is inherently modality-agnostic, operating purely in feature space without relying on input-specific assumptions. This makes it naturally extendable to unimodal scenarios. We implement a unimodal variant by splitting the video feature embedding into two halves and randomly swapping N dimensions between them, effectively synthesizing outliers within a single

Method	FPR@95↓	AUROC↑	ACC↑
Baseline	69.22	72.39	73.13
NP-Mix	62.69	74.95	71.46
Feature Mixing	**61.38**	**77.23**	**73.69**

(a) EPIC-Kitchens (Tri-modal)

Method	FPR@95↓	AUROC↑	ACC↑
Baseline	64.05	83.14	86.93
NP-Mix	62.53	83.62	87.15
Feature Mixing	**57.30**	**84.41**	**88.89**

(b) HMDB51 (Unimodal)

Table 4: OOD detection results with Feature Mixing in tri-modal and unimodal settings.

modality. Evaluated on the HMDB51 dataset, this approach outperforms the NP-Mix baseline across all metrics, confirming FM's applicability even in the absence of multiple modalities (Tab. 4b).

Different Classes as OOD. For multimodal OOD segmentation, we follow [17] and designate all vehicle classes as OOD. Here, we experiment with different OOD category assignments: "ground" (road, sidewalk, parking, other-ground), "structure" (building, other-structure), and "nature" (vegetation, trunk, terrain). As shown in Tab. 5, Feature Mixing remains robust across all splits, consistently outperforming the baseline by a significant margin.

Method	FPR@95↓	AUROC↑	AUPR↑	mIoU$_c$↑
"ground" classes as OOD				
A2D [17]	71.15	74.92	69.13	66.57
A2D + NP-Mix	53.60	94.71	95.30	65.07
A2D + FM (ours)	36.30	95.89	96.04	65.88
"structure" classes as OOD				
A2D [17]	23.50	95.20	75.23	61.38
A2D + NP-Mix	22.14	95.41	76.36	60.54
A2D + FM (ours)	18.05	96.09	79.85	61.79
"nature" classes as OOD				
A2D [17]	37.97	92.74	90.22	62.76
A2D + NP-Mix	30.88	95.18	92.51	61.21
A2D + FM (ours)	20.60	96.13	94.08	62.67

Table 5: Ablation on different classes as OOD on SemanticKITTI.

5 Conclusion

In this work, we introduce Feature Mixing, an extremely simple and fast method for multimodal outlier synthesis with theoretical support. Feature Mixing is modality-agnostic and applicable to various modality combinations. Moreover, its lightweight design achieves a $10\times$ to $370\times$ speedup over existing methods while maintaining strong OOD performance. To mitigate overconfidence, outlier features are optimized via entropy maximization within our framework. Additionally, we present CARLA-OOD, a challenging multimodal dataset featuring synthetic OOD objects captured under diverse scenes and weather conditions. Extensive experiments across eight datasets and four modalities validate the versatility and effectiveness of Feature Mixing and our proposed framework.

Limitations and future work. Feature Mixing uses a random selection of feature dimensions to swap between modalities. While this is highly efficient and theoretically grounded, it may not always target the most informative feature regions for outlier synthesis. Future work could explore *adaptive or learnable selection mechanisms* that dynamically identify features that maximize OOD separability.

Societal impact. In our work, Feature Mixing advances multimodal OOD detection and segmentation by enabling more effective and efficient outlier exposure during training. This contributes directly to the reliability and safety of autonomous systems, including self-driving cars, by helping models handle unfamiliar or unexpected scenarios in open-world environments. Beyond autonomous driving, the method also has broader societal impacts in other safety-critical domains such as healthcare and security, where robustness to unknown or anomalous data is crucial. By improving model reliability in the presence of distribution shifts, our approach supports the deployment of AI systems in dynamic, high-stakes environments where failure could have significant consequences.

Acknowledgments

The authors acknowledge that this work was supported by the Federal Ministry of Education and Research (BMBF) as part of MANNHEIM-AutoDevSafeOps (reference number 01IS22087E). It was also funded by the Bavarian Ministry for Economic Affairs, Regional Development and Energy as part of a project to support the thematic development of the Institute for Cognitive Systems. The authors also acknowledge the support of "In-service diagnostics of the catenary/pantograph and wheelset axle systems through intelligent algorithms" (SENTINEL) project, supported by the ETH Mobility Initiative.

References

[1] Bao, J., Sun, H., Deng, H., He, Y., Zhang, Z., Li, X.: Bmad: Benchmarks for medical anomaly detection. In: CVPR (2024)

[2] Baur, C., Wiestler, B., Albarqouni, S., Navab, N.: Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. arXiv preprint arXiv:1804.044886 (2018)

[3] Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: ICCV (2019)

[4] Berman, M., Triki, A.R., Blaschko, M.B.: The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In: CVPR (2018)

[5] Blum, H., Sarlin, P.E., Nieto, J., Siegwart, R., Cadena, C.: The fishyscapes benchmark: Measuring blind spots in semantic segmentation. International Journal of Computer Vision **129**(11), 3119–3135 (2021)

[6] Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)

[7] Cen, J., Yun, P., Zhang, S., Cai, J., Luan, D., Tang, M., Liu, M., Yu Wang, M.: Open-world semantic segmentation for lidar point clouds. In: ECCV (2022)

[8] Choi, H., Jeong, H., Choi, J.Y.: Balanced energy regularization loss for out-of-distribution detection. In: CVPR (2023)

[9] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)

[10] Cortinhal, T., Tzelepis, G., Aksoy, E.E.: Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds for autonomous driving. arXiv preprint arXiv:2003.03653 (2020)

[11] Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: ECCV (2018)

[12] Deschaud, J.E.: KITTI-CARLA: a KITTI-like dataset generated by CARLA Simulator. arXiv preprint arXiv:2109.00892 (2021)

[13] Dhamija, A.R., Günther, M., Boult, T.: Reducing network agnostophobia. In: NeurIPS (2018)

[14] Dong, H., Chatzi, E., Fink, O.: Towards multimodal open-set domain generalization and adaptation through self-supervision. arXiv preprint arXiv:2407.01518 (2024)

[15] Dong, H., Liu, M., Zhou, K., Chatzi, E., Kannala, J., Stachniss, C., Fink, O.: Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. arXiv preprint arXiv:2501.18592 (2025)

[16] Dong, H., Nejjar, I., Sun, H., Chatzi, E., Fink, O.: SimMMDG: A simple and effective framework for multi-modal domain generalization. In: NeurIPS (2023)

[17] Dong, H., Zhao, Y., Chatzi, E., Fink, O.: Multiood: Scaling out-of-distribution detection for multiple modalities. In: NeurIPS (2024)

[18] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V.: CARLA: An open urban driving simulator. In: CoRL (2017)

[19] Du, X., Wang, Z., Cai, M., Li, Y.: Vos: Learning what you don't know by virtual outlier synthesis. In: ICLR (2022)

[20] Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: ICCV (2019)

[21] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)

[22] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

[23] Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. arXiv preprint arXiv:1911.11132 (2019)

[24] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)

[25] Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. arXiv preprint arXiv:1812.04606 (2018)

[26] Jaritz, M., Vu, T.H., Charette, R.d., Wirbel, E., Pérez, P.: xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In: CVPR (2020)

[27] Jung, S., Lee, J., Gwak, D., Choi, S., Choo, J.: Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In: ICCV (2021)

[28] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)

[29] Kingma, D.P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[30] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: ICCV (2011)

[31] Lee, J., Oh, S.J., Yun, S., Choe, J., Kim, E., Yoon, S.: Weakly supervised semantic segmentation using out-of-distribution data. In: CVPR (2022)

[32] Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS (2018)

[33] Li, J., Dong, Q.: Open-set semantic segmentation for point clouds via adversarial prototype framework. In: CVPR (2023)

[34] Li, S., Gong, H., Dong, H., Yang, T., Tu, Z., Zhao, Y.: Dpu: Dynamic prototype updating for multimodal out-of-distribution detection. arXiv preprint arXiv:2411.08227 (2024)

[35] Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. arXiv preprint arXiv:1708.02002 (2017)

[36] Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. In: NeurIPS (2020)

[37] Liu, X., Lochman, Y., Zach, C.: Gen: Pushing the limits of softmax-based out-of-distribution detection. In: CVPR. pp. 23946–23955 (2023)

[38] Liu, Y., Ding, C., Tian, Y., Pang, G., Belagiannis, V., Reid, I., Carneiro, G.: Residual pattern learning for pixel-wise out-of-distribution detection in semantic segmentation. In: ICCV (2023)

[39] Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J., Mei, T.: Learning to localize actions from moments. In: ECCV (2020)

[40] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)

[41] Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)

[42] Masuda, M., Hachiuma, R., Fujii, R., Saito, H., Sekikawa, Y.: Toward unsupervised 3d point cloud anomaly detection using variational autoencoder. In: ICIP (2021)

[43] Nejjar, I., Dong, H., Fink, O.: Recall and refine: A simple but effective source-free open-set domain adaptation framework. arXiv preprint arXiv:2411.12558 (2024)

[44] Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: CVPR (2015)

[45] Soomro, K.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)

[46] Sun, H., Cao, Y., Fink, O.: Cut: A controllable, universal, and training-free visual anomaly generation framework. arXiv preprint arXiv:2406.01078 (2024)

[47] Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. ICML (2022)

[48] Tao, L., Du, X., Zhu, X., Li, Y.: Non-parametric outlier synthesis. arXiv preprint arXiv:2303.02966 (2023)

[49] Tian, Y., Liu, Y., Pang, G., Liu, F., Chen, Y., Carneiro, G.: Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In: ECCV (2022)

[50] Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: CVPR (2022)

[51] Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: ICML (2022)

[52] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: ICLR (2018)

[53] Zhuang, Z., Li, R., Jia, K., Wang, Q., Li, Y., Tan, M.: Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In: ICCV (2021)

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction clearly state the claims made.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the Appendix, we discussed limitations and future work.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The paper provides the full set of assumptions and a complete proof in the Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our paper discloses all the information needed to reproduce the main experimental results. Our source code and benchmark datasets will also be made publicly available upon acceptance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Our source code and benchmark datasets will also be made publicly available upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify all the training and test details in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We run all experiments three times and report the average value.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details in implementation details part.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Our paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: They are properly credited in our paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new assets introduced in the paper are well documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

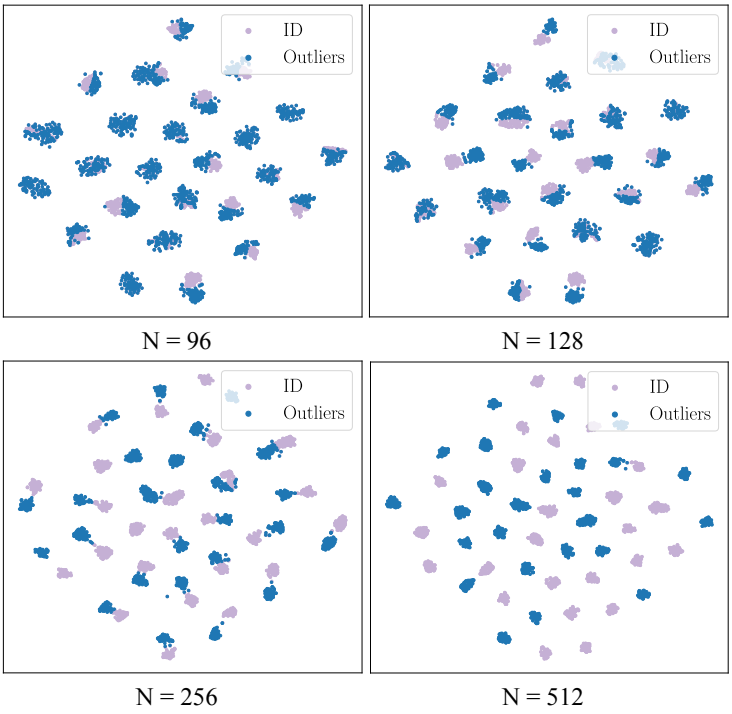


Figure 9: Feature Mixing introduces bias that shifts $\mathbf{F}_o$ away from the ID distribution $\mathbf{F}$, which is proportional to N. A large N promotes outliers to lie in low-likelihood regions of the ID feature distribution. Therefore, N can be adjusted to generate targeted outliers.

A Theoretical Insights for Feature Mixing

Problem Setup. Let $\mathbf{F} = \begin{bmatrix} \mathbf{F}_c \\ \mathbf{F}_l \end{bmatrix} \in \mathbb{R}^{2d}$ be the concatenated in-distribution (ID) features from two modalities, where $\mathbf{F}_c \sim P$ with mean $\boldsymbol{\mu}_c$ and covariance $\boldsymbol{\Sigma}_c$, $\mathbf{F}_l \sim Q$ with mean $\boldsymbol{\mu}_l$ and covariance $\boldsymbol{\Sigma}_l$, $\boldsymbol{\mu}_c \neq \boldsymbol{\mu}_l$. The joint distribution of $\mathbf{F}$ has mean $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_c \\ \boldsymbol{\mu}_l \end{bmatrix}$ and covariance $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_c & \boldsymbol{\Sigma}_{cl} \\ \boldsymbol{\Sigma}_{cl}^T & \boldsymbol{\Sigma}_l \end{bmatrix}$, where $\boldsymbol{\Sigma}_{cl}$ encodes cross-modal dependencies.

Feature Mixing swaps N features between $\mathbf{F}_c$ and $\mathbf{F}_l$ to generate perturbed features $\widetilde{\mathbf{F}}_c$ and $\widetilde{\mathbf{F}}_l$, then concatenates them to form $\mathbf{F}_o = \begin{bmatrix} \widetilde{\mathbf{F}}_c \\ \widetilde{\mathbf{F}}_l \end{bmatrix}$, which can be written as:

$$\widetilde{\mathbf{F}}_c = \mathbf{F}_c \odot (1 - \mathbf{M_1}) + \mathbf{F}_l \odot \mathbf{M_1}, \tag{8}$$

$$\widetilde{\mathbf{F}}_l = \mathbf{F}_l \odot (1 - \mathbf{M_2}) + \mathbf{F}_c \odot \mathbf{M_2}, \tag{9}$$

where $\mathbf{M_1}, \mathbf{M_2} \in \{0, 1\}^d$ is a binary mask with N ones.

Theorem 1 *Outliers $\mathbf{F}_o$ synthesized by Feature Mixing lie in low-likelihood regions of the distribution of the ID features $\mathbf{F}$, complying with the criterion for real outliers.*

Proof 1: After swapping N features, the feature of each modality has a $\frac{N}{d}$ probability of being swapped from the other modality. Therefore, the mean of $\widetilde{\mathbf{F}}_c$ is a weighted average of $\boldsymbol{\mu}_c$ and $\boldsymbol{\mu}_l$, and similarly for $\widetilde{\mathbf{F}}_l$:

$$\mathbb{E}[\widetilde{\mathbf{F}}_c] = \left(1 - \frac{N}{d}\right) \boldsymbol{\mu}_c + \frac{N}{d} \boldsymbol{\mu}_l, \tag{10}$$

$$\mathbb{E}[\widetilde{\mathbf{F}}_l] = \left(1 - \frac{N}{d}\right) \boldsymbol{\mu}_l + \frac{N}{d} \boldsymbol{\mu}_c. \tag{11}$$

The perturbed mean $\boldsymbol{\mu}_o$ of $\mathbf{F}_o$ becomes a weighted combination of the original modality means:

$$\boldsymbol{\mu}_o = \mathbb{E}[\mathbf{F}_o] = \begin{bmatrix} \mathbb{E}[\widetilde{\mathbf{F}}_c] \\ \mathbb{E}[\widetilde{\mathbf{F}}_l] \end{bmatrix} = \begin{bmatrix} \left(1 - \frac{N}{d}\right) \boldsymbol{\mu}_c + \frac{N}{d} \boldsymbol{\mu}_l \\ \left(1 - \frac{N}{d}\right) \boldsymbol{\mu}_l + \frac{N}{d} \boldsymbol{\mu}_c \end{bmatrix}. \tag{12}$$

The deviation from the original mean $\boldsymbol{\mu}$ becomes:

$$\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_o - \boldsymbol{\mu} = \frac{N}{d} \begin{bmatrix} \boldsymbol{\mu}_l - \boldsymbol{\mu}_c \\ \boldsymbol{\mu}_c - \boldsymbol{\mu}_l \end{bmatrix}. \tag{13}$$

Since $\boldsymbol{\mu}_c \neq \boldsymbol{\mu}_l$, $\Delta\boldsymbol{\mu} \neq 0$, introducing a bias that shifts $\mathbf{F}_o$ away from the ID distribution proportional to N (Fig. 9) and $|\boldsymbol{\mu}_c - \boldsymbol{\mu}_l|$. The Mahalanobis distance measures how far $\mathbf{F}_o$ deviates from the mean $\boldsymbol{\mu}$ of the original joint distribution $\mathbf{F}$, weighted by the inverse covariance $\boldsymbol{\Sigma}^{-1}$:

$$D^2(\mathbf{F}_o) = (\mathbf{F}_o - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{F}_o - \boldsymbol{\mu}). \tag{14}$$

By defining $\mathbf{F}_o - \boldsymbol{\mu}$ as $(\mathbf{F}_o - \boldsymbol{\mu}_o) + (\boldsymbol{\mu}_o - \boldsymbol{\mu})$, we get:

$$D^2(\mathbf{F}_o) = (\mathbf{F}_o - \boldsymbol{\mu}_o + \Delta\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{F}_o - \boldsymbol{\mu}_o + \Delta\boldsymbol{\mu}). \tag{15}$$

After expanding the quadratic form, we get:

$$\begin{aligned} D^2(\mathbf{F}_o) = (\mathbf{F}_o - \boldsymbol{\mu}_o)^T \boldsymbol{\Sigma}^{-1} (\mathbf{F}_o - \boldsymbol{\mu}_o) + (\Delta\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \Delta\boldsymbol{\mu} + \\ 2(\Delta\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{F}_o - \boldsymbol{\mu}_o). \end{aligned} \tag{16}$$

The first term captures the deviation of $\mathbf{F}_o$ from its perturbed mean $\boldsymbol{\mu}_o$, weighted by $\boldsymbol{\Sigma}^{-1}$. The second term is the bias from the mean shift $\Delta\boldsymbol{\mu}$, which grows with N and $|\boldsymbol{\mu}_c - \boldsymbol{\mu}_l|$. The last term is the cross-term that involves both perturbation noise and mean shift.

The original covariance $\boldsymbol{\Sigma}$ encodes intra- and cross-modal correlations. After swapping, $\mathrm{Cov}(\widetilde{\mathbf{F}}_c)$ becomes a mix of $\boldsymbol{\Sigma}_c$ and $\boldsymbol{\Sigma}_l$, similarly for $\mathrm{Cov}(\widetilde{\mathbf{F}}_l)$. Besides, swapped features disrupt dependencies between $\mathbf{F}_c$ and $\mathbf{F}_l$, invalidating $\boldsymbol{\Sigma}_{cl}$. Therefore, the perturbed features $\mathbf{F}_o$ have a new covariance structure $\boldsymbol{\Sigma}_o \neq \boldsymbol{\Sigma}$ and this mismatch inflates the first term in Eq. (16). $\mathbf{F}_o - \boldsymbol{\mu}_o$ represents deviations under the perturbed distribution $\boldsymbol{\Sigma}_o$, which are not aligned with the original covariance structure $\boldsymbol{\Sigma}$. This misalignment causes $\boldsymbol{\Sigma}^{-1}$ to assign incorrect weights to the deviations, leading to larger values in the quadratic form. Besides, the mean shift $\Delta\boldsymbol{\mu}$ in the second term and the last cross-term can also lead to large values for $D^2(\mathbf{F}_o)$. For Gaussian-distributed $\mathbf{F}$, the likelihood of $\mathbf{F}_o$ decays exponentially with $D^2(\mathbf{F}_o)$:

$$p(\mathbf{F}_o) \propto \exp\left(-\frac{1}{2} D^2(\mathbf{F}_o)\right). \tag{17}$$

Therefore, the inflated $D^2(\mathbf{F}_o)$ from covariance mismatch, mean shift, and cross-term forces $p(\mathbf{F}_o)$ to be small, satisfying the low-likelihood criterion for outliers.

Here, we show mathematically that the expected Mahalanobis distance of $\mathbf{F}_o$ exceeds that of ID samples $\mathbf{F}$:

$$\mathbb{E}[D^2(\mathbf{F}_o)] > \mathbb{E}[D^2(\mathbf{F})], \tag{18}$$

where $D^2(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$. For ID samples $\mathbf{F}$, the squared Mahalanobis distance follows a chi-squared distribution with $2d$ degrees of freedom with expectation:

$$\begin{aligned} \mathbb{E}[D^2(\mathbf{F})] &= \mathbb{E}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1} (\mathbf{F} - \boldsymbol{\mu})(\mathbf{F} - \boldsymbol{\mu})^T\right)\right] \\ &= \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}) = \mathrm{Tr}(I_{2d}) = 2d, \end{aligned} \tag{19}$$

where $\mathrm{Tr}(\cdot)$ is the trace of a matrix. For outliers $\mathbf{F}_o$, from Eq. (16) we know $\mathbb{E}[D^2(\mathbf{F}_o)]$ is the sum of the expectation of three terms. Since $\mathbb{E}[\mathbf{F}_o - \boldsymbol{\mu}_o] = 0$, the expectation of the last term is 0 and:

$$\mathbb{E}[D^2(\mathbf{F}_o)] = \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_o) + (\boldsymbol{\mu}_o - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_o - \boldsymbol{\mu}). \tag{20}$$

Let $\Delta\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_o - \boldsymbol{\Sigma}$, the trace term becomes:

$$\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_o) = \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} + \Delta\boldsymbol{\Sigma})) = 2d + \mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\Sigma}). \tag{21}$$

Now we prove that Feature Mixing ensures $\mathrm{Tr}(\boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\Sigma}) \geq 0$. Assume $\boldsymbol{\Sigma}$ is diagonal (without loss of generality via eigendecomposition) with $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{2d}^2)$. Feature Mixing increases variances in dimensions with low σ_i^2 and decreases them in dimensions with high σ_i^2. Let:

$$\Delta\boldsymbol{\Sigma} = \mathrm{diag}(\Delta_1, \ldots, \Delta_{2d}), \quad \Delta_i = \sigma_{o,i}^2 - \sigma_i^2. \tag{22}$$

- For $\sigma_i^2 \leq \sigma_{o,i}^2$: $\Delta_i \geq 0$, and $\frac{\Delta_i}{\sigma_i^2}$ is large.
- For $\sigma_i^2 > \sigma_{o,i}^2$: $\Delta_i \leq 0$, and $\frac{\Delta_i}{\sigma_i^2}$ is small in magnitude.

The trace becomes:

$$\text{Tr}(\boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\Sigma}) = \sum_{i=1}^{2d} \frac{\Delta_i}{\sigma_i^2}. \tag{23}$$

The positive terms dominate because $\boldsymbol{\Sigma}^{-1}$ weights low-variance dimensions more heavily. Thus, $\text{Tr}(\boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\Sigma}) \geq 0$. For example, if $\sigma_i^2 = a$ increases to $\sigma_{o,i}^2 = b$ and $\sigma_j^2 = b$ decreases to $\sigma_{o,j}^2 = a$, where $0 < a \leq b$:

$$\begin{aligned} &\frac{\Delta_i}{\sigma_i^2} = \frac{b-a}{a}, \quad \frac{\Delta_j}{\sigma_j^2} = \frac{a-b}{b} \\ \Rightarrow &\frac{\Delta_i}{\sigma_i^2} + \frac{\Delta_j}{\sigma_j^2} = \frac{(b-a)^2}{ab} \geq 0. \end{aligned} \tag{24}$$

Since Feature Mixing randomly swaps features from two modalities, the probability of increasing or decreasing σ_i^2 is the same, and therefore $\text{Tr}(\boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\Sigma}) \geq 0$. The second term in Eq. (20) is strictly positive for $\boldsymbol{\mu}_o \neq \boldsymbol{\mu}$:

$$(\boldsymbol{\mu}_o - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_o - \boldsymbol{\mu}) > 0. \tag{25}$$

Combining all terms:

$$\begin{aligned} \mathbb{E}[D^2(\mathbf{F}_o)] = 2d + \underbrace{\text{Tr}(\boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\Sigma})}_{\geq 0} + \underbrace{(\boldsymbol{\mu}_o - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_o - \boldsymbol{\mu})}_{>0} \\ > 2d = \mathbb{E}[D^2(\mathbf{F})]. \end{aligned} \tag{26}$$

Since for Gaussian-distributed $\mathbf{F}$, the likelihood of $\mathbf{F}_o$ decays exponentially with $D^2(\mathbf{F}_o)$, $p(\mathbf{F}_o)$ is much smaller than $p(\mathbf{F})$ and therefore $\mathbf{F}_o$ lie in low-likelihood regions.

Theorem 2 *Outliers $\mathbf{F}_o$ are bounded in their deviation from $\mathbf{F}$, such that $|\mathbf{F}_o - \mathbf{F}|_2 \leq \sqrt{2N} \cdot \delta$, where $\delta = \max_{i,j} \left| \mathbf{F}_c^{(i)} - \mathbf{F}_l^{(j)} \right|$.*

Proof 2: While $\mathbf{F}_o$ is statistically anomalous, it remains geometrically proximate to $\mathbf{F}$ and is bounded in their deviation from $\mathbf{F}$. The Euclidean distance between $\mathbf{F}_o$ and $\mathbf{F}$ is:

$$|\mathbf{F}_o - \mathbf{F}|_2 = \sqrt{|\widetilde{\mathbf{F}}_c - \mathbf{F}_c|_2^2 + |\widetilde{\mathbf{F}}_l - \mathbf{F}_l|_2^2}. \tag{27}$$

For each modality, the deviation is bounded by the maximum feature difference $\delta = \max_{i,j} \left| \mathbf{F}_c^{(i)} - \mathbf{F}_l^{(j)} \right|$:

$$|\widetilde{\mathbf{F}}_c - \mathbf{F}_c|_2 = |\mathbf{M} \odot (\mathbf{F}_l - \mathbf{F}_c)|_2 \leq \sqrt{N} \cdot \delta, \tag{28}$$

$$|\widetilde{\mathbf{F}}_l - \mathbf{F}_l|_2 = |\mathbf{M} \odot (\mathbf{F}_c - \mathbf{F}_l)|_2 \leq \sqrt{N} \cdot \delta. \tag{29}$$

Therefore:

$$|\mathbf{F}_o - \mathbf{F}|_2 \leq \sqrt{(\sqrt{N} \cdot \delta)^2 + (\sqrt{N} \cdot \delta)^2} = \sqrt{2N} \cdot \delta. \tag{30}$$

Since $N \ll d$, $\sqrt{2N} \cdot \delta$ remains small, ensuring $\mathbf{F}_o$ stays near $\mathbf{F}$ and preserves *semantic consistency*. In conclusion, both $\mathbf{F}_o$ and $\mathbf{F}$ share the same embedding space, but $\mathbf{F}_o$ lies in low-likelihood regions of the distribution of $\mathbf{F}$.

B Additional Visualization Results

Fig. 10 to Fig. 12 present visualizations of multimodal OOD segmentation results across different datasets, showcasing the RGB image, 3D semantic ground truth, and predictions from various baselines. The baseline method struggles to identify OOD objects, whereas our method accurately segments OOD objects with minimal noise.

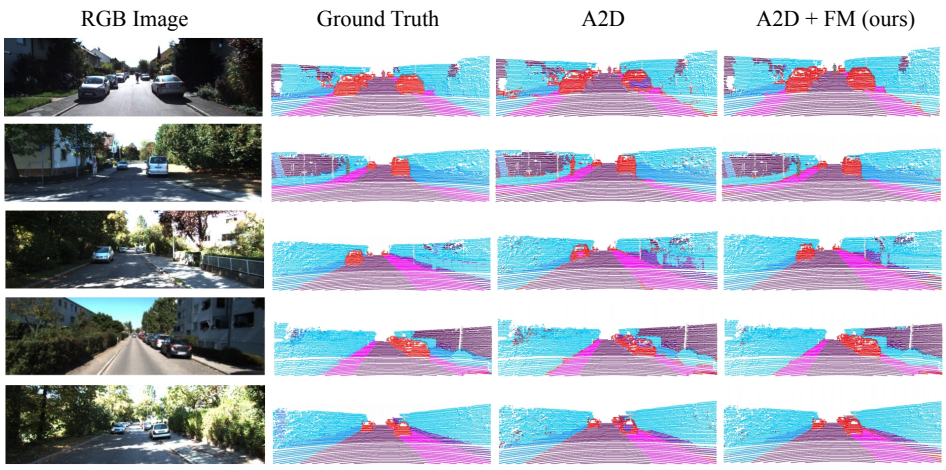


Figure 10: More visualizations on SemanticKITTI dataset. Points with red color are OOD objects. Our method accurately segments OOD objects, outperforming the baseline method.

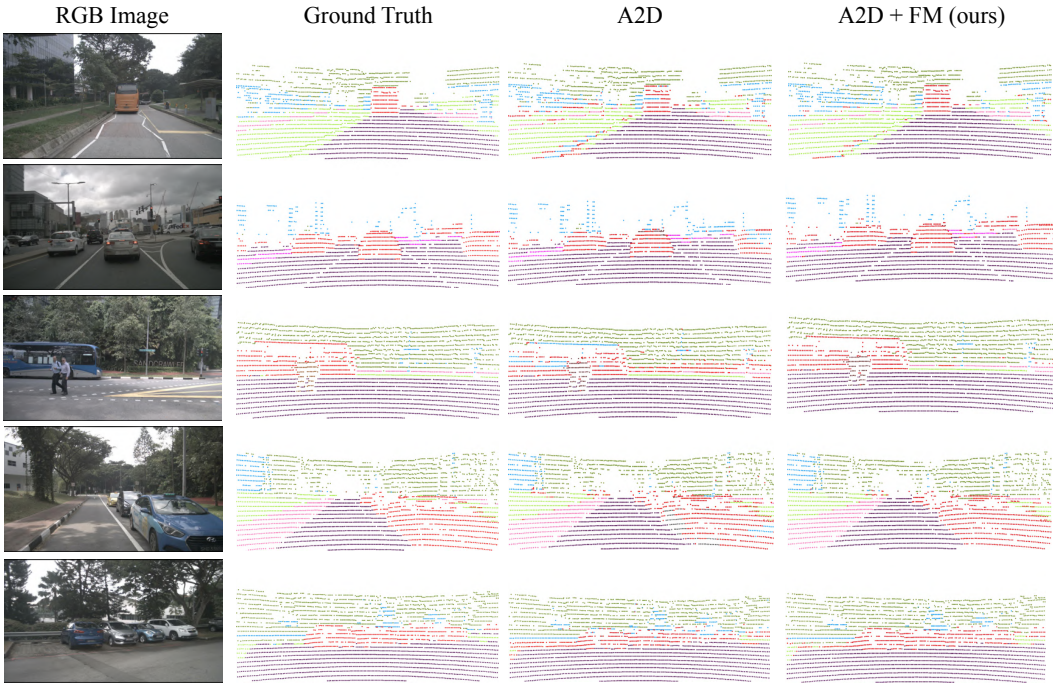


Figure 11: More visualizations on nuScenes dataset. Points with red color are OOD objects. Our method accurately segments OOD objects, outperforming the baseline method.

C More Details on Datasets

C.1 Realistic Datasets

We evaluate our approach on two widely-used autonomous driving datasets, including nuScenes [6] and SemanticKITTI [3]. Both datasets provide paired LiDAR point cloud and RGB image, along with point-level semantic annotations. The nuScenes dataset contains $28,130$ training frames and $6,019$ validation frames, annotated with 16 semantic classes. SemanticKITTI consists of $21,000$ frames from sequences 00-10 for training and validation, annotated with 19 semantic classes. Following [53, 7], we use sequence 08 for validation. The remaining sequences (00-07 and 09-10) are used for training. In our OOD segmentation setting, we follow [17] to map all *vehicle* categories to a single *unknown* class to represent out-of-distribution (OOD) objects in both datasets. Specifically, in

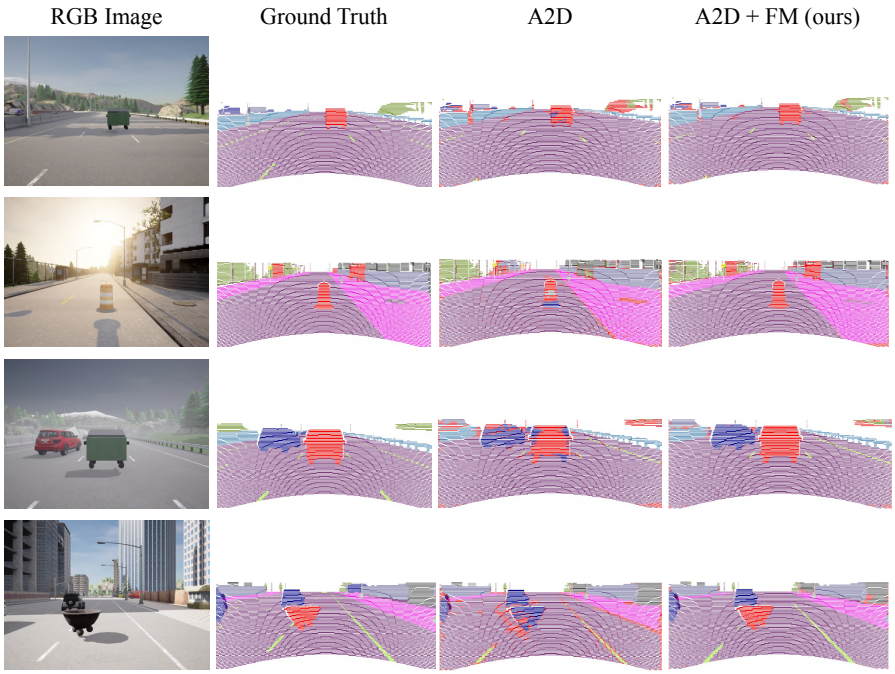


Figure 12: More visualizations on CARLA-OOD dataset. Points with red color are OOD objects. Our method accurately segments OOD objects, outperforming the baseline method.

SemanticKITTI, the categories *{car, bicycle, motorcycle, truck, and other-vehicle}* are remapped, while in nuScenes, the remapped categories include *{bicycle, bus, car, construction_vehicle, trailer, truck, and motorcycle}*. All other categories are retained as in-distribution (ID) classes and follow the standard segmentation settings. During training, we set the labels of OOD classes to void and ignore them. During inference, we aim to segment the ID classes with high Intersection over Union (IoU) and detect OOD classes as *unknown*.

C.2 Synthetic Dataset

Limitations of the realistic datasets. The evaluation of multimodal OOD segmentation on datasets like nuScenes [6] and SemanticKITTI [3] often relies on manually remapping existing categories to simulate OOD classes. This methodology, while prevalent, suffers from two significant drawbacks. Firstly, the categories designated as OOD are often common objects (e.g., vehicles, ground, structures) that may not faithfully represent the characteristics of genuine, unseen anomalies encountered in real-world scenarios. Secondly, despite these designated OOD classes being nominally ignored during training (e.g., by excluding them from loss computation for known classes), the model is nevertheless exposed to these objects within the training data. This creates a substantial risk of data leakage, as the model may implicitly learn characteristics of these supposedly 'unseen' OOD classes.

Inspired by the development of existing 2D OOD segmentation benchmarks, where models are trained on Cityscapes [9] and tested on Fishyscapes [5] with the same class setup but additional synthetic OOD objects, we create the CARLA-OOD dataset for multimodal OOD segmentation task. We use the KITTI-CARLA dataset [12] to train the base model, which is generated using the CARLA simulator [18] with paired LiDAR and camera data. The CARLA-OOD dataset aligns with the KITTI-CARLA sensor configurations but incorporates randomly placed OOD objects in diverse scenes and weather conditions for testing. The KITTI-CARLA dataset consists of 7 sequences, each containing $5,000$ frames captured from distinct CARLA maps and annotated with 22 classes for the LiDAR point cloud. We select 1,000 evenly sampled frames from each sequence, resulting in a total of $7,000$ frames for training and validation. The dataset is split into a training set (Town01, Town03–Town07) and a validation set (Town02), with testing performed on our CARLA-OOD dataset. The CARLA-OOD dataset consists of 245 paired LiDAR and camera samples captured across 5 CARLA maps and 6 weather conditions, each sample containing at least one OOD object.

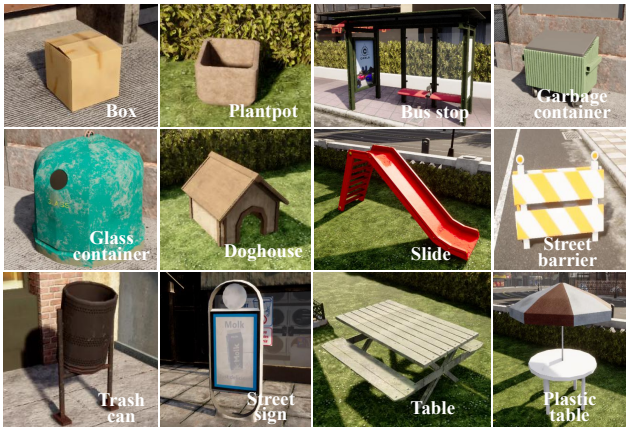


Figure 13: Example of OOD objects in our CARLA-OOD dataset.

Sensor	Position (x, y, z) [meters]	Configurations
LiDAR	(0, 0, 1.80)	Channels: 64 Range: 80.0 meters Upper FOV: 2 degrees Lower FOV: -24.8 degrees
RGB Camera	(0.30, 0, 1.70)	FOV: 72 degrees
Semantic Camera	(0.30, 0, 1.70)	FOV: 72 degrees

Table 6: Sensor configurations for CARLA-OOD dataset.

To avoid class overlap with KITTI-CARLA, 34 OOD objects are selected and randomly placed within the scenes during dataset generation. The dataset is annotated with 22 classes aligned with KITTI-CARLA, along with an additional *unknown* class for OOD objects.

C.3 Generation of CARLA-OOD Dataset

The CARLA-OOD dataset is created using the CARLA simulator, with a sensor setup aligned to the KITTI-CARLA dataset, consisting of a camera and a LiDAR on the ego-vehicle. Detailed sensor configurations are provided in Tab. 6, with positions defined relative to the ego-vehicle. Thirty-four obstacles from CARLA's dynamic and static classes are randomly placed in front of the ego-vehicle at varying distances as OOD objects (Fig. 13). The simulation spans diverse scenes (Town01, Town02, Town04, Town05, Town10) and weather conditions (e.g., clear, wet, foggy, sunshine, overcast), capturing both semantic and covariate shifts. The dataset includes RGB images with a resolution of $1392 * 1024$ pixels, LiDAR point cloud, point-level semantic labels, and transformation matrices between sensors.

C.4 MultiOOD benchmark

MultiOOD [17] is the first benchmark designed for Multimodal OOD Detection, comprising five action recognition datasets (EPIC-Kitchens [11], HAC [16], HMDB51 [30], UCF101 [45], and Kinetics-600 [28]) with over $85,000$ video clips, where video, optical flow, and audio are used as different types of modalities. Fig. 14 shows an example of the Far-OOD setup in MultiOOD. This setup considers an entire dataset as in-distribution (ID) and further collects datasets, which comprise similar tasks but are disconnected from any ID categories, as OOD datasets. In this scenario, both semantic and domain shifts are present between the ID and OOD samples. We follow the

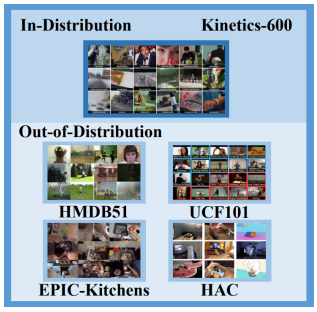


Figure 14: Multimodal Far-OOD setup in MultiOOD, where Kinetics-600 is the ID dataset and the other four datasets are OOD.

same setup and framework as proposed in MultiOOD for experiments. More details on the MultiOOD benchmark are in [17].

D Implementation Details

For the **Multimodal OOD Segmentation** task, we follow [17] to adopt the fusion framework from PMF [53], modifying it by adding an additional segmentation head to the combined features from the camera and LiDAR streams. We use ResNet-34 [22] as the backbone for the camera stream and SalsaNext [10] for the LiDAR stream. For optimization, we use SGD with Nesterov [41] for the camera stream and Adam [29] for the LiDAR stream. The networks are trained for 50 epochs with a batch size of 4, starting with a learning rate of 0.0005 and with a cosine schedule. To prevent overfitting, we apply various data augmentation techniques, including random horizontal flipping, random scaling, color jitter, 2D random rotation, and random cropping. For hyperparameters, we set N in Feature Mixing to 10 and γ_1 in loss to 3.0. For A2D, we set γ_2 to 1.0. For xMUDA, we set γ_2 to 0.5. For the **Multimodal OOD Detection** task, we conduct experiments across video and optical flow modalities using the MultiOOD benchmark [17]. We use the SlowFast network [20] to encode video data and the SlowFast network's slow-only pathway for optical flow. The models are pre-trained on each dataset's training set using standard cross-entropy loss. The Adam optimizer [29] is employed with a learning rate of 0.0001 and a batch size of 16. For hyperparameters, we set N in Feature Mixing to 512.

E Compatible with Cross-modal Training Techniques

Our proposed framework not only supports the basic late-fusion strategy, but is also compatible with advanced cross-modal training techniques that promote interaction across modalities. To demonstrate its versatility, we show how to integrate A2D [17] and xMUDA [26] into the framework.

Agree-to-Disagree (A2D), designed for multimodal OOD detection, aims to amplify the modality prediction discrepancy during training. It assumes additional outputs $\mathbf{O}^c$ and $\mathbf{O}^l$ from each modality. By removing the c-th value from $\mathbf{O}^c$ and $\mathbf{O}^l$, A2D derives new prediction probabilities without ground-truth classes, denoted as $\bar{\mathbf{O}}^c$ and $\bar{\mathbf{O}}^l \in \mathbb{R}^{M\times(C-1)}$. A2D then seeks to maximize the discrepancy between $\bar{\mathbf{O}}^c$ and $\bar{\mathbf{O}}^l$, which is defined as:

$$\mathcal{L}_{A2D} = -\frac{1}{M}\sum_{m=1}^{M} D(\bar{\mathbf{O}}^c_m, \bar{\mathbf{O}}^l_m), \tag{31}$$

where $D(\cdot)$ is a distance metric quantifying the similarity between two probability distributions. By integrating A2D into the framework, the final loss function becomes:

$$\mathcal{L} = \mathcal{L}_{foc} + \mathcal{L}_{lov} + \gamma_1\mathcal{L}_{ent} + \gamma_2\mathcal{L}_{A2D}. \tag{32}$$

xMUDA facilitates cross-modal learning by encouraging information exchange between modalities, allowing them to learn from each other. xMUDA also assumes additional outputs $\mathbf{O}^c$ and $\mathbf{O}^l$ from each modality and define cross-modal loss as:

$$\mathcal{L}_{\text{xM}} = \mathbf{D}_{KL}(\mathbf{O}^c||\mathbf{O}) + \mathbf{D}_{KL}(\mathbf{O}^l||\mathbf{O}), \tag{33}$$

where $\mathbf{D}_{KL}$ is the Kullback–Leibler divergence and the final loss function in this case is:

$$\mathcal{L} = \mathcal{L}_{foc} + \mathcal{L}_{lov} + \gamma_1\mathcal{L}_{ent} + \gamma_2\mathcal{L}_{\text{xM}}. \tag{34}$$

F More Ablation Studies

Hyperparameter Sensitivity. We evaluate the sensitivity of γ_1 in the loss function on the SemanticKITTI dataset. Our findings, as illustrated in Fig. 15, demonstrate that training with multimodal outlier generation and optimization consistently outperforms the baseline across all parameter settings. These ablations suggest that our approach is robust and exhibits minimal sensitivity to variations in hyperparameter choices.

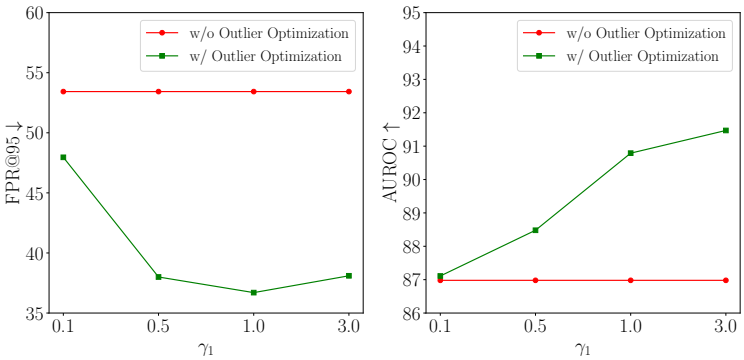


Figure 15: Ablation of γ_1 in the loss function on SemanticKITTI.

OOD Scores	FPR@95↓	AUROC↑	AUPR↑
MaxLogit [23]	31.76	92.83	61.99
MSP [24]	32.93	91.57	51.68
Energy [36]	32.05	92.87	64.87
Entropy [13]	33.00	92.50	59.11
GEN [37]	32.63	93.00	64.43

Table 7: Ablation of OOD Scores on SemanticKITTI.

Impact of Various OOD Scores. We evaluate the impact of different commonly used OOD scores by replacing the OOD detection module in our framework with MaxLogit [23] (our default), MSP [24], Energy [36], Entropy [13], and GEN [37]. As shown in Appendix F, the FPR@95 and AUROC show minimal fluctuation (less than 2%) across different OOD scores, further demonstrating the adaptability of our framework to various design choices.

Multimodal OOD Detection Results on Kinetics-600. Tab. 8 presents the multimodal OOD detection results using Kinetics-600 as the ID dataset. Feature Mixing outperforms other outlier generation methods in most cases, achieving the lowest FPR@95 of 54.75% and the highest AUROC of 79.23% on average when using HMDB51 as ID. These results further demonstrate the effectiveness of Feature Mixing in improving OOD detection across diverse tasks and modalities, while maintaining negligible impact on ID accuracy.

Feature Mixing in Unimodal Settings for OOD Segmentation. We further extend the unimodal FM variant to the OOD segmentation task using LiDAR-only input. The same strategy used for unimodal classification is directly applied to LiDAR point cloud features. As shown in Tab. 9a, our method achieves competitive performance compared to NP-Mix and baseline methods, demonstrating the scalability of FM to segmentation tasks in unimodal settings with minimal modification.

Sensitivity of Parameter N on SemanticKITTI. We further investigate the sensitivity of FM to the mixing parameter N (i.e., the number of swapped feature dimensions) on the SemanticKITTI dataset. As shown in Tab. 9b, FM achieves stable performance across different values of N, supporting the robustness of the method. We use N=10 by default in our main experiments.

G Further Discussions on Feature Mixing

Challenges of extending existing solutions to multimodal setting. While methods such as Mixup and VOS have shown success in unimodal tasks, directly applying them to multimodal data is non-trivial due to the following key challenges:

(1) *Modality heterogeneity*: Multimodal data often involves inputs with fundamentally different structures and distributions—e.g., dense 2D images, sparse 3D point clouds, temporal audio signals, etc. Pixel-level interpolation techniques like Mixup are inherently incompatible across heterogeneous modalities, making it difficult to define meaningful cross-modal augmentations. Moreover, Mixup's

Methods	OOD Datasets										ID ACC ↑
	HMDB51		**UCF101**		**EPIC-Kitchens**		**HAC**		Average		
	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	
Baseline	72.64	71.75	70.12	71.49	43.66	82.05	61.50	74.99	61.98	75.07	73.14
Mixup [52]	66.53	70.33	68.36	69.53	39.68	86.62	56.35	77.95	57.73	76.11	73.40
VOS [19]	65.23	71.48	68.29	73.97	38.12	86.05	56.11	79.02	56.94	77.63	73.07
NPOS [48]	65.72	70.93	68.29	73.97	35.13	86.78	55.89	80.49	56.26	78.04	73.49
NP-Mix [17]	63.27	74.17	67.20	74.50	34.07	87.49	56.69	80.20	55.31	79.09	73.67
Feature Mixing (ours)	62.86	74.32	67.74	74.38	33.51	87.64	54.89	80.58	**54.75**	**79.23**	73.67

Table 8: Multimodal OOD Detection using video and optical flow, with **Kinetics-600** as ID. Energy is used as the OOD score.

Method	**FPR@95↓**	**AUROC↑**	**AUPR↑**	**mIoU↑**
Baseline	47.03	85.82	36.06	59.81
NP-Mix	48.26	86.79	45.73	59.52
Feature Mixing	**46.40**	**88.81**	**47.67**	**60.04**

(a) Unimodal feature mixing for OOD segmentation on SemanticKITTI (LiDAR only).

N	**FPR@95↓**	**AUROC↑**	**AUPR↑**	**mIoU↑**
8	38.58	91.14	53.45	**61.59**
10	**38.10**	**91.47**	**58.74**	61.18
12	37.17	90.90	55.19	60.75

(b) Ablation of N for OOD segmentation on SemanticKITTI.

Table 9: More ablations on the OOD segmentation task.

random linear blending can produce ambiguous or noisy samples that may lie near or within the in-distribution (ID) manifold, reducing its effectiveness for OOD detection.

(2) *Computational inefficiency*: Methods like VOS and NP-Mix rely on distribution estimation, sampling, or searching in high-dimensional feature spaces, which becomes computationally expensive when extended to multimodal settings—especially for dense prediction tasks such as segmentation, where per-pixel or per-point processing is required.

Our Feature Mixing addresses these challenges by (i) operating entirely in the feature space, where modality-specific structures have already been abstracted, (ii) introducing a lightweight, swap-based perturbation strategy to synthesize outliers efficiently, avoiding costly sampling or searching, (iii) being theoretically grounded (Theorems 1 and 2) and empirically validated to maintain ID-OOD separation while improving efficiency and scalability.

Clarification on the properties of Feature Mixing and their benefit to OOD detection. Theoretical results in Theorems 1 and 2 establish two key properties of Feature Mixing: (1) the synthesized outliers have low likelihood under the ID distribution, and (2) their deviation from the ID distribution is bounded. These properties are important for OOD detection and segmentation, as they ensure that *FM produces challenging but plausible outliers that populate the low-density regions near the boundary of the ID distribution*. This supports effective decision boundary regularization and reduces model overconfidence on unseen data. As shown in Fig. 4, the FM-generated outliers are well-separated from the ID features in the embedding space. This separation helps the model to better distinguish between in-distribution and out-of-distribution regions. These visualizations provide intuitive, empirical evidence that our method generates meaningful and bounded outliers, as predicted by our theoretical analysis.

More details on outlier optimization. The core idea is that the feature mixing branch generates pseudo-OOD features. We apply an entropy maximization loss in Eq. (5) on these features to explicitly encourage the model to produce uncertain (i.e., high-entropy) predictions for them. This discourages the model from making overconfident predictions on ambiguous or unfamiliar inputs—behavior that is characteristic of OOD samples. In contrast, the classification loss encourages low-entropy (i.e., high-confidence) predictions on in-distribution (ID) samples. Together, these complementary objectives train the model to develop a confidence-based decision boundary: ID samples are associated with confident predictions, while pseudo-OOD samples—introduced through feature mixing—are explicitly pushed toward uncertain predictions. As shown in Fig. 2, this training objective increases the separation in predictive confidence between ID and OOD inputs, which is crucial for effective OOD detection.