Benchmarking Vision Foundation Models for Input Monitoring in Autonomous Driving

Mert Keser[1,2]
mert.keser@tum.de
Halil Ibrahim Orhan[2]
ibrahim.orhan@tum.de
Niki Amini-Naieni[3]
niki.amini-naieni@eng.ox.ac.uk
Gesina Schwalbe[4]
gesina.schwalbe@uni-luebeck.de
Alois Knoll[2]
k@tum.de
Matthias Rottmann[5]
matthias.rottmann@uos.de

[1] Continental AG, Germany
[2] Technical University of Munich, Germany
[3] University of Oxford, United Kingdom
[4] University of Lübeck, Germany
[5] Osnabrück University, Germany

Abstract

Deep neural networks (DNNs) remain challenged by distribution shifts in complex open-world domains like automated driving (AD): Robustness against yet unknown novel objects (semantic shift) or styles like lighting conditions (covariate shift) cannot be guaranteed. Hence, reliable operation-time monitors for identification of out-of-training-data-distribution (OOD) scenarios are imperative. Current approaches for OOD classification are untested for complex domains like AD, are limited in the kinds of shifts they detect, or even require supervision with OOD samples. We instead establish a framework around a principled, unsupervised and model-agnostic method that unifies detection of semantic and covariate shifts: Find a full model of the training data's feature distribution, to then use its density at new points as in-distribution (ID) score. To implement this, we propose to combine Vision Foundation Models (VFMs) as feature extractors with density modeling techniques. Through a comprehensive benchmark of 4 VFMs with different backbone architectures and 5 density-modeling techniques against established baselines, we provide the first systematic evaluation of OOD classification capabilities of VFMs across diverse conditions. VFMs embeddings with density estimation outperform existing approaches in identifying OOD inputs. Additionally, we show that our method detects high-risk inputs likely to cause errors in downstream tasks, thereby improving overall performance. Overall, VFMs, when coupled with robust density modeling techniques, are promising to realize model-agnostic, unsupervised, reliable safety monitors in complex vision tasks.

1 Introduction

In recent decades, deep neural networks (DNNs) have become widely used for environmental perception due to their superior performance compared to traditional computer vision

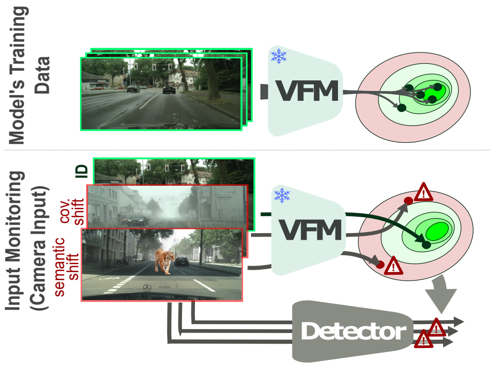

Figure 1: An input monitoring system for autonomous driving using a pre-trained Vision Foundation Model as an image encoder to detect distribution shifts. The system compares incoming images (bottom row) against the model's training distribution (top row) to identify whether they fall within the in-distribution (ID, green) or exhibit covariate (*e.g.*, fog) or semantic (*e.g.*, unseen objects like animals) shifts. Red dots in the embedding space indicate out-of-data-distribution inputs. The resulting flags inform the downstream detector to assess potential failure risks.

methods [12, 14, 29, 31, 38, 45, 54]. However, in complex open-world domains such as autonomous driving (AD) this performance is inevitably challenged by the long tail of the input distribution, i.e., unforeseeable out-of-distribution (OOD) scenarios for which DNN generalization cannot be guaranteed. AD spans multiple operational levels, from driver assistance (L1-L2) to fully autonomous operation (L4-L5), with conditional automation (L3) bridging the transition [42]. Across these levels, camera-based perception systems support critical functions, from lane-keeping and traffic sign recognition to full scene understanding [8]. As these systems operate in increasingly diverse environments, a fundamental question emerges: *How can we ensure that perception models function reliably in complex, open-world scenarios, particularly when encountering inputs that deviate from their training distribution?* To mitigate these risks and ensure safe operation, reliable monitors that can detect and respond to such deviations have become mandatory [16, 25]. An important approach for OOD detection is input monitoring (fig. 1), which identifies when input data fed to the system deviates from the training data distribution [41].

Existing binary OOD classification methods suffer from critical limitations for AD applications: they often require direct access to model internals, are specialized to specific architectures, or need exposure to anticipated shift types during training [23, 30, 33, 47, 53]. These constraints compromise reliability when encountering truly unforeseen distribution shifts in open-world environments Furthermore, accessing internal model layers becomes impractical in AD systems. A more principled approach would model the complete training data distribution in feature space, using density values or likelihoods as in-distribution (ID) scores for new samples. The challenge for complex visual domains lies in extracting sufficiently powerful features—standard DNNs and autoencoders (AE) trained on limited datasets typically fail to capture the necessary representational richness.

This is precisely where VFMs offer a compelling solution. VFMs encode comprehensive semantic contexts and subtle visual distinctions within their latent representations [9, 36], making them particularly well-suited for distribution modeling. Their appeal for input monitoring in AD stems from demonstrated strong generalization capabilities across diverse visual tasks and novel scenarios [24, 35], potentially enabling robust detection of unforeseen distribution shifts without requiring model-specific adaptations or prior exposure to OOD examples.

In this study we investigate: *How suitable are VFM image encodings for modeling a data distribution for OOD scenario detection in complex vision domains like AD?* In particular: *How do they compare to standard DNN encodings and SoTA baseline binary-OOD detection methods? Is this OOD of use for improving overall system reliability in downstream applications?* Also, while most existing works focuses on either semantic OOD detection on object level [4, 10, 11] or on covariate shifts detection [5, 21], we are interested in monitoring capabilities for *both* semantic, covariate, and combined shift detection on *scene* level. Our **contributions** are:

(i) We propose a **modular pipeline** for **unsupervised**, **model-agnostic** detection of **distribution shifts** in input images (see fig. 1), in particular both **semantic, covariate, and joint shifts**. At its core we use rich VFM feature encodings for distribution density modeling.

(ii) We for the first time **benchmark** VFMs for OOD on the challenging open-world domain of **AD**. Our extensive tests compare 4 VFMs on 12 both synthetic and real datasets, and 7 OOD methods as baselines. Our results show the clear superiority of VFM encodings in unified distribution shift detection.

(iii) We demonstrate **practical usability** on a downstream AD perception application, showcasing that, despite being model-agnostic, VFMs successfully **identify OOD samples with higher risk of errors**.

(iv) We establish a comprehensive **benchmarking framework** for evaluating binary OOD classification methods in AD by providing diverse experimental scenarios combining different types of distribution shifts, datasets, and evaluation protocols.

2 Related Work

2.1 Binary-OOD Detection Methods

Binary OOD detection methods aim to distinguish between ID and OOD data. Existing methods can be categorized based on their detection mechanisms. Feature-based methods like GradNorm [23] analyze network activations and gradients to identify distribution shifts. Logit-based approaches, including Energy [30], Entropy [33], and MaxLogit [53], examine output characteristics of the neural network to detect anomalies. Confidence-based methods such as DICE [46] leverage confidence scores derived from the model's predictions. Distance-based approaches like KNN [47] assess feature space similarities to determine whether inputs deviate from the training distribution. ODIN [27] enhances OOD detection through temperature scaling and input preprocessing.

These methods have primarily been evaluated on standard classification datasets with artificial perturbations (*e.g.* Gaussian noise), but rarely tested on complex real-world problems with natural distribution shifts like those encountered in AD [3]. Most approaches require direct access to model internals [23, 30, 47]. This architectural dependency limits their applicability in AD systems where downstream network architectures may be proprietary or inaccessible. Additionally, these approaches require supervision with anticipated shift types during training [3], which compromises reliability when encountering truly unforeseen distribution shifts in open-world environments.

2.2 Vision Foundation Models

FMs are large-scale machine learning models trained on diverse data, allowing them to adapt to varied tasks beyond their original objective [26, 50]. The field includes models like CLIP [40], trained contrastively on image-text pairs; DINO [9], using self-supervised learning with self-distillation; DINOv2 [36], enhancing DINO with masked image modeling; Grounding-DINO [29], extending DETR-based detection [52] with text guidance; InternImage [49], utilizing deformable convolutions for adaptive receptive fields; Florence-2 [51], employing hierarchical visual tokenization; and LLaVA [28], connecting vision encoders to large language models through visual instruction tuning. This study focuses on the first four VFMs, selected for their architectural diversity (spanning Convolutional Networks, Vision Transformers, and Swin Transformers) and absence of test data leakage (see Supplementary). For a comprehensive overview of FMs, we refer to [2, 6], while [53] provides a detailed review of VFMs specifically in AD. Note that, while CLIP [40] and GroundingDINO [29] may also accept language inputs, we here for now only consider their image encodings.

3 Shift Detection Methodology & Evaluation

3.1 Shift Detection Methodology: Overview

We propose a method to detect distribution shift in image data using a DNN $F: x \mapsto f$ for feature extraction (a VFM, standard DNN, or AE). Concretely, we aim to classify whether a new image x^* is an outlier with respect to the distribution of a given (monitor training) dataset $\mathcal{X}_{\text{monitor}}$ (the ID data). In our case we choose $\mathcal{X}_{\text{monitor}} = \mathcal{X}_{\text{perception}}$ (e.g., Cityscapes train set [15]), the dataset the perception system is trained on. During monitor **training**, the monitor is then parametrized on the (fixed) features $\mathcal{F}_{\text{monitor}} := F(\mathcal{X}_{\text{monitor}})$ to capture the distribution $\mathcal{F}_{\text{monitor}}$. From the distribution model we then derive a function c that provides an ID score $c(x^*) = \text{score}(f^*)$ for any new input x^*, based on its similarity to or density in the monitor data feature distribution. For **testing**, we combine an ID dataset (e.g., Cityscapes validation set [15]) with an OOD dataset (e.g., Foggy Cityscapes [44]), both containing samples not seen by the monitor previously: $\mathcal{X}_{\text{test}} = \mathcal{X}_{\text{ID}} \cup \mathcal{X}_{\text{OOD}}$. The OOD samples differ from $\mathcal{X}_{\text{monitor}}$ through either semantic shifts (e.g., novel object classes) or covariate shifts (e.g., varying lighting conditions or adverse weather).

We consider the following **choices for the monitor** distribution modeling further detailed in the following in order of increasing complexity: 1) instance-based: averaged pairwise similarities to the training examples and its computationally efficient derivation using similarity to the mean feature vector, 2) a one-class support vector machine, 3) a Gaussian mixture model and 4) a normalizing flow.

Average Pairwise Similarity (APS) and Mean Feature Similarity (MFS): We use cosine similarity to measure how closely a test image's feature vector aligns with those from the training set $\mathcal{X}_{\text{monitor}}$. Cosine similarity is well-suited for high-dimensional, normalized feature spaces, as it (i) prioritizes direction over magnitude, (ii) remains invariant to scaling, and (iii) aligns with the latent geometry of many VFMs.

Given a test sample x^* with feature vector f^*, APS computes the average similarity to all n training vectors:

$$c_{\text{APS}}(x^*) = \frac{1}{n}\sum_{i=1}^{n}\cos(f^*, f_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{f^* \cdot f_i}{\|f^*\|\,\|f_i\|} \tag{1}$$

To reduce computational overhead, we introduce MFS, which compares f^* to the mean training vector $\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$:

$$c_{\text{MFS}}(x^*) = \cos(f^*, \bar{f}) = \frac{f^* \cdot \bar{f}}{\|f^*\| \, \|\bar{f}\|} \tag{2}$$

This reduces inference cost from $O(n)$ to $O(1)$, making MFS efficient for real-time monitoring.

One-Class Support Vector Machine (OC-SVM): OC-SVM [1, 7] is a classical unsupervised method for novelty detection, effective when abundant data is available from a single (in-distribution) class. We use OC-SVM to model feature vectors $\mathcal{F}_{\text{monitor}} = \{f_1, \dots, f_n\}$ extracted from $\mathcal{X}_{\text{monitor}}$ and score new samples from $\mathcal{X}_{\text{ID}} \cup \mathcal{X}_{\text{OOD}}$.

The model learns a hyperplane in a high-dimensional space (via a kernel map ϕ) that separates the origin from the training data with maximal margin, solving:

$$\min_{w,\xi,\rho} \quad \frac{1}{2}\|w\|^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \xi_i - \rho \tag{3}$$

$$\text{s.t.} \quad (w \cdot \phi(f_i)) \geq \rho - \xi_i, \quad \xi_i \geq 0 \tag{4}$$

Here, w defines the hyperplane, ρ the offset, ξ_i slack variables, and $\nu \in (0, 1]$ adjusts tolerance to outliers.

Given a test sample x^* with feature f^*, we compute:

$$c_{\text{OC-SVM}}(x^*) = (w \cdot \phi(f^*)) - \rho \tag{5}$$

Lower values of $c_{\text{OC-SVM}}$ indicate greater deviation from the training distribution, i.e., higher likelihood of being OOD.

Gaussian Mixture Model (GMM): We model the feature distribution $\mathcal{F}_{\text{monitor}}$ using a Gaussian Mixture Model (GMM) [34] with K components:

$$p(f_i \mid \Theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(f_i \mid \mu_k, \Sigma_k) \tag{6}$$

Here, $\Theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ includes mixing weights π_k (with $\sum_k \pi_k = 1$), means μ_k, and covariances Σ_k for each Gaussian.

Parameters are estimated via the Expectation-Maximization (EM) algorithm. To select K, we apply the Akaike Information Criterion (AIC) over candidate models; details are provided in the supplementary.

Given a test sample x^* with feature f^*, we compute its log-likelihood under the trained GMM:

$$c_{\text{GMM}}(x^*) = \log p(f^* \mid \Theta) = \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(f^* \mid \mu_k, \Sigma_k)\right) \tag{7}$$

Higher scores indicate stronger conformity to the in-distribution density.

Normalizing Flows (NF): Normalizing Flows [18, 37] enable tractable density estimation by learning an invertible transformation g_θ that maps feature vectors $f \in \mathcal{F}_{\text{monitor}}$ to a standard normal distribution:

$$P_F \circ g_\theta^{-1} \approx \mathcal{N}(\cdot \mid 0, I) \tag{8}$$

The density at a test point f^* is computed via change-of-variables:

$$p(f^*) = \mathcal{N}(g_\theta(f^*) \mid 0, I) \cdot \left|\det\left(J_f g_\theta(f^*)\right)\right| \tag{9}$$

where $J_f g_\theta$ is the Jacobian of g_θ at f^*.

We implement g_θ using Real-NVP [18], an architecture with affine coupling layers. Each layer splits input into x_1 and x_2 via a binary mask and applies:

$$y_1 = x_1 \qquad\qquad y_2 = x_2 \odot \exp(s(x_1)) + t(x_1) \tag{10}$$

Here, $s(\cdot)$ and $t(\cdot)$ are neural networks for scaling and translation, and $\odot$ denotes element-wise multiplication. The log-determinant simplifies to $\log|\det(J)| = \sum_i s(x_1)_i$, enabling efficient likelihood evaluation.

To score test samples, we normalize the negative log-likelihoods to the [0, 1] range—higher values indicate increased OOD likelihood.

4 Experiments

This section presents a comprehensive empirical study evaluating the ability of VFMs to detect distributional shifts in the autonomous driving (AD) domain. We systematically assess unsupervised shift detection methods—based on modeling the distribution of $\mathcal{F}_{\text{monitor}}$—across semantic, covariate, and combined shift scenarios (see Supplementary for combined results). Our approach is benchmarked against state-of-the-art OOD detection methods and ImageNet-trained counterparts with identical architectures. Finally, we demonstrate downstream applicability through integration with a semantic segmentation task.

Metrics: For each test image $x^* \in \mathcal{X} = \mathcal{X}_{\text{ID}} \cup \mathcal{X}_{\text{OOD}}$, we compute a confidence score $c(x^*)$ using one of the models outlined above. Higher values of $c(x^*)$ indicate a greater likelihood that x^* belongs to the ID data represented by the set $\mathcal{X}_{\text{ID}}$. To evaluate the model's performance in distinguishing between ID and OOD samples, we consider the following binary classification metrics:

Area Under the Receiver Operating Characteristic Curve, AUROC: The ROC curve plots true against false positive rate for varying decision thresholds. An AUROC value of 1 indicates perfect classification, while 0.5 corresponds to random guessing.

Area Under the Precision-Recall Curve, AUPR: The AUPR is particularly useful for performance evaluation on class-imbalanced test datasets.

False Positive Rate (FPR) at 95% True Positive Rate (TPR), FPR95: This metric measures the FPR for the threshold at which the TPR is 95%. Specifically, we determine the threshold t that achieves a 95% TPR for ID samples and compute the corresponding FPR for OOD samples at t. Lower FPR95 values mean fewer OOD samples are misclassified as ID, i.e., better OOD detection performance.

Datasets: To investigate semantic shifts, we use the Cityscapes train set [15] as $\mathcal{X}_{\text{monitor}}$, and its validation set as $\mathcal{X}_{\text{ID}}$. OOD data comes from: (1) Lost and Found [39], which captures real-world road hazards in environmental conditions matching Cityscapes, and (2) bravo-synobjs [32], which provides synthetic anomalies through generative inpainting of Cityscapes. For covariate shifts, we maintain semantic consistency while varying visual characteristics. Using the same Cityscapes configuration, we evaluate on: Foggy Cityscapes [43] (synthetic fog), bravo-synrain/synflare [32] (synthetic rain/lens flare), and ACDC [44] (fog, night, rain, snow conditions). For semantic and covariate shifts, we use SegmentMeIfYouCan (SMIYC) [10] and the Indian Driving Dataset (IDD) [48] which contain both novel objects and environmental variations. See Supplementary for examples and visualizations.

Pre-trained VFMs and Implementation Details: We employ a diverse selection of pre-trained VFMs, including CLIP [40], DINO [9], DINOv2 [36], and Grounding DINO [29], each representing a distinct paradigm in visual representation learning.

For CLIP [40] we include both ResNet variants (RN50, RN101, RN50x4/x16/x64) and Vision Transformers (ViT-B/32, ViT-B/16, ViT-L/14)[1]. DINO is tested with multiple ViT configurations (ViT-S/16, ViT-S/8, ViT-B/16, ViT-B/8) and ResNet (RN50)[2]. DINOv2 [36] is represented by Vision Transformer variants of increasing model capacity (Dinov2-ViT-S/14, Dinov2-ViT-B/14, Dinov2-ViT-L/14, Dinov2-ViT-G/14)[3]. Grounding DINO [29] uses Swin Transformer [31] architecture, providing comparison with the ViT-based approaches[4].

For baseline comparisons, we include ImageNet-trained counterparts [17]: ResNets (RN50, RN101, RN152) [20], ViTs (ViT-B/16, ViT-L/16) [19], and Swin Transformers (SwinB, SwinT)[5] [31]. Additionally, we trained autoencoders (AE) on $\mathcal{X}_{\text{monitor}}$ dataset to serve as feature extractors for further comparison. Architectural specifications, detailed comparisons with ImageNet-trained counterparts, and implementation considerations are presented in the supplementary material.

4.1 Semantic Shift

Table 1 shows the two top-performing backbones of each model for semantic shift detection (find the complete comparison in the Supplementary). Observations are discussed in below.

Architecture-Specific Modeling: Grounding DINO with a Swin Transformer backbone achieves strong performance when paired with normalizing flows (e.g., Lost and Found: FPR95 0.16; Bravo-Synobj: 0.02). Its hierarchical attention enables richer semantic encoding—but only when supported by appropriate modeling. Simpler methods (APS, MFS, OC-SVM) perform poorly with this architecture (FPR95: 0.47–0.54), highlighting the need to align modeling strategy with architecture design.

Real vs. Synthetic Robustness: CLIP's RN50x64 shows stable performance with GMM and OC-SVM on real-world anomalies (FPR95: 0.23–0.24 on Lost and Found) but degrades on synthetic inputs (0.53–0.54 on Bravo-Synobj). In contrast, Grounding DINO with NF generalizes well across both, indicating robustness to the nature of the shift source.

[1]CLIP: https://github.com/openai/CLIP
[2]DINO: https://github.com/facebookresearch/dino
[3]DINOv2: https://github.com/facebookresearch/dinov2
[4]Grounding DINO: https://github.com/IDEA-Research/GroundingDINO
[5]ImageNet-trained ResNets: https://github.com/pytorch/pytorch; ViT and Swin Transformers: https://github.com/huggingface/pytorch-image-models

Model	Backbone	Latent Space Dim.	Lost and Found														
			APS			MFS			OC-SVM			GMM			NF		
			AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓
CLIP	RN50x64	1024	**0.94**	**0.93**	**0.24**	**0.93**	**0.93**	**0.27**	**0.94**	**0.93**	**0.24**	**0.94**	**0.93**	**0.23**	0.88	0.84	0.37
	ViT-B/16	512	0.88	0.86	0.38	0.86	0.85	0.41	0.89	0.88	0.43	0.91	0.89	0.33	0.91	0.88	0.34
DINO	ViT-B/16	768	0.79	0.75	0.52	0.72	0.64	0.60	0.79	0.76	0.59	0.84	0.82	0.55	0.75	0.71	0.76
	ViT-B/8	768	0.81	0.76	0.49	0.73	0.64	0.59	0.77	0.75	0.56	0.87	0.86	0.45	0.82	0.81	0.59
DINOv2	ViT-L/14	1024	0.89	0.87	0.43	0.91	0.88	0.37	0.84	0.82	0.68	0.82	0.77	0.52	0.83	0.78	0.54
	ViT-G/14	1536	0.90	0.87	0.30	0.89	0.86	0.37	0.85	0.82	0.63	0.86	0.78	0.39	0.83	0.75	0.46
Grounding DINO	SwinB	768	0.88	0.84	0.47	0.82	0.76	0.52	0.84	0.80	0.54	0.91	0.89	0.32	0.95	0.94	**0.16**
	SwinT	768	0.85	0.80	0.53	0.79	0.73	0.60	0.83	0.79	0.59	0.92	0.90	0.28	**0.96**	**0.95**	**0.16**
			Bravo-Synobj														
CLIP	RN50x64	1024	**0.93**	**0.94**	0.54	**0.93**	**0.94**	0.54	**0.93**	**0.95**	**0.54**	**0.94**	**0.95**	0.53	0.94	0.96	0.43
	ViT-B/16	512	0.91	0.93	**0.50**	0.91	0.93	**0.50**	0.91	0.93	**0.54**	0.92	0.94	**0.48**	0.93	0.95	0.49
DINO	ViT-B/16	768	0.81	0.82	0.69	0.81	0.82	0.69	0.83	0.85	0.75	0.86	0.88	0.60	0.84	0.87	0.80
	ViT-B/8	768	0.82	0.84	0.64	0.82	0.84	0.64	0.83	0.86	0.70	0.86	0.89	0.59	0.88	0.91	0.65
DINOv2	ViT-L/14	1024	0.78	0.79	0.77	0.78	0.79	0.77	0.78	0.81	0.80	0.82	0.85	0.74	0.86	0.88	0.63
	ViT-G/14	1536	0.79	0.80	0.73	0.79	0.80	0.73	0.81	0.83	0.69	0.82	0.85	0.71	0.86	0.89	0.69
Grounding DINO	SwinB	768	0.60	0.56	0.83	0.61	0.56	0.82	0.58	0.54	0.85	0.76	0.68	0.58	**0.99**	**0.99**	**0.02**
	SwinT	768	0.57	0.54	0.89	0.58	0.54	0.90	0.57	0.54	0.91	0.64	0.58	0.78	0.91	0.84	0.20

Table 1: Comparison of VFMs with various backbones on the Lost and Found [39] and Bravo-Synobj[32] datasets. focusing on semantic shift sensitivity. Results show AUROC↑. AUPR↑ (higher is better) and FPR95↓ (lower is better) metrics. Best values per metric in bold.

4.2 Covariate Shift

As previously, Table 2 present results for only the two top-performing backbones per model in terms of FPR95 scores, focusing on GMM and NF methods due to their superior detection capabilities and FPR95's sensitivity in distinguishing architectural performance.

Model	Backbone	Latent Space Dim.	Foggy Cityscapes (i=0.05)		Foggy Cityscapes (i = 0.2)		Bravo-Synflare		Bravo-Synrain		ACDC Fog		ACDC Night		ACDC Rain		ACDC Snow	
			FPR95↓		FPR95↓		FPR95↓		FPR95↓		FPR95↓		FPR95↓		FPR95↓		FPR95↓	
			GMM	NF	GMM	NF	GMM	NF	GMM	NF	GMM	NF	GMM	NF	GMM	NF	GMM	NF
CLIP	RN50x64	1024	0.76	0.67	0.26	0.21	0.62	0.47	0.01	0.04	**0.06**	**0.04**	0.13	0.23	**0.26**	**0.22**	**0.05**	**0.03**
	ViT-B/16	512	**0.57**	0.48	0.18	0.10	0.43	0.35	0.20	0.12	0.07	0.07	0.34	0.16	0.27	0.25	0.05	0.05
DINO	ViT-B/16	768	0.72	0.74	0.31	0.34	0.37	0.57	0.17	0.03	0.43	0.50	0.33	0.23	0.53	0.68	0.43	0.48
	ViT-B/8	768	0.72	0.52	0.35	0.09	0.51	0.33	0.10	**0**	0.35	0.35	0.48	0.27	0.52	0.58	0.38	0.34
DINOv2	ViT-L/14	1024	0.92	0.88	0.84	0.70	0.88	0.80	0.56	0.37	0.69	0.59	0.48	0.39	0.83	0.67	0.62	0.38
	ViT-G/14	1536	0.89	0.83	0.73	0.54	0.88	0.80	0.61	0.35	0.63	0.46	0.31	0.31	0.71	0.58	0.49	0.28
Grounding DINO	SwinB	768	0.79	**0.36**	0.17	**0**	0.26	**0.05**	**0**	**0**	0.16	0.08	0.10	**0.02**	0.60	0.35	0.24	0.04
	SwinT	768	0.72	0.42	**0.10**	0.02	**0.25**	0.07	**0**	**0**	0.08	0.13	**0.04**	0.08	0.52	0.34	0.24	0.07

Table 2: Comparison of VFMs with various backbones on covariate shift scenarios using GMM and NF methods. Best values for each dataset are highlighted in bold.

Perturbation Intensity: Detection improves with increasing severity of covariate shifts. For example, on Foggy Cityscapes [43], FPR95 drops from 0.67 ($i = 0.05$) to 0.21 ($i = 0.2$), indicating sensitivity to perturbation magnitude.

Architecture and Detection: Grounding DINO with Swin Transformer and NF achieves near-perfect detection (FPR95: 0) under various conditions. However, performance varies with rain type: it excels on Bravo-Synrain, which features lens-level artifacts, but degrades on ACDC Rain (FPR95: 0.34–0.35), which captures environmental rain (see Appendix). CLIP RN50x64 shows similar behavior—performing poorly on synthetic distortions (fog: 0.67, flare: 0.47) but well under real conditions (ACDC Snow: 0.03). These suggest greater sensitivity to sensor-level anomalies than to scene-level context changes.

4.3 Benchmarking Against OOD Methods and ImageNet-Trained Counterparts

We benchmark our VFM-based approach against established OOD detection methods [3] and ImageNet-pretrained counterparts using the same architectures. All baseline methods are implemented within a DeepLabV3+ segmentation model with ResNet50 backbone[6], trained

[6] https://github.com/open-mmlab/mmsegmentation/tree/main/configs/deeplabv3plus

on Cityscapes [15]. Our evaluation includes GradNorm [23], Energy [30], Entropy [33], MaxLogit [53], DICE [46], KNN [47], and ODIN [27].

We use Cityscapes' training set as $\mathcal{X}_{\text{monitor}}$ and its validation set as $\mathcal{X}_{\text{ID}}$, testing across diverse OOD conditions: real obstacles (Lost and Found [39]), synthetic anomalies (Bravo-SynObj [32]), and weather-based covariate shifts (Foggy Cityscapes [43], Bravo-SynRain [32]). Table 3 summarizes the results.

	Method	Lost and Found		Bravo-Synobj		Foggy Cityscapes (i=0.2)		Bravo-Synrain	
		AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓	AUROC↑	FPR95↓
OOD Methods	**DICE [46]**	0.50	0.94	0.57	0.91	0.69	0.92	0.77	0.80
	Energy [30]	0.50	0.97	0.50	0.97	0.49	0.99	0.50	0.99
	Entropy [33]	0.60	0.88	0.61	0.86	0.64	0.90	0.70	0.86
	GradNorm [23]	0.80	0.43	0.45	0.94	0.98	0.06	0.99	**0.00**
	KNN [47]	0.68	0.78	0.49	0.91	0.57	0.88	0.42	0.94
	MaxLogit [53]	0.56	0.60	0.56	0.95	0.63	0.95	0.62	0.97
	ODIN [27]	0.56	0.94	0.56	0.95	0.59	0.93	0.58	0.97
Image-Net Trained Counterparts	**RN50-GMM**	0.76	0.62	0.84	0.75	0.92	0.28	0.94	0.23
	RN50-NF	0.75	0.63	0.87	0.74	0.93	0.25	0.97	0.12
	SwinB-NF	0.86	0.44	0.91	0.53	0.92	0.35	0.99	0.05
VFMs	**Ours (RN50-GMM)**	0.90	0.36	0.89	0.66	0.85	0.47	0.85	0.43
	Ours (RN50-NF)	0.82	0.52	0.92	0.57	0.86	0.53	0.95	0.16
	Ours (SwinB-NF)	**0.95**	**0.16**	**0.99**	**0.02**	**1.00**	**0.00**	**1.00**	0.02

Table 3: Comparison of OOD detection between baseline methods and our VFM-based approach across semantic and covariate shifts Best results per dataset in bold.

Confidence-based (DICE [46]) and distance-based (KNN [47]) methods show moderate results on Lost and Found (AUROC: 0.50–0.68) and Bravo-Synobj (0.49–0.57), reflecting limited effectiveness in capturing complex semantic deviations. This is likely due to their reliance on model confidence scores or local feature similarity, which are unreliable under high semantic variability.

Logit-based approaches (Energy [30], Entropy [33], MaxLogit [53]) perform more consistently (AUROC: 0.50–0.70) but offer limited discriminative power. While they separate ID and OOD samples based on logit magnitude, they tend to miss fine-grained or localized anomalies that do not strongly affect output logits.

Feature-based methods such as GradNorm [23] excel on covariate shifts (e.g., Foggy Cityscapes AUROC: 0.98, FPR95: 0.06; Bravo-Synrain: 0.99, 0.00) and perform reasonably on real semantic shifts (Lost and Found: 0.80, 0.43), but struggle with synthetic anomalies (Bravo-Synobj: 0.45, 0.94). This suggests a higher sensitivity to global distortions than to object-level insertions, which may not elicit strong gradient responses.

ImageNet-trained models perform competitively on covariate shifts—e.g., SwinB-NF on SynRain (AUROC: 0.99, FPR95: 0.05)—but fall short on semantic shifts. For example, on Bravo-Synobj, ImageNet SwinB-NF yields FPR95: 0.53, while VFM SwinB-NF achieves 0.02. This highlights the benefit of diverse, multimodal pretraining in VFMs for capturing rare or out-of-domain semantics.

4.4 Input Monitoring

To evaluate practical applicability, we applied our input monitoring approach to the downstream task of semantic segmentation using DeepLabv3+ [13] with ResNet-50, ResNet-101 [20], and MobileNetV3 [22], trained on Cityscapes [15] and tested on the IDD dataset [48]. We modeled the Cityscapes training distribution using Grounding DINO's SwinB backbone with normalizing flows, and computed log-likelihoods for each IDD test image.

To quantify the utility of input monitoring, we evaluated segmentation performance on the IDD [48] test set and after applying increasing levels of input filtering based on log-likelihood scores. Specifically, we retained the top 75% (low), 50% (medium), and 25%

(high) of IDD images most similar to the training distribution. Segmentation performance was measured using mean Intersection over Union (mIoU) and mean Accuracy (mAcc).

Table 4 shows that filtering consistently improves segmentation across all architectures. A key advantage of our method is its model-agnostic nature: without accessing the segmentation model, we can identify inputs likely to cause errors—supporting fallback strategies, confidence calibration, or selective processing in safety-critical perception systems.

Filter Strength	**ResNet50**		**ResNet101**		**MobileNetV3**	
	mIoU↑	**mAcc↑**	**mIoU↑**	**mAcc↑**	**mIoU↑**	**mAcc↑**
None	43.58	62.84	45.85	65.10	34.22	52.25
Low	44.73	63.62	46.81	65.81	35.40	53.32
Medium	46.86	64.82	48.35	66.54	36.59	54.11
High	49.38	65.19	50.52	66.42	38.36	54.02

Table 4: Semantic segmentation on IDD under varying input filtering, with DeepLabv3+ backbones.

5 Conclusion

We proposed an unsupervised, model-agnostic OOD detection framework that leverages density modeling of VFM feature distributions to detect semantic and covariate shifts. To our knowledge, this is the first comprehensive evaluation of such an approach in the context of AD. Our experiments show that this method reliably detects diverse distribution shifts—including rare semantic anomalies and subtle covariate changes—while outperforming classical OOD techniques and non-foundation model baselines. We further demonstrate that input filtering based on VFM likelihoods improves downstream semantic segmentation without accessing the task model, providing a practical signal for identifying high-risk inputs and supporting robust, shift-aware perception in real-world systems.

Acknowledgment

M.K. and H.I.O. acknowledge support from the "NXT GEN AI Methods" research project, funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK). Furthermore, M.R. acknowledges funding from the "REFRAME" project (grant no. 01IS24073C), funded by the German Federal Ministry of Research, Technology and Space (BMFTR).

References

[1] Shamshe Alam, Sanjay Kumar Sonbhadra, Sonali Agarwal, and P Nagabhushan. One-class support vector classifiers: A survey. *Knowledge-Based Systems*, 196:105754, 2020.

[2] Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[3] Muqsit Azeem, Marta Grobelna, Sudeep Kanav, Jan Křetínský, Stefanie Mohr, and Sabine Rieder. Monitizer: Automating design and evaluation of neural network mon-

itors. In *International Conference on Computer Aided Verification*, pages 265–279. Springer, 2024.

[4] Petra Bevandić, Ivan Krešo, Marin Oršić, and Siniša Šegvić. Dense outlier detection and open-set recognition based on training with noisy negative images. *arXiv preprint arXiv:2101.09193*, 2021.

[5] Jan-Aike Bolte, Markus Kamp, Antonia Breuer, Silviu Homoceanu, Peter Schlicht, Fabian Huger, Daniel Lipinski, and Tim Fingscheidt. Unsupervised domain adaptation to improve image segmentation quality both in the source and target domain. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[6] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[7] Abdenour Bounsiar and Michael G Madden. One-class support vector machines revisited. In *2014 International Conference on Information Science & Applications (ICISA)*, pages 1–4. IEEE, 2014.

[8] CarADAS. Understanding ADAS: Lane Keep Assist, 2024. URL https://caradas.com/understanding-adas-lane-keep-assist/. Accessed: 4 March 2025.

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

[10] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. *arXiv preprint arXiv:2104.14812*, 2021.

[11] Robin Chan, Matthias Rottmann, and Hanno Gottschalk. Entropy maximization and meta classification for out-of-distribution detection in semantic segmentation. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 5128–5137, 2021.

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[14] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024.

[15] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[16] Council of the European Union. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts - analysis of the final compromise text with a view to agreement. https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf, 2024. Accessed: 2024-03-23.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[18] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Franz Hell, Gereon Hinz, Feng Liu, Sakshi Goyal, Ke Pei, Tetiana Lytvynenko, Alois Knoll, and Chen Yiqiang. Monitoring perception reliability in autonomous driving: Distributional shift detection for estimating the impact of input data on prediction accuracy. In *Proceedings of the 5th ACM Computer Science in Cars Symposium*, pages 1–9, 2021.

[22] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[23] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

[24] Sadia Ilyas, Ido Freeman, and Matthias Rottmann. On the potential of open-vocabulary models for object detection in unusual street scenes. *arXiv preprint arXiv:2408.11221*, 2024.

[25] International Organization for Standardization. ISO/PAS 8800:2024 – Road Vehicles – Safety and Artificial Intelligence, 2024. URL https://www.iso.org/standard/83303.html. Accessed: 4 March 2025.

[26] Mert Keser, Sinan Simsek, Deniz Karacor, and Alois Knoll. Subset selection for autonomous driving datasets via fine-tuning vision foundation models. In *2025 IEEE Intelligent Vehicles Symposium (IV)*, pages 2251–2258. IEEE, 2025.

[27] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[29] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[30] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[32] Thibaut Loiseau, Tuan-Hung Vu, Mickael Chen, Patrick Pérez, and Matthieu Cord. Reliability in semantic segmentation: Can we use synthetic data? *arXiv preprint arXiv:2312.09231*, 2023.

[33] David Macêdo, Tsang Ing Ren, Cleber Zanchettin, Adriano LI Oliveira, and Teresa Ludermir. Entropic out-of-distribution detection. In *2021 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2021.

[34] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, 2 edition, 2021. ISBN 978-0-262-04466-0.

[35] Alexey Nekrasov, Rui Zhou, Miriam Ackermann, Alexander Hermans, Bastian Leibe, and Matthias Rottmann. Oodis: Anomaly instance segmentation benchmark. *arXiv preprint arXiv:2406.11835*, 2024.

[36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[37] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[38] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing visual perception into llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27124–27133, 2024.

[39] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1099–1106. IEEE, 2016.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[41] Quazi Marufur Rahman, Peter Corke, and Feras Dayoub. Run-time monitoring of machine learning for robotic perception: A survey of emerging trends. *IEEE Access*, 9:20067–20075, 2021.

[42] SAE International. Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles (SAE J3016), 2021. URL https://www.sae.org/standards/content/j3016_202104/. Accessed: 4 March 2025.

[43] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126:973–992, 2018.

[44] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021.

[45] Dai Shi. Transnext: Robust foveal visual perception for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17773–17783, 2024.

[46] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, pages 691–708. Springer, 2022.

[47] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

[48] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1743–1751. IEEE, 2019.

[49] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14408–14419, 2023.

[50] Jianhua Wu, Bingzhao Gao, Jincheng Gao, Jianhao Yu, Hongqing Chu, Qiankun Yu, Xun Gong, Yi Chang, H Eric Tseng, Hong Chen, et al. Prospective role of foundation models in advancing autonomous vehicles. *Research*, 7:0399, 2024.

[51] Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4818–4829, 2024.

[52] H Zhang, F Li, S Liu, L Zhang, H Su, J Zhu, LM Ni, and HY Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arxiv 2022. *arXiv preprint arXiv:2203.03605*, 5, 2022.

[53] Zihan Zhang and Xiang Xiang. Decoupling maxlogit for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3388–3397, 2023.

[54] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.

[55] Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. Vision language models in autonomous driving: A survey and outlook. *IEEE Transactions on Intelligent Vehicles*, 2024.