Simplifying Traffic Anomaly Detection with Video Foundation Models

Svetlana Orlova Tommie Kerssies Brunó B. Englert Gijs Dubbelman
Eindhoven University of Technology
{s.orlova, t.kerssies, b.b.englert, g.dubbelman}@tue.nl

Abstract

Recent methods for ego-centric Traffic Anomaly Detection (TAD) often rely on complex multi-stage or multi-representation fusion architectures, yet it remains unclear whether such complexity is necessary. Recent findings in visual perception suggest that foundation models, enabled by advanced pre-training, allow simple yet flexible architectures to outperform specialized designs. Therefore, in this work, we investigate an architecturally simple encoder-only approach using plain Video Vision Transformers (Video ViTs) and study how pre-training enables strong TAD performance. We find that: (i) advanced pre-training enables simple encoder-only models to match or even surpass the performance of specialized state-of-the-art TAD methods, while also being significantly more efficient; (ii) although weakly- and fully-supervised pre-training are advantageous on standard benchmarks, we find them less effective for TAD. Instead, self-supervised Masked Video Modeling (MVM) provides the strongest signal; and (iii) Domain-Adaptive Pre-Training (DAPT) on unlabeled driving videos further improves downstream performance, without requiring anomalous examples. Our findings highlight the importance of pre-training and show that effective, efficient, and scalable TAD models can be built with minimal architectural complexity. We release our code, domain-adapted encoders, and fine-tuned models to support future work: https://github.com/tue-mps/simple-tad.

1. Introduction

Traffic risk estimation is fundamental to safe driving, as failures to anticipate danger can lead to life-threatening consequences. Autonomous vehicles must therefore assess potential hazards in real time, even under uncertain, dynamic, and unfamiliar conditions. A common formulation for this problem is the ego-centric traffic anomaly detection (TAD) task [10, 54], which aims to identify abnormal or dangerous events in a video stream captured by a vehicle-mounted camera. Analyzing the top-performing TAD methods [8, 24, 25, 35, 37, 57], we find that they rely on specialized, complex architectures illustrated in Fig. 2, b and c: *two-stage* approaches [8, 37, 57], which combine a vision encoder with a temporal component, and *multi-representation fusion* approaches [24, 25, 35], which fuse additional representations, often generated by separate deep neural networks or model-based algorithms. While these complex designs have improved performance, their impact on efficiency has not been evaluated, even though this is crucial for TAD, where rapid detection is needed to enable timely action and prevent accidents.

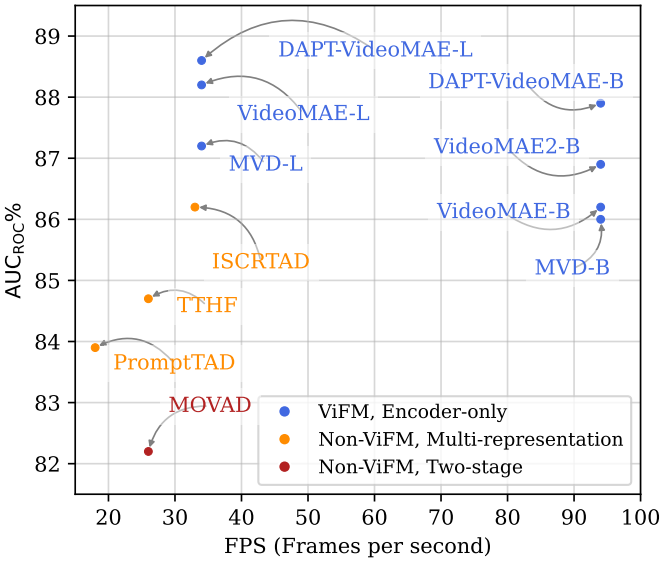

Figure 1. **Traffic Anomaly Detection (TAD) performance on DADA-2000** [10]. Simple encoder-only models with advanced pre-training (blue) are faster and more accurate than recent multi-component architectures (orange and red), see Tabs. 2 and 3.

Considering that TAD is, in essence, a binary classification task on video, we turn to recent methods for general video classification. On standard video classification benchmarks [6, 12, 17, 40], Video Foundation Models (ViFMs) achieve state-of-the-art performance, predominantly Video Vision Transformers (ViTs) [1, 9], which rely on large-scale self- and weakly-supervised pre-training to learn ex-

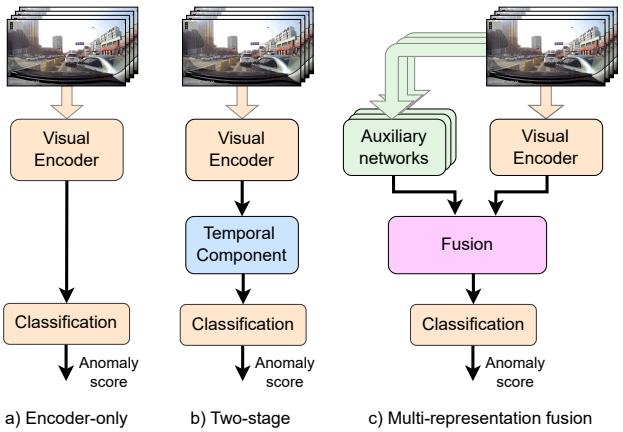

Figure 2. Types of model architectures for TAD: simple encoder-only (a), two-stage (b), and multi-representation fusion (c).

pressive and transferable spatio-temporal representations, rather than on architectural inductive biases. Recent work in related visual perception tasks shows that advanced pre-training reduces the need for downstream task-specific components [18, 19, 47]. We hypothesize that the same applies to TAD, such that a simple ViT-based ViFM can be applied to this task and match or even outperform complex architectures. Since TAD depends on motion understanding, we investigate whether pre-training strategies that capture spatio-temporal structure are particularly effective.

To test our hypotheses, we evaluate multiple ViT-based ViFMs, sharing the same plain Video ViT architecture but different pre-training, on two TAD datasets: DoTA [54] and DADA-2000 [10]. We adopt an *encoder-only* design, with a single linear layer on top of the Video ViT model, as illustrated in Fig. 2, a. While prior work has attempted the encoder-only design [8], it was found to be inferior [24, 35, 37]. We revisit this design using stronger pre-training. Similar to prior work, we assess in-domain and out-of-domain generalization, and unlike prior work, we assess the computational efficiency of the models and compare to the best-performing specialized TAD methods.

We confirm our hypothesis by showing that advanced pre-training enables a plain Video ViT, used as the encoder in an encoder-only design for TAD, to match and even surpass state-of-the-art methods while also being significantly more efficient, as shown in Fig. 1 and Tab. 2. Interestingly, performance on standard video classification benchmarks does not correlate with TAD. Our comparison of pre-trained models shows that weak supervision from language and full supervision from class labels are effective on standard benchmarks but less so for TAD, likely because they promote appearance-centric features and semantics that generalize poorly to anomalous motion [51]. In contrast, self-supervised learning with Masked Video Modeling (MVM), which trains the model to reconstruct missing spatio-temporal regions using both spatial structure and temporal continuity, proves most effective for TAD. Models pre-trained with this objective achieve state-of-the-art performance at their respective model size, as shown in Tab. 3.

Next, motivated by the scarcity of labeled data for TAD and the abundance of unlabeled driving videos, we explore whether we can leverage self-supervised learning to better adapt an off-the-shelf ViFM to the downstream driving domain. Specifically, we apply *Domain-Adaptive Pre-Training* (DAPT) [13] using the Video Masked Autoencoding (MAE) [44] approach on unlabeled driving videos. We find that MAE-based DAPT, even when applied at a relatively small scale compared to the preceding generic pre-training, significantly boosts the performance even further, particularly for smaller models, as shown in Fig. 3. Importantly, we find that including abnormal driving examples for DAPT is not necessary, as shown in Tab. 4, which is valuable given the difficulty of collecting them at scale.

We summarize our main contributions as follows:

1. We show that a plain encoder-only Video ViT, when equipped with advanced pre-training, outperforms all prior specialized architectures for TAD, while also being significantly more efficient.
2. We compare pre-training strategies and find that self-supervised learning with a Masked Video Modeling (MVM) objective is most effective, outperforming both weakly- and fully-supervised alternatives.
3. We demonstrate that Domain-Adaptive Pre-training with MVM leads to measurable performance gains on TAD, even when applied at a small scale to domain-relevant but anomaly-free data.

Moreover, since every TAD pipeline relies on its visual encoder, our results offer clear guidance for selecting effective pre-training strategies, enabling future methods to detect traffic anomalies more accurately and robustly.

2. Related work

2.1. Traffic Anomaly Detection

Traffic Anomaly Detection (TAD) is typically framed as a binary video classification task, where the goal is to detect potentially dangerous or abnormal events in traffic scenarios from the egocentric viewpoint of a vehicle-mounted camera. While related to the broader field of Video Anomaly Detection, which commonly targets static-camera surveillance settings, TAD poses unique challenges due to ego-motion, frequent occlusions, and dynamic interactions of agents in complex driving environments. Before the availability of labeled datasets for TAD, earlier approaches often relied on unsupervised reconstruction or prediction of video frames to flag anomalies using temporal autoencoders [30, 49], future frame prediction with spatial, tem-

poral, and adversarial losses [26], or spatio-temporal tubelet modeling with ensemble scoring [54]. Some works also explored the use of synthetic data [20, 39]. The introduction of large-scale driving anomaly datasets [10, 54] with comprehensive annotations has enabled more active development of fully-supervised methods and led to substantial improvements in detection performance. As shown in Fig. 2, we categorize TAD methods into three classes based on their architectural complexity, which we detail below:

Encoder-only design (Fig. 2, a). Since TAD is a binary classification task, a minimal solution consists of a feature encoder followed by a linear classifier, without any task-specific modules. Prior work [8] shows that such designs can be effective in related tasks, where R(2+1)D [45] and ViViT [1] demonstrate considerably strong performance. However, recent studies report underwhelming results for encoder-only ablation variants of their methods [24, 35, 37].

Two-stage design (Fig. 2, b). Two-stage methods combine a visual encoder with a separate temporal module. VidNeXt [8] pairs a ConvNeXt [27] backbone with a non-stationary transformer (NST) to model both stable and dynamic temporal patterns, evaluating and introducing a new dataset CycleCrash [8] for the related task of collision prediction. Its ablations, ConvNeXt+VT and ResNet+NST, also yield strong results. MOVAD [37] uses a VideoSwin Transformer [28] as a short-term memory encoder over several frames, followed by an LSTM-based long-term module, achieving state-of-the-art performance for TAD.

Multi-representation fusion design (Fig. 2, c). Fusion-based models, which currently report state-of-the-art performance in TAD, explicitly combine multiple information sources. TTHF [24] augments the CLIP [36] framework with a high-frequency motion encoder and a cross-modal fusion module to align motion features with textual prompts. PromptTAD [35] extends MOVAD [37] by incorporating bounding box prompts via instance- and relation-level attention, enhancing object-centric anomaly localization. ISCRTAD [25] integrates agent features (e.g., appearance, trajectory, depth) using graph-based modeling and fuses them with scene context through contrastive multimodal alignment for robust anomaly detection.

In this work, we follow the simple encoder-only design and explore whether advanced pre-training can compensate for the lack of task-specific architectural inductive biases. We demonstrate that, when equipped with rich priors from large-scale self-supervised pre-training, such models can achieve state-of-the-art performance, while remaining architecturally simple and highly efficient.

2.2. Video Foundation Models

While early 3D CNN-based architectures can be referred to as foundation models [31], the term foundation model today typically refers to Transformer-based [46] models that leverage large-scale pre-training. For vision, these models typically adopt the Vision Transformer [9] (ViT) architecture. Although other architectures such as recurrent, hybrid, and state-space models are active research areas [23, 34, 53], ViT-based ViFMs are currently the dominant paradigm due to their strong performance, scalability, versatile pre-training strategies, and widespread optimization support (*e.g.*, FlashAttention [7], optimized libraries, hardware acceleration) designed around the plain Transformer architecture. These qualities, along with the growing availability of pre-trained Video ViTs, make them particularly promising for tasks like TAD, where generalization, robustness, and efficiency are critical.

Unlike convolutional architectures, which impose strong spatial and temporal inductive biases by design, ViTs must learn such priors directly from data during pre-training [9]. As a result, their downstream performance depends heavily on the scale and quality of the pre-training dataset, as well as the choice of training objective [56]. Pre-training strategies differ in supervision strength and scalability:

Fully-supervised learning (FSL) methods rely on manually annotated labels, providing precise semantic guidance but limited scalability. First Video ViTs exploited supervised classification as their pre-training method. ViViT [1] adopted the ViT architecture to video by introducing spatio-temporal 3D cubes called tubelets instead of 2D patches used for images. While this demonstrated that attention-based models can handle video inputs, the method struggled to balance accuracy and efficiency.

Weakly-supervised learning (WSL) methods use natural language or metadata as training signals. Though less curated, they offer rich semantic structure and broader concept coverage from web-scale data. HowTo100M [32] introduced large-scale WSL pre-training by aligning video instructions with narrations using contrastive loss, which MIL-NCE [33] extended with Multiple Instance Learning (MIL) to handle temporally imprecise supervision. UMT [22] and SMILE [43] leverage feature distillation from a vision-language model; UMT adopts an unmasked token alignment, while SMILE introduces motion-guided masking for better semantic and temporal understanding.

Self-supervised learning (SSL) methods learn from the data itself without any annotations, enabling large-scale pre-training and highly transferable representations. Unlike FSL and WSL, which introduce annotation noise and bias, SSL provides denser and more unbiased training signals [4, 14]. VideoMAE [44] adopted masked autoencoding (MAE) [14], a type of Masked Video Modeling (MVM), as an effective and efficient self-supervised pre-training strategy for plain video ViTs. Its tube masking, when applied to a large fraction of input patches, forces the model to infer spatio-temporal structure from limited visible content.

Year	Model	Stage	Type	Objective	Supervision
2021	ViViT [1]	1	FSL	Classification	Class labels
2022	VideoMAE [44]	1	SSL	Masked autoencoding	Video frame pixels
2023	MVD [51]	1	SSL	Masked feature distillation	High-level features of VideoMAE and ImageMAE teachers
2023	VideoMAE2 [50]	1 2 3	SSL FSL FSL	Dual masked autoencoding Classification Logit distillation	Video frame pixels Class labels Logits of larger VideoMAE2 after stage 2
2023	UMT [22]	1	WSL	Unmasked feature distillation	Features of a vision-language encoder
2024	InternVideo2 [52]	1 2 3	WSL+SSL WSL WSL+SSL	Unmasked feature distillation Contrastive Feature distillation	Features of VideoMAE2 and a vision-language encoder Video-text and audio-text pairs Features of InternVideo2 after stage 2 across multiple depths
2025	SMILE [43]	1,2	WSL	Masked feature distillation	Features of a vision-language encoder

Table 1. **Overview of ViFMs.** For models trained via distillation, we denote the supervision type(s) used for the teacher. **FSL**: fully-supervised, **WSL**: weakly-supervised, **SSL**: self-supervised learning.

This approach yielded strong results while maintaining architectural simplicity and has since inspired a series of ViT-based Video Foundation Models (ViFMs) that employ self-supervised pre-training [15, 38, 41, 43, 50, 51].

To our knowledge, we are the first to research which type of video pre-training is most effective for the TAD task, and hypothesize that self-supervised pre-training with an MVM objective, with its emphasis on learning patch-dense and temporally-aware representations, is particularly well-suited for this task.

3. Methodology

We fine-tune multiple Video Foundation Models (ViFMs) for TAD. We follow the encoder-only design and attach a single linear layer as a classification head to the output of the encoder. This minimal design ensures that performance primarily reflects the effectiveness of the ViFM backbone in capturing patterns relevant for traffic anomaly detection.

Given a video input $X_t \in \mathbb{R}^{T\times H\times W\times 3}$ at time t, we extract spatio-temporal feature vector $F_t \in \mathbb{R}^E$ using the encoder (ViFM), where T, H, and W are the number of frames, height, and width, and E is the feature dimension. Then, a linear layer maps F_t to classification logits $L_t \in \mathbb{R}^C$, with $C = 2$. We fine-tune using cross-entropy loss and apply softmax during inference to obtain anomaly class probability as the anomaly score, which can then be binarized with a decision threshold (typically 0.5). More details in Appendix A.

We investigate (i) whether a simple Video ViT model, pre-trained at scale, can achieve state-of-the-art performance on TAD, (ii) whether better ViFMs for general video recognition are also better for TAD, and what type of pre-training is more effective, and, finally, (iii) whether small-scale domain-adaptive pre-training (DAPT) is feasible and effective for adapting Video ViTs to the driving domain.

3.1. Task definition

We formulate Traffic Anomaly Detection (TAD) as a binary classification task, specifically focusing on frame-level, ego-centric anomaly classification, where each frame captured from a moving vehicle-mounted camera is assigned an anomaly label.

Let $X_t = \{I_{t-\tau+1}, I_{t-\tau+2}, \ldots, I_t\}$ denote a time-ordered sliding window of τ consecutive video frames captured from a vehicle-mounted camera up to time t. Each I_k represents an RGB frame at time step k, from the egocentric viewpoint of the vehicle.

The task is to learn a function f_θ that maps an input window X_t to a prediction A_t at each timestep t:

$$A_t = f_\theta(X_t), \quad t = T, T+1, \ldots$$

where $A_t \in \{0, 1\}$ is a binary label that indicates whether an anomaly is observed at time t.

In the general case, τ can be 1 and f_θ may use recurrence or autoregressive conditioning on past predictions.

3.2. Evaluation Procedure

Prior work typically reports the *Area Under the Receiver Operating Characteristic Curve* (AUC_{ROC}) as the primary evaluation metric for TAD [24, 25, 35, 54, 57], and we adopt it as well for comparison with previous methods. However, because handling data imbalance is especially important in TAD, we follow related work [16, 42, 48] and also report the Matthews Correlation Coefficient (MCC). MCC considers all entries of the confusion matrix, including true negatives, and better reflects overall performance under class imbalance [5]. MCC at a given threshold is defined as:

$$\text{MCC} = \frac{\text{TP}\cdot\text{TN} - \text{FP}\cdot\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}, \quad (1)$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively. Note

that MCC ranges from -1 (inverse prediction) to 1 (perfect prediction), with 0 indicating random performance, but we show it in the range -100, 100 to improve readability.

To assess discriminative ability independently of the decision threshold, we compute MCC across thresholds in the range $[0, 1]$ and report the area under this curve, referred to as the *Area Under the MCC Curve* (AUC_{MCC}). We also report MCC at a fixed threshold of 0.5 (MCC@0.5).

Beyond metric design, we adopt a broader evaluation protocol that emphasizes both generalization and efficiency. We assess in-domain and out-of-domain performance, as well as efficiency based on parameter count, peak GPU memory usage, and frames per second (FPS).

3.3. Pre-trained Encoders

We select a range of recent Video ViT-based ViFMs that represent various pre-training strategies, and apply them to the TAD task; see Tab. 1 for an overview of their pre-training strategies. When possible, we select variants pre-trained on Kinetics-400 [17] for consistency. For completeness, we also include R(2+1)D [45], a fully-convolutional encoder, motivated by recent studies showing its competitive performance in the related task of collision anticipation [8].

We include ViViT [1] and fully-convolutional R(2+1)D [45] to represent fully-supervised pre-training. VideoMAE [44], MVD [51], and VideoMAE2 [50] are selected to evaluate progressively stronger variants of self-supervised pre-training from videos. Among weakly-supervised methods, we assess UMT [22] and SMILE [43], both of which align video tokens with CLIP [36] supervision. Finally, InternVideo2 [52] combines self-supervised learning from videos and weakly-supervised learning from multiple modalities, and is one of the leading models across numerous video benchmarks. Together, these models cover a diverse range of pre-training strategies.

3.4. Domain-Adaptive Pre-Training

To better align the Video ViT encoder with the driving domain, we adopt the *Domain-Adaptive Pre-Training* (DAPT) strategy, a simple method originally proposed in the field of natural language processing [13]. DAPT introduces an additional pre-training stage of smaller scale between generic pre-training and downstream fine-tuning, using unlabeled data from the target domain.

We apply VideoMAE-based [44] DAPT as follows:

- **Step 1: Generic pre-training.** As before, we initialize the encoder with an off-the-shelf VideoMAE model pre-trained on large-scale generic video data, mostly unrelated to the driving domain.
- **Step 2: Domain-Adaptive Pre-training (DAPT).** We continue pre-training the same model on a medium-sized dataset of unlabeled driving videos using the exact same VideoMAE reconstruction objective:

$$\mathcal{L}_{\text{DAPT}} = \|M \odot (x - f_\theta(x_{\text{masked}}))\|_2^2$$

  where x is the input video, x_{masked} is the masked input, f_θ is the encoder-decoder VideoMAE model, M is the binary mask, and $\odot$ is element-wise multiplication.
- **Step 3: Fine-tuning on TAD.** As before, we fine-tune the encoder-only model on TAD datasets using the same configuration with a simple linear classification head.

The intermediate DAPT step (Step 2) specializes the model towards the driving domain without requiring any labels. It introduces no additional parameters, preserves model efficiency, and remains fully compatible with standard VideoMAE pipelines.

4. Experiments

4.1. Experimental setup

Datasets. We evaluate on DoTA [54] and DADA-2000 [10], two large-scale real-world driving anomaly datasets with temporal and frame-level annotations. We also manually refine 1% of annotations of the DoTA dataset and refer to this variant as DoTA_{ref}, more details in Appendix D. DAPT uses Kinetics-700 [3], BDD100K [55], and CAP-DATA [11].

Model input. All Video ViTs and R(2+1)D are trained on sliding windows of size 224×224×16 at 10 FPS (1.5s temporal context) by default. For InternVideo2 and UMT, which use tubelets of size 1, we use 224×224×8 at 5 FPS to match the same duration. MOVAD processes videos frame-by-frame at resolution 640×480.

Fine-tuning. With all Video ViTs, R(2+1)D, and VidNeXt variants, we closely follow the VideoMAE fine-tuning recipe for HMDB51. We train for 50 epochs (5 warmup), with 50K randomly sampled examples per epoch and a batch size of 56. For MOVAD, we follow the original training settings.

Domain-adaptive pre-training (DAPT). We apply the VideoMAE pre-training strategy [44], masking 75% of tokens, using MSE loss on masked tokens only. Training uses a batch size of 800 and 1M samples per epoch, with 12 epochs. We explore DAPT on three domains: (a) Kinetics-700, (b) BDD100K (normal driving), and (c) BDD100K + CAP-DATA (a mix of normal and abnormal driving). More details can be found in Appendix A.

4.2. Can an encoder-only model outperform specialized TAD methods?

To answer this question, we evaluate models along three critical axes: classification performance, generalization, and efficiency, as shown in Tab. 2. We select a range of recent top-performing methods proposed for the TAD task.

Method	**DoTA** AUC_{ROC}, %		**D2K** AUC_{ROC}, %		# Param	Peak GPU	FPS
	DoTA→DoTA	D2K→DoTA	D2K→D2K	DoTA→D2K			
Two-stage TAD methods							
VidNeXt [8]	73.9	69.3	70.1	72.4	125 M	0.78 GB	27
ConvNeXt+VT [8]	73.1	61.2	66.8	67.3	125 M	0.77 GB	27
ResNet+NST [8]	74.0	70.1	71.2	72.3	24 M	0.19 GB	124
MOVAD [37]	82.2	77.6	77.0	75.2	153 M	1.10 GB	26
Multi-representation fusion TAD methods							
TTHF [24]	84.7†	–	–	71.7†	140 M	0.80 GB	26
PromptTAD [35]	83.9†	–	–	74.6†	106 M	1.88 GB	18
ISCRTAD [25]	86.2†	–	–	82.7†	359 M‡	1.51 GB‡	33‡
Encoder-only models							
R(2+1)D [45]	81.5	76.4	78.8	78.4	27 M	0.27 GB	104
DAPT-VideoMAE-S (ours)	**86.4**	**81.7**	**85.6**	**84.3**	22 M	0.16 GB	95
DAPT-VideoMAE-B (ours)	**87.9**	**83.5**	**87.6**	**85.8**	86 M	0.54 GB	94
DAPT-VideoMAE-L (ours)	**88.4**	**85.0**	**88.5**	**86.6**	304 M	1.80 GB	34

Table 2. **Traffic Anomaly Detection (TAD) performance and efficiency.** Video ViT-based encoder-only models set a new state of the art on both datasets, while being significantly more efficient than top-performing specialized methods. FPS measured using NVIDIA A100 MIG, $\frac{1}{2}$ GPU. †From prior work. ‡Optimistic estimates using publicly available components of the model. "**A**→**B**": trained on A, tested on B; **D2K**: DADA-2000.

Among encoder-only models, we apply the R(2+1)D [45] model and different sizes of VideoMAE pre-trained Video ViTs (with DAPT, see Sec. 4.4).

The results show that these Video ViTs consistently strike a strong balance, demonstrating a good combination of predictive accuracy, generalization across datasets, and computational efficiency. Notably, Video ViTs with advanced pre-training achieve the highest AUC_{ROC} scores across both DoTA and DADA-2000 datasets, both in-domain and in cross-dataset evaluation, while being highly efficient with a low memory footprint. In contrast, specialized TAD-specific models not only demonstrate lower classification performance but also incur substantially higher computational costs and latency. R(2+1)D and ResNet+NST, while being highly efficient, fall short in predictive quality. This confirms that we can outperform specialized, multi-component TAD methods with a simple encoder-only model by applying a Video ViT with advanced pre-training. A more detailed comparison is provided in Appendix C.

4.3. What pre-training is better for TAD?

We evaluate a range of publicly available Video ViT models of several sizes, using their Top-1 accuracy on the general benchmarks Kinetics-400 [17] and Something-SomethingV2 [12] alongside AUC_{MCC} and MCC@0.5 on the TAD benchmarks DoTA [54] and DADA-2000 [10]. Results are shown in Tab. 3, from which we observe three key trends.

First, we find that for TAD, self-supervised pre-training, *i.e.* the Masked Video Modeling (MVM) objective dominates: models, pre-trained with the Masked Autoencoding (MAE) approach (VideoMAE), and their distilled variants (VideoMAE2, MVD) achieve the highest AUC_{MCC} within each size tier. Additional experiments on different MVM objectives can be found in Appendix B. Second, models that employ weakly-supervised pre-training (UMT, SMILE and InternVideo2) perform worse on TAD, indicating that vision-language supervision may not transfer well to fine-grained temporal anomaly understanding. Finally, classification accuracy on general benchmarks is not representative of TAD performance: ViViT-Base, despite matching VideoMAE on Kinetics-400, demonstrates significantly lower AUC_{MCC}, and InternVideo2, state-of-the-art on general benchmarks, clearly underperforms on TAD. This suggests that the representations that are beneficial for general video classification may not align well with those needed for TAD. In particular, TAD appears to benefit more from dense representations that emphasize fine-grained temporal irregularities rather than the coarse semantic categories typically targeted by general video recognition models.

The overall top-ranking model on TAD is VideoMAE2, which incorporates dual masking, an additional pre-training step with distillation from a larger model, and ∼6 times larger-scale pre-training datasets, compared to other MVM pre-trained models. This confirms that both the scale of pre-training and the choice of objectives significantly impact the transferability of ViFMs to TAD.

Model	Variant	Type	Top-1 accuracy		MCC@0.5		AUC_{MCC}	
			K400	SthSthV2	$DoTA_{ref}$	D2K	$DoTA_{ref}$	D2K
$VideoMAE_{1600}$ [44]	Small	SSL	79.0	66.8	55.5	49.5	52.1	48.1
MVD_{fromB} [51]	Small	SSL	80.6	70.7	56.2	49.8	50.0	48.1
MVD_{fromL} [51]	Small	SSL	81.0	70.9	56.5	51.1	50.2	49.1
VideoMAE2 [50]	Small	SSL+FSL	83.7	–	**56.8**	**51.6**	**55.2**	**50.3**
InternVideo2 [52]	Small	WSL+SSL	**85.4**	**71.6**	51.6	44.5	49.7	43.7
ViViT [1]	Base	FSL	79.9	–	30.7	27.6	28.9	26.7
$VideoMAE_{800}$ [44]	Base	SSL	80.0	–	58.0	52.0	54.5	51.2
$VideoMAE_{1600}$ [44]	Base	SSL	81.0	69.7	58.7	52.6	56.0	52.2
MVD_{fromB} [51]	Base	SSL	82.7	72.5	57.8	51.6	56.0	50.9
MVD_{fromL} [51]	Base	SSL	83.4	73.7	**59.2**	52.1	**57.0**	51.0
SMILE [43]	Base	WSL	83.4	72.5	56.7	48.6	54.8	49.8
VideoMAE2 [50]	Base	SSL+FSL	86.6	**75.0**	58.4	**54.8**	56.5	**53.4**
UMT [22]	Base	WSL	87.4	70.8	48.4	40.2	46.3	38.1
InternVideo2 [52]	Base	WSL+SSL	**88.4**	73.5	52.2	44.2	50.0	43.1
$VideoMAE_{1600}$ [44]	Large	SSL	85.2	74.3	**61.6**	**56.9**	**59.7**	**55.4**
MVD_{fromL} [51]	Large	SSL	**86.0**	**76.1**	60.5	54.6	59.0	53.7

Table 3. **Comparing Video ViT pre-training strategies.** In contrast to general video classification benchmarks (K400, SthSthV2), fully and weakly supervised pre-training are less effective for TAD ($DoTA_{ref}$, D2K), while self-supervised pre-training yields the best performance. **FSL**: fully-supervised; **WSL**: weakly-supervised; **SSL**: self-supervised learning; **K400**: Kinetics-400 [17]; **SthSthV2**: Something-SomethingV2 [12]; **D2K**: DADA-2000 [10].

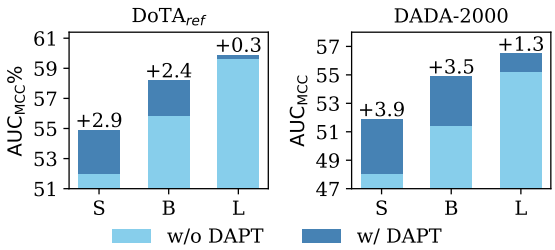

Figure 3. **DAPT scaling across different model sizes.** Smaller models benefit more. **S**: Small, **B**: Base, **L**: Large variants of the Video ViT.

Method	$DoTA_{ref}$, AUC_{MCC}		D2K, AUC_{MCC}	
	D→D	D2K→D	D2K→D2K	D→D2K
w/o DAPT	52.1	43.8	48.1	46.6
Generic DAPT	51.6 -0.5	43.8	48.5 +0.4	46.2 -0.4
Driving DAPT	54.9 +2.8	47.0 +3.2	52.1 +4.0	49.7 +3.1
↳ + anomalies	55.0 +2.9	47.1 +3.3	52.0 +3.9	49.8 +3.2

Table 4. **DAPT ablation.** Comparing generic (Kinetics-700 [17]), driving (BDD100K [55]), and driving + anomaly (CAP-DATA [11]) domains shows that driving videos improve performance without requiring anomalies. Using Video ViT-Small. "**A→B**": trained on A, tested on B; **D**: $DoTA_{ref}$; **D2K**: DADA-2000.

4.4. Domain-Adaptive Pre-Training (DAPT)

Larger ViFMs can be pre-trained on a larger scale and, as a result, exhibit better out-of-the-box generalization across domains, while smaller models have shown to benefit less from longer pre-training due to their limited capacity and faster saturation [56]. Therefore, we expect that domain adaptation can help better utilize the capacity of smaller, yet more efficient and faster models. Given that MAE pre-training proves especially effective for TAD, and unlabeled driving data is available in abundance, we investigate whether small-scale self-supervised DAPT with MAE can be an effective and efficient way to scale the performance of smaller models. We initialize a ViT model with VideoMAE pre-trained weights and perform several epochs of additional pre-training with the VideoMAE objective on in-domain data. Compared to the original ∼192K training steps with batch size 2048, we use only 15K steps with batch size 800.

As shown in Fig. 3, DAPT via MAE brings clear improvements for Small and Base VideoMAE pre-trained models. As expected, given our small-scale DAPT protocol, the Large model sees less improvement.

To disentangle the impact of domain relevance from that of additional pre-training, we conduct an ablation study on the data used for DAPT. Specifically, we adapt a VideoMAE model, pre-trained on a general human activity dataset, using three types of unlabeled video data: (1) the original, general pre-training domain, which is not related to the TAD task, (2) normal ego-centric driving, and (3) ego-

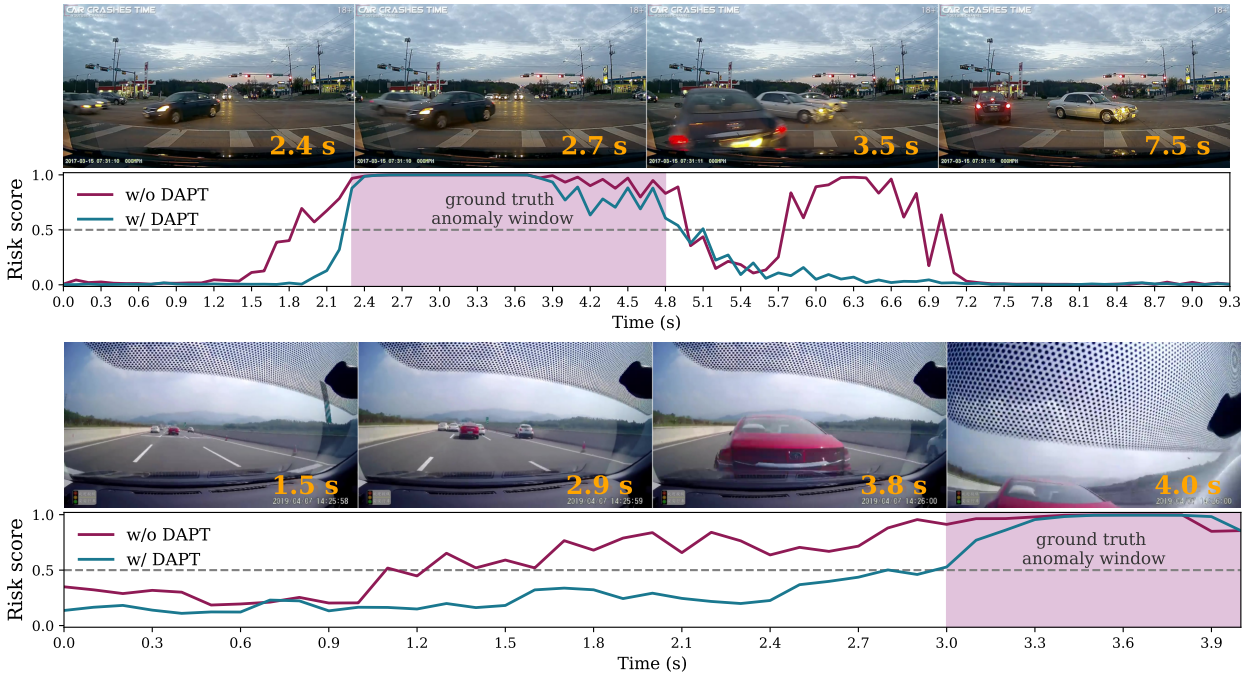

Figure 4. **Qualitative examples for the effect of DAPT.** Predicted anomaly-scores of VideoMAE with and without DAPT (top: Video ViT-Small, bottom: Video ViT-Base).

centric driving mixed with anomalies. This setup allows us to evaluate whether the observed improvements stem from domain alignment or from simply continuing pre-training. As shown in Tab. 4, adaptation with domain-relevant data (both normal and abnormal driving) consistently improves in-domain performance and generalization, while additional pre-training on the original, generic domain yields no notable gains. Interestingly, pre-training on normal driving videos is sufficient, and mixing in data with driving anomalies does not provide any further improvement. We conclude that small-scale self-supervised DAPT is a simple and effective way to improve the performance and generalization of smaller Video ViTs for TAD, without the need to collect rare anomalous examples.

Finally, in Fig. 4 we also include some qualitative examples which clearly demonstrate the positive effect of DAPT.

5. Discussion

In this work, we show that with advanced pre-training, an encoder-only Video Vision Transformer outperforms all prior Traffic Anomaly Detection models while also being significantly more efficient. However, it remains an open question whether the additional components introduced in earlier work become redundant as pre-training scales, as shown in related perception tasks [19], or whether they still provide complementary benefits.

6. Conclusion

Ego-centric Traffic Anomaly Detection (TAD) is a challenging task that requires modeling motion dynamics and agent interactions. While most recent methods for TAD rely on complex, multi-component architectures, we show that a simple encoder-only design using a plain Video Vision Transformer with strong self-supervised pre-training is not only more efficient, but also more effective and generalizable. Building on this, Domain-Adaptive Pre-Training offers a label-free and compute-efficient way to further boost performance, particularly for smaller models. These findings highlight the strength of learned inductive biases from large-scale pre-training as an alternative to manually crafted architectural complexity, a principle to which TAD is no exception.

Our experiments demonstrate that Masked Video Modeling is the most effective pre-training strategy for TAD, in contrast to general video classification tasks. This suggests that different video tasks may benefit from different pre-training objectives tailored to their downstream requirements. While TAD is a crucial task in autonomous driving (AD), other AD tasks may align more closely with conventional action recognition. This motivates further research into a universally effective video pre-training strategy, evaluated by its generalization across diverse AD tasks. We hope our findings provide a foundation for future work in this direction.

Acknowledgements

This work was funded by the Horizon Europe programme of the European Union, under grant agreement 101076754 (project AITHENA).

We also acknowledge the Dutch national e-infrastructure with the support of the SURF Cooperative, grant agreement no. EINF-10314, financed by the Dutch Research Council (NWO), for the availability of high-performance computing resources and support.

Disclaimer

Views and opinions expressed here are those of the authors only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 1, 3, 4, 5, 7

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 12

[3] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 5

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020. 3

[5] Davide Chicco and Giuseppe Jurman. The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification. *BioData Mining*, 16(1):4, 2023. 4

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 1

[7] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359, 2022. 3

[8] Nishq Poorav Desai, Ali Etemad, and Michael Greenspan. Cyclecrash: A dataset of bicycle collision videos for collision prediction and analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025. 1, 2, 3, 5, 6, 12

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3

[10] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE transactions on intelligent transportation systems*, 23(6):4959–4971, 2021. 1, 2, 3, 5, 6, 7, 13, 14, 15

[11] Jianwu Fang, Lei-Lei Li, Kuan Yang, Zhedong Zheng, Jianru Xue, and Tat-Seng Chua. Cognitive accident prediction in driving scenes: A multimodality benchmark. *CoRR*, abs/2212.09381, 2022. 5, 7, 12

[12] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 1, 6, 7

[13] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. 2, 5

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[15] Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, and Limin Wang. Mgmae: Motion guided masking for video masked autoencoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13493–13504, 2023. 4, 13

[16] P Rajesh Kanna, S Vanithamani, P Karunakaran, P Pandiaraja, N Tamilarasi, and P Nithin. An enhanced traffic incident detection using factor analysis and weighted random forest algorithm. In *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, pages 1355–1361. IEEE, 2024. 4

[17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 5, 6, 7, 12, 13

[18] Tommie Kerssies, Daan de Geus, and Gijs Dubbelman. First Place Solution to the ECCV 2024 BRAVO Challenge: Evaluating Robustness of Vision Foundation Models for Semantic Segmentation. *arXiv preprint arXiv:2409.17208*, 2024. 2

[19] Tommie Kerssies, Niccolo Cavagnero, Alexander Hermans, Narges Norouzi, Giuseppe Averta, Bastian Leibe, Gijs Dubbelman, and Daan de Geus. Your vit is secretly an image segmentation model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25303–25313, 2025. 2, 8

[20] Hoon Kim, Kangwook Lee, Gyeongjo Hwang, and Changho Suh. Crash to not crash: Learn to identify dangerous vehicles using a simulator. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 978–985, 2019. 3

[21] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 12

[22] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19948–19960, 2023. 3, 4, 5, 7

[23] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2024. 3

[24] Rongqin Liang, Yuanman Li, Jiantao Zhou, and Xia Li. Text-driven traffic anomaly detection with temporal high-frequency modeling in driving videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 1, 2, 3, 4, 6, 14

[25] Rongqin Liang, Yuanman Li, Zhenyu Wu, and Xia Li. An interaction-scene collaborative representation framework for detecting traffic anomalies in driving videos. *IEEE Transactions on Intelligent Transportation Systems*, 2025. 1, 3, 4, 6, 13, 14

[26] Wen Liu, Weixin Luo, Dongze Lian, and Shenghua Gao. Future frame prediction for anomaly detection–a new baseline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6536–6545, 2018. 3

[27] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 3

[28] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 3

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 12

[30] Weixin Luo, Wen Liu, and Shenghua Gao. Remembering history with convolutional lstm for anomaly detection. In *2017 IEEE International conference on multimedia and expo (ICME)*, pages 439–444. IEEE, 2017. 2

[31] Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. Foundation models for video understanding: A survey. *Authorea Preprints*, 2024. 3

[32] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 3

[33] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9879–9889, 2020. 3

[34] Viorica Pătrăucean, Xu Owen He, Joseph Heyward, Chuhan Zhang, Mehdi S. M. Sajjadi, George-Cristian Muraru, Artem Zholus, Mahdi Karami, Ross Goroshin, Yutian Chen, Simon Osindero, João Carreira, and Razvan Pascanu. Trecvit: A recurrent video transformer. *arXiv preprint arXiv:2412.14294*, 2024. 3

[35] Hao Qiu, Xiaobo Yang, and Xiaojin Gong. Prompttad: Object-prompt enhanced traffic anomaly detection. *IEEE Robotics and Automation Letters*, 2025. 1, 2, 3, 4, 6, 13, 14

[36] 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. 3, 5

[37] Leonardo Rossi, Vittorio Bernuzzi, Tomaso Fontanini, Massimo Bertozzi, and Andrea Prati. Memory-augmented online video anomaly detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6590–6594. IEEE, 2024. 1, 2, 3, 6, 12, 14

[38] Mohammadreza Salehi, Michael Dorkenwald, Fida Mohammad Thoker, Efstratios Gavves, Cees GM Snoek, and Yuki M Asano. Sigma: Sinkhorn-guided masked video modeling. In *European Conference on Computer Vision*, pages 293–312. Springer, 2024. 4, 12, 13

[39] Tim J Schoonbeek, Fabrizio J Piva, Hamid R Abdolhay, and Gijs Dubbelman. Learning to predict collision risk from simulated video data. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 943–951. IEEE, 2022. 3

[40] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[41] Xinyu Sun, Peihao Chen, Liangwei Chen, Changhao Li, Thomas H Li, Mingkui Tan, and Chuang Gan. Masked motion encoding for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2235–2245, 2023. 4, 12, 13

[42] Tiago Tamagusko, Matheus Gomes Correia, Minh Anh Huynh, and Adelino Ferreira. Deep learning applied to road accident detection with transfer learning and synthetic images. *Transportation research procedia*, 64:90–97, 2022. 4

[43] Fida Mohammad Thoker, Letian Jiang, Chen Zhao, and Bernard Ghanem. Smile: Infusing spatial and motion semantics in masked video learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8438–8449, 2025. 3, 4, 5, 7

[44] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2, 3, 4, 5, 7, 12, 13, 14

[45] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 3, 5, 6, 12

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[47] Tuan-Hung Vu, Eduardo Valle, Andrei Bursuc, Tommie Kerssies, Daan de Geus, Gijs Dubbelman, Long Qian, Bingke Zhu, Yingying Chen, Ming Tang, Jinqiao Wang, Tomáš Vojíř, Jan Šochman, Jiří Matas, Michael Smith, Frank Ferrie, Shamik Basu, Christos Sakaridis, and Luc Van Gool. The BRAVO Semantic Segmentation Challenge Results in UNCV2024. 2024. 2

[48] Junyao Wang, Arnav Vaibhav Malawade, Junhong Zhou, Shih-Yuan Yu, and Mohammad Abdullah Al Faruque. Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7493–7502, 2024. 4

[49] Lin Wang, Fuqiang Zhou, Zuoxin Li, Wangxia Zuo, and Haishu Tan. Abnormal event detection in videos using hybrid spatio-temporal autoencoder. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2276–2280. IEEE, 2018. 2

[50] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14549–14560, 2023. 4, 5, 7

[51] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, and Yu-Gang Jiang. Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6312–6322, 2023. 2, 4, 5, 7

[52] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 4, 5, 7, 12

[53] Jiewen Yang, Xingbo Dong, Liujun Liu, Chao Zhang, Jiajun Shen, and Dahai Yu. Recurring the transformer for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14063–14073, 2022. 3

[54] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Yuchen Wang, Ella Atkins, and David Crandall. Dota: unsupervised detection of traffic anomaly in driving videos. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 1, 2, 3, 4, 5, 6, 12, 13, 14

[55] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 5, 7, 12

[56] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022. 3, 7

[57] Zhili Zhou, Xiaohua Dong, Zhetao Li, Keping Yu, Chun Ding, and Yimin Yang. Spatio-temporal feature encoding for traffic accident detection in vanet environment. *IEEE Transactions on Intelligent Transportation Systems*, 23(10): 19772–19781, 2022. 1, 4, 14