Self-Supervised Method for Risky Situation Detection in Road Traffic Sequences Using Video Masked Autoencoder

Abdelhafid Berroukham[1], Mohammed Lahraichi[2], Khalid Housni[3]
Department of Computer Science, Faculty of Science, Ibn Tofail University, Kenitra, Morocco[1, 2, 3]
CRMEF Casablanca-Settat, Casablanca, Morocco[2]

***Abstract*—Road traffic accidents are a significant public health issue, particularly in developing nations, where infrastructure and traffic monitoring systems may be limited. Risky situations including sudden stopping, lane switching, and near-misses can lead to accidents. In this study, we present an original approach for recognizing risky situations in road traffic sequences using Video Masked Autoencoder (VideoMAE), a self-supervised deep learning model built upon Vision Transformer architecture. By applying a pre-trained VideoMAE on a large dataset of videos and fine-tuning it on labeled traffic sequences categorized as risky or non-risky, our model learns spatiotemporal features without requiring extensive manual labeling. The method achieves high accuracy on testing data, demonstrating strong potential for high-risk detection with an accuracy of 95%. This study highlights the promise of self-supervised video representation learning for real-world safety applications and paves the way for the development of intelligent traffic monitoring and crash prevention tools.**

***Keywords*—Video processing; risk detection; VideoMAE; vision transformer; deep learning; computer vision**

I. Introduction

Road accidents are a significant public health concern, with an estimated 1.24 million lives lost to road traffic accidents annually, according to the World Health Organization (WHO)[1]. Hazardous road traffic situations, including rapid braking, lane changes, and near-misses, can lead to accidents. Therefore, it is crucial to develop techniques for detecting and preventing such situations.

Computer vision and artificial intelligence have advanced significantly in recent years, particularly in applications requiring automated interpretation of visual scenes. Traffic scene analysis is one such domain, playing a critical role in road safety and intelligent transportation systems. While notable progress has been made, detecting and predicting hazardous situations—such as sudden braking, lane changes, or near-crash events—remains challenging due to the dynamic and complex nature of real-world traffic environments.

Despite modern advances, existing approaches exhibit several limitations. Most rely on supervised learning, which requires large, labeled datasets that are both costly and time-consuming to produce. Furthermore, many methods focus primarily on object recognition (e.g., identifying pedestrians or vehicles) while overlooking the temporal dynamics and interactions that are crucial for assessing risk in time-varying scenes. As a result, there is a gap in the literature regarding models capable of learning meaningful patterns from unlabeled video data or operating effectively with limited annotated datasets.

To address these challenges, self-supervised learning approaches offer a promising alternative. By reducing or eliminating the need for manual labeling, they present a more efficient and scalable solution for training models in complex traffic scenarios. Video Masked Autoencoder (VideoMAE) is one of the self-supervised learning techniques that has proved to be useful for many different tasks [2], including motion estimation, object detection, and video classification. VideoMAE learns video representations by predicting masked patches in the input videos. These learned representations prove valuable for various downstream tasks, including detection, classification, and segmentation.

In this study, we propose a method based on VideoMAE for detecting risky situations in road traffic sequences. By leveraging VideoMAE's ability to learn rich spatiotemporal features from video data. We train the model on a set of road traffic sequences, which are labeled as risky or non-risky. After the model is trained, it can be employed to find risky conditions in new road traffic sequences. This method not only reduces the reliance on labeled datasets but also improves generalization across diverse environments.

Our contributions are as follows:

- We propose a novel method to detect risky situations in road traffic sequences using VideoMAE.
- We train a VideoMAE model on a small dataset of road traffic sequences (fine-tuning).
- We evaluate our model on a held-out test set.

The rest of this study is organized as follows: In Section II (Related Work), we review existing approaches used for detecting risky situations in road traffic and highlight their limitations; Section III (Method) introduces the VideoMAE model, detailing its architecture and the preprocessing steps applied to the video data; Section IV (Experiment and Results) presents the experimental setup as well as the quantitative and qualitative results; Section V (Discussion) provides an analysis of the results; and Section VI (Conclusion) summarizes the key findings of our research and suggests areas for further

exploration to enhance the model's performance and applicability in real-world traffic safety applications.

II. Related Work

There are a number of existing methods for detecting risky situations in road traffic [3], [4], [5], as well as others that consider road accident classification a pivotal field of study [6], while some studies have focused on driving behavior [7], [8]. These methods can be broadly divided into two categories: traditional methods and machine learning methods.

The classical approach to detecting hazardous road traffic situations relies on handcrafted features, such as vehicle speed, position, and acceleration [9]. These features are then used to train a classifier model to predict the risk level of a given situation.

Machine learning algorithms for detecting dangerous situations on the road often employ deep learning models that automatically learn features from the data [5], [10]. These models are typically trained on small datasets of traffic monitoring videos [11]. The most widely used machine learning techniques for this task are Convolutional Neural Networks (CNNs)[11] and Recurrent Neural Networks (RNNs) [12]. CNNs are deep learning models particularly well suited for processing images and videos [13]. They have proven effective in various road traffic safety applications, such as vehicle detection, road sign recognition, and lane detection[11]. RNNs, on the other hand, are better suited for tasks involving sequential data, such as traffic flow forecasting and vehicle tracking [14].

Recently, there has been growing evidence of the success of self-supervised learning in a variety of computer vision tasks [15], [16], including object detection, motion estimation, and video classification. Self-supervised learning techniques automatically learn representations from data without the need for human-labeled annotations. One of the most widely used self-supervised learning techniques is VideoMAE[2]. VideoMAE learns to represent videos by predicting masked-out patches within the video frames. These learned representations can then be used for several downstream tasks, including classification, detection, and segmentation.

To the best of our knowledge, VideoMAE has not yet been applied to detect unsafe conditions in road traffic sequences. However, its demonstrated success in other domains suggests that it could be an effective model for identifying hazardous situations in traffic scenes.

In this study, we explore the use of VideoMAE for detecting risky situations in road traffic sequences. We fine-tune the VideoMAE model on a dataset of road traffic sequences labeled as either "risky" or "non-risky". Once trained, the model can be used to detect hazardous conditions in previously unseen traffic sequences. We believe that this approach could become a valuable tool for developing innovative safety solutions in the domain of road traffic management.

III. Method

In this section, we present our proposed approach, which involves fine-tuning the VideoMAE model using a dataset of risky road traffic situations. The model achieved an accuracy of 0.95 on both the validation and testing datasets. Fig. 4 illustrates the diagram of the proposed technique. First, we review VideoMAE[2], which is based on ImageMAE[17].

A. Masked Autoencoder (ImageMAE)

The Masked Autoencoder (MAE) [17] is a neural network architecture designed to learn meaningful latent representations from data and map them to a high-dimensional space using large datasets. It operates on the principle of randomly masking portions of the input image and then reconstructing the missing regions. The model follows a key design principle: an asymmetric encoder–decoder architecture. The encoder processes only the visible (unmasked) patches, while the lightweight decoder reconstructs the complete image using the latent representations along with the mask tokens.

It randomly masks out 75% of the grid patches for reconstruction, creating a challenging and meaningful self-supervised learning task. Masked autoencoders have proven to be scalable self-supervised models for a wide range of computer vision applications. This strategy is widely adopted because of its simplicity and effectiveness.

B. Video Masked Autoencoder (VideoMAE)

Video Masked Autoencoder (VideoMAE) [2] is an efficient self-supervised learning framework for video pre-training tasks that promotes the extraction of more informative video representations. Remarkably, it delivers strong performance even when trained on limited datasets, without relying on additional external data (i.e., fully self-supervised). The model is inspired by ImageMAE [17].

VideoMAE performs video self-supervised learning using a straightforward Vision Transformer (ViT) [18] backbone combined with a basic masked autoencoder. Due to its extremely high masking ratio, VideoMAE achieves a significant reduction in pre-training time compared to contrastive learning techniques, approximately a 3.2× speedup.

The architecture of VideoMAE includes three key components: an encoder network, a decoder network, and a masking mechanism. The encoder processes input video frames, extracting hierarchical features that capture both spatial and temporal dependencies within the sequences. These encoded features are then passed to the decoder, which reconstructs the original video frames while minimizing reconstruction error. Fig. 1 illustrates the model architecture based on VideoMAE. This self-supervised approach enables VideoMAE to effectively capture nuanced behaviors and subtle indicators of risky situations in traffic videos, such as sudden braking, lane changes, wrong-direction driving, or accidents. Moreover, its ability to learn from unlabeled data reduces the dependency on large annotated datasets, making it scalable and adaptable to various traffic scenarios.

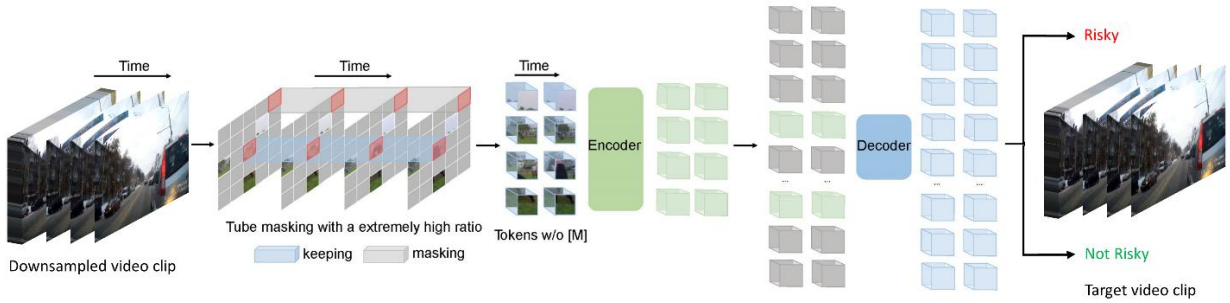

Fig. 1. Used model architecture based on video masked autoencoder.

Given these advantages, we assert that VideoMAE has strong potential as a powerful tool for detecting risky situations in road traffic sequences.

C. Data collection (Datasets)

To validate the effectiveness of the proposed model, we conduct experiments on large-scale datasets. Due to the limited availability and quality of publicly available datasets related to risky situations in road traffic, we collected and created our own data from various sources. This dataset comprises a combination of publicly available traffic video datasets. Public datasets such as the Car Crash Dataset (CCD), which is designed for traffic accident anticipation [14]; XD-Violence[19], a large-scale audio-visual violence detection dataset; and CADP [20], a dataset for accident analysis based on CCTV traffic cameras, provide a rich collection of annotated video sequences captured from various traffic scenarios. These datasets offer diverse examples of road conditions, vehicle interactions, and environmental factors. Moreover, the cameras continuously record traffic flow, capturing various risky situations such as sudden lane changes, speeding, and near-miss collisions. Fig. 2 shows some samples of risky situations from the dataset used.

Fig. 2. Samples of various risky situations on traffic roads from the used dataset.

D. Preprocessing

The collected video data undergoes several preprocessing steps, including frame extraction, resolution adjustment, and noise reduction, to ensure high-quality inputs for the VideoMAE model. This combination of diverse and comprehensive data sources enables a robust evaluation of the model's capability to detect risky situations in various traffic contexts.

The dataset contains videos with variable input frame sizes. Therefore, all videos are resized to a target resolution of 224×224 pixels. For the VideoMAE model, each input clip consists of sixteen frames. We use a frame rate of four, which defines the stride or interval between selected frames. From each video, we extract a single clip.

For training data transformations, we apply random cropping, pixel normalization, uniform temporal subsampling, and random horizontal flipping with a probability of 0.5. For the validation and evaluation datasets, we apply the same transformation pipeline, excluding random cropping and horizontal flipping, to ensure consistency during evaluation.

All preprocessing and data augmentation transformations are implemented using TorchVision's transforms module.

E. Training and Validation

The training process of the model begins with the preparation and preprocessing of the dataset, which includes frame extraction, resizing, and normalization. Typically, the dataset is divided into three subsets: training, validation, and test sets. The training set is used to train the VideoMAE model and contains a diverse range of traffic sequences, allowing the model to learn, reconstruct, and understand the underlying patterns in the data.

The validation set plays a crucial role in adjusting hyperparameters and monitoring the model's performance during training. The test set, which remains completely separate from the training and validation sets, serves as the final benchmark to assess the model's generalization ability to unseen traffic scenarios and its accuracy in detecting risky situations. Fig. 3 shows the distribution of data across each class for the three subsets of the dataset.

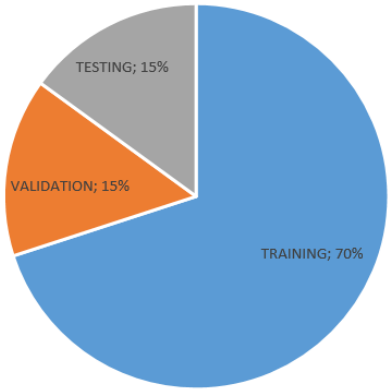

Fig. 3. Distribution of the used dataset.

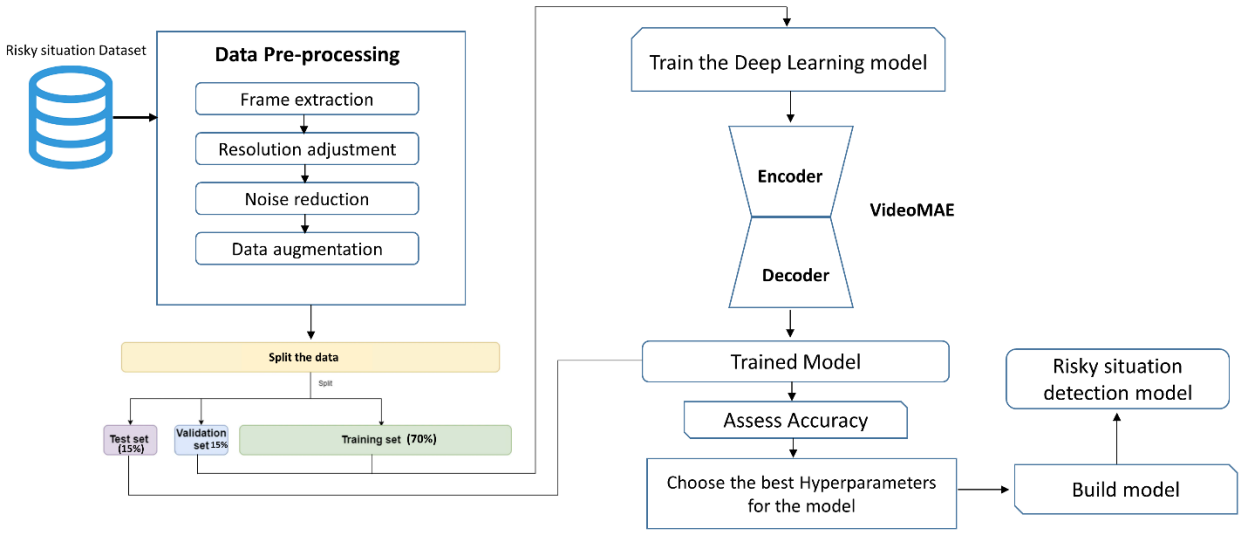

Fig. 4. Methodology diagram.

F. Evaluation Metrics

In this section, we present the metrics used to assess our model and analyze the implemented method.

Accuracy: It is an evaluation metric that measures the ratio of correctly classified sequences to the total number of sequences. It provides an assessment of how well the model's predictions align with the ground-truth annotations. In the context of video classification, accuracy is defined as the proportion of video sequences that are correctly classified into their actual categories. It reflects the precision with which the predicted labels match the true labels across the video dataset.

$$Accuracy = \frac{Total\ of\ accurate\ predictions}{Total\ of\ predictions\ made} \quad (1)$$

Precision: It is the ratio of true positive predictions to the total number of positive predictions made by the model. It indicates how many of the instances predicted as positive are actually correct, providing insight into the reliability of the model's positive classifications.

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \quad (2)$$

Recall: It is a measuring metric that is the ratio of the true positives that were detected to the total actual number of the same class. It shows how good the model is at detecting instances of all categories.

$$Recall = \frac{True\ positive}{True\ positive + False\ negative} \quad (3)$$

F1-score: It is the harmonic mean of precision and recall. It provides a balanced measure of a model's performance. It is especially useful when there is an uneven class distribution or when both false positives and false negatives need to be considered equally important.

$$F1 - Score = 2\ x\ \frac{recall\ x\ precision}{recall + precision} \quad (4)$$

Confusion Matrix: The confusion matrix is an effective tool for evaluating classification tasks, as it provides a detailed summary of the model's performance. It records the number of true positives, true negatives, false positives, and false negatives, offering a comprehensive view of how well the model performs across different classes. This information enables the straightforward calculation of key evaluation metrics such as accuracy, precision, recall, and F1-score.

IV. Experiment and Results

A. Experiment and Results

The training of MAE models, especially in the context of video recognition, presents a substantial computational challenge, limiting access for researchers with limited resources. For instance, training a standard VideoMAE model on Kinetics-400 [21] requires a significant investment of up to 5.6 days with the support of 64 GPUs [22]. Therefore, we fine-tune a pre-trained VideoMAE model, which was pre-trained on the Kinetics-400 dataset for approximately 1600 epochs without any additional data. We add a randomly initialized classification head on top of the pre-trained encoder and fine-tune the model on a labeled dataset of risky situations. Additionally, we initialize the feature extractor associated with the model.

As a risk detection network, we aim to detect risky situations in road traffic sequences using VideoMAE, which is based on Vision Transformer [18]. The performance evaluation of the model on the Risky Road Traffic dataset showed the best detection performance, achieving an accuracy of 0.9 with the masking strategy.

We train our model using the PyTorch framework with an environment of 12GB of RAM and a GPU (Nvidia Tesla T4) to expedite computation and handle the complexity of video data processing.

The hyperparameter configuration for the VideoMAE model was carefully selected to optimize performance while maintaining computational efficiency. The Adam optimizer [23] was chosen for its ability to adaptively adjust learning rates during training. The learning rate was initially set to 5e-5, a value found to provide an optimal balance between fast

learning and avoiding overfitting. A weight decay of 0.05 was implemented to regularize the model and reduce overfitting potential by penalizing large weights and encouraging generalization. The batch size was set to two to prevent out-of-memory errors while maintaining a balance between computational resource limitations and gradient quality requirements. The model was trained for ten epochs, a duration sufficient for learning useful patterns without overfitting risk. A summary of these hyperparameters is provided in Table I.

TABLE I. THE HYPERPARAMETERS OF THE MODEL

Hyperparameter	Value
Optimizer	Adam
Weight decay	0.05
Learning rate	5e-5
Batch size	2
Epochs	10

This configuration enabled effective learning from traffic video sequences, achieving high accuracy in risky situation detection while maintaining robustness across diverse traffic scenarios.

B. Results

The experimental results on benchmark datasets demonstrate that our approach, which has minimal computational cost and efficient data, performs well in the unsafe situation identification task.

Epoch	Training Loss	Validation Loss	Accuracy
0	0.420000	1.151754	0.695652
1	0.422200	2.008217	0.619565
2	0.831100	0.158780	0.956522
3	0.290500	0.398978	0.923913
4	0.000300	0.573320	0.913043
5	0.001500	0.335427	0.956522
6	0.190500	0.357622	0.956522
7	0.000100	0.346179	0.945652
8	0.000100	0.304265	0.956522
9	0.000100	0.342376	0.956522

Fig. 5. Overview of the model's training process.

Fig. 5 presents an extensive description of the training of the model by tabulating the training loss, validation loss, and validation accuracy at each of the ten epochs. The training loss column follows the model's performance on the training data, showing an overall decline as the model improves at reconstructing the frames of the videos and recognizing risky situations. The validation loss column reports the model's validation set performance, which indicates how the model generalizes to new data after each pass through the epochs. The validation accuracy column reports the ratio of correct identification of risky situations on the validation set and is an unambiguous metric of the success of the model. Fig. 5 provides an overall trend of increasing validation accuracy as training advances, which flattens towards the end epochs, indicating the model's success at generalization of learned features. These metrics, together, provide a complete perspective on the learning and performance of the model during the training period.

Table II shows an overall assessment of the model's performance through various important metrics: Precision, Recall, F1-score, and Accuracy.

TABLE II. EVALUATION MODEL

Metric	Value
Precision	0.960
F1-score	0.955
Recall	0.954
Accuracy	0.95

Precision at 0.960 shows the model's capacity to accurately detect true instances of risky situations and hints at high accuracy with few false positives.

Recall at 0.954 indicates the model's capacity to detect the majority of actual risky situations and shows that it is efficient enough to capture most of the dangerous events in the traffic sequences.

The F1-score, determined as the harmonic mean of Precision and Recall, is at 0.955 and offers a balanced evaluation of the model's precision and recall performance.

Lastly, accuracy at 0.950 is the overall ratio of correctly classified instances, whether risky or non-risky. This high accuracy is an attestation to the model being consistent regardless of different scenarios.

All these metrics together illustrate the good performance of the model in detecting and classifying risky situations on the road successfully and verify its usability for real-world traffic safety improvement.

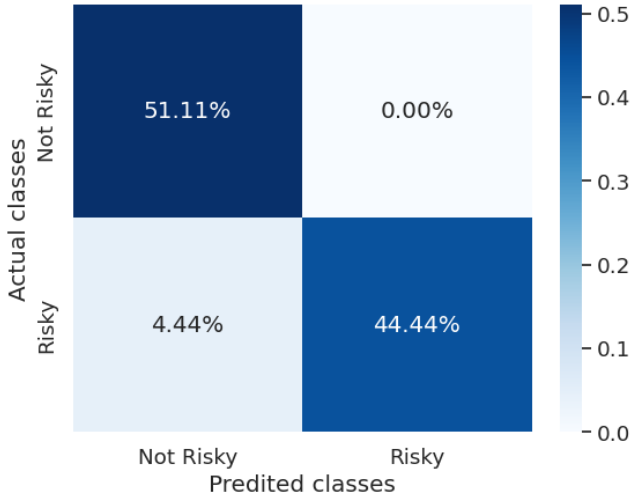

Fig. 6. Model evaluation using confusion matrix.

Fig. 6 indicates that the confusion matrix offers an insightful analysis of the model's performance by classifying the test results into true positives, true negatives, false positives, and false negatives.

The matrix shows that:

- The model successfully identified 44% of risky situations as true positives, demonstrating its strong capability to detect risky conditions in traffic sequences.
- The model correctly classified 51% of non-risky situations as true negatives, showing its good capability to identify safe conditions without false alarms.
- Notably, the model produced 0% false positives, meaning it never incorrectly labeled safe conditions as risky - a crucial factor for building confidence in the system's alerts.
- There were 4.44% false negatives, representing instances where the model failed to detect actual risky situations.

The trained model was tested on several video sequences to evaluate the proposed approach, and the results were collected and analyzed. The method demonstrates successful identification of dangerous traffic conditions in surveillance videos with 95% accuracy.

Fig. 7 illustrates the detection accuracy across multiple video sequences, showing that the model predicts correct situations with high probability.

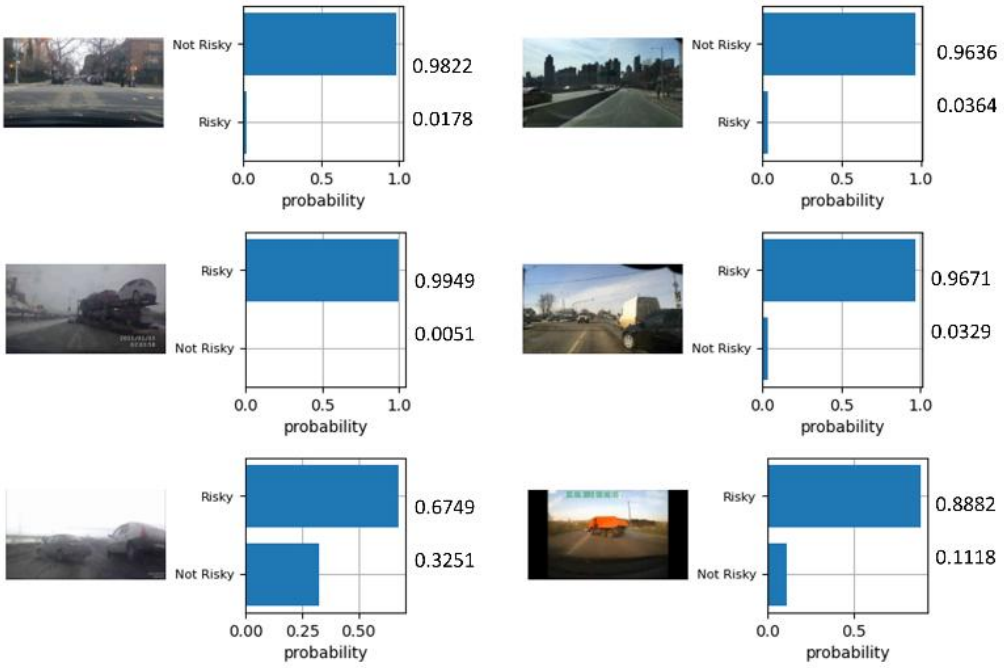

Fig. 7. Illustrations of many detection examples accompanied by the probability.

V. Discussion

The results obtained from the model highlight its strong potential to enhance traffic safety through the accurate detection of risky situations in road traffic sequences. The model's high Precision of 0.960 indicates its ability to correctly identify hazardous situations with minimal false alarms, which is critical in preventing unnecessary interventions. The Recall of 0.954 reflects the model's effectiveness in capturing nearly all actual risky situations, ensuring that critical events are not overlooked. The F1-score of 0.955 demonstrates a well-balanced performance between Precision and Recall, confirming that the model maintains consistency in identifying and classifying risks. Additionally, the accuracy of 0.95 indicates that the model performs reliably across different traffic scenarios. The confusion matrix also confirms these results with 44% true positives and no false positives. The absence of false alarms is essential for real-time traffic monitoring applications, since false alarms can cause driver desensitization or unjustified intervention. However, the model reported 4.44% false negatives, where it did not detect actual risky situations. This indicates that although the model is very accurate for the situations it does detect, there is further scope to improve its sensitivity to detect all potential danger areas. These results indicate the model's resilience but also the necessity for further improvement, specifically to increase its capability to detect all risky situations so that the occurrence of false negatives is minimized and overall traffic safety is comprehensive.

When compared to other methodologies for risky situation detection in road traffic, the model provides various improvements over previous approaches. The conventional approach tends to employ Convolutional Neural Networks (CNNs) to analyze single frames from traffic videos. Although CNNs are good at learning spatial features, they are weak at accounting for temporal dependencies between frames, since risky situations are often realized through unfolding events over time. To address the weakness of CNNs, Recurrent Neural Networks such as LSTM[24] networks are sometimes

employed together with CNNs to consider the temporal aspect. However, these models tend to perform poorly with long-range dependencies and are prone to vanishing gradients, thus unable to adequately understand intricate, multi-frame traffic situations.

By contrast, VideoMAE is transformer-based and specifically designed to model both spatial and temporal dependencies simultaneously. This enables the model to capture sequences of events in traffic, like an abrupt stop of a car or the appearance of a pedestrian on the roadway, which may not be detected or well-represented using standard CNN or RNN-based approaches.

In further testing the efficacy of our approach, we compared the model's performance with that of a Vision Transformer (ViT) model that has also been used for detecting risky situations from videos. As evident from Table III, our model performs significantly better than the ViT-based approach and other deep learning based methods, achieving 95% accuracy compared to 92% for the ViT model[25]. In addition, the work presented in [26], which aimed to predict traffic accidents using an RNN network, achieved an accuracy of 71%. In [27], the authors used a combination of models, including a multi-layer perceptron (MLP), and achieved an accuracy of 72%. Moreover, another study focused on road accident prediction by combining machine learning and deep learning models, specifically the RFCNN model [28], and achieved an accuracy of 81%. These values demonstrates VideoMAE's superior capability to handle both spatial and temporal dependencies in traffic sequences.

TABLE III. MODEL COMPARISON

Model	Accuracy
RNN[26]	71%
MLP[27]	72%
RFCNN[28]	81%
FN-ViT[25]	92%
Ours	**95%**

VI. CONCLUSION

In summary, the findings of this study demonstrate the capability of the VideoMAE-based model to detect risky conditions in road traffic streams and its potential contribution to enhancing traffic safety through advanced video data processing. The strong performance of the model, reflected by high Precision, Recall, and F1-score values, highlights its ability to accurately identify dangerous situations.

While the system demonstrates strong potential and high accuracy, a few drawbacks or areas for improvement have been identified. First, the training process is computationally intensive. Second, there is the occurrence of false negatives, where some risky situations go undetected, highlighting the need for improved sensitivity, particularly in complex traffic scenes. Nevertheless, integrating the model into real-time traffic monitoring systems holds significant promise for improving road safety by enabling early warnings and reducing the risk of accidents.

Overall, this work offers valuable insights into the application of deep learning for traffic safety and paves the way for future advancements in intelligent transportation systems.

REFERENCES

[1] A. A. Mohammed, K. Ambak, A. M. Mosa, et D. Syamsunur, « A Review of the Traffic Accidents and Related Practices Worldwide », Open Transp. J., vol. 13, no 1, p. 65-83, juin 2019, doi: 10.2174/1874447801913010065.

[2] Z. Tong, Y. Song, J. Wang, et L. Wang, « VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training », 18 octobre 2022, arXiv: arXiv:2203.12602. Consulté le: 16 septembre 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2203.12602

[3] S. K. Kumaran, D. P. Dogra, et P. P. Roy, « Anomaly Detection in Road Traffic Using Visual Surveillance: A Survey », ACM Comput. Surv., vol. 53, no 6, p. 1-26, nov. 2021, doi: 10.1145/3417989.

[4] Y. Ma, J. Xu, C. Gao, M. Mu, G. E, et C. Gu, « Review of Research on Road Traffic Operation Risk Prevention and Control », Int. J. Environ. Res. Public. Health, vol. 19, no 19, p. 12115, sept. 2022, doi: 10.3390/ijerph191912115.

[5] A. Berroukham, K. Housni, M. Lahraichi, et I. Boulfrifi, « Deep learning-based methods for anomaly detection in video surveillance: a review », Bull. Electr. Eng. Inform., vol. 12, no 1, p. 314-327, févr. 2023, doi: 10.11591/eei.v12i1.3944.

[6] M. Sobhana, G. S. S. Venkatesh Mendu, N. Vemulapalli, et K. Kumar Chintakayala, « Optimized feature selection approaches for accident classification to enhance road safety », IAES Int. J. Artif. Intell. IJ-AI, vol. 13, no 3, p. 3283, sept. 2024, doi: 10.11591/ijai.v13.i3.pp3283-3290.

[7] S. Bouhsissin, N. Sael, et F. Benabbou, « Driver Behavior Classification: A Systematic Literature Review », IEEE Access, vol. 11, p. 14128-14153, 2023, doi: 10.1109/ACCESS.2023.3243865.

[8] A. Salbi, M. A. Gadi, T. Bouganssa, A. Eloudrhiri Hassani, et A. Lasfar, « Design and implementation of a driving safety assistant system based on driver behavior », IAES Int. J. Artif. Intell. IJ-AI, vol. 13, no 3, p. 2603, sept. 2024, doi: 10.11591/ijai.v13.i3.pp2603-2613.

[9] J. Zhao et al., « Unsupervised Traffic Anomaly Detection Using Trajectories », p. 8.

[10] L. Tišljarić, S. Fernandes, T. Carić, et J. Gama, « Spatiotemporal Road Traffic Anomaly Detection: A Tensor-Based Approach », Appl. Sci., vol. 11, no 24, p. 12017, déc. 2021, doi: 10.3390/app112412017.

[11] S. W. Khan et al., « Anomaly Detection in Traffic Surveillance Videos Using Deep Learning », Sensors, vol. 22, no 17, p. 6563, août 2022, doi: 10.3390/s22176563.

[12] R. M. Schmidt, « Recurrent Neural Networks (RNNs): A gentle Introduction and Overview », 23 novembre 2019, arXiv: arXiv:1912.05911. Consulté le: 4 août 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1912.05911

[13] K. O'Shea et R. Nash, « An Introduction to Convolutional Neural Networks », 2 décembre 2015, arXiv: arXiv:1511.08458. Consulté le: 22 mai 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1511.08458

[14] W. Bao, Q. Yu, et Y. Kong, « Uncertainty-based Traffic Accident Anticipation with Spatio-Temporal Relational Learning », in Proceedings of the 28th ACM International Conference on Multimedia, oct. 2020, p. 2682-2690. doi: 10.1145/3394171.3413827.

[15] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, et M. Shah, « Anomaly Detection in Video via Self-Supervised and Multi-Task Learning », 10 septembre 2021, arXiv: arXiv:2011.07491. Consulté le: 22 février 2023. [En ligne]. Disponible sur: http://arxiv.org/abs/2011.07491

[16] N. Madan et F. S. Khan, « Self-Supervised Masked Convolutional Transformer Block for Anomaly Detection », IEEE Trans. PATTERN Anal. Mach. Intell., vol. 14, no 8, 2022.

[17] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, et R. Girshick, « Masked Autoencoders Are Scalable Vision Learners », in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New

Orleans, LA, USA: IEEE, juin 2022, p. 15979-15988. doi: 10.1109/CVPR52688.2022.01553.

[18] A. Dosovitskiy et al., « An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale », 3 juin 2021, arXiv: arXiv:2010.11929. Consulté le: 22 mai 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/2010.11929

[19] P. Wu et al., « Not only Look, But Also Listen: Learning Multimodal Violence Detection Under Weak Supervision », in Computer Vision – ECCV 2020, vol. 12375, A. Vedaldi, H. Bischof, T. Brox, et J.-M. Frahm, Éd., in Lecture Notes in Computer Science, vol. 12375. , Cham: Springer International Publishing, 2020, p. 322-339. doi: 10.1007/978-3-030-58577-8_20.

[20] A. P. Shah, J.-B. Lamare, T. Nguyen-Anh, et A. Hauptmann, « CADP: A Novel Dataset for CCTV Traffic Camera based Accident Analysis », in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand: IEEE, nov. 2018, p. 1-9. doi: 10.1109/AVSS.2018.8639160.

[21] W. Kay et al., « The Kinetics Human Action Video Dataset », 19 mai 2017, arXiv: arXiv:1705.06950. Consulté le: 31 juillet 2024. [En ligne]. Disponible sur: http://arxiv.org/abs/1705.06950

[22] Xianhang Li, Peng Wang, Xinyu Li, Heng Wang and Cihang Xie, « Efficient VideoMAE via Temporal Progressive Training », 2024. [En ligne]. Disponible sur: https://openreview.net/forum?id=vex1yNHNFL

[23] D. P. Kingma et J. Ba, « Adam: A Method for Stochastic Optimization », 29 janvier 2017, arXiv: arXiv:1412.6980. Consulté le: 12 juin 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1412.6980

[24] S. Hochreiter et J. Schmidhuber, « Long Short-Term Memory », Neural Comput., vol. 9, no 8, p. 1735-1780, nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[25] A. Berroukham, K. Housni, et M. Lahraichi, « Fine-Tuning Pre-trained Vision Transformer Model for Anomaly Detection in Video Sequences », in Proceedings of the 6th International Conference on Big Data and Internet of Things, vol. 625, M. Lazaar, E. M. En-Naimi, A. Zouhair, M. Al Achhab, et O. Mahboub, Éd., in Lecture Notes in Networks and Systems, vol. 625. , Cham: Springer International Publishing, 2023, p. 279-289. doi: 10.1007/978-3-031-28387-1_24.

[26] M. Sameen et B. Pradhan, « Severity Prediction of Traffic Accidents with Recurrent Neural Networks », Appl. Sci., vol. 7, no 6, p. 476, juin 2017, doi: 10.3390/app7060476.

[27] L. Wahab et H. Jiang, « Severity prediction of motorcycle crashes with machine learning methods », Int. J. Crashworthiness, vol. 25, no 5, p. 485-492, sept. 2020, doi: 10.1080/13588265.2019.1616885.

[28] M. Manzoor et al., « RFCNN: Traffic Accident Severity Prediction Based on Decision Level Fusion of Machine and Deep Learning Model », IEEE Access, vol. 9, p. 128359-128371, 2021, doi: 10.1109/ACCESS.2021.3112546.