sciendo

Transport and Telecommunication, 2024, volume 25, no. 1, 20–30
Transport and Telecommunication Institute, Lomonosova 1, Riga, LV-1019, Latvia
DOI 10.2478/ttj-2024-0003

EFFICIENT ROAD TRAFFIC VIDEO CONGESTION CLASSIFICATION BASED ON THE MULTI-HEAD SELF-ATTENTION VISION TRANSFORMER MODEL

Sofiane Abdelkrim Khalladi[1*], Asmâa Ouessai[2], Nadir Kamel Benamara[3], Mokhtar Keche[4]

[1, 2,3,4]Signals and images laboratory, Faculty of Electrical Engineering, Department of Electronics, University of Sciences and Technology of Oran Mohamed Boudiaf USTO-MB, B.P. 1505, El Mnaouar-Bir el Djir-Oran, Algeria
[2]Faculty of Technology, Department of Telecommunications, Dr. Tahar Moulay University, Saida, Algeria
[1]Sofianeabdelkrim.khalladi@univ-usto.dz
[2]asmaa.ouessai@univ-saida.dz
[3]nadirkamel.benamara@univ-usto.dz
[3]mokhtar.keche@univ-usto.dz

Due to rapid population growth, traffic congestion has become one of the major issues in urban areas. The utilization of technology may help to address this issue. This paper proposes a new Multi-head Self-attention Vision Transformer (MSViT) based macroscopic approach, for road traffic congestion classification. To evaluate this approach, we use the UCSD (University of California San Diego) dataset that includes different weather conditions (clear, overcast and rainy) and different traffic scenarios (light, medium and heavy). The classification accuracy reached a high level of 99.76% with this dataset and 99.37% when night-mode frames are added to it. The proposed MSViT based method outperforms the state-of-the-art macroscopic and microscopic methods that have been evaluated using the same UCSD dataset, which makes it an efficient solution for traffic congestion prediction.

Keywords: Road traffic classification, macroscopic approach, vision transformers, multi-head self-attention, deep learning

1. Introduction

Road traffic congestion is a common problem in urban areas; it is characterized by slow or stopped traffic on roads. A variety of factors can contribute to this problem such as, increased vehicle ownership, population growth and limited road infrastructure. Traffic congestion results in prolonged travel times, air pollution and reduced efficiency of the transportation system. Several approaches can be implemented to alleviate road traffic congestion. Therefore, intelligent Transportation Systems (ITS) are increasingly being used to improve traffic flow through actions, such as traffic signal optimization, incidents management, and advanced traveler information systems. Effective transportation planning and management are crucial in reducing traffic congestion and its negative impacts. Road traffic classification is used to categorize vehicles based on their characteristics such as size, weight, and usage. This technique is used in transportation planning and engineering to help analyze the traffic flow on a particular road or on a network. Several methods have been used for road traffic classification; the most common one is based on visual observation (Bull & CEPAL, 2003). In this method, a trained observer records the type and number of vehicles passing through a specific point over a period of time. Electronic systems such as loop detectors can also be used to classify vehicles based on their characteristics. Road traffic congestion classification involves estimating the traffic state, based on some selected features related to vehicle motion, such as speed, traveled distance, and the number of vehicles. These variables are referred to as microscopic parameters of the traffic. On the other hand, the macroscopic parameters are selected based on an overall analysis of the scene. Road traffic congestion classification can be affected by several factors, the most common of which are weather conditions. Inclement weather, such as rain or snow can alter the behavior and movement of vehicles on the road, resulting in changes in traffic flow and speeds. The reduced visibility during bad weather can result in increased braking and slower speeds, leading to more congested traffic. Heavy rainfall can also lead to water accumulation on roads, making it challenging for vehicles to keep traction. In order to mitigate the above issues in traffic congestion estimation, we propose, in this paper, a new traffic classification method that is based solely on the Multi-head Self-attention Vision Transformer

Model. This method is evaluated using UCSD dataset images. Employing images in deep learning is crucial because they provide a rich and detailed source of information, comprising features such as patterns, textures, and spatial connections. Deep learning models are ideal for capturing and comprehending this information. Furthermore, the hierarchical feature learning capabilities enables them to automatically detect and extract significant features from raw picture data. This is a significant benefit since it eliminates the need for manual feature engineering, and improves their adaptability to varied tasks without requiring explicit human interaction and can detect significant patterns and representations on their own, which is particularly useful when dealing with big datasets and complicated visual information.
Our contributions throughout this study are listed below:

(1) We classify traffic congestion based on videos recorded under various weather conditions and illumination changes, including nighttime scenarios.
(2) We classify traffic congestion using MSViT without the need for vehicles' monitoring for parameters' extraction.
(3) We reduce the image size from 320 x 240 pixels to 16 x 16 and reach high classification accuracy without allocating a large RAM memory.
(4) The proposed method is suitable for real-time video streaming applications.
(5) Using the public dataset UCSD (Chan & Vasconcelos, 2004), we evaluate the proposed method and show that it outperforms the state- of-the-art microscopic and macroscopic methods.

The rest of the paper is organized as follows: Section 2 presents the previous methods proposed for traffic classification, with a brief discussion of the flaws of each method. Section 3 describes the proposed method. It starts with a presentation of the proposed system's structure, for road traffic classification, and then it explains the principal stage, which includes the encoder and the decoder. Section 4 is dedicated to the presentation and discussion of the obtained classification results. Finally, the main conclusions are summerised in Section 5.

2. Related works

Several studies have addressed the problem of classifying road traffic congestion based on the macroscopic approach. (Asmaa *et al.*, 2013) presented a comparison study of macroscopic and microscopic approaches for estimating road traffic density. The classification task was performed using three classifiers: The K Nearest Neighbor (KNN) classifier, the Learning Vector Quantization (LVQ) classifier, and the SVM classifier, in order to classify the road traffic into three categories: light, medium, and heavy. The goal of this research was to develop a high-precision algorithm for estimating road traffic density. It was found that the SVM with the RBF kernel and one versus one scheme produced the highest classification rate (96.37%) when the macroscopic parameters were used. However, the adopted block-matching method in this macroscopic approach can be sensitive to noise, illumination changes, and occlusions, which can affect the quality of motion estimation. This can result in incorrect motion vectors, which produce misclassifications of traffic flow. (Gonçalves *et al.,* 2015) applied the technique of dynamic textures recognition to traffic classification. This technique models the object's motion with a graph in which the nodes are represented by the objects (vehicles), and the edges define the motion relationships between the objects. The results showed that it can accurately classify different types of traffic scenarios and that it outperformed several existing dynamic texture recognition methods. Using the UCSD dataset and the SVM classifier, the method achieved a correct classification rate of 97.24%. (Sobral *et al.,* 2013) presented a method for classifying highway traffic congestion levels using holistic properties of traffic videos. The method first extracts holistic features from the video, such as motion energy, color, and texture, then classifies the traffic congestion level. The results showed that the proposed method can accurately classify different levels of traffic congestion. When applied to the 254 highway traffic videos of the UCSD dataset, the features classification with neural networks gave 94.50% accuracy. Nerveless, changes over time, such as changes in lighting or object movement, can affect the holistic properties of a scene and represent a challenge for video classification algorithms.

(Chan *et al.,* 2005) proposed a method for classifying and retrieving traffic video data using Auto-Regressive Stochastic Processes (ARSP). The traffic was first modeled as an ARSP and then classified using the SVM classifier. The obtained results, using real-world traffic video data, indicated that the proposed method can accurately classify the traffic congestion, as light, medium, or heavy. However, the method requires significant computational resources to extract and process the ARSPs, which may limit its use in real-time traffic monitoring systems. (Derpanis *et al.,* 2011) presented a method for categorizing recordings of traffic jams, according to the observed visual dynamics. Traffic flow identification is viewed as an instance of dynamic texture classification at the core of the proposed system. This method does not

require segmentation, tracking, or motion estimation. Matching space-time orientation structure distributions (or histograms) serves as the basis for classification. An empirical study, using a publicly accessible data set, demonstrated strong classification performance and robustness to typical environmental circumstances (e.g., variable lighting). Unfortunately, some factors like the frame rate of the video sequence, the speed of the objects, and the presence of noise can affect the accuracy of the spatiotemporal orientation analysis, making it difficult to produce trustworthy results in some cases. (Luo *et al.,* 2015) explored the potential of traffic monitoring in the absence of any motion features. The proposed approach was designed to process ultra-low frame rate films or videos for which accurate motion characteristics cannot be computed. The assessment of traffic state as fluid, dense, or congested, was performed using 2D spatial characteristics and a machine learning algorithm. To evaluate the accuracy of this approach several 2D textures attributes and machine learning techniques were adopted. The experimental findings using the UCSD traffic dataset showed that this approach generalized well across different weather and lighting conditions. It even performed better than cutting-edge spatiotemporal feature-based traffic analysis techniques. (Luo *et al*., 2016) investigated the monitoring of highway traffic using movies with a frame rate that is too slow to reliably infer motion features. Instead of measuring traffic conditions as is typically done, the suggested method seeks to identify them. This approach takes advantage of the highly repetitive nature of top-down traffic scenes. The limited variety of items seen in traffic scenes allows one to focus on elements that are unique to these kinds of photographs. Without the use of motion characteristics, this method can segment traffic photos, categorize traffic scenes and calculate traffic density. Distinct Convolutional Neural Networks (CNN) models are suggested to categorize traffic photographs into separate groups (empty, fluid, heavy, jam), segment traffic images into three different classes (road, automobile, and background), and estimate traffic density. However, this approach may not perform well in non-ideal conditions, such as significant variations in lighting or camera angle, which can affect the quality of the video. (Riaz *et al.,* 2013) provided a motion vector statistical method for categorizing highway traffic congestion using the pyramidal Kanada-Lucas-Tomasi (KLT) tracker algorithm. Following the extraction of motion data, the traffic patterns were divided into three groups: light, medium, and heavy. The classification performed using neural networks demonstrated robustness to environmental factors, like fluctuating luminance, but no results were provided regarding occlusions and sensitivity to noise. (Wang *et al.,* 2023) developed a new traffic congestion prediction model called the spatio-temporal transformer (STTF), based on a deep learning model transformer and a traffic congestion index. The model takes into account a wide range of factors, including the topology of the road network, the spatio-temporal relationships between individual road segments, etc. In order to increase the precision of traffic congestion prediction, the authors embedded temporal and spatial information into the model through the embedding layer for learning and used the spatio-temporal attention module to extract the spatio- temporal information hidden within the data. However, the proposed method was evaluated using only a vehicle GPS dataset. (Chen *et al.,* 2012) proposed a system for classifying traffic congestion in surveillance footage taken at night. The framework was divided into three parts. Firstly, the headlights were identified using three key attributes. Secondly, by analyzing their relationships, headlamps were divided into various cars. Thirdly, a virtual detection line was used to collect traffic data for evaluating traffic congestion. In real-time, the amount of traffic was divided into five levels: jam, heavy, medium, moderate, and low. Satisfactory experimental findings confirmed the efficiency of the suggested framework. (Adnan *et al.,* 2022) proposed a novel method for classifying traffic videos that uses color coding before training a deep convolutional neural network with traffic data. The You Only Look Once (YOLO) (Jiang *et al.,* 2022) algorithm was used to accomplish vehicle recognition after the conversion of the video data into imagery data. The imagery dataset was afterwards converted into a binary image dataset using a color-coded system. Finally, these binary pictures were used to feed a Deep Convolutional Neural Network. A classification accuracy of 98.2% was achieved. (Ramana *et al*., 2023), the authors developed an effective traffic prediction system by using Vision Transformers (VTs) in conjunction with CNNs to predict the traffic flow. A traffic image is first fed into the CNN to produce feature maps, which are then delivered to the VT where they undergo a tokenization followed by a projection. This results in vision information, which is then passed to a LSTM (Long Short-Term Memory). The outcomes of the experiments carried out prove the accuracy of the vision transformer prediction method. Indeed, in terms of precision, accuracy, and recall, the suggested technique outperforms conventional approaches and saves energy. Road traffic classification using vision transformer has become an important research area in recent years. The vision transformer was first extremely effective in natural language processing field (Dosovitskiy *et al.,* 2020), and then it has been adopted into other domains, such as computer vision (Khan *et al*., 2022). It uses the self-attention mechanism, which allows the model to focus on important parts of the input image, making it more robust to scene and scale changes. This is particularly useful in the context of traffic classification, as the vehicles can appear in different

positions, scales, and orientations in the image. Although Vision Transformers (ViTs) have been employed in a number of image recognition tasks across various fields, their usage in traffic classification remains limited. Hence, we were motivated to incorporate them into our work. Unlike in reference (Ramana *et al.,* 2023), in the present work the ViTs are used without incorporating a CNN, which makes our approach simpler and more suitable for real-time applications.

3. Description of the proposed method

As stated previously, the proposed method is based on the use of a vision transformer to determine the status of the road traffic from images. Before describing this transformer, we defined, firstly, how its inputs are constructed. As illustrated in the left part of Figure 1, the image is first reshaped into a sequence of flattened 2D patches. By analogy with textual data, this process is called tokenization and the resulting patches are referred to as tokens. Then, the tokens are converted to their corresponding embeddings with a trainable linear projection.

3.1. The transformer network architecture

A transformer network is a special type of deep learning architecture, which gets rid of recurrences that make the optimization procedure time-consuming. It was introduced by (Vaswani *et al*., 2017) as a revolutionary shift in artificial networks.

The network architecture consists of two components: an encoder and a decoder. In both, the important part is the block that comes after the positional encoding, as shown in the right part of Figure 1, where the left and right halves represent the encoder and the decoder, respectively. This block contains two modules, a multi head attention module and a feed forward neural network module. The multi head attention module is the core of the transformer network. The inputs to this module are first split into several heads and then in each head, self-attention is used to compute a weighted sum of the input embeddings that is used as input to the next layer of the network. The weights are determined by the evaluation of the similarities between each pair of input embeddings. This is done by computing a dot product between each pair of embeddings and applying a Softmax function.

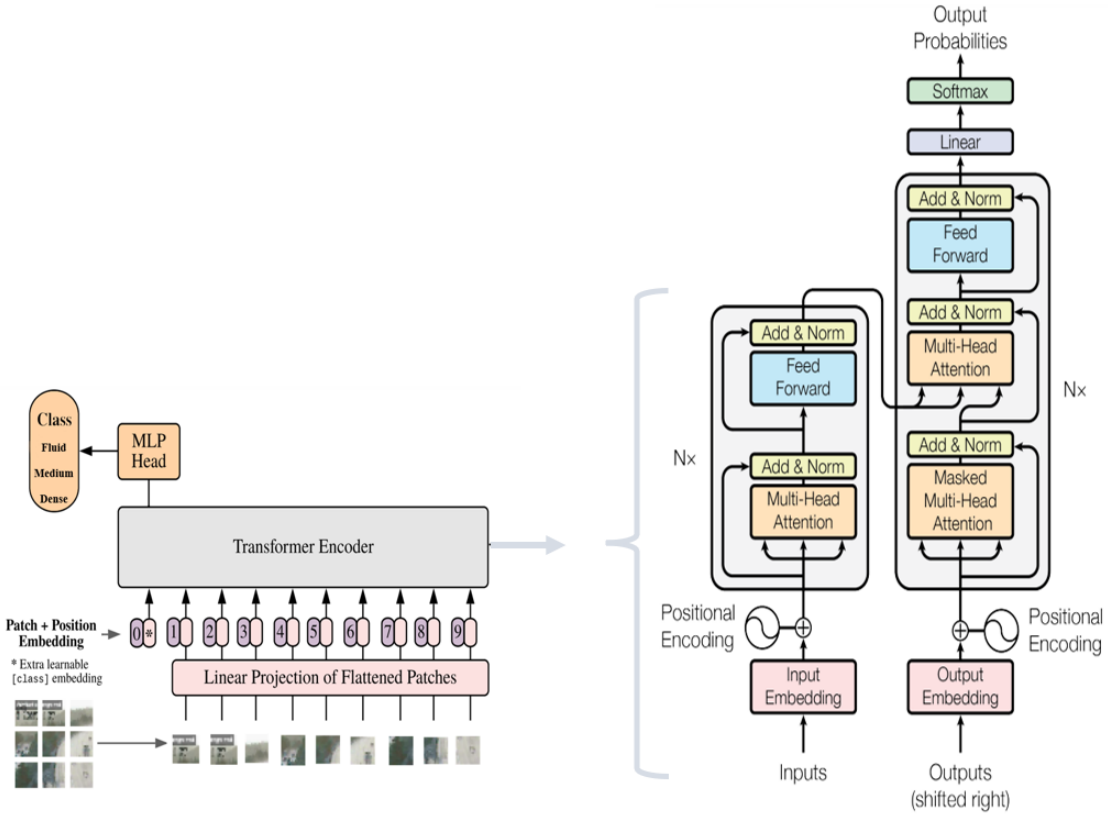


Figure 1. Left: Architecture of the vision transformer from (Dosovitskiy *et al.,* 2020). Right: detailed diagram of the transformer encoder and decoder used for the traffic congestion classification

To ensure that the attention scores are not too large or too small, these weights are normalized using a process called "batch normalization." This involves subtracting the mean of the attention scores across all inputs in a batch and dividing by the standard deviation. This ensures that the attention scores have zero mean and unit variance, which can help stabilizing the training process and improving the performance of the network. The results from the different heads are then concatenated and passed to a feed-forward neural network, which applies a non-linear transformation. This process is repeated for each layer in the encoder, with the output of one layer serving as an input to the next one. The final output of the encoder are the embeddings for each token in the input sequence. These embeddings are then passed to the decoder, which generates the output sequence.

The decoder is not only responsible for embedding some output but also for producing it. Hence, in the decoder block, the output of the previous timestep is fed into a masked multi-head attention block along with the current input. The masked attention mechanism ensures that each position can only attend to its preceding positions in the sequence, which prevents it from attending to positions that have not been predicted yet. Next, the output of the masked multi-head attention block is combined with the encoder outputs through another multi-head attention block. This allows the decoder to attend to relevant parts of the input sequence and generate context-aware representations for the output sequence. Finally, the combined output is passed through a feed-forward neural network block to produce the final output for the current timestep. This process is repeated N times, where N is the number of decoder blocks in the architecture. The Softmax function is used to compute the probability distribution over the output classes for each image. This allows the model to generate the most likely output class for each timestep, given the input sequence and the context learned by the model.

3.2. The self-attention mechanism

The self-attention mechanism is defined by three main components: the Query (Q), Key (K), and Value (V).

Given an input sequence of tokens, $X = \{x_1, x_2, \ldots, x_n\}$, the query, the key, and the value for each token are calculated using linear projections:

$$Q = W_q \cdot X, \tag{1}$$

$$K = W_k \cdot X, \tag{2}$$

$$V = W_v \cdot X, \tag{3}$$

where W_q, W_k, and Wv are weight matrices. The attention scores for each token are then calculated using the dot-product of the query and the key:

$$Attention = soft\max\left(\frac{Q^T \cdot K}{\sqrt{d_q}}\right), \tag{4}$$

where d_q is the dimension of the key vectors. The dot-product is scaled by the square root of the dimension to prevent large values. Finally, the attention scores are used to weight the values:

$$Output = Attention \cdot V. \tag{5}$$

This output is then passed to a feed-forward neural network, which applies a non-linear transformation. This process is repeated for each layer in the encoder. The final output of the encoder is the set of embeddings of different tokens in the input sequence. These embeddings capture the meaning and context of the input sequence and are used by the decoder to generate the output sequence.

3.3. Multi-head attention

Mathematically, the multi head attention function is described as follows:

$$multihead(Q,K,V) = W^0 concat(head_1, head_2, ..., head_h). \tag{6}$$

$$head_i = attention(W_i^Q Q, W_i^K K, W_i^V V). \tag{7}$$

$$attention(Q,K,V) = soft\max\left(\frac{Q^T K}{\sqrt{d_q}}\right). \tag{8}$$

The general idea behind the attention mechanism is the existence of similar key values (K) in a database to which a query (Q) is compared. The keys that are the most similar are assigned the highest weights used in a weighted combination of the corresponding values (V) to produce an output. Subsequently, V, K, and Q are fed into a linear layer to compute a scaled dot product attention, for each head. The outputs corresponding to different heads are then concatenated and input to a linear layer that provides the multi head attention output, as shown in Figure 2.

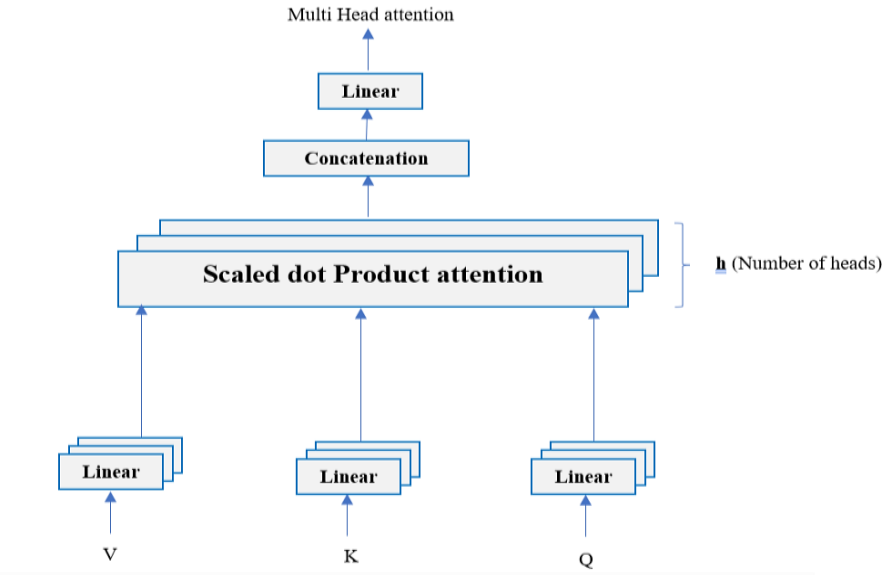


Figure 2. Multi-head attention

3.4. Masked multi head attention

The masked attention is calculated as follows:

$$maskedAttention(Q,K,V) = soft\max\left(\frac{Q^T K+M}{\sqrt{d_k}}\right), \tag{9}$$

where M is a mask matrix of zeros and -∞'s.

Masked multi head attention is a part of the decoder block; it aims to mask some of the values by nullifying their probabilities to prevent them from being selected for combination. In other words, when generating outputs in the decoder, it makes sense for actual outputs to depend on previous outputs; however, it is not logical that they depend on future outputs. For this reason, future output values are masked to prevent creating dependencies on outputs that have not yet been generated. Mathematically, in the multi head attention function, a variable M is added to effectively assign null probabilities to the unwanted future terms.

4. Experimental results

Before discussing the results of our road traffic classification approach, we will first provide an overview of the used dataset.

4.1. Dataset preparation

We evaluated our approach and compared it to the state of art methods by using a publicly available dataset that has been used in several previous studies. This dataset, called UCSD, is composed of two parts. The first part includes 254 video sequences of highway traffic captured during daytime, in Seattle, over a two-day period. The videos were recorded by a stationary traffic camera in different weather conditions, such as overcast, rainy, and sunny, with some videos having drops on the lens. Each video in the dataset contains between 42 to 52 frames recorded at 10 fps, with a 320 x 240 pixels resolution. In our work, these videos were sampled to get a total of 13435 frames. These frames were labeled manually, based on the perceived level of traffic congestion, and grouped into three categories: 8731 frames of light traffic, 2378 frames of medium traffic, and 2326 frames of heavy traffic. The second part of the dataset contains 19

highway traffic videos that were captured during nighttime, using the same camera, in different weather conditions such as clear, rainy, and overcast. Each video has a duration of about 5 seconds and is composed of between 45 to 50 frames. In total, we have 1112 frames in this part of the dataset. Some frames from the daytime and nighttime parts of the UCSD dataset are presented in Figure 3.

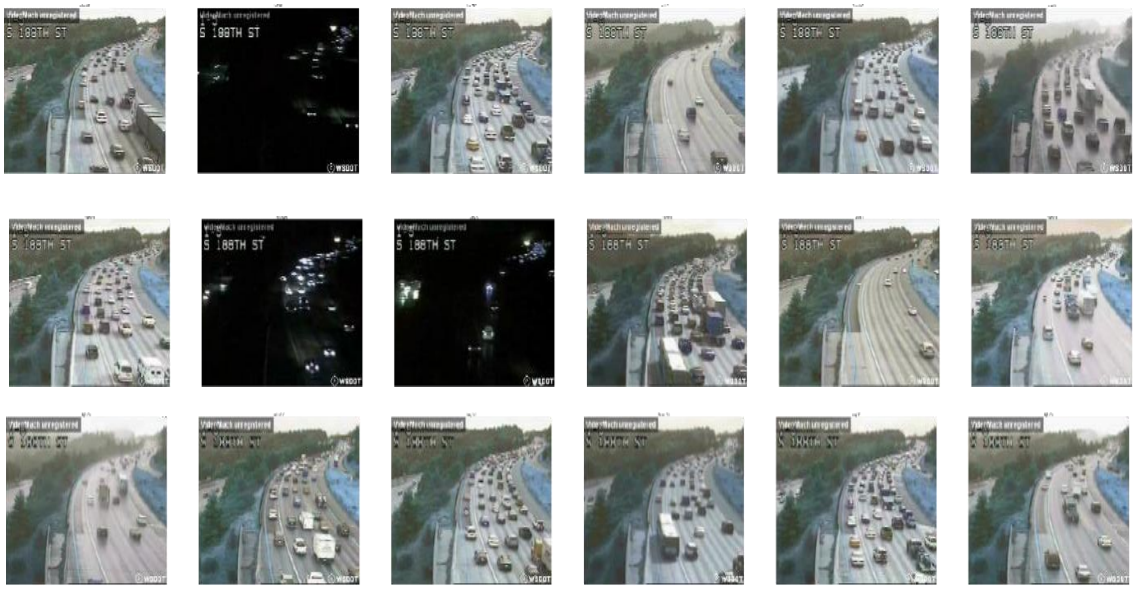


Figure 3. Examples of frames from the daytime and nighttime parts of the UCSD dataset

The traffic in the UCSD dataset was classified into three categories: light, medium, and heavy. The dataset was split into 75% of videos for training and 25% for testing. In the first experiment, we trained our model using only daytime frames, with 100 epochs and a batch size of 256. In the second experiment, we included the nighttime frames and retrained the model with the same parameters as in the first experiment.

4.2. Used evaluation metrics

Evaluating a deep learning model involves multiple steps, and the usage of metrics to assess the model's performance on a specific task. In order to evaluate our transformer model, we used three metrics: recall, precision, F1 score, and accuracy. Before presenting these metrics, we define the quantities True Positive (TP), False Positive, and False Negative (FN) that intervene in their calculation.

TP: Number of observations that are actually positive and that are predicted to be positive.

FN: Number of observations that are actually positive and that are predicted to be negative.

TN: Number of observations that are actually negative and that are predicted to be negative.

FP: Number of observations that are actually negative and that are predicted to be positive.

These definitions are applicable to binary classification scenarios. For multi-class classification TP, FN, TN, and FP are defined for each class. The class of interest is considered as the positive class, and the remaining classes form together the negative class.

4.2.1. Recall

Recall (Sensitivity or True Positive Rate) measures the proportion of positive cases that are correctly classified by the model. Mathematically, it is defined as:

$$\text{Recall} = \frac{\sum TP}{\sum TP + FN}. \tag{10}$$

4.2.2. Precision

Precision measures the proportion of positive predictions that are actually correct. It is defined by the ratio of true positive predictions (TP) to the total number of positive predictions made by the model:

$$\text{Precision} = \frac{\sum TP}{\sum TP + FP}. \tag{11}$$

4.2.3. F1 score

F1 Score (Harmonic Mean of Precision and Recall) is a weighted average of precision and recall that gives equal weight to both precision and recall. It is a single metric that balances both precision and recall. Mathematically, it is defined as:

$$\text{F1-Score} = \frac{2*(Precision*Re\ call)}{Precision+Re\ call}. \tag{12}$$

4.2.4. Accuracy

Accuracy represents the percentage of correct classifications (TP and TN) made by the model. It is calculated by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{13}$$

For a multi-class classification, each class is in turn considered as the positive class and its metrics are calculated by considering all the other classes as the negative class. The metrics of different classes are then averaged with or without weighting, to obtain the global metrics. When no weighting is applied, the average is called macro- average.

4.3. Classification results

For the first experiment, the obtained rates of the metrics defined above, using macro-average, are given in Table 1.

Table 1. Metrics values for the first experiment

Metric	Recall	Precision	F1 Score	Accuracy
Value	99.59%	99.63%	99.61%	99.76%

It can be observed from the results presented in Table 1 that our model performs well and has high accuracy in classifying the positive cases. Based on these metrics, the model appears to have a low number of false negatives and false positives, as well as a good balance between precision and recall.

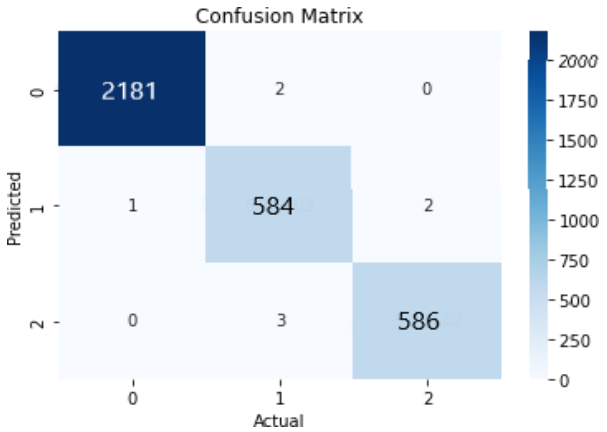


Figure 4. Confusion matrix of the results of the first experiment (0-light, 1-medium, 2-heavy)

The metrics given in Table 1 were deduced from the confusion matrix, presented in Figure 4. The diagonal elements of this matrix represent the number of correct predictions, with 2181, 584, and 586 correct predictions, for the light, the medium, and the heavy classes, respectively. The off-diagonal elements represent the number of incorrect predictions, with only one frame from the light class predicted as medium,

five frames from the medium class misclassified (2 as light and 3 as heavy), and only two frames from the heavy class predicted as medium. We notice that the misclassifications occur only between neighboring classes.

In the second experiment, we added nighttime frames. The following table lists the obtained values of different metrics used to evaluate the model in this experiment.

Table 2. Metrics values for the second experiment

Metric	**Recall**	**Precision**	**F1 Score**	**Accuracy**
Value	98.94%	99.38%	99.15%	99.37%

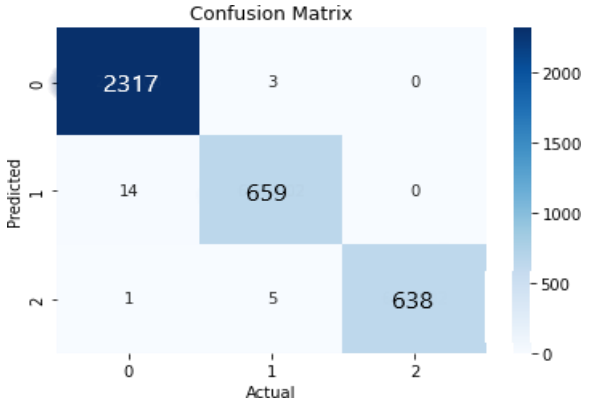


Figure 5. Confusion matrix of the results of the second experiment (0-light, 1-medium, 2-heavy)

The classification accuracy in the second experiment decreased by only 0.3% in comparison with the first experiment; this means that the adopted model is robust towards illumination changes as well as different weather conditions.

From the confusion matrix depicted in Figure 5, The confusion matrix shows that 6 heavy frames are misclassified (5 as medium and 1 as light). The prediction remains good enough for the light and medium classes, with only 23 frames misclassified over 3637 frames from both classes. A comparison between the results obtained with the proposed method and those obtained with the state of art methods, which used the daytime part of the UCSD dataset, is presented in Table 3. It should be noticed that none of the cited works has taken up the challenge of classifying road traffic during nighttime.

Table 3. Comparison between the classification accuracies of our method and other methods

	Work	Accuracy (%)
Previous works	Chan and Vasconcelos, 2005	94.50
	Sobral *et al*., 2013	94.50
	Derpanis and Wildes, 2011	95.30
	Asmaa *et al*., 2013	96.37
	Gonçalves, 2015	97.24
	Luo *et al.*, 2016	97.64
	Adnan, M. F. *et al.*, 2022	98.20
	Ribeiro, M. V. L *et al*., 2020	98.82
Our Proposed Approach	First experiment	99.76
	Second experiment	99.37

By reaching an accuracy of 99.76% in the first experiment with daytime frames, our method outperforms the best state of the art method (Ribeiro *et al.,* 2020) by 0.94%. When we added the nighttime frames, we obtained an accuracy of 99.37%, which remains higher than the daytime accuracies of these methods, despite the presence of challenging traffic visibility in these frames. This means that the transformer model is more robust in classifying road traffic congestion. Our proposed method is based on the transformer model, which gives it some solid advantages. As the transformer model is based solely on

the self-attention mechanism, it provides the ability to capture dependencies between different parts of an image, which improves the performance. Another advantage of the transformer model's architecture is the multi-head self-attention mechanism that enables the model to learn multiple aspects of the data and generalize better to newer unseen patterns. This improved the model's robustness to noise in images and makes it suitable for the traffic classification task. Furthermore, this task can be accomplished in real-time, thanks to the parallel structure of the transformer, which allows for fast computation. The vision transformer architecture offers several advantages over traditional convolutional neural networks (CNNs). First, the transformer architecture is designed to handle sequential data, which makes it well suitable for tasks that involve image sequence processing, such as video classification or optical flow estimation. Second, vision transformers allow greater flexibility in architecture design, as they can be modified to incorporate additional layers or attention mechanisms that are appropriate for specific tasks or data sets. Additionally, the self-attention mechanism in vision transformers allows easy interpretation of the most important regions of the image, which can be difficult to interpret in traditional CNNs because of their complex and abstract feature maps. Finally, vision processors can learn to capture more abstract and general representations of data compared to CNNs, which can lead to improved generalization and better performance on unseen data. In addition to being better in terms of accuracy than the best of the art microscopic method (Ribeiro *et al.,* 2020), as shown in Table 3, our proposed method is a macroscopic approach, which does not require vehicles detection and tracking, therefore, it is simpler and more suitable for real time applications.

Moreover, our model achieved an accuracy of 99.07% on the nighttime UCSD Dataset, which outperforms the accuracy of the method proposed in (Asmaa *et al*., 2013), as shown in Table 4. This is the only method to which we could compare our method, since none of the methods cited in Table 3, presented results regarding nighttime video classification.

Table 4. Comparison between the proposed method and the motion vector method

	Motion vectors method (Asmaa *et al.,*2013)	Proposed method
Accuracy	89.47%	99.07%

5. Conclusion

In this paper, a novel macroscopic approach for classifying road traffic flow was proposed. The method uses the multi-head attention mechanism transformer model to classify road traffic into three classes: light, medium, and heavy. The performance of this method, in terms of accuracy, was evaluated and compared to that of the state- of-the-art methods using the public UCSD dataset, which includes road traffic videos recorded during daytime under varying weather conditions (rainy, overcast, and clear). With a 99.76 % accuracy, the proposed macroscopic approach outperforms the best method known to us so far, which is a microscopic method that requires vehicles' monitoring, i.e., vehicles detection and tracking as well as traffic parameters extraction. Moreover, the proposed method achieved a remarkable accuracy of 99.37% when we included nighttime frames in the dataset. It should be highlighted that none of the state-of-the-art known methods took the challenge to classify the road traffic in nighttime videos. The outstanding performance of the proposed method is achieved at a lower computational cost, due to its macroscopic nature, which does not require the time-consuming vehicle monitoring step.

The MSViT approach for road traffic classification is suitable for versatile applications. It can be integrated seamlessly into traffic management systems to optimize traffic flow and reduce travel delays. It can assist the concerned authorities in intelligent urban traffic planning, by dynamically adjusting road traffic signalisation and rerouting traffic, based on real-time congestion estimation. Last, but not least it can operate under different weather conditions, which makes it a potential competitor to existing commercial systems, such as INRIX (INRIX, 2022).

References

1. Adnan, M. F., Ahmed, N., Ishraque, I., Al Amin, M. S. and Hasan, M. S. (2022) Traffic congestion prediction using deep convolutional neural networks: A color-coding approach. In: *2022 International Conference on Engineering and Emerging Technologies (ICEET)*, 1-5. IEEE. DOI: 10.1109/ICEET56468.2022.10007425.
2. Asmaa, O., Mokhtar, K. and Abdelaziz, O. (2013) Road traffic density estimation using microscopic and macroscopic parameters. *Image and Vision Computing*, 31(11), 887-894. DOI: 10.1016/j.imavis.2013.09.006.
3. Bull, A., & CEPAL, N. (2003) Traffic congestion: the problem and how to deal with It. ECLAC.

4. Chan, A. B. and Vasconcelos, N. (2005) Classification and retrieval of traffic video using auto-regressive stochastic processes. In: *IEEE Proceedings. Intelligent Vehicles Symposium*, 771-776. IEEE. DOI: 10.1109/IVS.2005.1505198.
5. Chan, A.B. and Vasconcelos, N. (2004) Efficient computation of the KL divergence between dynamic textures. technical report SVCL-TR2004-02, http://www.svcl.ucsd.edu/, November 2004.
6. Chen, H. T., Tsai, L. W., Gu, H. Z., Lee, S. Y. and Lin, B. S. P. (2012) Traffic congestion classification for nighttime surveillance videos. In: *2012 IEEE International Conference on Multimedia and Expo Workshops*, 169-174. IEEE. DOI: 10.1109/ICMEW.2012.36.
7. Derpanis, K. G. and Wildes, R. P. (2011) Classification of traffic video based on a spatiotemporal orientation analysis. In: *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, 606-613. IEEE. DOI: 10.1109/WACV.2011.5711560.
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, Th., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv*: 2010.11929.
9. Gonçalves, W. N., Machado, B. B. and Bruno, O. M. (2015) A complex network approach for dynamic texture recognition. *Neurocomputing*, 153, 211-220. DOI: 10.1016/j.neucom.2014.11.034.
10. INRIX. (2022) *INRIX Launches New Artificial Intelligence Traffic Solution - INRIX*. Inrix. https://inrix.com/press-releases/inrix-launches-new-artificial-intelligence-traffic-solution/
11. Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B. (2022) A Review of Yolo Algorithm Developments. *Procedia Computer Science*, 199, 1066-1073.
12. Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022) Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s), 1-41. DOI: 10.1145/3505244.
13. Luo, Z., Jodoin, P. M., Li, S. Z. and Su, S. Z. (2015) Traffic analysis without motion features. In: *2015 IEEE International Conference on Image Processing (ICIP)*, 3290-3294. IEEE. DOI: 10.1109/ICIP.2015.7351412.
14. Luo, Z., Jodoin, P. M., Su, S. Z., Li, S. Z. and Larochelle, H. (2016) Traffic analytics with low-frame-rate videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(4), 878-891. DOI: 10.1109/TCSVT.2016.2632439.
15. Ramana, K., Srivastava, G., Kumar, M. R., Gadekallu, T. R., Lin, J. C. W., Alazab, M. and Iwendi, C. (2023) A vision transformer approach for traffic congestion prediction in urban areas. *IEEE Transactions on Intelligent Transportation Systems*. DOI: 10.1109/TITS.2022.3233801.
16. Riaz, A. and Khan, S. A. (2013) Traffic congestion classification using motion vector statistical features. In: *Sixth International Conference on Machine Vision (ICMV 2013)*, 9067, 245-251. SPIE. DOI: 10.1117/12.2051463.
17. Ribeiro, M. V. L., Samatelo, J. L. A., and Bazzan, A. L. C. (2020) A new microscopic approach to traffic flow classification using a convolutional neural network object detector and a multi- tracker algorithm. *IEEE Transactions on Intelligent Transportation Systems*. DOI: 10.1109/TITS.2020.3040594.
18. Sobral, A. L. O., Schnitman, L. and De Souza, F. (2013) Highway traffic congestion classification using holistic properties. In: *10th IASTED International Conference on Signal Processing, Pattern Recognition and Applications.*
19. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017) Attention is all you need. *Advances in neural information processing systems*, 30, (NIPS 2017).
20. Wang, X., Zeng, R., Zou, F., Liao, L. and Huang, F. (2023) STTF: An efficient transformer model for traffic congestion prediction. *International Journal of Computational Intelligence Systems*, 16(1), 2.