IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCES IN MACHINE LEARNING AND COGNITIVE COMPUTING FOR INDUSTRY APPLICATIONS

Received September 26, 2019, accepted October 22, 2019, date of publication October 29, 2019, date of current version November 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2950162

A Lightweight Moving Vehicle Classification System Through Attention-Based Method and Deep Learning

NASARUDDIN NASARUDDIN[1], (Member, IEEE), KAHLIL MUCHTAR[1,2], (Member, IEEE), AND AFDHAL AFDHAL[1], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, Syiah Kuala University, Aceh CO 23111, Indonesia
[2]Nodeflux, Jakarta CO 12730, Indonesia

Corresponding author: Kahlil Muchtar (kahlil@unsyiah.ac.id)

This work was supported by the Ministry of Research, Technology and Higher Education, Indonesia, under Grant 93/UN11.2/PP/SP3/2019.

ABSTRACT The convolutional neural network (CNN) has shown excellent benefits in the classification of objects in the latest years. An important job in the context of intelligent transportation is to properly identify and classify vehicles from videos into various kinds (e.g., car, truck, bus, etc.). For monitoring, tracking and counting purposes, the classified vehicles can be further evaluated. At least two major difficulties stay, however; excluding the uninteresting area (e.g., swinging movement, noise, etc.) and designing an effective and precise system. In order to obviously differentiate the interesting region (moving car) from the un-interesting region (the rest of the area), we introduce a novel attention-based approach. Finally, to significantly increase the classification efficiency, we feed the deep CNN with the respective interesting region. We use several challenging outdoor sequences from the CDNET 2014 (baseline, bad weather and camera jitter classes), and our own dataset to assess the proposed approach. Experimental results show that it costs around ∼85 fps in GPU (and ∼50 fps in CPU) to classify moving vehicles and maintaining a highly accurate rate. Compared with other state-of-the-art object detection approaches, our method obtains a competitive detection accuracy. In addition, we also verify the result of the proposed approach by comparing with recent 3D CNN method, called saliency tubes.

INDEX TERMS Attention approach, convolutional neural network (CNN), smart transportation, vehicle type classification.

I. INTRODUCTION

In the previous occasions, under the highways are mounted the detectors, inductive ground loops or laser scanners to classify the vehicle type [1]. Due to a latest advance in an integrated surveillance system, the image dataset of vehicles on the highway is commonly accessible at low price. This system provides well-integrated CCTV and built-in communication. Therefore, it is highly practical to provide an automatic vehicle type classification system using a computer vision method. Earlier researches related to the image classification tend to use a well-known model and image features, such as Bayesian [2], support vector machine [3]–[5], LBP (local binary pattern), SIFT [6], and etc. Recently, the convolutional neural network (CNN) has been used extensively to address issues with image classification, including face recognition, activity recognition and sort on [7], [8]. Especially for vehicle classification, the traditional approach combined Histograms of Oriented Gradient with Support Vector Machine (HOG + SVM) in order to localize and recognize the vehicles [9]. However, with high consistency, CNN can attain outstanding validation precision compared to traditional image classification models. To date, several CNN-based methods have been proposed in order to classify vehicle type, such as [1], [9]–[11].

Since classification of the vehicle type is an important component of the intelligent transportation system, the method with high precision, less interfering and effective, are inevitable. However, complicated lighting, bad illumination, bad weather, swaying movement are still challenging

The associate editor coordinating the review of this manuscript and approving it for publication was Min Xia.

issues in real-world settings. Therefore, in this paper, we introduce an attention-based approach in order to focus only on the moving vehicle region in an input frame. In other words, we pay less attention to the uninteresting region such as static region, swinging movement, intermittent noise, and so on. Finally, extracted moving vehicle region on each frame is fed to the deep CNN in determining the vehicle type. Compared with the common CNN classification that processes the whole frame, our attention-based approach can achieve considerable performance improvements. It is worth noting that the most relevant research is proposed by Zhao [12], which highlight key areas of an image via deep reinforcement learning. Instead of utilizing the deep approach, our work proposes a completely distinct yet light approach by introducing a robust attention-based moving object detection in video sequences. This is an extended version of a preliminary conference report that was published in [13].

The **contributions of this paper** are the following: a) We demonstrate that an effective detection of static surveillance cameras can significantly enhance the efficiency of the CNN classification, b) Rather than the full frame dimension, the fine-grained classification of the vehicle only inferred interesting area (the region of moving vehicles), c) We have gathered a particular vehicle sample dataset that is appropriate for Indonesia areas. It is made up of four groups: car, bus, truck, and motorcycle. As a result, as one of the world's emerging market economies, our model is very robust to the integrated surveillance system in Indonesia.

II. PROPOSED METHODOLOGY

We describe a novel attention-based detection to cope with swaying movement, camera jitter and bad weather that usually happens in outdoor scenes. Our proposed idea uses bilateral texturing to construct the robust model and produces an attention region (moving vehicle areas) as complete as possible.

We then feed the region to the classification module as a grid input. The module will finally output the class map of probability and the respective final detections. Our classification problem comprises of 4 classes with 49,652 annotated training data (car, truck, bus, and motorcycle).

Fig. 1 illustrates the overview of our system workflow, the details of which are discussed in detail in the following parts. More specifically, we divide the section into two primary parts; attention-based detection and lightweight fine-grains classification.

A. ATTENTION-BASED DETECTION

Many moving object detection methods did not fully exclude swinging movement and mechanical vibration, such as swinging trees, ripple water, camera jitter and so on. Our proposed idea relies on robust detection in order to completely extracting the moving vehicle regions. This enables the following module to concentrate on the interesting region while increasing the frame rate at the same moment.

Recently, some state-of-the-art techniques produce texture data from a frame instead of directly modeling incoming pixel values and construct the corresponding BG model. As mentioned in detail in [14], [15], this texture generation and modeling are able to resist to abrupt illumination changes and shadow interference. Furthermore, a texture-based approach is very effective compared to earlier pixel-based methods. However, as illustrated in Fig. 2 (CDNET 2014 dataset [16]) below, occasionally, the swinging movement is misclassified

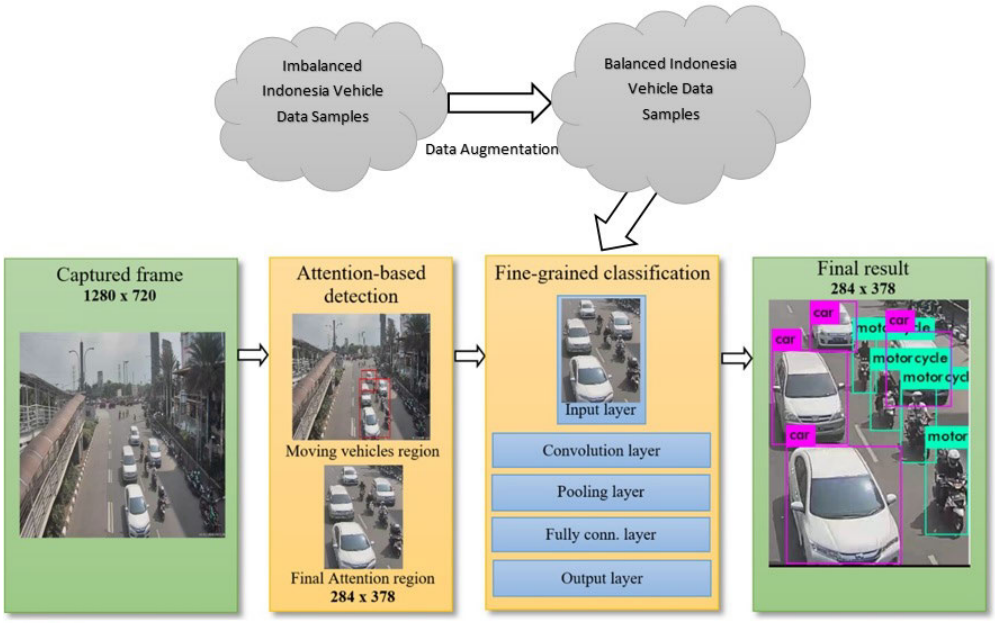


FIGURE 1. System workflow of our approach.

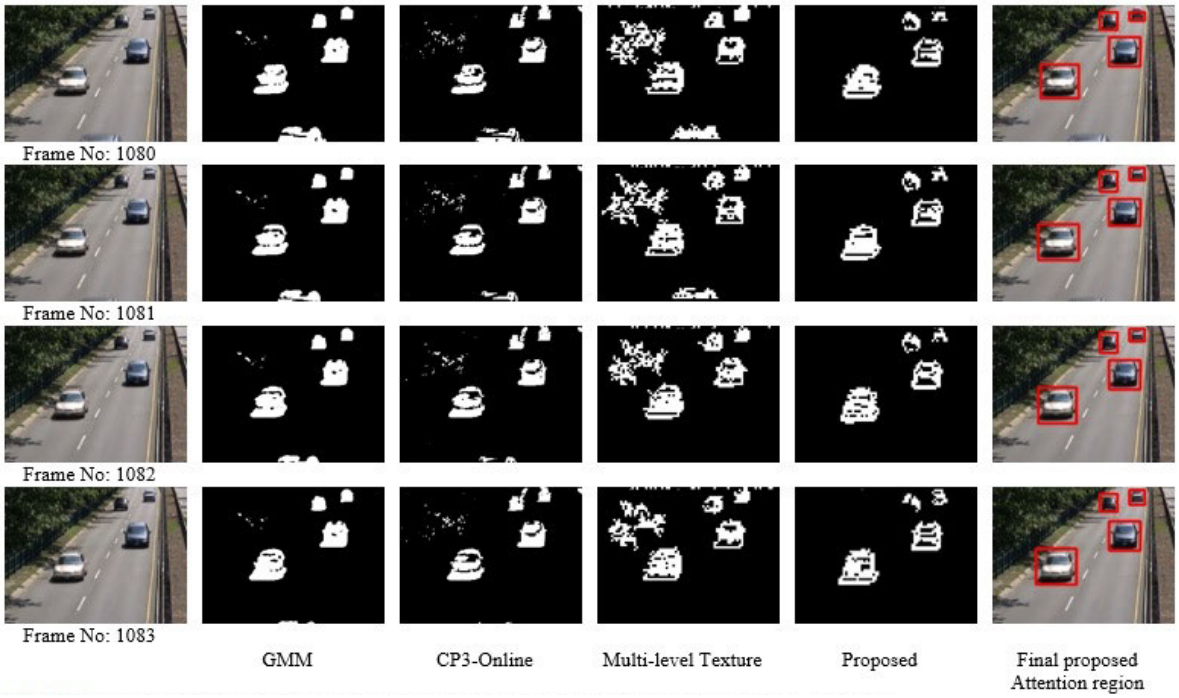


FIGURE 2. Illustration of successive frames that contain swaying trees (Dataset: CDNET 2014-highway).

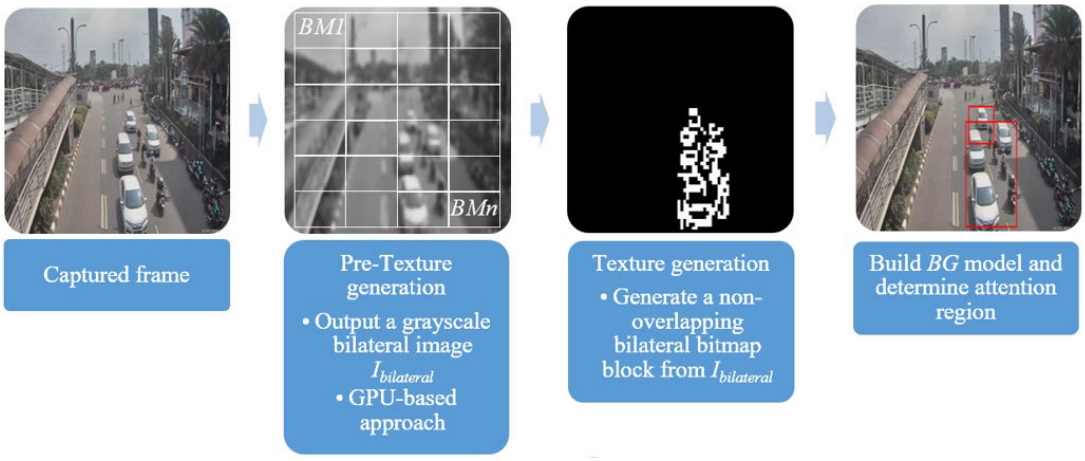


FIGURE 3. Step-by-step of finding attention region.

as a foreground region. Although the misclassification is adaptively fixed after several frames (when the movement slows down or lastly static), it is hard to construct a coherent attention-based system for efficient vehicle classification purposes. Therefore, a bilateral texturing method is introduced that effectively decreases noise and preserves the edges of observed areas. In our case, such a technique is capable of constructing a stable BG model to obviously see the region of the moving vehicle as an area of visual attention and the remainder as an uninterested area. As will be discussed later, our proposed bilateral texturing model can achieve ∼85 fps and ∼50 fps for 720p (1280×720) input format in GPU and CPU environment, respectively.

In Fig. 3, the step-by-step generation of texture information is described. First, we use bilateral filtering [17] to an input frame I, and denoted the greyscale output image as $I_{bilateral}$. The $I_{bilateral}$ is used to generate a non-overlapping block-based texture. More specifically, the $I_{bilateral}$ is divided into blocks of sizes n x n pixels. In our setting, the n is set to 4. Then we calculate each block's mean and use it to create a binary bitmap. The bitmap BM_{bil} is obtained by comparing the mean with each pixel value in a block. If the value of the pixel is below the mean, the binary value is 0, and vice versa. Finally, the BM_{bil} of each block is used to build the initial BG model BM_{mod}, and becoming a reference when the new incoming frame exists. Our current BG model update rule

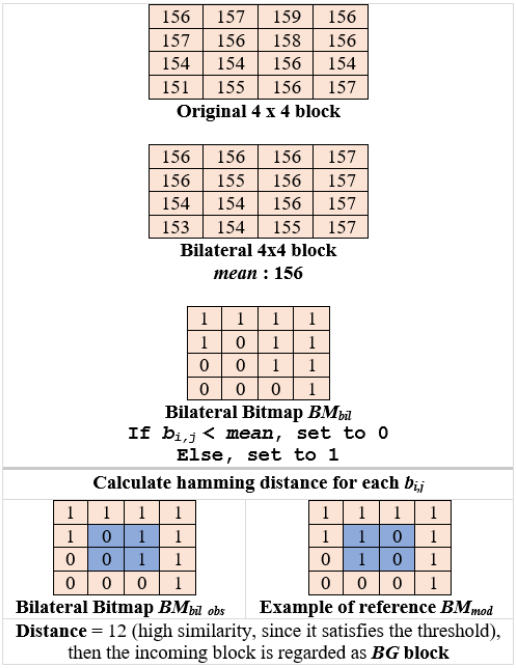


FIGURE 4. An example of calculating distance between incoming block BM_{bil_obs} and reference block model BM_{mod}.

and its appropriate learning rate are similar to our previous method [18].

In principle, when a new frame arrives, we simply perform a hamming distance for each block to determine whether the observed block BM_{bil_obs} is regarded as BG block or moving region block. Note that, the b_{ij} indicates the corresponding bit value in i, j position of a block. In Fig. 4, we provide an example of calculating the distance between BM_{bil_obs} and BM_{mod}.

$$Dist\left(BM_{bil\_obs}, BM_{mod}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\left(b_{ij}^{bil\_obs} \oplus b_{ij}^{mod}\right) \quad (1)$$

While keeping the edges of the moving vehicle fairly sharp, compared to most filters, the bilateral filter is very slow. We therefore use the texture memory of a CUDA instead of using global memory to process an input frame and perform a bilateral GPU-based filter [17]. It is noteworthy that on the experimental section we use the CPU-based filter to fairly compare with other state-of-the-art methods. To weaken the un-interesting region and noise, we need to carefully set the three important parameters in bilateral filtering. This will also ensure that we maintain the edges of moving vehicles and provide our BG model with good texture characteristics. In experiments, we set the kernel size, sigma color and sigma spatial to 8, 150 and 150, respectively.

B. EFFICIENT FINE-GRAINED CLASSIFICATION

Our fine grain classification is based on YOLOv3 for effectiveness reasons [19]. YOLOv3 is a deep convolutional neural network architecture which predicts bounding boxes by using anchor boxes which is originally introduced in [20]. YOLOv3 calculates these anchor boxes from the ground truth bounding boxes. It applies a specialized k-means clustering on training sets bounding boxes which specialized on IOU (Intersection over Union) metric. Finally, it will select the best k centroids from the clustering method resulting in the largest IOU and nominate them as anchor boxes. The obtained attention region will be processed through this pipeline, and output the number of detection vehicles in a frame.

Although YOLOv3 has provided 80 common multi-class models, we are building our particular model and are more appropriate for the problem of vehicle classification. As briefly discussed in the previous section, we prepare our own dataset by providing 49,652 annotated training frames. Note that, an annotated frame may include more than one class of vehicle and multiple numbers of the vehicle. The detail of class distribution can be found in Table 1 below:

Since our class distribution is imbalanced, the model will overfit to the classes which have more data. We first perform an image augmentation in order to avoid deep learning model

FIGURE 5. The result of proposed workflow.

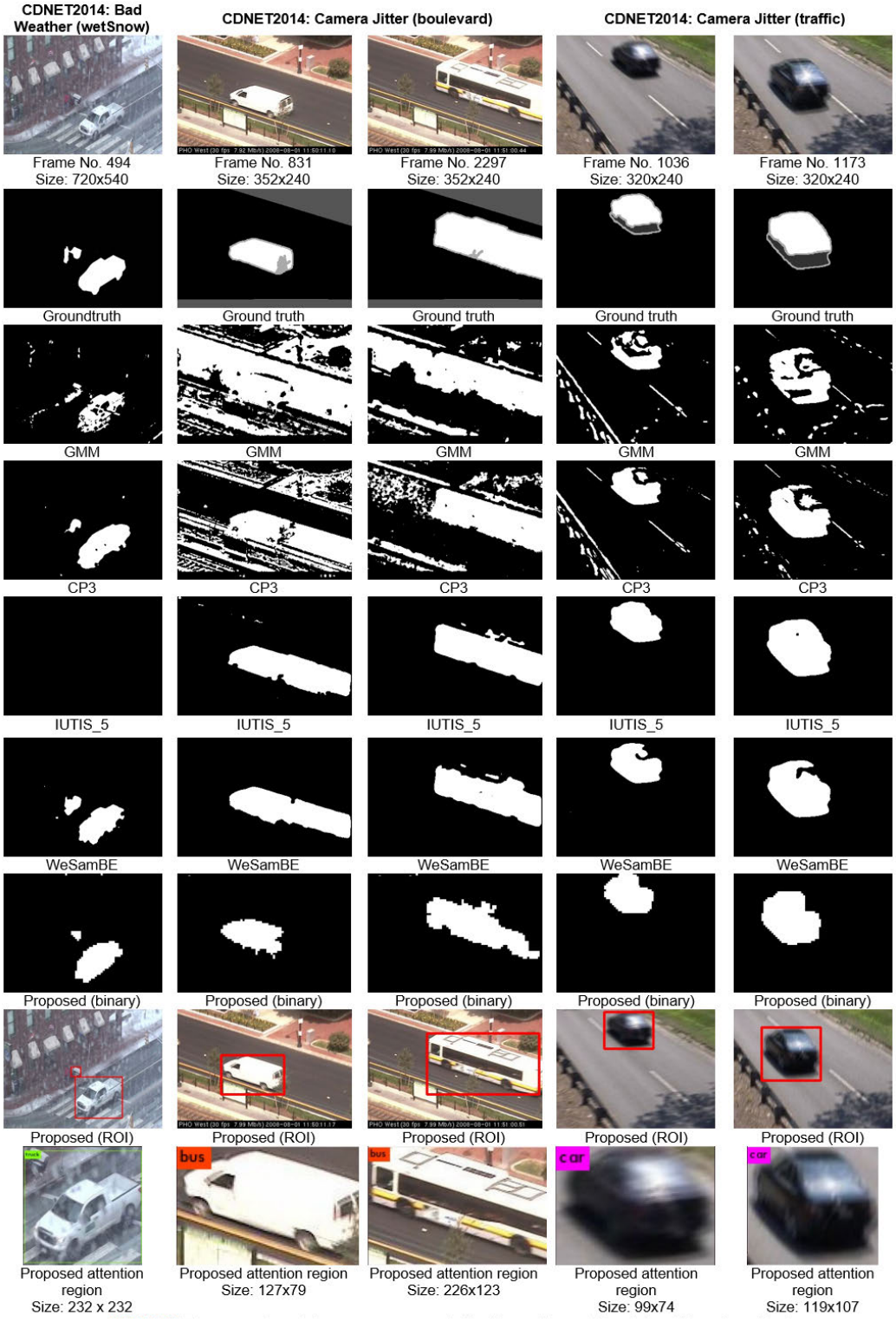


FIGURE 6. The comparison between our proposed attention region and the state-of-the-arts method.

bias. In this paper, the open-sourced tool called **imgaug**[1] is chosen due to its efficiency.

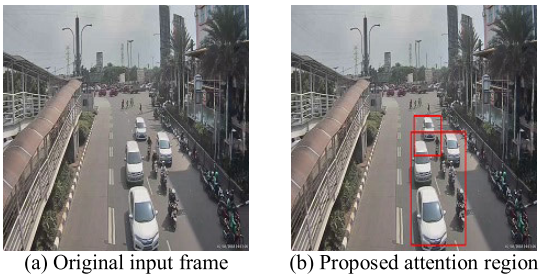
[1]https://github.com/aleju/imgaug

III. RESULTS AND DISCUSSIONS

In this section, to demonstrate the robustness of our proposed idea, we assess the qualitative and quantitative measurements. We provide two distinct scenes in qualitative measurements that contain swaying trees in the highway

TABLE 1. The class distribution of dataset.

Class distribution	No. of vehicle type
Car	124,742
Bus	16,585
Truck	19,047
Motorcycle	21,969



(a) Original input frame (b) Proposed attention region

(c) Visual attention of saliency tube

FIGURE 7. The comparison between our unsupervised attention region and 3D CNN Saliency method.

		Ground truth		Total
		Foreground	Background	
Detection	Foreground	TP	FP	P'
	Background	FN	TN	N'
Total		P	N	

FIGURE 8. A confusion matrix for binary classification.

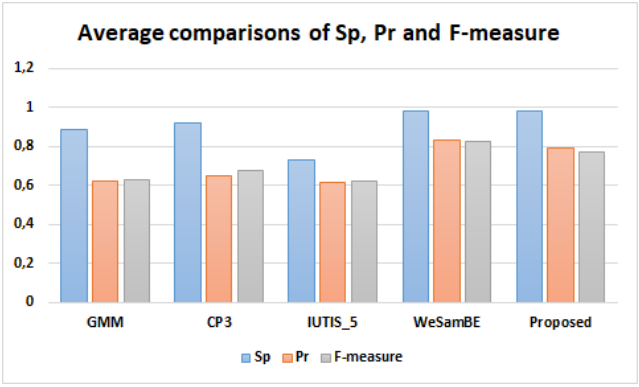


FIGURE 9. A group bar chart of average comparisons of Sp, Pr and F-measures (higher scores are better).

outdoor environment. In addition, several challenging outdoor traffic scenes on CDNET2014 are thoroughly evaluated, including baseline, camera jitter, and bad weather classes.

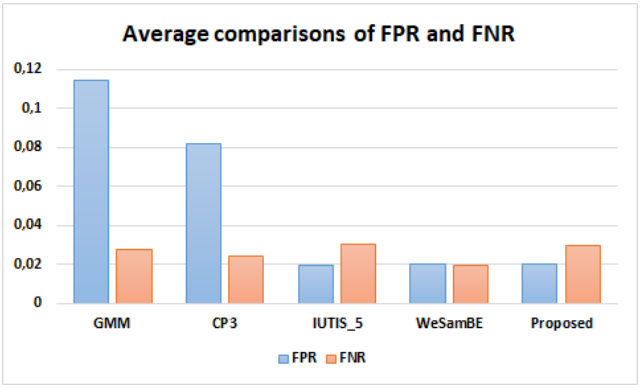


FIGURE 10. A group bar chart of average comparisons of FPR and FNR (lower scores are better).

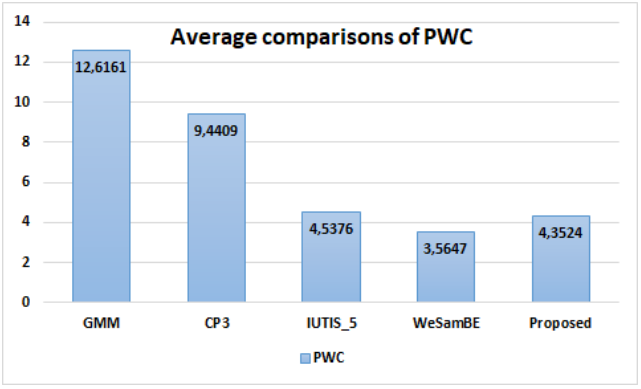


FIGURE 11. A group bar chart of average comparisons of PWC (lower scores are better).

We compare our attention-based detection with other well-known methods; GMM [21], CP3-Online [22], IUTIS_5 [23], and WeSamBE [24]. To be specific, we show the specificity (Sp), false-positive rate (FPR), false-negative rate (FNR), percentage of wrong classification (PWC), precision (Pr), and f-measure scores for all compared methods. Finally, the confidence score and frame rate evaluation are given in the last subsection. The tested environment is equipped with Intel i7-7700HQ processor, 16 GB of memory, and NVIDIA GeForce GTX 1050 Ti 4 GB.

A. QUALITATIVE MEASUREMENT FOR ATTENTION-BASED DETECTION

In Fig. 5, we provide the result of the proposed workflow in two challenging videos. The data sets are acquired from the Jakarta, Indonesia's capital city. All videos have swaying trees with arbitrary motion. Our proposed approach excludes this region and effectively feeds the obtained attention region to the fine-grained classification module. Besides swaying motion, the camera jitter and bad weather often happen in outdoor scenes and presents a major challenge in moving object detection. As depicted in Fig. 6, the camera jitter leads to a very critical false detection, especially on GMM and CP3 methods. We find that our bilateral texturing-based models can well handle the camera jitter. The obtained bilateral

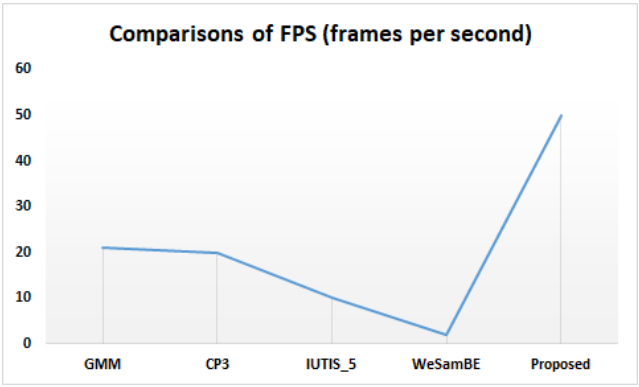
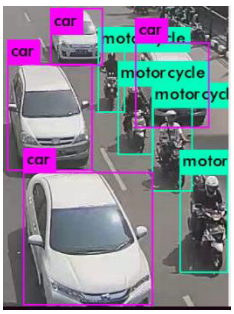


FIGURE 12. A comparisons of FPS in CPU (frames per second).

Attention Region (Frame No. 2100)



motorcycle: 99%, motorcycle: 98%, motorcycle: 87%, motorcycle: 65% car: 100%, car: 100%, car: 100%, car: 99%

FIGURE 13. An illustration of obtained confidence score of each input region.

bitmap is able to preserve the edge of moving object while weakening other static/unimportant regions at the same time. Besides, we also verify the result of our lightweight approach with recent 3D CNN method. To be specific, we perform a saliency tubes [25] which proposes an approach to find the main focus points of the network (visual attention region from each video frame). As depicted in Fig. 7, compared with this state-of-the-art method, in general, our proposed idea can extract a similar attention region.

B. QUANTITATIVE MEASUREMENT FOR ATTENTION-BASED DETECTION

In this section, the performance of all methods is measured. Some common indicators are explained further in Eqs. (2) to (7), which are measured based on a binary pixel-wise approach. Note that TP, FP, FN, and TN are true positive, false positive, false negative and true negative, respectively (see Fig. 8).

The Specificity measures the proportion of actual negatives that are correctly identified, it is also called true negative rate. The formula is defined as follows:

$$\text{Specificity(Sp)} = \frac{TN}{TN + FP} \tag{2}$$

TABLE 2. Vehicle detection: Comparison of quantitative measurements (CDNET2014 Dataset "Bad Weather-wetSnow", Frame No. 494).

	GMM [21]	CP3 [22]	IUTIS_5 [23]	WeSamBE [24]	Proposed
Sp	0,9826	0,9829	N/A	**0,9914**	**0,9864**
FPR	0,0173	0,0170	N/A	**0,0085**	**0,0135**
FNR	0,0221	**0,0063**	0,0684	**0,0065**	0,0123
PWC	3,7005	**2,1901**	6,4075	**1,4127**	2,4231
Pr	0,7273	0,7844	N/A	**0,8790**	**0,8055**
F-measure	0,7006	**0,8415**	N/A	**0,8912**	0,8125

TABLE 3. Vehicle detection: Comparison of quantitative measurements (CDNET2014 Dataset "Camera Jitter-Boulevard", Frame No. 831 & 2297).

	GMM	CP3	IUTIS_5	WeSamBE	Proposed
Sp	0,6217	0,7532	0,9567	**0,9616**	**0,9844**
FPR	0,3782	0,2466	0,0432	**0,0382**	**0,0154**
FNR	**0,0239**	**0,0238**	0,0298	0,0338	0,0579
PWC	35,3736	23,697	6,3964	**6,2226**	**6,0670**
Pr	0,3135	0,3868	0,7364	**0,7563**	**0,8570**
F-measure	0,425	0,4997	0,7568	**0,7662**	**0,7584**

TABLE 4. Vehicle detection: Comparison of quantitative measurements (CDNET2014 Dataset "Camera Jitter-Traffic", Frame No. 1036 & 1173).

	GMM	CP3	IUTIS_5	WeSamBE	Proposed
Sp	0,9434	0,9476	0,9684	**0,9692**	**0,9687**
FPR	0,0564	0,0522	0,0314	**0,0307**	**0,0311**
FNR	0,0347	0,0297	**0,0055**	**0,0146**	0,0204
PWC	8,3059	7,4498	**3,3495**	**4,1028**	4,6672
Pr	0,5159	0,5643	**0,7545**	**0,7408**	0,7171
F-measure	0,5664	0,6236	**0,8378**	**0,7924**	0,7548

TABLE 5. Vehicle detection: Comparison of quantitative measurements (CDNET2014 Dataset "Baseline-Highway", Frame No. 1083).

	GMM	CP3	IUTIS_5	WeSamBE	Proposed
Sp	0,9949	0,9896	**0,9959**	**0,9956**	0,9801
FPR	0,0050	0,0103	**0,0040**	**0,0043**	0,0198
FNR	0,0289	0,0384	**0,0179**	**0,0234**	0,0270
PWC	3,0846	4,4270	**1,9973**	**2,5208**	4,2526
Pr	0,9354	0,8613	**0,9544**	**0,9482**	0,7921
F-measure	0,8124	0,7242	**0,8848**	**0,8505**	0,7630

FP Rate (FPR) is the ratio of FP to FP and TN.

$$\text{FPR} = \frac{FP}{FP + TN} \tag{3}$$

FN Rate (FNR) is the ratio of FN to FN and TP. Both FPR and FNR, a low value are desired.

$$\text{FNR} = \frac{FN}{FN + TP} \tag{4}$$

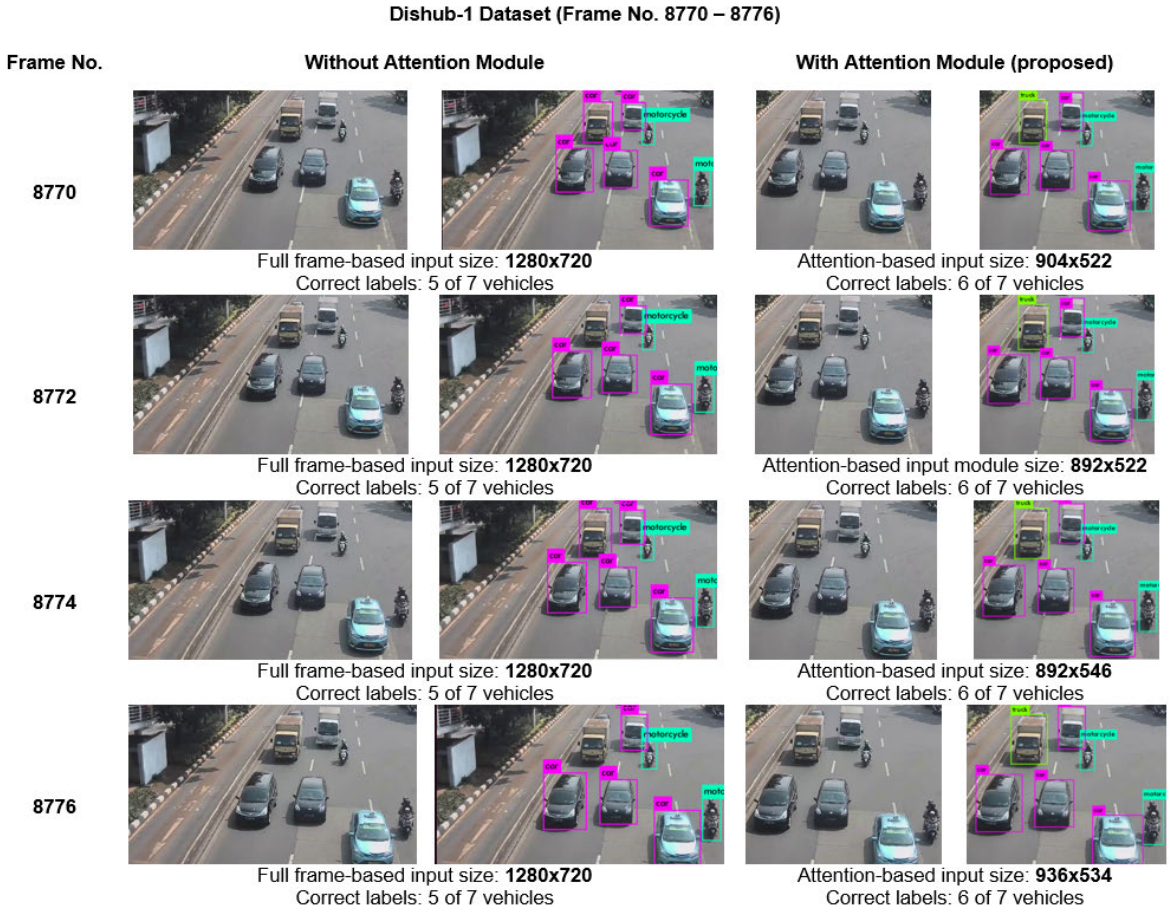


FIGURE 14. Illustration of classification accuracy with and without attention module in successive frames.

Percentage of wrong classification (PWC) is 100 times the ratio of FP and FN to all detected pixels. Therefore, lower PWC value, better detection will be.

$$\text{PWC} = \frac{100 * (FP + FN)}{TP + TN + FP + FN} \tag{5}$$

Precision is a measure of the accuracy of a foreground category being predicted:

$$\text{Precision(Pr)} = \frac{TP}{TP + FP} \tag{6}$$

The f-measure is the harmonic mean of the precision rate and the recall rate. The formula is defined as follows:

$$\text{F} - \text{meas.} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} \tag{7}$$

As shown in Table 2 through Table 5, the specific assessment demonstrates that the proposed technique can be very competitive, particularly with swaying movement and noise in a difficult frame. Apart from the ''baseline'' class, we also evaluate the ''bad weather'' and ''camera jitter'' classes that are very common problems for the outdoor integrated surveillance system. Finally, in Table 6, we provide the average scores of all compared methods. Note that, the scores that highlighted in red and blue color are the best and the second-best, respectively. The corresponding group bar charts are also provided to clearly visualize the competitiveness of the proposed method in terms of accuracy and frame rates.

TABLE 6. Vehicle detection: Comparison of average quantitative measurements and frame rates.

	GMM	CP3	IUTIS_5	WeSamBE	Proposed
Avg. Sp	0,8856	0,9183	0,7302	**0,9794**	**0,9799**
Avg. FPR	0,1142	0,0815	**0,0196**	0,0204	**0,0199**
Avg. FNR	0,0274	**0,0245**	0,0304	**0,0195**	0,0294
Avg. PWC	12,6161	9,4409	4,5376	**3,5647**	**4,3524**
Avg. Pr	0,6230	0,6492	0,6113	**0,8310**	**0,7929**
Avg. F-meas	0,6261	0,6722	0,6198	**0,8250**	**0,7721**
FPS in CPU	**~21 fps**	~20 fps	~10 fps	~2 fps	**~50 fps**

C. QUANTITATIVE MEASUREMENT FOR ATTENTION-BASED CLASSIFICATION

Our attention-based detection can achieve ~85 fps for input of 720p (1280×720) format in GPU (and ~50 fps in CPU).

Next, the obtained attention region of each frame is sent to the classification module and processed asynchronously. For the classification module, it takes about ∼0.074649 seconds in order to predict a region. We also provide the classification accuracy in successive frames with and without attention-based detection. We find that the obtained attention-based region can increase the accuracy of classification. For example, as shown in Fig. 14, the "truck" is incorrectly labeled in full-frame classification but can be successfully labeled using obtained attention-based input. Moreover, our approach can also obtain better frame rate not only using our trained model but also on pre-trained YOLO model (as depicted in Table 7).

TABLE 7. Vehicle classification: The comparison of accuracies and frame rates (Dishub-1 Dataset, Frame No. 8770-8776).

	Without Attention Module	With Attention Module
Average Correct Labels (accuracy)	71,4%	**85,7%**
FPS using Own Model	9 frames/sec	**11** frames/sec
FPS using pre-trained YOLO Model	24 frames/sec	**50** frames/sec;

IV. CONCLUSION

This article presents a technique of bilateral texturing to find each frame's region of attention. In particular, it seeks to robustly exclude the unimportant region (e.g., swaying region, noise, and etc.) and extract the region of visual attention. The region will then be supplied to the fine-grained classification part. Finally, the noise-free region with the classified vehicle is obtained. Our proposed method is simple, efficient, and accurate. As shown in the experimental section, the qualitative and quantitative measurements yield a very competitive result and can be applied in a real-time environment.

REFERENCES

[1] H. Huttunen, F. S. Yancheshmeh, and K. Chen, "Car type recognition with deep neural networks," in *Proc. IEEE Intell. Vehicles Symp.*, Gothenburg, Sweden, Jun. 2016, pp. 1115–1120.
[2] D. Preotiuc-Pietro and F. Hristea, "Unsupervised word sense disambiguation with N-Gram features," *Artif. Intell. Rev.*, vol. 41, no. 2, pp. 241–260, Feb. 2014.
[3] O. Amayri and N. Bouguila, "A study of spam filtering using support vector machines," *Artif. Intell. Rev.*, vol. 34, no. 1, pp. 73–108, Jun. 2010.
[4] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "Incremental support vector learning for ordinal regression," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 7, pp. 1403–1416, Jul. 2015.
[5] B. Gu, V. S. Sheng, Z. Wang, D. Ho, S. Osman, and S. Li, "Incremental learning for v-support vector regression," *Neural Netw.*, vol. 67, pp. 140–150, Jul. 2015.
[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[7] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6479–6488.
[8] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 1891–1898.
[9] X. Wang, W. Zhang, X. Wu, L. Xiao, Y. Qian, and Z. Fang, "Real-time vehicle type classification with deep convolutional neural networks," *J. Real-Time Image Process.*, vol. 16, no. 1, pp. 5–14, Feb. 2019.
[10] W. Liu, M. Zhang, Z. Luo, and Y. Cai, "An ensemble deep learning method for vehicle type classification on visual traffic surveillance sensors," *IEEE Access*, vol. 5, pp. 24417–24425, 2017.
[11] J. Sochor, A. Herout, and J. Havel, "BoxCars: 3D boxes as CNN Input for improved fine-grained vehicle recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3006–3015.
[12] D. Zhao, Y. Chen, and L. Lv, "Deep reinforcement learning with visual attention for vehicle classification," *IEEE Trans. Cogn. Develop. Syst.*, vol. 9, no. 4, pp. 356–367, Dec. 2017.
[13] K. Muchtar, Nasaruddin, Afdhal, and I. Nugraha, "Attention-based approach for efficient moving vehicle classification," in *Proc. 4th Int. Conf. Comput. Sci. Comput. Intell. (ICCSCI)*, Yogyakarta, Indonesia, 2019, pp. 683–690.
[14] M. Jian, K.-M. Lam, and J. Dong, "Illumination-insensitive texture discrimination based on illumination compensation and enhancement," *Inf. Sci.*, vol. 269, pp. 60–72, Jun. 2014.
[15] C.-Y. Lin, K. Muchtar, W.-Y. Lin, and Z.-Y. Jian, "Moving object detection through image bit-planes representation without thresholding," *IEEE Trans. Intell. Transp. Syst.*, to be published.
[16] Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar, "CDnet 2014: An expanded change detection benchmark dataset," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2014, pp. 393–400.
[17] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Bombay, India, Jan. 1998, p. 2.
[18] C.-H. Yeh, C.-Y. Lin, K. Muchtar, and L.-W. Kang, "Real-time background modeling based on a multi-level texture description," *Inf. Sci.*, vol. 269, pp. 106–127, Jun. 2014.
[19] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*, [Online]. Available: https://arxiv.org/abs/1804.02767
[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[21] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 1999, pp. 246–252.
[22] D. Liang, S. Kaneko, M. Hashimoto, K. Iwata, and X. Zhao, "Co-occurrence probability-based pixel pairs background model for robust object detection in dynamic scenes," *Pattern Recognit.*, vol. 48, pp. 1374–1390, Apr. 2015.
[23] S. Bianco, G. Ciocca, and R. Schettini, "Combination of video change detection algorithms by genetic programming," *IEEE Trans. Evol. Comput.*, vol. 21, no. 6, pp. 914–928, Dec. 2017.
[24] S. Jiang and X. Lu, "WeSamBE: A weight-sample-based method for background subtraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2105–2115, Sep. 2018.
[25] A. Stergiou, G. Kapidis, G. E. Kalliatakis, C. Chrysoulas, R. Veltkamp, and R. Poppe, "Saliency Tubes: Visual explanations for spatio-temporal convolutions," in *Proc. IEEE ICIP*, Taipei, Taiwan, May 2019, pp. 1–5.

NASARUDDIN NASARUDDIN (M'13) received the B.Eng. degree in electrical engineering from the Sepuluh Nopember Institute of Technology, Surabaya, Indonesia, in 1997, and the M.Eng. and D.Eng degrees in physical electronics and informatics from the Graduate School of Engineering, Osaka City University, Japan, in 2006 and 2009, respectively. He was the Head of master of electrical engineering programme with the Graduate School, Syiah Kuala University. He is currently the Head of the Electrical and Computer Engineering Department, Faculty of Engineering, Syiah Kuala University, where he is also a Full Professor with the Electrical Engineering Department. His research interests include digital communications, information theory, and computer and communication networks.

KAHLIL MUCHTAR (M'17) received the B.S. degree in information technology from the School for Engineering of PLN's Foundation (STT-PLN), Jakarta, Indonesia, in 2007, the M.S. degree in computer science and information engineering from Asia University, Taichung, Taiwan, in 2012, and the Ph.D. degree in electrical engineering from National Sun Yat-sen University (NSYSU), Kaohsiung City, Taiwan. He is currently a Lecturer and Researcher with the Department of Electrical and Computer Engineering, Syiah Kuala University, Banda Aceh, Indonesia. He is also an AI Research Scientist with Nodeflux, Jakarta, Indonesia. His research interests include computer vision and image processing. He served as the Publication Chair of the IEEE ICELTICs 2017 and 2018. He received the 2014 IEEE GCCE Outstanding Poster Award and IICM Taiwan 2017 The Best of Ph.D. Dissertation Award.

AFDHAL AFDHAL (M'17) received the bachelor's degree (S.T.) in computer system engineering from the Department of Electrical Engineering, Faculty of Engineering, Syiah Kuala University, Banda Aceh, Indonesia, in 2003, and the M.Sc. degree in distributed computing and networks from the School of Computer Sciences, Universiti Sains Malaysia, Penang, Malaysia, in 2013. He has been a Lecturer of computer engineering program with the Department of Electrical and Computer Engineering, Faculty of Engineering, Syiah Kuala University, since 2004, where he is currently appointed as the Coordinator of computer engineering undergraduate program with the Department of Electrical and Computer Engineering. His research interests include intelligent transportation systems, data and communication systems, parallel and distributed systems, mobile ad-hoc networks, vehicular ad-hoc networks, network security, the Internet of Things (IoT), and the Internet of vehicles (IoV).