Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 8, No. 4, December 2022, pp. 587-598
ISSN: 2338-3070, DOI: 10.26555/jiteki.v8i4.25120

A Deep Neural Network Model for Realtime Semantic-Segmentation Video Processing supported to Autonomous Vehicles

Trung-Nguyen Bui [1,2], Hanh Phan-Xuan [1,2], Thuong Le-Tien [1,2*]
[1] Electrical & Electronics Engineering Department HoChiMinh City University of Technology, HCMUT, HoChiMinh city - 70000, Vietnam
[2] Vietnam National University, VNU HoChiMinh city, Vietnam

ARTICLE INFO

Article history:

Received October 30, 2022
Revised December 10, 2022
Published December 27, 2022

Keywords:

Traffic density;
semantic segmentation;
mean Intersection over Union;
F1 metric;
Saigon Aerial and UAVid data set

ABSTRACT

Traffic congestion has been a huge problem, especially in urban area during peak hours, which causes a major problem for any unmanned/autonomous vehicles and also accumulate environmental pollution. The solutions for managing and monitoring the traffic flow is challenging that not only asks for performing accurately and flexibly on routes but also requires the lowest installation costs. In this paper, we propose a synthetic method that uses deep learning-based video processing to derive density of traffic object over infrastructure which can support usefull information for autonomous vehicles in a smart control system. The idea is using the semantic segmentation, which is the process of linking each pixel in an image to a class label to produce masked map that support collecting class distribution among each frame. Moreover, an aerial dataset named Saigon Aerial with more than 110 samples is also created in this paper to support unique observation in a biggest city in Vietnam, HoChiMinh city. To present our idea, we evaluated different semantic segmentation models on 2 datasets: Saigon Aerial and UAVid. Also to track our model's performance, F1 and Mean Intersection over Union metrics are also taken into account. The code and dataset are uploaded to Github and Kaggle repository respectively as follow: Saigon Aerial Code, Saigon Aerial dataset.

Corresponding Author:

Thuong Le-Tien, Electrical & Electronics Engineering Department HoChiMinh City University of Technology, HoChiMinh city - 70000, Vietnam -Vietnam National University, VNU, HoChiMinh city, Vietnam
Email: thuongle@hcmut.edu.vn

1. INTRODUCTION

Traffic congestion is a condition that occurs frequently in many route system in the world. Characterized by long cumulated queue of vehicles moving in extremely low speed, traffic congestion has become a root of negative health issues occurs in traffic participants, accounting both mental and physical sides. Moreover, the passive flow in traffic jam generally discontinues transportation in which postpone the economic activities and affect productivity of some certain group of workers. To encounter this problem, many individual and systematic solutions have been suggested. One that has been discussed in [1] by Naoki Shibata et al. about using Inter-Vehicle Communication based on standard IEEE 802.11 to measure cars activities in network and generate the statistic that relevant to traffic jam situation in routes. More specifically, authors calculated the time that each car moves through 2 ends of the road and then compare with the results received from other cars which pass the same road to estimate the time that vehicle can arrive to destination. Although showing good idea of detecting traffic flow, this method however, requires a broad samples with the same network installed while unable to resolve the core reason which is traffic jam. Another approach that has been proposed in [2] where Yinli Jin et al. suggested using a surveillance system with high resolution for segmentation task to detect traffic congestion. With this approach Yinli Jin has proposed using pixel-wise

detection with existing infrastructure CCTV (Closed Circuit Television), which is both potential and simple to launch in large scale.

Using Unmanned Aerial Vehicles (UAVs) as a replacement for fixed cameras offers a superior in terms of flexibility and cost-efficiency. It is evident that the drawback of static cameras are the limitation of the frames and the constant existence of dead zones outside of the cameras' range. In reality, there are several studies that have applied UAVs to monitor the traffic situation [3][4][5][6]. In reality, there are several studies that have applied UAVs to monitor the traffic situation. The paper "A UAV-Based Traffic Monitoring System" [3] by Haoran Niu and his team implements a prototype with a quadcopter, an onboard camera with video and data processing algorithms, and a web application. The study based on the Haar cascade model and a frame-by-frame tracking stage. However, the results of the study just detect a single vehicle in ideal condition so it is not suitable for traffic density monitoring. Another study "Monitoring road traffic with a UAV-based system" [4] by M. Elloumi and his team employs the method of collecting various information, namely speed, vehicle position and movement direction for UAV trajectories computation with the aim to monitor the optimal amount of vehicles within the widest time frame possible, which is an attempt to identify congestion basing on the detection of one single low-speed vehicle. Nevertheless, relying on vehicle's speed to detect traffic density is not really optimal, especially when the vehicles temporarily stop for traffic light signal.

Image segmentation is one of the most sophisticated technologies in image processing which its roles can be found in multiple applications from anatomical segmentation [7][8][9] to aerial imagery segmentation [10][11][12]. In pixel-level, we can apply this technique to acquired higher analysis ability as well as understand the contextual meaning within considered problem. In this paper, we propose a theoretical model for detecting traffic density using UAV with deep-neural network, this would help the estimated tracking of autonomous vehicles. Moreover a small local aerial dataset is also created to support our research, we call it "Saigon Aerial". With the wider scene provided from UAV, we expect to form a mask from semantic segmentation application with 4 different classes that will help us estimate overall traffic capacity in urban route. The pixel-by-pixel processing is highly effective in analyzing and identifying traffic density based on each frame. We recommend 2 typical models that have been verified in many previous studies which is DeepLabV3plus model applying Atrous convolution [13] and Unet applying up-convolution for multi-class semantic segmentation [9]. Throughout this research, we also analyze our Saigon Aerial dataset and point out what factors that can affect prediction ability. At method section, we will introduce briefly about Saigon Aerial dataset, the architectures that we use and the technique to specify traffic density index. Experiment section demonstrates the process and tools that we utilize in this study. Finally, in evaluation we compare our result with another aerial dataset named "UAVid" to yield the final performance of our methods as well as to discuss about further development that we can improve the study.

2. METHODS

2.1. Saigon Aerial Dataset

In this paper, we introduce an aerial dataset which captured multiple view angles of route system in Ho Chi Minh city. Especially, according to [14] and [15], up to 2nd Quarter 2022, motorcycles contribute about 65,917,000 registered vehicles in Vietnam, proving that this private vehicles play an important role in Vietnam's transportation. Thus we focus on motorcycle class in *Saigon Aerial dataset* with the hope to create an adaptable information that fit with Vietnamese transportation routine. From 31 different videos collected with 4K resolution and 30 fps we split and resized them into 114 images with size 1920×1080.

After clarifing images, we labeled the objects in frame with help of **CVAT (Computer Vision Annotation Tool)** [16] tool which is a free, open-source, web-based image annotation tool designed for image labelling purpose. In this dataset, we defined 4 classes which are crucial for our application: *'motorbike'*, *'car'*, *'road'* and *'background clutters'*. Especially, *'car', 'motorbike'* and *'road'* play an important role on specifying *traffic density* which is another essential index that we will discuss later in Section 2.3. We also performed analysis on Saigon Aerial to check the distribution on each class in this dataset. As Fig. 1 has shown, in total the classes that possess the most number of pixels are *'background cluster'* and *'road'* as they account for 97% of class contribution in the dataset. Even with the image that having the most excited traffic activities, we still recorded the strictly unbalanced between infrastruction class type and vehicles class type. This can be an unpleasant matter for the model in the effort of learning the distinction between different groups later on. On the other hand, labeling processing is a quite expensive process as it required at least 2 hours to annotate every object in 1 frame of Saigon Aerial dataset to their truth labels (Fig. 2). From the labeled dataset, we divided into 2 sets of images. Specifically, 91 images are used for training, the remaining 23 images are used for validation. Unet architecture illustration show in Fig. 3.

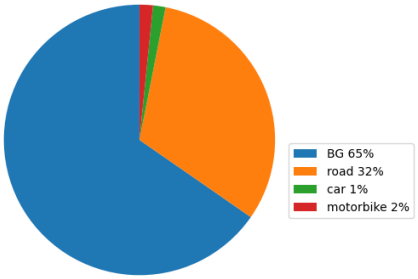

Fig. 1. Class distribution in Saigon Aerial dataset. In the general scale, while unnecessary class 'background' yields the enormous amount in Saigon Aerial, other core classes take the much lower contribution on the dataset, expecially vehicle class as we expect the ratio of road and vehicle is approximate to each other [14][15]

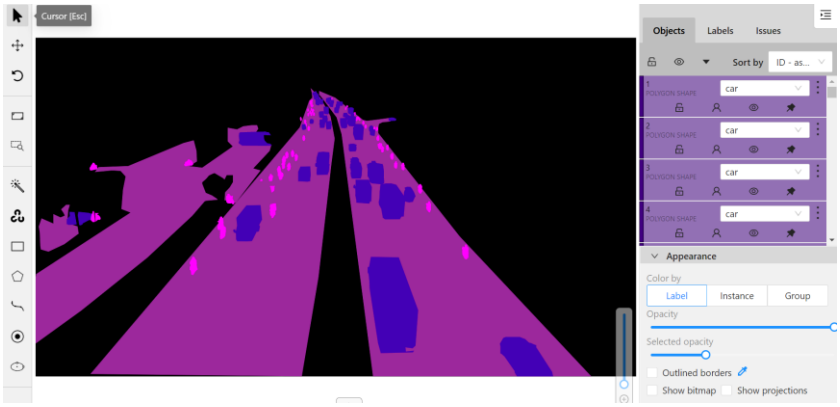

Fig. 2. Saigon Aerial dataset on CVAT tool [16]

Class categories are depicted in the following RGB colors:

Classes	***RGB colors***
• motorbikes	(192, 0, 192)
• cars	(64, 0, 128)
• road	(128, 64, 128)
• background clutter	(0, 0, 0)

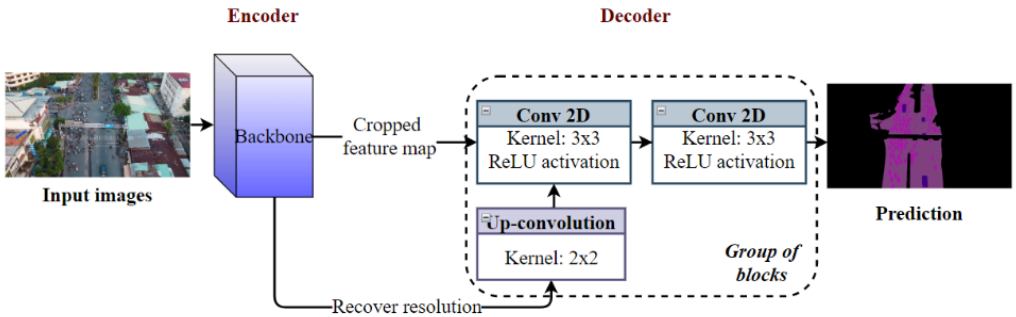

Fig. 3. Unet architecture illustration. Backbone network is chosen from state-of-art model, plays a role as encoder to exploit feature in multi-scale. Decoder consist of up-convolution and 2D convolution recover original resolution of feature map

2.2. Segmentation Model

2.2.1. Unet

In several years, Unet [6] and its variants [17][18][19] have been used in medical semantic segmentation application for simple and accuracy reasons. Moreover, Unet is depicted as an elegant model that does

not require much training images while still produce good segmentation result. For the traditional Unet, the network architecture consists of 2 important modules: Encoder and Decoder.

Inside Encoder is multiple of 3×3 unpadded Convolution layers for feature extraction operation, supported by 2×2 Max Pooling layer for down sampling operation. In contrast with Encoder, at Decoder part, feature resolution will be recover by using 2×2 up sampling convolution layers and at the end, go through 1×1 convolution where feature components of previous layers are mapped to the classes that we want to predicted. On the other hand, Encoder and Decoder process is connected by a so-called "cropped feature map" from Encoder and concatenate with corresponding block at Decoder. This contracting path is crucial as in every convolution step, pixel's border will be loss, thus cropping option provides additional feature to gain a better performance at output segmentation map.

2.2.2. DeeplabV3plus

DeeplabV3plus was first introduced in 2019, it is actually an updated version from previous work on DeeplabV3 model [21][22], where the problem of multi scale object existence was resolved by applying Spatial Pyramid Pooling (SPP) [23][24][25]. On the other hand, DeeplabV3plus was also successfully applied to many aerial segmentation datasets as well as achieved realiable output prediction on street's revelant context [26][27]. The encoder module of DeeplabV3plus is attached with Atrous convolution [21], replace for deep convolutional neural network (DCNN) to control the resolution of output feature map. The benefit of Atrous convolution was approved in [13][21] where applying consecutive Atrous convolution with rate shows a better field-of-view in output rather than the effect of striding in DCNN. Notably, the structure of DeeplabV3plus still utilizes a great validity usage of Atrous Spatial Pyramid Pooling [21] from DeeplabV3 to perform multi-rate resampling. Specifically, in DeeplabV3plus a group of rates *(1,6,12,18)*, Atrous convolution is derived parallelly while later on their outputs along with pooling layer's output are concatenated before going through a 1×1 convolution to produce output of Encoder module.

Moreover, originally DeeplabV3plus used a special operator named depthwise separable convolution [28] helps to reduce computational and resources cost while maintaining the same efficiency (Fig. 4). To accomplish this, a modified version of Xception network [29][30][31] in which all max pooling layers are replaced by depthwise separable convolution, is use with the aim to both allow atrous separable convolution involve extract feature map in arbitrary resolution and lessen operations that the model have to calculate.

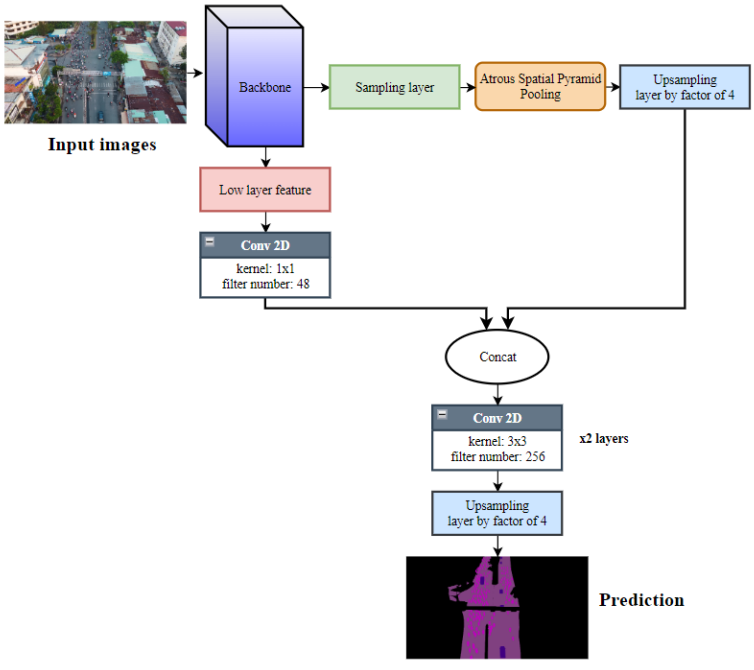

Fig. 4. DeeplabV3plus architecture. Atrous Spatial Pyramid Pooling is a sequence of parallel Atrous convolutions with different rates. By adusting filter's rate, multi-scale information is expoilted significantly

2.3. Traffic Density Analysis

There are 2 factors that can account to determine whether a sub-road is dense or not: *the number of vehicles moving in road surface* and *the area of road surface*. In this paper, we want to apply these 2 parameters under pixel's quantities form to forecast the density of a road in 1 frame, derived by below equation:

$$Traffic\ density\ = \frac{Number\ of\ vehicles's\ pixels}{Number\ of\ road\ surface's\ pixels} \quad (1)$$

In Saigon Aerial dataset, we calculate *traffic density* for each image to classify degree coverage categories of vehicles over road surface. As has presented in Fig. 5, for total 114 images in scope of Saigon Aerial, we find that all of them have traffic density index less than 0.25. It means that when regardless the time parameter, the traffic scenario that we have captured are all sparse. The highest traffic density value that we have recorded in set of Saigon Aerial dataset is 0.1882, derived from picture 1 of folder 8 as we have presented in Fig. 6. Intuitively, the Fig. 6 shows the scene at T-junction with traffic light where we can observe all vehicles are just starting to move. This explains the reason why traffic density of this image is higher than others as the cumulative vehicles stop for red light create small portion that account to cover the road surface. Thus for real-time process application, adding time parameter is a good decision to avoid situation where temporary stoppage at crossroads.

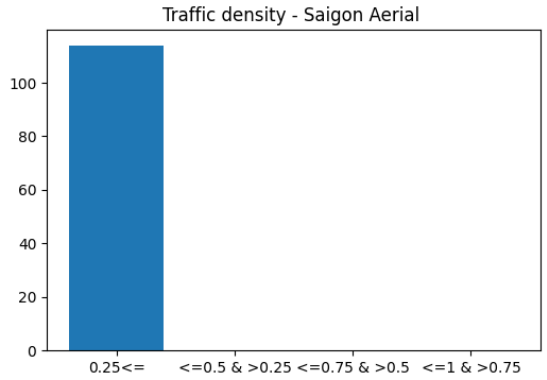

Fig. 5. Traffic density results at of Saigon Aerial. The absolute range of traffic density less than 0.25 refelects the fact that the traffic flow is under sparse conditions in our dataset scope

Fig. 6. Masked image in Saigon Aerial dataset. The capturing scene was at T-junction after red light goes off where the amount of accumulated vehicles are, although noticeable, it easy see that the free space on the road is still considerablely large

2.4. Experiment

2.4.1. Hardware

Normally, comprehensive sensors are the key to control autonomous vehicles. The raw data from the sensors are analyzed and processed in order to make informed decisions. The vehicle then plans to track a near-range trajectory autonomously. In the proposed model, Fig. 7, we propose a long range tracking with the support of UAVs by prediting traffic density combined to various types of common sensors such as cameras, LIDARs, RADARs, SONARs, IMU, GPS (or any navigation systems), etc. The information about the

traffic processed by an UAV will be sent to the control centre or directly to autonomous vehicles for estimating the moving time or slecting an optimum path. In the practical experiment, the hardware for image data collection, we use the DJI Mini 2 drone for obtaining video. With its compact design and equipped with a camera that can capture videos in full HD quality, so that it helps to collect data effectively. For the simulation, we use Google Colab-Pro to train the image dataset. It allows us to get access to GPUs P100 which is impressively fast. Also, Google Colab Pro supports 32 GB of RAM and runtimes are up to 24 hours. The framework that we chose to work with is Tensorflow [32] and its sub-library Keras as both Keras and Tensorflow support convienient API that make creating and testing 2 proposed models easier.

Fig. 7. The proposed model for the link with UAV-based traffic segmentation for helping autonomous vehicles (DJI Mini 2 drone [33])

3. RESULTS AND DISCUSSION

3.1. Training Model

To present our idea, we use 2 segmentation models which are DeeplabV3plus and Unet, combines with multiple pretrained networks like Resnet [34], Inception [35] and Xception [28] to get more effective result. The training process does not require too many time as with 512×512 image resolution and on the same platform Google Colab Pro, all models took totally more than 1 hour to complete training. Importantly, we notice that the Saigon Aerial dataset is critically imbalanced, which leads to the fact that inference results just display predominantly 2/4 classes. In order to resolve this problem, we decide to combine 2 strategies:
(1): Use Dice loss [36] and Focal loss [37] to focus on detecting important minor classes in the training process.
(2): Add class weights to prevent the model tends to predict more common classes in imbalanced dataset. Specifically, we set class weight according to the statistics in Fig. 1, thus the rate of each class is [1,2,8,8] corresponding to *background-road-car-motorbike*.
Formula of Dice loss is defined as:

$$DL(y,\hat{y}) = \frac{y + \hat{y} - 2y\hat{y}}{y + \hat{y} + \alpha} \quad (2)$$

where α is a constant coefficient to prevent the denominator equal to 0.
Formula of Focal loss is defined as:

$$FL = \; \lVert -\alpha(1 - p\_i) \rVert \; \hat{} \gamma \; log(p\_i) \quad (3)$$

where α is a weighting factor in balanced cross entropy, γ is focusing parameter for modulating factor $(1 - p_i)$, and p_i is frequency of class i appears in prediction mask.

In training process, we recorded that the parameters were updated very smoothly however overfitting still happened with validation process when the degradation of loss function seemly fluctuate although we have applied regularizer term. In another aspect, the performances evaluated by IoU metrics then showed an active trend on prediction when training and validation mean IoU both surpassed 80%. As we have expected the noise variation in training set are much more than validation set, resulting IoU values of validation set outperform over training set. Remind that the 2 important classes *'car'* and *'motorbike'* account for a very low percentage in Saigon Aerial dataset, thus high metric results can still open the probability of incorrect predictions on these 2 classes. Loss function in training and validation set of Unet + Inception model show in Fig. 8. mIOU metrics in training and validation set of Unet + Inception model show in Fig. 9.

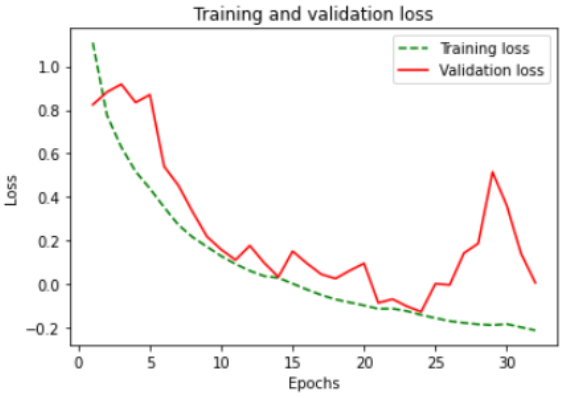

Fig. 8. Loss function in training and validation set of Unet + Inception model

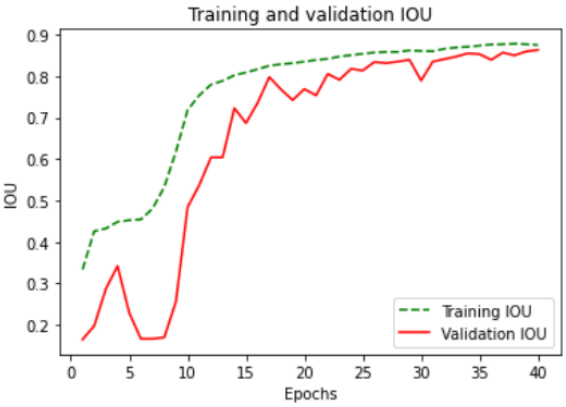

Fig. 9. mIOU metrics in training and validation set of Unet + Inception model

3.2. Semantic Segmentation Result

One of the efficiency way to evaluate segmentation model nowadays is using mean Intersection Over Union metric (mIoU) [38], presented by following equation:

$$IoU = TP/(TP + FP + FN) \tag{4}$$

where TP, FP, FN stand for True Positive, False Positive and False Negative respectively. This metrics will tell how our prediction mask matches with the image's label. It can be said that, the higher IoU score is, the closer prediction similar to the ground truth.

Another metrics that is really effective with imbalanced data that we also applied in our works is F1-score [39]. Different with IoU, F1-score using 2 basic metrics which are Precision and Recall to calculate for each class then systhesize the final measurement as F1-score. Using F1-score help to evaluate the overall performance of segmentation while IoU tend to penetrate more on single false prediction.

$$\begin{aligned} F1\ score &= 2 \times (Precision \times Recall)/(Precision + Recall) \\ &= (2 \times TP)/(2 \times TP + FP + FN) \end{aligned} \tag{5}$$

In the inference, we compare the result of our model on Saigon Aerial and UAVid dataset [40] to evaluate the efficiency of proposal method. However, the final outputs show that Unet architecture is more appropriate with Saigon Aerial dataset than DeeplabV3plus. We think this phenomenon occurs because the amount of data in Saigon Aerial is too small to applied such complex structure like DeeplabV3plus, leads to the fact that DeeplabV3plus has under average mIOU index. This problem also happened with UAVid when we applied DeeplabV3plus with this dataset. On the other hand, we present the amount of parameters of each model has in Table 1, showing the slight differences in parameters of these model. In fact the weights of these models mostly come from pretrained backbone networks where low-level feature extracting is accomplished.

Moreover, to measure the mean time needed to produce the predicted mask, we fed 3 random images into each model. In general, we notice that Resnet-based architectures took more time to produce outputs when both used structures consisting of Resnet50 and Resnet101 spend the longest time respectively. Inception-based architecture instead had the shortest time to predict as this model also has the least weight compared with other models. Mean IOU metrics validation set of on 2 datasets show in Table 2. Mean prediction time on SG Aerial dataset show in Table 3.

Table 1. Number of parameters for each architecture

Architecture	No. Parameters
DeeplabV3plus + Resnet101	30 millions
DeeplabV3plus + Xception	35 millions
Unet + Resnet50	32 millions
Unet + Inception	29 millions

Table 2. Mean IOU metrics validation set of on 2 datasets

Dataset	Architectures	mIoU	F1
SG Aerial	DeeplabV3plus + Resnet101	63.66 %	55.55 %
	DeeplabV3plus + Xception	68.43 %	54.33 %
	Unet + Resnet50	58.48 %	54.30 %
	Unet + Inception	**86.28 %**	80.75 %
UAVid	DeeplabV3plus + Resnet50	43.97 %	-
	DeeplabV3plus + Xception	35.59 %	-
	Unet + Resnet50	38.93%	-
	Unet + Inception	**46.03%**	-

Table 3. Mean prediction time on SG Aerial dataset

Architectures	Mean recognition time
DeeplabV3plus + Resnet101	0.7306 s
DeeplabV3plus + Xception	0.6419 s
Unet + Resnet 50	0.8790 s
Unet + Inception	0.6070 s

After doing multiple tests with Unet and DeeplabV3plus model, we consider the combination between Unet + Inception backbone brings the best result that closely matches our expectations. As Fig. 10 presents the predictions which were accomplished by Unet + Inception network, showing that this model can specify significantly the boundary between road class and background class as well as points out most of the motor-cycles in the frame. As Fig. 11 present the prediction mask produced by Unet + Inception backbone on UAVid dataset.

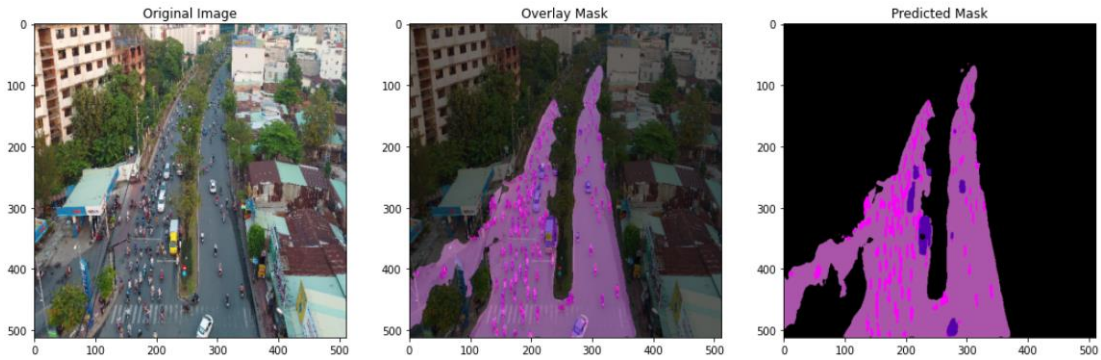

Fig. 10. Prediction mask produced by Unet + Inception backbone on Saigon Aerial dataset

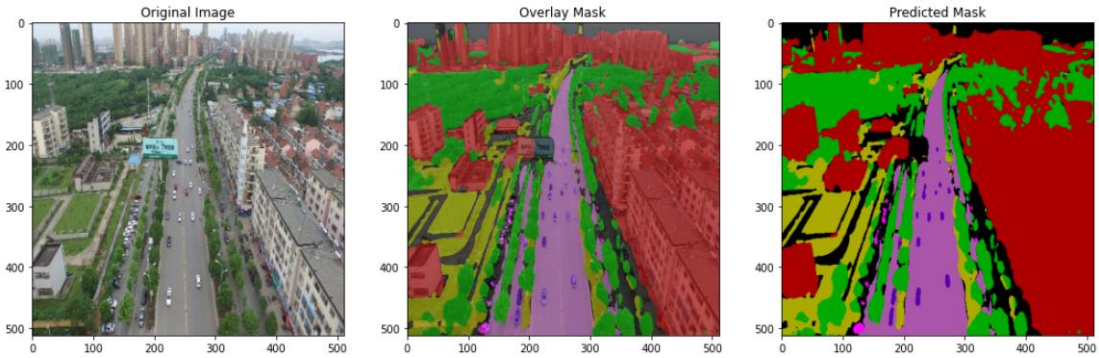

Fig. 11. Prediction mask produced by Unet + Inception backbone on UAVid dataset

Along with the statistic in Fig. 1, our prediction in Fig. 10 shows a similar trend where the number of motorbike class are much lower than the road surface class with a small difference when the ratio between road and vehicles is approxiamate *0.166* instead of 0.09 like in Fig. 1. In the Fig. 10, the image in the test set display the percentage of pixel that belong to *'road', 'car'* and *'motorbike'* are 22.84 %, 0.74% and 3.06% respectively, thus giving **traffic density index = 0.166**. As we can see, the traffic condition in test image in Fig. 10 is an ideal one when all vehicles have a lot of space to move, thus the traffic density that the model returns should be consider as an acceptable one. Distribution of labels in image show in Fig. 12.

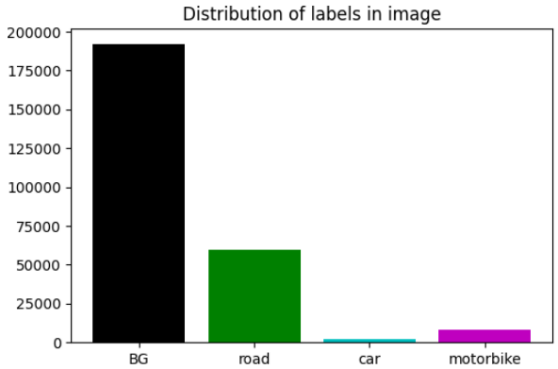

Fig. 12. Class distribution of predicted test image in Fig. 10. The ratio between the road and motorbike is reasonable to compare with the context of the test image in Fig. 10 as it is easy to notice that traffic density is sparse

On the other hand, when we try to figure out the model, we found that even though the quantity of motorbikes/cars can be resolved by adding more samples, there is the fact that the size of motorbike object displayed on our dataset is pretty small in such a broad road surface and this can somehow cause ambiguity when model tries to predict motorbike class with the similar shape of the road. Fig. 13 displays the cropped images representing the portion that we truly want to analyze and segment. By eliminating background, we believe that there is a higher chance that a single vehicle will have more attention from the deep learning model and from that avoid the phenomenon of the model predicting vehicle class mimics road class without knowing the distinction between these 2 classes.

Fig. 13. Examples for the cropped frames that we want to focus on. However, the size of motorbikes can be a critical problem for the model to predict in large scale

3.3. Discussion

In this paper, we have evaluated our method to identify traffic conditions with the segmentation method using UAV for supporting autonomous vehicles. The final result largely depends on the segmenting ability of deep learning models thus multiple additional techniques have been proposed to increase the efficiency of predictions. With the test set, we have tried to run several cases and also achieved very good outcomes alt-

hough with the shortage in date, the model still returns its labeling masks with explicitly segmentation between classes. The only problem is that when we try to use another source as an input (with different scenarios, e.g the street is extremely crowded) then the model quickly shows the unfamiliar with the context and derive miss segmentation on vehicle classes.

Through this paper, researchers and stakeholders can enhance a new method to supervise traffic conditions quickly and flexibly at every location in the city. However, to dive into a model that can detect local regions objects, demands a specific dataset to adapt to habit and cultural signature of that area. In this case, we implement the proposed method in Ho Chi Minh city streets. Unfortunately, labeling manually images that we have collected may consume a lot more cost and time. This can be completed in the future but in this paper, we only extract masks that we have manually labeled to exhibit the proposed method.

We realize that there are still some cautions that can be further evaluated while doing this method. Firstly, by using a UAV to observe traffic conditions, we need to establish a certain altitude degree that allows the UAV to obtain the aggregate details of the street while still ensuring safety standards. If the drone operates under an unsuitable (too low or too high) altitude, one can achieve a noisy result and abate the accuracy of prediction. Secondly, weathers conditions like rain can probably affect to operating and recording ability of a drone in reality

4. CONCLUSION

Along with developing proposed method, a combination of traffic flow measurement and real-time processing can upgrade this method to a new direction of UAV-based supporting systems for improving smart controls of autonomous vehicles and also solving urban traffic problems. Compared with previous methods, this model has significant improvements while utilize the flexibility of a drone to wriggle through the out-of-range position of a CCTV. Moreover, using segmentation even though demands more computations, the extracted information it return are more in details and covered in overall scale which is unaffected by unusual behavior of a few individuals. Furthermore, traffic density data may also support authority/infrastructure stakeholders to analyze critical constraints of current system and from that improve a more efficient approach to control traffic flow during rush hours. Although there are few points that our method struggles with, we believe that these problems are able to settle in future, showing that this method still have a high potential to deploy as a practical application. We hope that our research will motivate more works in the future that can complete our method as well as develop a protocol that real-time processing traffic density analysis can be accomplished under mobile device.

ACKNOWLEDGMENT

This study is partially supported as a research grant, To-ĐĐT-2021-01 by University of Technology (HCMUT), National University of Ho Chi Minh City (VNU), VIETNAM,

REFERENCES

[1] N. Shibata *et al*., "A Method for Sharing Traffic Jam Information using Inter-Vehicle Communication," *2006 3rd Annual International Conference on Mobile and Ubiquitous Systems – Workshops*, pp. 1-7, 2006, https://doi.org/10.1109/MOBIQW.2006.361760.

[2] Y. Jin, W. Hao, P. Wang and J. Wang, "Fast Detection of Traffic Congestion from Ultra-low Frame Rate Image based on Semantic Segmentation," *2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 528-532, 2019, https://doi.org/10.1109/ICIEA.2019.8834159.

[3] H. Niu, N. Gonzalez-Prelcic and R. W. Heath, "A UAV-Based Traffic Monitoring System - Invited Paper," *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pp. 1-5, 2018, https://doi.org/10.1109/VTCSpring.2018.8417546.

[4] M. Elloumi, R. Dhaou, B. Escrig, H. Idoudi and L. A. Saidane, "Monitoring road traffic with a UAVbased system," *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1-6, 2018, https://doi.org/10.1109/WCNC.2018.8377077.

[5] R. Aissaoui, J.-C. Deneuville, C. Guerber, and A. Pirovano, "UAV Traffic Management : A Survey On Communication Security," *arXiv preprint arXiv*, 2022, https://doi.org/10.48550/arXiv.2211.05640.

[6] J. Y. J. Chow, "Dynamic UAV-based traffic monitoring under uncertainty as a stochastic arc-inventory routing policy," *International Journal of Transportation Science and Technology*, vol. 5, no. 3, pp. 167–185, 2016, https://doi.org/10.48550/arXiv.1609.03201.

[7] J. Chen et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv*, 2021, https://doi.org/10.48550/arXiv.2102.04306.

[8] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, "Self-Supervision with Superpixels: Training Few-shot Medical Image Segmentation without Annotation,", *arXiv*, pp. 762-780, 2020, https://doi.org/10.48550/arXiv.2007.09886.

[9] O. Ronneberger, *et al.*, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *arXiv*, pp. 234-241, 2015, https://doi.org/10.48550/ARXIV.1505.04597.
[10] M. Tschannen, L. Cavigelli, F. Mentzer, T. Wiatowski, and L. Benini, "Deep Structured Features for Semantic Segmentation," *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 61-65, 2017, https://doi.org/10.48550/arXiv.1609.07916.
[11] Z. Zhang, Q. Liu, and Y. Wang, "Road Extraction by Deep Residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, May 2018, https://doi.org/10.48550/arXiv.1711.10684.
[12] L. Huang, Q. Dong, L. Wu, J. Zhang, J. Bian, and T.-Y. Liu, "AF2: Adaptive Focus Framework for Aerial Imagery Segmentation," *arXiv preprint arXiv*, 2022, https://doi.org/10.48550/arXiv.2202.10322.
[13] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," *European Conference on Computer Vision*, pp. 833–851, 2018, https://doi.org/10.48550/arXiv.1802.02611.
[14] Vietnam Association of Motorcycle Manufacturers. (n.d.), "Sales data," *VAMM*. Retrieved July 15, 2022, from https://vamm.vn/sales-data/.
[15] ASEAN Statistics Division. (n.d.). "Total number of registered motorcycles (in Thousand)," Aseanstatsdataportal. Retrieved October 7, 2022, from https://data.aseanstats.org/indicator/ASE.TRP.ROD.B.011.
[16] B. Sekachev *et al.*, opencv/cvat: v1.1.0. Zenodo, 2020. doi: 10.5281/zenodo.4009388.
[17] A. G.-U. Juarez, H. A. W. M. Tiddens, and M. de Bruijne, "Automatic Airway Segmentation in chest CT using Convolutional Neural Networks," *In Image analysis for moving organ, breast, and thoracic images*, pp. 238-250, 2018, https://doi.org/10.48550/arXiv.1808.04576.
[18] N. Ibtehaz and M. S. Rahman, "MultiResUNet : Rethinking the U-Net architecture for multimodal biomedical image segmentation," *Neural Networks*, vol. 121, pp. 74–87, 2020, https://doi.org/10.48550/arXiv.1902.04049.
[19] N. Siddique, S. Paheding, C. P. Elkin, and V. Devabhaktuni, "U-Net and Its Variants for Medical Image Segmentation: A Review of Theory and Applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021, https://doi.org/10.48550/arXiv.2011.01118.
[20] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, 2014, https://doi.org/10.48550/arXiv.1409.0575.
[21] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 99, 2016, https://doi.org/10.48550/arXiv.1606.00915.
[22] Liang-Chieh Chen, George Papandreou, Florian Schroff, Hartwig Adam. "Rethinking Atrous Convolution for Semantic Image Segmentation," *Computer Vision and Pattern Recognition*, 2017, https://doi.org/10.48550/arXiv.1706.05587.
[23] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *Computer Vision – ECCV 2014, Springer International Publishing*, pp. 346–361, 2014, https://doi.org/10.48550/arXiv.1406.4729.
[24] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," *Tenth IEEE International Conference on Computer Vision (ICCV'05)*, pp. 1458-1465, 2005, https://doi.org/10.1109/ICCV.2005.239.
[25] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2881-2890, 2016, https://doi.org/10.48550/arXiv.1612.01105.
[26] T. Yasuno, J. Fujii, H. Sugawara, and M. Amakata, "Road Surface Translation Under Snow-covered and Semantic Segmentation for Snow Hazard Index," *arXiv*, pp. 81-93, 2021, https://doi.org/10.48550/arXiv.2101.05616.
[27] M. R. Heffels and J. Vanschoren, "Aerial Imagery Pixel-level Segmentation," *arXiv*, 2020, https://doi.org/10.48550/arXiv.2012.02024.
[28] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices," *arXiv*, 2017, https://doi.org/10.48550/arXiv.1707.01083.
[29] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1800-1807, https://doi.org/10.1109/CVPR.2017.195.
[30] J. Dai *et al.*, "Deformable Convolutional Networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 764-773, 2017, https://doi.org/10.48550/arXiv.1703.06211.
[31] A. G. Howard *et al.*, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv*, 2017, https://doi.org/10.48550/arXiv.1704.04861.
[32] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M.& others, "TensorFlow: A System for Large-Scale Machine Learning," *12th USENIX Symposium on Operating Design and Implementation (OSDI)*, pp. 265-283, 2016, https://doi.org/10.5281/zenodo.6574269.
[33] DJI Official, "DJI Mini 2 – Video" - DJI. (n.d.). Retrieved October 27, 2022, from https://www.dji.com/mini-2/video.
[34] He, K., Zhang, X., Ren, S., & Sun, J. , "Deep Residual Learning for Image Recognition," *arXiv*, 2015, https://doi.org/10.48550/ARXIV.1512.03385.

[35] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision" *arXiv*, 2015, https://doi.org/10.1109/CVPR.2016.308.
[36] C.H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. J. Cardoso, "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (DLMIA ML-CDS 2017) ,Lecture Notes in Computer Science, Springer, Cham*, vol. 10553, pp. 240-248, 2017, https://doi.org/10.1007/978-3-319-67558-9_28.
[37] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *arXiv*, pp. 2980-2988, 2017, https://doi.org/10.1109/ICCV.2017.324.
[38] A. A. Taha, and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC medical imaging*, vol. 15, no. 1, pp. 1-28, 2015, https://doi.org/10.1186/s12880-015-0068-x.
[39] Z. C. Lipton, C. Elkan, and B. Naryanaswamy, "Optimal Thresholding of Classifiers to Maximize F1 Measure," *Machine learning and knowledge discovery in databases : European Conference, ECML PKDD: proceedings. ECML PKDD (Conference)*, pp. 225–239, 2014. https://doi.org/10.1007/978-3-662-44851-9_15.
[40] Y. Lyu, G. Vosselman, G. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A Semantic Segmentation Dataset for UAV Imagery," *arXiv*, vol. 165, pp. 108-119, 2018, https://doi.org/10.48550/arXiv.1810.10438.

BIOGRAPHY OF AUTHORS

Trung-Nguyen Bui was born in 2000 and currently studying Electrical & Electronic Engineering in his final year at Ho Chi Minh city University of Technology.
+ Field of research: digital image processing, deep learning
+ Email: nguyentrungbui45@gmail.com

Hanh Phan-Xuan got the B.Eng and M.Eng Degrees from the HoChiMinh city University of Technology (HCMUT), Vietnam. His research relates to image signal processing, neural networks and deep learning techniques to solve problems in computer vision, image forgery detection, biometrics signal processing and autonomous robotics. Currently, he is a Ph.D student at the EEE department of HCMUT.
Email: phantyp@gmail.com

Thuong Le-Tien received the B.Eng. (1981), M.Eng. from the HoChiMinh city University of Technology (HCMUT), Vietnam National University, VNU, Vietnam and Ph.D at University of Tasmania, Australia, all in Electronics-Communications Engineering. He currently is a full Professor, at the EEE Department of HCMUT. His interests include Signal Processing, Electronics Circuits, and Machine Learning.
Email: thuongle@hcmut.edu.vn
ORCID : ID 0000-0002-5917-4270