DEEP LEARNING FOR SEMANTIC SEGMENTATION OF UAV VIDEOS

Yiwen Wang[1], Ye Lyu[1], Yanpeng Cao[2], Michael Ying Yang[1]

[1]Scene Understanding Group, University of Twente, The Netherlands
[2]School of Mechanical Engineering, Zhejiang University, China

ABSTRACT

As one of the key problems in both remote sensing and computer vision, video semantic segmentation has been attracting increasing amounts of attention. Using video segmentation technique for Unmanned Aerial Vehicle (UAV) data processing is also a popular application. Previous methods extended single image segmentation approaches to multiple frames. The temporal dependencies are ignored in these methods. This paper proposes a novel segmentation method to solve this problem. Combining the fully convolutional networks (FCN) and the Convolution Long Short Term Memory (Conv-LSTM) together, we segment the sequence of the video frames instead of segmenting each individual frame separately. FCN serves as the frame-based segmentation method. Conv-LSTM makes use of the temporal information between consecutive frames. Experimental results show the superiority of this method especially in some classes compared to the single image segmentation model using video dataset from UAV.

Index Terms— FCN, Conv-LSTM, video semantic segmentation, UAV

1. INTRODUCTION

In recent years, with exploding researches in deep learning, video semantic segmentation achieves remarkable advances [1] and has been used in a wide range of applications, including autonomous driving, indoor navigation, surveillance, action recognition and many other academic and real-world applications. Video segmentation is a spatiotemporal foreground segmentation problem. Using video segmentation technique for Unmanned Aerial Vehicle (UAV) data processing is also a popular application. UAVs could obtain images and videos from the dangerous and inaccessible areas where the manned vehicle cannot reach. The largest challenge for video segmentation from UAV is the significant change of object appearance over the video frames which mainly caused by the viewpoint changes of UAV.

The previous researches for the video segmentation from UAV are mainly focused on the frame-based method [2]. Many state-of-the-art deep learning methods such as convolutional neural networks (CNN) have been successfully applied in this topic. The most popular one is fully convolutional networks (FCN) [3], which could provide pixel level segmentation result and take the arbitrary size of input with efficient inference and learning. However, these traditional methods are just extending single image segmentation approaches to multiple frames. Thus, the temporal consistency of the result is poor.

In this paper, we propose an FCN + Conv_LSTM framework for semantic video segmentation. The proposed algorithm tries to combine the FCN model and the Convolution Long Short Term Memory (Conv-LSTM) model [6] together. In this algorithm, the FCN model serves as the frame-based segmentation method which is used to segment each frame individually. The output of this part is segmentation result of each frame. The Conv_LSTM model serves as the post-processing method which makes use of the temporal information between consecutive frames. The inputs of this part are sequences formed by the output segmentation results from FCN model. Conv_LSTM learn the temporal information of these sequences and output the final segmentation results. Experimental results show the superiority of this method especially in some specific classes compared to the single image segmentation model using video dataset from UAV.

2. METHOD

The flowchart of the proposed framework is shown in Fig.1, which describes the whole process of the semantic video segmentation from UAV.

2.1 Dataset

The dataset used in this method is 10 videos captured by UAVs. These videos were captured in Wuhan, China from June to August 2017 and in Gronau, Germany in May 2017. We extracted 30 image sequences from the UAV videos. The extraction

interval is 150 frames. These images were annotated into 5 classes including 4 foreground classes (building, road, cars, vegetation) and 1 background class (clutter).

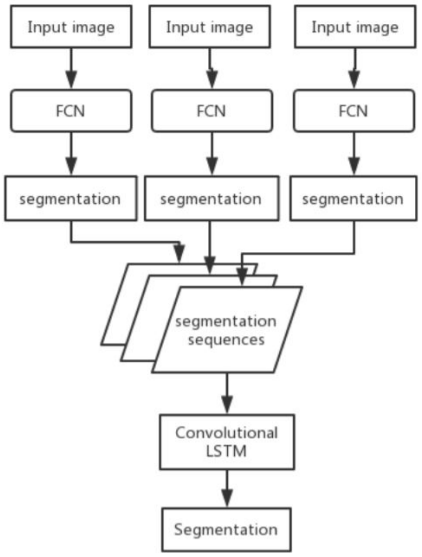

Fig. 1: Framework of the proposed method.

2.2 FCN

In this paper, we choose FCN as the frame-based segmentation method to get the segmentation of each frame individually. This method takes advantage of existing CNNs as powerful visual models that are able to learn hierarchies of features [4]. Compared to the CNN model, it could take the arbitrary size of input and produce correspondingly-sized pixel level output with efficient inference and learning, the typical FCN structure is shown in Fig. 2 [3].

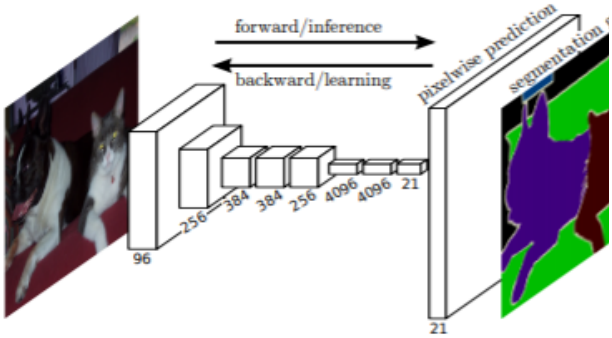

Fig. 2 FCN architecture in [3].

Specifically, there are several different FCN architectures, in this work, we accept skip architecture for better segmentation. We choose FCN-8s which added a skip from pool3 at stride 8 to get more information from the global structure. The outputs of this part are segmentation results of each frame.

2.3 Conv_LSTM

After the frame-based segmentation, we propose to apply Conv-LSTM to make use of temporal information. Conv-LSTM is a kind of Recurrent Neural Networks (RNNs). This neural network contains a looping structure with feedback session which enables it to have a memory of previous states and to learn temporal patterns in data [5]. Compared to traditional RNN, Conv-LSTM could overcome the vanishing gradient problem. It has convolutional structures in both the input-to-state transition and the state-to-state transition [6]. And could help to keep the temporal consistency. Fig. 3 shows the structure of Conv-LSTM.

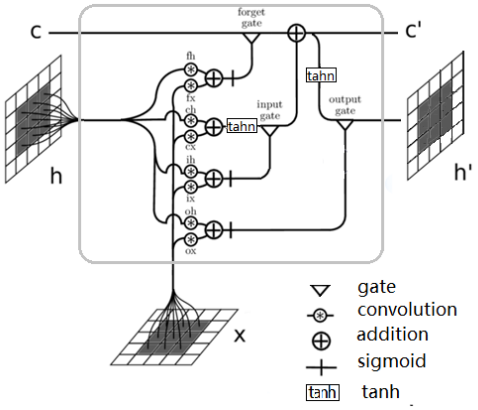

Fig3. Structure of Conv-LSTM.

In this work, the inputs of this part are sequences formed by the output segmentation results from FCN model. Conv_LSTM learn the temporal information of these sequences and output the final segmentation results.

3. EXPERIMENTS

3.1 Experimental settings

The proposed method contains two networks, we first train the FCN model with the extracted images from UAV videos. The sizes of the original images are 3840×2160 pixels or 4096×2160 pixels. To save the GPU memory, each image is clipped to smaller patches, the size of these patches is 2048×1024 pixels. The ground truth images are also clipped to the same size as the corresponding original images. This model is implemented in TensorFlow.

We train the Conv-LSTM model with the sequences formed by the output segmentation results of FCN model. Due to the limitation of the GPU memory, the segmentation results are resized to 512*288 pixels and these sequences are clipped into several blocks. The length of each block is 4 frames. The overlap of the consecutive blocks is 3 frames. The ground truth images are also formed in the same way. Fig. 4 demonstrates the formation of the blocks and the sequence.

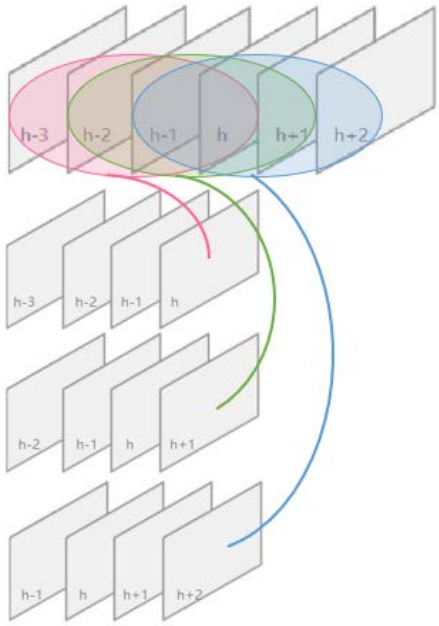

Fig. 4 Formation of blocks and sequence.

The Conv-LSTM model is implemented in Keras and use Tensorflow as backbone.

3.2 Segmentation results

To evaluate the experimental results, Intersection over Union (IoU) score is used. It is defined as follows

$$\text{IoU} = \frac{TP}{TP+FP+FN}$$

where TP denotes *true positive*, FP denotes *false negative* and FN denotes *false negative.*

We calculate the IoU of 5 specific classes and the mean IoU of all classes to compare the segmentation results of the FCN model and the segmentation results of the FCN + Conv-LSTM model. The IoUs are shown in Table 1. The segmentation results of these two models are shown in Fig. 5. The proposed FCN + Conv-LSTM method provides better mean IoU than the FCN method. The IoU of cars decreases 6%. This may be caused by the resize processing on the input of Conv-LSTM model which due to the limitation of computation ability. The IoU of the road increases 12 percent and could provide better visualization in the segmentation.

4. CONCLUSION

In this paper, we proposed the FCN + Conv-LSTM framework for semantic video segmentation. The proposed method improves the overall segmentation results especially on the road class. For future work, we will extend this method to improve its accuracy and overcome the problems caused by the influence of the noisy information from the previous frames.

5. REFERENCE

[1] W.-D. Jang and C.-S. Kim, "Online Video Object Segmentation via Convolutional Trident Network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5849–5858.

[2] M. A. Mohammad, I. Kaloskampis, and Y. Hicks, "New Method for Evaluation of Video Segmentation Quality," in *Proceedings of the 10th International Conference on Computer Vision Theory and Applications*, 2015, pp. 523–530.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[4] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Appl. Soft Comput.*, vol. 70, pp. 41–65, 2018.

[5] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. S. Kruthiventi, and R. V. Babu, "A Taxonomy of Deep Convolutional Neural Nets for Computer Vision," *Front. Robot. AI*, vol. 2, 2016.

[6] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," *Proc. 28th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 802–810.

[7] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1647–1655.

Table 1. IoU scores for different models

Model	Building	Vegetation	Road	Car	Clutter	Mean IoU
FCN	75.6	83.9	37.7	11.9	49.6	51.7
FCN+Conv-LSTM	74.7	82.1	50.0	5.6	50.3	52.6

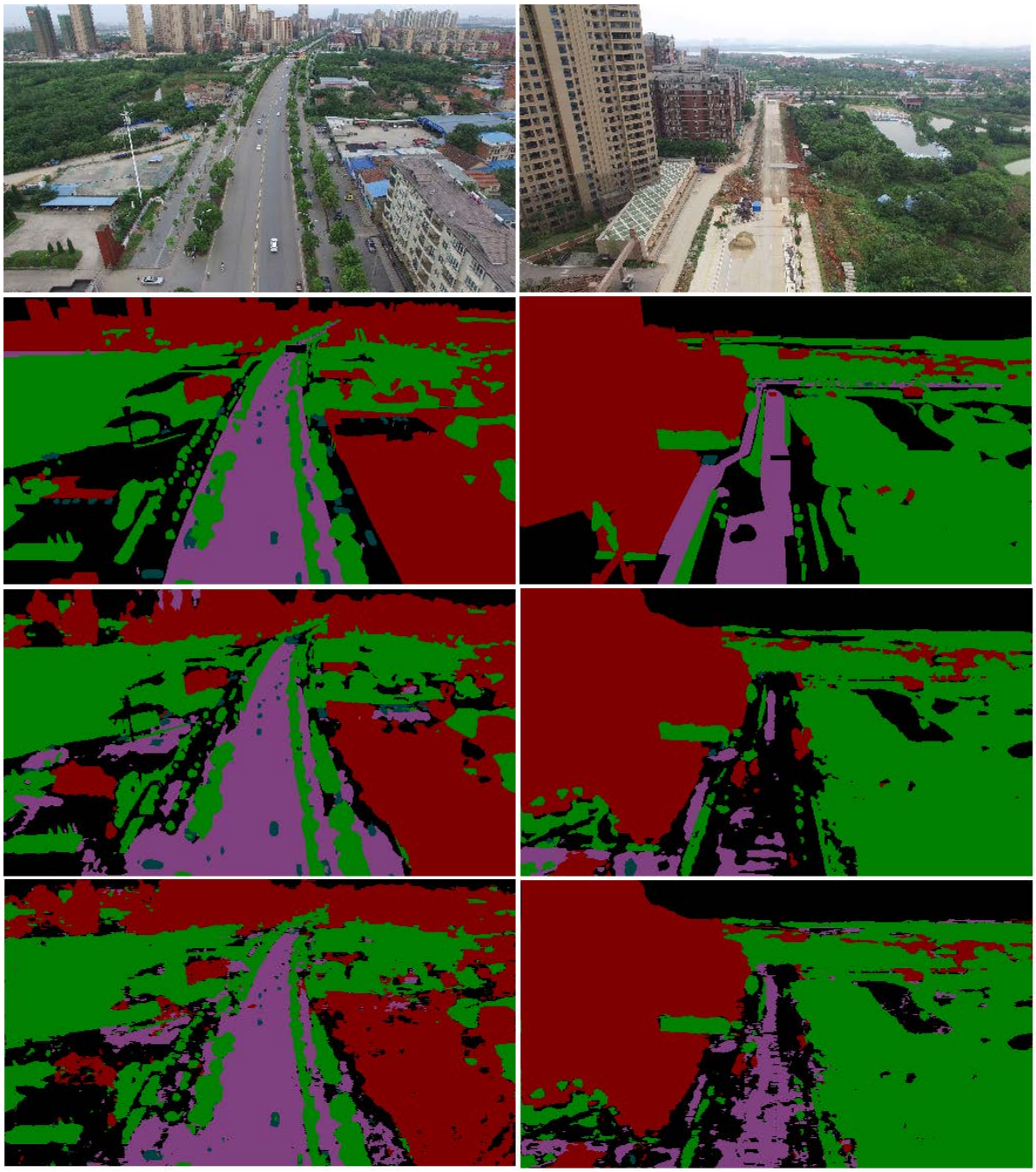

Fig 5. Segmentation results of the different model. The first row shows the original image extracted from the video. The second row shows the corresponding ground truth images. The Third row shows the segmentation results of the FCN model. The fourth row shows the segmentation results of the FCN + Conv-LSTM model.