Moving Object Proposals with Deep Learned Optical Flow for Video Object Segmentation

Ge Shi
Umass Amherst
140 Governors Dr
geshi@umass.cs.edu

Zhili Yang
Umass Amherst
140 Governors Dr
zhiliyang@umass.cs.edu

Abstract

Dynamic scene understanding is one of the most conspicuous field of interest among computer vision community. In order to enhance dynamic scene understanding, pixel-wise segmentation with neural networks is widely accepted. The latest researches on pixel-wise segmentation combined semantic and motion information and produced good performance. In this work, we propose a state of art architecture of neural networks to accurately and efficiently get the moving object proposals (MOP [12]). We first train an unsupervised convolutional neural network (UnFlow [11]) to generate optical flow estimation. Then we render the output of optical flow net to a fully convolutional SegNet model. The main contribution of our work is (1) Fine-tuning the pretrained optical flow model on the brand new DAVIS Dataset [14]; (2) Leveraging fully convolutional neural networks with Encoder-Decoder architecture to segment objects. We developed the codes with TensorFlow, and executed the training and evaluation processes on an AWS EC2 instance.

1. Introduction

The Convolutional Neural Networks (CNN [13]) has achieved great success in learning image features. Through Encoder-Decoder model, the convolutional neural networks can preserve spatial information when extracting general feature maps. This model greatly promoted the performance of convolution neural networks when handling pixel-wise segmentation [1] and object detection [15] tasks. Pixel-wise segmentation [6] is to assign each pixel an class based on its semantic meaning or which object it belongs to.

Video object segmentation [5, 8] is the task of extracting spatial regions corresponding to objects from a sequence of image frames (Figure 1). This is not just a process to apply image segmentation algorithm to all frames in the
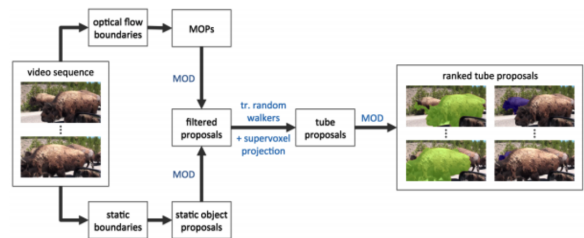

Figure 1. Fusion static object information and motion information to segment images.

video, but the variational information with respect to time base also facilitates better understanding of object independence. The traditional methods solving this problem [22, 3] relies on hand-crafted features without leveraging CNN to learn video representation. Recently, proposal of regions that likely contains independent objects and using CNN currently dominate object detection in static images. Object segmentation [17] is on the basis of proposed regions of interest object. When it comes to segmentation of moving objects in videos, [7] came up with a paradigm for detecting moving objects in videos by introducing moving object proposals (MOP). MOP [12] is a motion based proposal method to generate a set of region proposals in each frame using segmentations on optical flow [19]. To get impressive results in objects segmentation in video, many advanced works [4, 10, 16, 11] unites optical flow boundaries coming from motion information and static boundaries coming from static object proposals.

SegNet [1] is a deep fully convolutional neural network architecture for semantic pixel-wise segmentation, for which the model can be divided into two parts, encoder and decoder (Figure 2). The encoder contains the 13 convolutional layers adapted from VGG16 [18] network, which is commonly used in many segmentation architectures; the decoder uses pooling indices computed in the max-pooling step of the corresponding encoder to perform non-linear up-
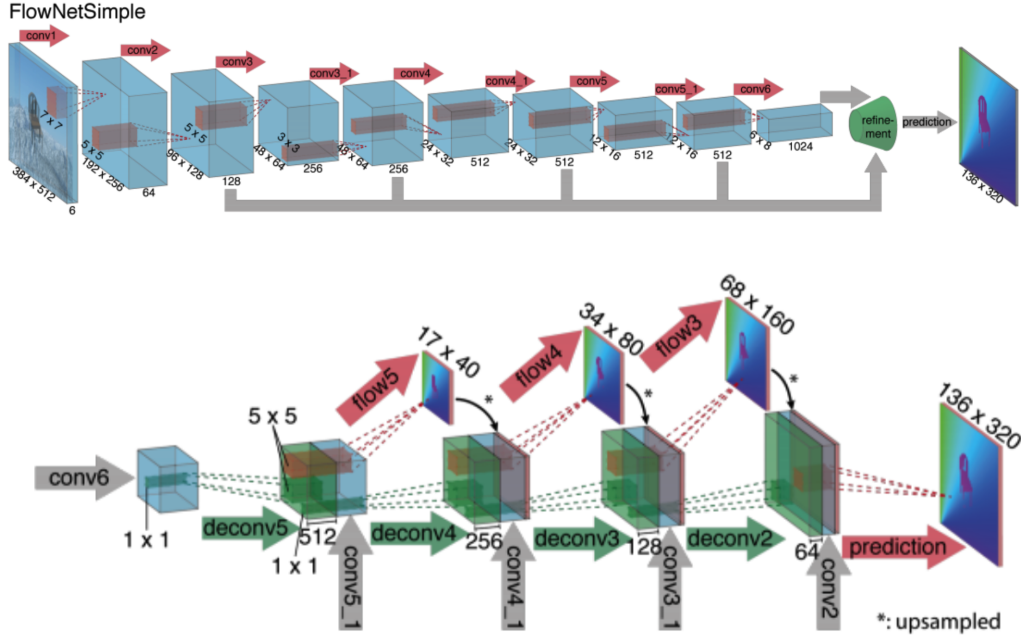
Figure 2. Convolution and deconvolution layers

sampling. In our project, instead of semantic segmentation, we adapted the source codes to achieve moving object segmentation in videos.

Optical flow is another traditional computer vision research area which estimates the relative motion of objects and edges in a visual scene. The algorithms utilizing neural networks to address the task have been invented recently. Researchers explored both innovative supervised and unsupervised structures of optical flow learning networks. FlowNet [4] and UnFlow [11] are representatives of them.

In this work, we decided to develop a neural networks to segment objects in video leveraging unsupervised learning neural networks to train optical flow and feed the outputs of optical flow estimation to into a SegNet model.

We choose DAVIS 2017 [14] as the dataset for training and evaluation, which contains 50 high quality video sequences and so as the corresponding ground truth for evaluation. Each video of the dataset is not limited in one specific object, instead it might contain multiple objects in one frame and each of them are annotated in the ground truth.

2. Related Works

2.1. Methods for Estimating Optical Flow

FlowNet [4]: End-to-end supervised learning of convolutional networks for optical flow was first introduced with FlowNet. The network takes two consecutive input images and outputs a dense optical flow map using an encoder-decoder architecture. For supervised training, a large synthetic dataset was generated from static background images and renderings of animated 3D chairs. A follow up work [10] introduced the more accurate, but also slower and more complex FlowNet2 family of supervised networks.

UnFlow: [11] built on the previous FlowNet-based networks and extend it unsupervised video segmentation settings. First, they designed a symmetric, occlusion-aware loss based on bidirectional (i.e., forward and backward) optical flow estimates. Second, they trained FlowNetC with their comprehensive unsupervised loss to estimate bidirectional flow. Third, they made use of iterative refinement by stacking multiple FlowNet networks. Optionally, they can also use a supervised loss for fine-tuning their networks on sparse ground truth data after unsupervised training.

2.2. Methods for Object Segmentation

PCM [2]: In this work, a likelihood function uses a novel combination of the angle and magnitude of the optical flow to maximize the information about the truth motions of objects is derived. It is used for assessing the probability of an optical flow vector given the 3D motion direction.

MP-Net [20]: MP-Net learns to recognize motion patterns in a flow field. Despite its decent performance, it is limited by its frame-based nature and also overlooks appearance features of objects.

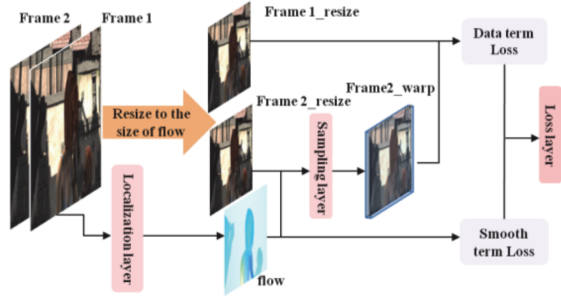**ConvGRU-Net [21]:** ConvGRU-Net is a novel approach

2

Figure 3. Unsupervised learning model based on variational constraint.

for video object segmentation. It combines two complementary sources of information: appearance and motion, with a visual memory module, realized as a bidirectional convolutional gated recurrent unit. The ConvGRU module encodes spatio-temporal evolution of objects in a video and uses this encoding to improve motion segmentation.

3. Technical Approach

Our method to segment moving objects can be separated as two steps: (1) training an unsupervised neural networks to learn optical flow; (2) feeding a SegNet model with the learned output from optical flow networks and generate moving object segmentation. The details are as following.

3.1. Optical Flow

Since suggested [9], most dense approaches estimating optical flow are based on the variational formulation. According to the optical flow constraint, the gray value of a moving pixel stays constant over time, i.e.

$$I(x, y, t) = I(x + u, y + v, t + 1)$$

where (u, v) is the optical flow vector of a pixel (x, y) from time t to time t +1. Using the first-order Taylor approximation, the linearized version of the image function can be derived.

$$I(x, y, t) \approx I(x, y, t + 1) + \nabla I(x, y, t + 1)^T (u, v)$$

$$0 = I(x, y, t + 1) - I(x, y, t) + \nabla I(x, y, t + 1)^T (u, v)$$

The partial derivatives of the image function can be denoted as I_t , I_x ,and I_y. The optical flow constraint equation has inherent dependency on the image position (x,y).

$$0 = I_t + I_x u + I_y v.$$

The smoothness penalty which penalizes the derivative of the optical flow field, was introduced by the variational

methods (Figure 3),

$$\min_{u(x), v(x)} \{ \int_\Omega (|\nabla u(x,y)|^2 + |\nabla v(x,y)|^2) d\Omega +$$

$$\lambda \int_\Omega (I_t + I_x u(x,y) + I_y v(x,y)) d\Omega \}$$

where Ω is the image domain. The improvement of basic variational energy function has been exhaustively studied by researchers. We adopt the the Charbonnier penalty $\Psi(x) = \sqrt{(x^2 + \epsilon^2)}$ and here ϵ is 0.001. Then we have two terms of loss function.

$$l_d = \lambda \int_\Omega \Psi(I_1(x,y) - I_2(x + u(x,y), y + v(x,y))) +$$

$$\Psi(\nabla I_x) + \Psi(\nabla I_y) d\Omega$$

$$l_s = \int_\Omega \Psi(u(x,y) - u(x+1,y)) + \Psi(u(x,y+1) - u(x,y)) +$$

$$\Psi(v(x,y) - v(x+1,y)) + \Psi(v(x,y+1) - v(x,y)) d\Omega$$

l_d take the warped first image and second image as input and minimize the difference. l_s

3.2. Moving Object Proposals

We propose a temporal stable moving object detector using SegNet architecture. As we can see from Figure 4, SegNet is separated into two different parts, encoder and decoder. Basically, the encoder uses the first 13 convolutional layers from the VGG16 networks that designed for object classification. And in the decoder part, each layer has a corresponding layer in the encoder, therefore, it also contains 13 layers as the encoder. At the end, the output of decoder is fed to a multi-class soft-max classifier to produce class probability for each pixel. In our project, since our goal is producing segmentation for moving objects rather than semantic segmentation; after getting the input images from the optical flow model, the SegNet model only needs to consider the foreground objects and the background in each image. Therefore, we only need two classes in soft-max function.
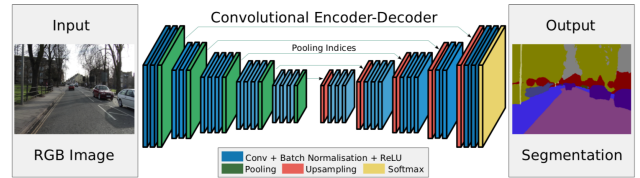

Figure 4. SegNet Architecture

In each encoder layer, it performs convolutional with a filter bank, and followed by batch normalization and ReLu, what's more, max-pooling is also used at the end to achieve

translation invariance over small spatial shifts in the input image. However, max-pooling causes boundary detail loses. As a solution, the SegNet model stores the max-pooling indices for each encoder layer, and in each decoder layer, it can use corresponding indices for upsampling.

In the training procedure, we train all the variants with Adam Optimizer and the corss-entropy loss function.

3.3. Expected results and Evaluation

We decide to evaluate our results by applying the intersection over union overlap (IoU) between the final results and the ground truth. We can compute IoU according to the formula below,

$$IoU = A \cap B / A \cup B$$

where A denotes the area covered on the expected outputs and B denotes the area annotated on ground truth. According to the formula, IoU simply means the proportion of the overlap area within the union area of the model's output and the corresponding ground truth.

Baselines: For the evaluation, we decide to compare our method with an advanced methods **PCM** [2].

4. Experiment

Our model consists of two parts: (1) Unflow net (Unsupervised Learning of Optical Flow); (2) SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. In our implementation, we combined the two models, and modified the source codes for training the dataset DAVIS 2017 [14]. The evaluation details are as following.

4.1. Unflow Experiment

In our implementation, we modified the input and training codes to import the DAVIS 2017 as our training and evaluation dataset. In the training process, only first two images in each classes (50 classes) were chosen to be trained, the training iterations for each class was set as 250.

4.1.1 Pre-processing

First, we divide the images in each class into a pair of lists, each pair of images are consecutive frames, therefore, they could be stacked and fed as inputs into the model. What's more, all images need to be reshaped to 448×832, then applied normalization.

4.1.2 Early attempts and failures

At the beginning, we decided to train our modified model from scratch. The *learning rate* was $0.25e^-5$, and the *save interval* = 250, so that a checkpoint would be generated in each 250 iterations. Figure 6. displays the losses along with the iterations. As we can see, the training loss dropped dramatically at the beginning, but it stopped reducing at 400. In the middle of the training process, we tried to reduce the *learning rate* to $0.25e^-6$ and even smaller, but the training loss still could not reduce. It might be due to the other hyperparameters that take significant roles in the model. But we did not get enough time and resource to have a further test.

4.1.3 Successful Approach

Instead of training everything from scratch, we finally decided to move to a pretrained model that has been finetuned. After modifying the codes of evaluation, we successfully adapted our DAVIS 2017 into the pretrained model.

4.1.4 Results Evaluation

Although, the pretrained model could provide us with relatively better outputs than our original model, the outputs still consist a lot of glitches that might affect the final results of the whole model. We summarize some common errors occur among those outputs in Figure 7.

Obstacles: As we can see from the example *Dog agility* in Figure 7, the unflow outputs includes the pillars as moving objects, which might make sense because while the dog is crossing those pillars, it might also cause the pillars keep vibrating in different frames. As a result, those vibrating pillars are also detected.

Reflections: Reflections are another common features in images, for example, in *flamingo*, the reflection of the flamingo on the water is also included in the output.

Edges: It is easy for the model to recognize the body of an object, but detecting edge's information always be a big problem for the model. We can find this problem almost in all examples in Figure 7, such as the missing arms and legs in *dance twirl* and *tennis*, the missing wheels in *bus* and *car shadow*.

4.2. SegNet Experiment

4.2.1 Training Details

We trained the SegNet with Adam Optimizer for 10000 iterations. $\beta1 = 0.9$, $\beta2 = 0.999$, the start learning rate is $5e^-4$, with a decay of 0.5 times every 2000 global steps. We fed the optical flow estimation from Unflow and ground truth segmentation into SegNet and the get the moving object segmentation.

4.2.2 Qualitative Analysis

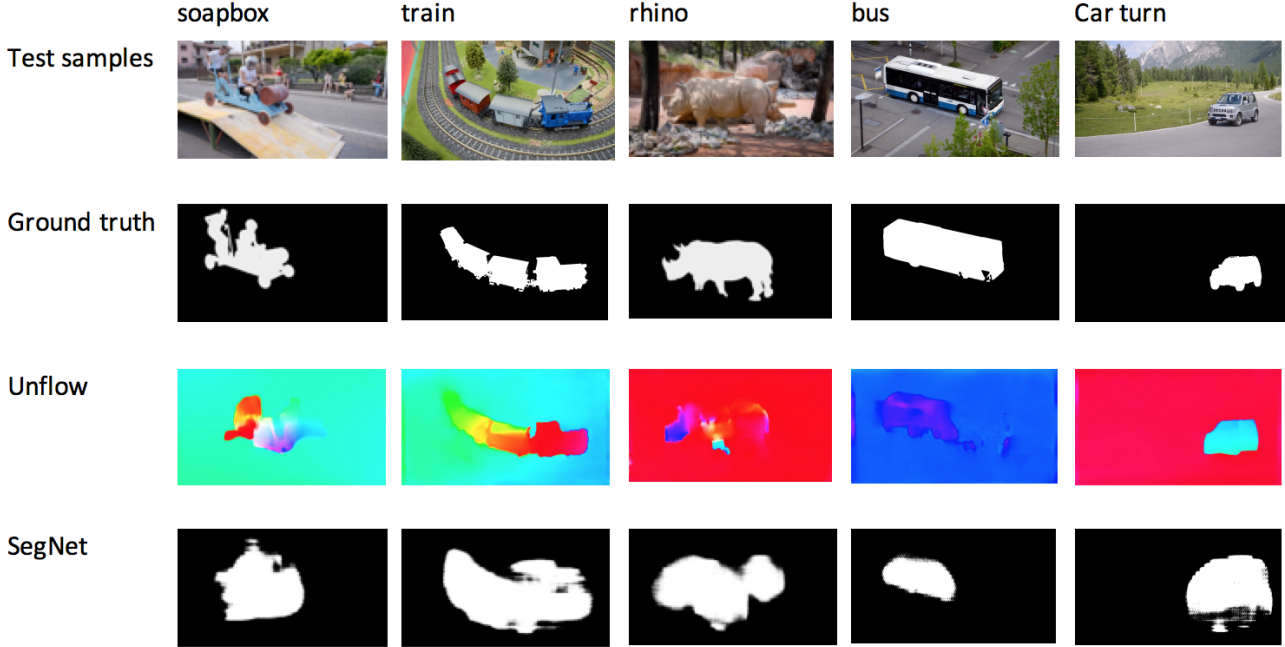We evaluate our results on the evaluation dataset of DAVIS for the lack of test data. The Figure 8 shows the loss de-

4

	soapbox	train	rhino	bus	Car turn
Test samples					
Ground truth					
Unflow					
SegNet					

Figure 5. SegNet and Unflow predictions of the samples from DAVIS 2017

Figure 6. Unflow training loss

Measure	PCM	CVOS	KEY	NLC	FST	Ours
Mean IOU	40.1	48.2	49.8	55.1	55.8	41.9

Table 1. Our method compared to other state-of-the-art methods on DAVIS.
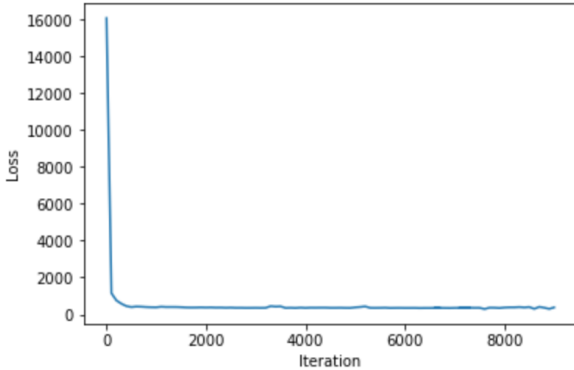
creasing with global steps growing. We compared out approach with Table 1 compares our approach to the state-of-the-art methods on DAVIS. The measure to evaluate results is mean IOU. Because our approach only consider the motion information without semantic information, the result of ours can only be a little better than PCM. When it is put on the test data, we assume it would be worse to some extent.

4.2.3 Quantitative Analysis

Based on the results of our experiment, we found our networks is more suitable for segment monolithic moving objects with gentle motion. This come from the intrinsic properties of Unflow networks. Also, we found in training process, the segmentation boundaries of moving object tends to shrink with the increasing of training steps.

5. Conclusion

5.1. Insights

This paper introduces a novel approach for video moving object segmentation. In general, we tried to combine Unflow and SegNet to get the motion segmentation and the basic idea is working. However, limited by time and resources, we cannot train the networks with a lot more global steps and elaborate parameters. In this work, we leveraged the state-of-the-art unsupervised optical flow estimation approach and fine-tuned it on the novel DAVIS 2017 dataset in which we learned the constrained optical flow formula with smoothness penalty. Besides, we studied the Encoder-Decoder architecture and the advantages of up-sampling and up-convolutional layers.
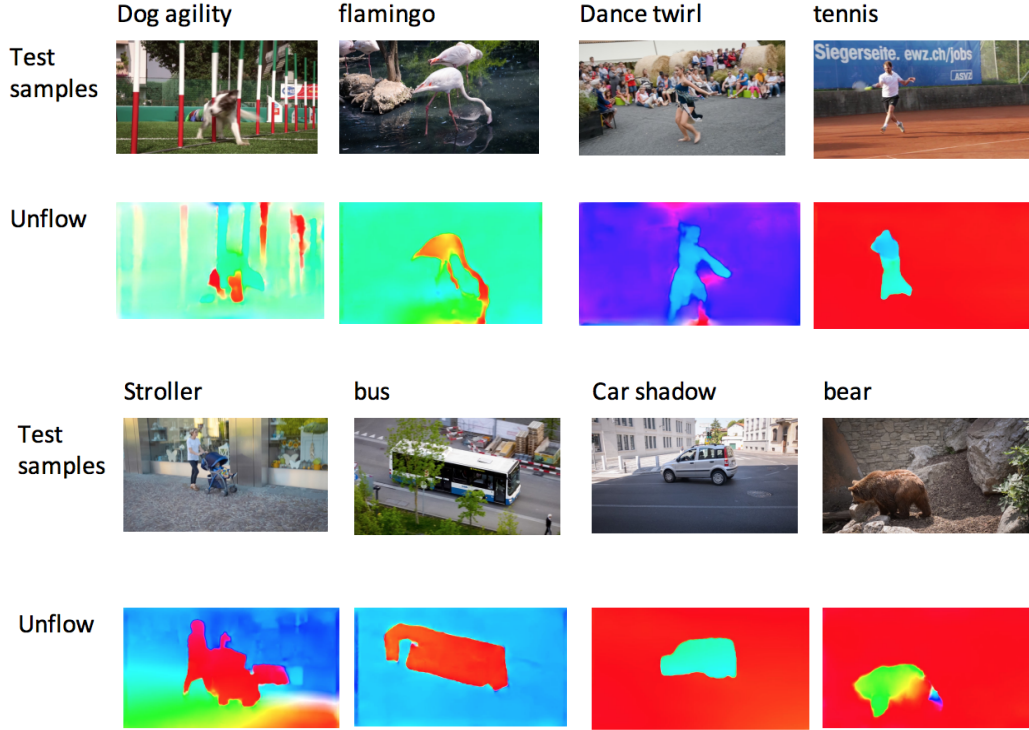
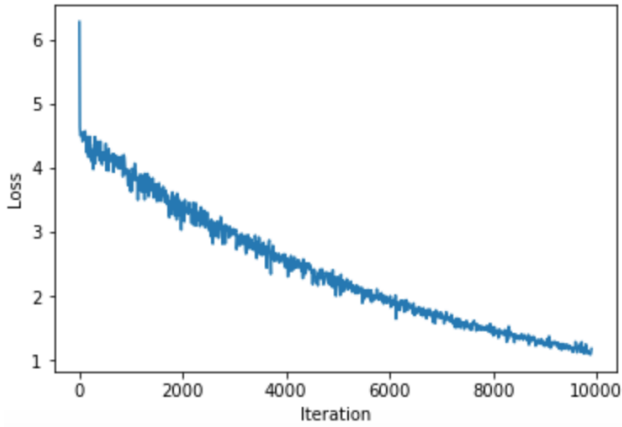Figure 7. The predicted outputs generated by fine-tuned model of Unflow

Figure 8. SegNet training loss

5.2. Future Improvement

The analysis of results inspires us to make improvement in the flowing ways.

5.2.1 Adding LSTM Model

Normal convolutional neural networks deal with independent images in an impressive way. However, in sequential video learning tasks, detached CNN cannot preserve the learned features from the previous frames. To solve this problem, ConvLSTM is developed with the motivation to unify CNN and LSTM in one model. ConvLSTM is one kind of LSTM which is specially designed to aid images sequence learning with convolutional layer as input processing cell. The most recent research adopted this model in video object segmentation and outperforms previous work. The convLSTM layer can be inserted between the Encoder and Decoder layer.

5.2.2 Involving Semantic Information

Currently, our work feeds the output of Unflow networks to SegNet, but the two network architecture can be simplified by unifying those two. We assume the input image can be rendered into two branches of local network. One is the Encoder part of semantic networks and the other is the dilated convolution layers of Unflow. Than take both tensors into the Decoder architecture. At the loss end, optical flow unsupervised loss and segmentation supervised loss are both considered. In this process, not only the networks can be simplified by removing redundancy but also the semantic information is taken into consider.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[2] P. Bideau and E. Learned-Miller. Its moving! a probabilistic model for causal motion segmentation in moving camera videos. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 433–449. Springer, 2016.

[3] H.-D. Cheng, X. H. Jiang, Y. Sun, and J. Wang. Color image segmentation: advances and prospects. *Pattern recognition*, 34(12):2259–2281, 2001.

[4] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. Van Der Smagt, D. Cremers, and T. Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[5] S. Dutt Jain, B. Xiong, and K. Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3664–3673, 2017.

[6] M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Fathy, R. Klette, and F. Huang. Stfcn: spatio-temporal fcn for semantic video segmentation. *arXiv preprint arXiv:1608.05971*, 2016.

[7] K. Fragkiadaki, P. Arbelaez, P. Felsen, and J. Malik. Learning to segment moving objects in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4083–4090, 2015.

[8] N. Haque, N. D. Reddy, and K. M. Krishna. Joint semantic and motion segmentation for dynamic scenes using deep convolutional networks. *arXiv preprint arXiv:1704.08331*, 2017.

[9] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[10] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[11] S. Meister, J. Hur, and S. Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[12] P. O. O Pinheiro, R. Collobert, and P. Dollár. Learning to segment object candidates. *Advances in neural information processing systems*, 28, 2015.

[13] K. O'Shea and R. Nash. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*, 2015.

[14] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[16] Z. Ren, J. Yan, B. Ni, B. Liu, X. Yang, and H. Zha. Unsupervised deep learning for optical flow estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[17] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[19] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 2432–2439. IEEE, 2010.

[20] P. Tokmakov, K. Alahari, and C. Schmid. Learning motion patterns in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3386–3394, 2017.

[21] P. Tokmakov, K. Alahari, and C. Schmid. Learning video object segmentation with visual memory. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4481–4490, 2017.

[22] Y. J. Zhang. A survey on evaluation methods for image segmentation. *Pattern recognition*, 29(8):1335–1346, 1996.