Weakly Supervised Semantic Segmentation Learning on UAV Video Sequences

Bianca-Cerasela-Zelia Blaga, Sergiu Nedevschi
Computer Science Department
Technical University of Cluj-Napoca
Cluj-Napoca, Romania
Email: {Zelia.Blaga, Sergiu.Nedevschi}@cs.utcluj.ro

***Abstract*—The domain of scene understanding from Unmanned Aerial Vehicles (UAVs) is of high interest for researchers in the computer vision domain, since it can be used for object detection and tracking in scenarios like deforestation monitoring, traffic surveillance, or for civil engineering tasks. However, the topic of dense video segmentation from drones has been insufficiently explored due to the lack of annotated ground truth data. We propose a solution based on a framework composed of a deep neural network for semantic segmentation and an optical flow generator, linked together by a spatio-temporal GRU component to efficiently solve the problem of weakly supervised semantic segmentation of video sequences recorded from UAVs. The novelty of our work comes from the employment of depthwise separable convolutions for the GRU component, which decrease the computation time and increase the segmentation accuracy. We test our methodology on the synthetic dataset Mid-Air, for low-altitude drone flight, and report results that prove the usefulness of the proposed system.**

***Keywords*—** *video semantic segmentation, weakly supervised learning, optical flow, unmanned aerial vehicles.*

I. Introduction

Unmanned Aerial Vehicles (UAVs) have seen a rise in popularity since the latest developments in the field, which allow them to perform a large range of tasks, such as aerial photography, surveying, and mapping. Many applications based on them were implemented to solve problems ranging from agriculture to civil engineering. For example, aerial photography was used for the ecological management of different tree species [1], and for monitoring tree heights in forestry [2]. Additionally, drones are employed to perform power line inspection, to maintain the reliability, availability, and sustainability of electricity supply, as presented in [3]. Using camera sensors, faults in power lines can be identified, like trees obstructing the lines, broken poles, or missing parts. UAVs can also be a first-response platform in the case of disaster management, for tasks like prediction, assessment, response, and recovery [4].

For the task of forest monitoring needed to assess the degree of deforestation in time, annotated ground truth data such as depth maps and semantic labels are needed. However, these are not always available since manual annotation is difficult and time-consuming. Solutions to this bottleneck can be represented by synthetic datasets that are obtained from game engine simulators [5], and by methods that rely on semi, self or unsupervised learning [6]. From these, we focus our attention on synthetic datasets and weakly supervised learning.

Furthermore, there is an increasing need for efficient algorithms that can solve tasks such as object detection, depth estimation, or autonomous navigation for UAV systems. We are particularly interested in the task of semantic segmentation from drones. In our previous work, we have addressed these issues by analyzing the performance of deep learning frameworks on various types of data collected from virtual engines or the real world [7]. However, the main drawback of the studied methodologies was the lack of spatial and temporal constraints which are needed since the input comes from video sequences that carry overlapping information between frames.

In this paper, we propose an improved methodology based on a recurrent deep learning framework that takes the semantic segmentation from the previous frame, uses the optical flow to warp it to the current timestamp, and feeds it together with the previous and current color images to a spatio-temporal GRU component that corrects and refines the semantic segmentation prediction. We approach this from a weakly supervised learning perspective and perform several experiments that allow us to obtain the optimal quantity of necessary ground truth data necessary for training. Additionally, we improve the GRU component by introducing depthwise separable convolutions, which aims at improving the accuracy and the speed needed for accurate video sequence labeling. In Section II, we present some recent developments in the fields of semantic segmentation and weakly supervised learning related to drone applications, followed by the presentation of the network components in Section III. The results are presented and discussed in Section IV.

II. Related Work

To solve tasks such as object detection and tracking, visual localization and mapping, and autonomous navigation and control, drones rely on sensor fusion between cameras, positioning systems, and laser scanners. For example, a 10-layer network is used in [8] for scene classification from videos recorded using a quadrotor, the end goal being autonomous forest trail navigation. The system learns to classify the input image into one of three actions: turn left, go straight, and turn right, which is further processed by a controller that determines the correct angle for navigation.

A topic worth exploring is related to spatio-temporal based methods, where information from previous frames is merged with the current segmentation in order to correct any mislabelings. In this area, methods rely on Recurrent Neural Networks (RNN) like Long short-term memory (LSTM) and Gated Recurrent Units (GRU). One such method is Spatio-Temporal Fully Convolutional Network (STFCN) [9], which uses LSTM to define temporal features, while spatial maps are used to infer future information. Similar approaches can be found in [10], [11], which enhance the networks by providing optical flow information, and by taking into account multiple views or even 3D data to better model pixel dependencies.

To deal with the redundancy encountered in videos, the authors of [12] introduce LERNet (Light, Efficient and Real-time network), based on feature propagation and holistic attention. The network is an encoder-decoder with residual connections, while the Temporal Holistic Attention (THA) module computes spatial correlations for consecutive frames. The experimental results showcase improvements in the network's prediction accuracy as compared to other state-of-the-art methods, through an increased class consistency maintaned over consecutive frames.

Self-attention and bi-directional GRU modules are combined in [13], to improve the results of semantic segmentation in dynamic scenes. The method relies on two parallel bi-directional GRUs, for the horizontal and vertical directions, in the context of the same image. Additionally, the self-attention component enriches the feature information extracted from a ResNet module. Spatial neural-attention for image captioning is presented in [14], where object regions can be identified and localized in two complementary attention maps. One focuses on generating pseudo-annotations for weakly-supervised semantic segmentation, and the other one predicts the discriminative parts.

A dual temporal memory network is proposed in [15], which is composed of two sub-networks: short-term memory - for modeling fine-grained spatio-temporal correlations between neighboring frames in a video based on a graph-based network, and long-term memory - for improving the segmentation by using a simplified GRU, thus accounting for occlusions and drift errors. To exploit prior network knowledge, the authors of [16] propose an end-to-end network that generates masks to identify semantic regions and then expands on the information they provide to classify and segment frames.

The work presented in [17] uses an RNN model with a spatial transformer component to improve the semantic segmentation prediction on video sequences. The segmentation labels and color images are warped according to the optical flow, to propagate the information to the next frame. Based on these, the $STGRU$ component can improve the prediction of the network. An end-to-end framework based on this method was presented in [18], which proves the best performance in an ideal case scenario, when the ground truth optical flow is provided. The system is able to obtain the same segmentation accuracy when using only 25% of the available data, compared to a static segmentation network.

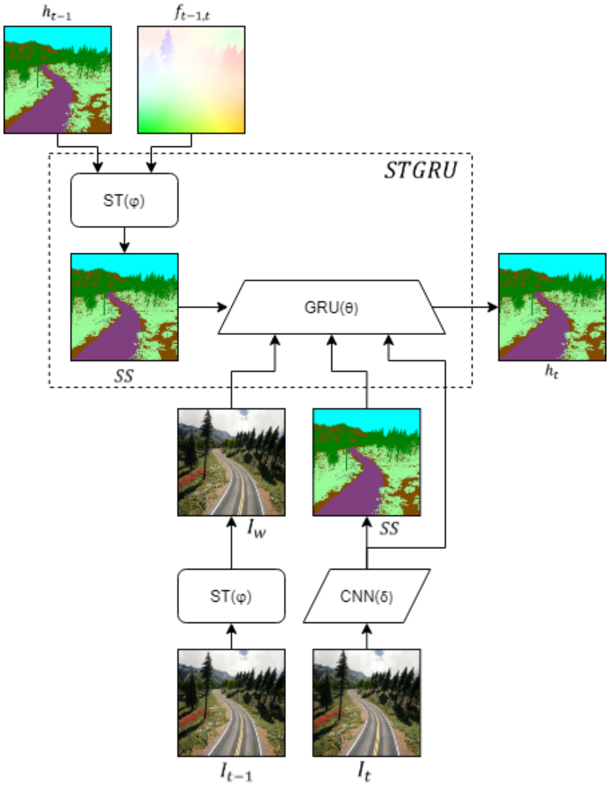

Fig. 1: Framework flow of the proposed methodology, where the input RGB frames together with the semantic segmentation and optical flow are fed to the $STGRU$ component to propagate and refine the labeling.

III. Methodology

We use a framework composed of two networks: ERFNet [19] - for the task of semantic segmentation, and PWC-Net [20] - to obtain the optical flow information, which extends on the works presented in [17] and [18]. These two networks are linked together by a spatio-temporal GRU component that aims to refine the labeling. The logical flow of the architecture can be seen in Fig. 1, where $f_{\mathrm{t-1,t}}$ is the optical flow, $I_{\mathrm{t-1}}$ and I_{t} are the color images from two consecutive timestamps, I_{w} is the warped RGB image, SS - the semantic segmentation, $h_{\mathrm{t-1}}$ and h_{t} - GRU input and ouput, and the operations of spatio-temporal warping - $ST(\varphi)$, semantic segmentation - $CNN(\delta)$ and GRU - $GRU(\theta)$.

ERFNet [19] is an encoder-decoder network with residual connections which takes as input the color image and outputs a labeled image in which each pixel corresponds to a class. For the optical flow module, we chose PWC-Net [20] which is based on pyramidal feature warping. It is a network that obtains more accurate results as compared to previously used architectures such as FlowNet2 [21]. Since this step was proven to be time-consuming, we only use this network to generate the ground truth for the weakly supervised setting. When training the network end-to-end, we employ VCN [22] which is a more lightweight model. We use the weights made available by the authors, which were obtained after training the model on FlyingChairs [23] and FlyingThings [24] datasets and finetuned on the real KITTI dataset [25].

The most important component of the framework is called $STGRU$ and its modules are ST - a Spatio-Temporal transformer, and GRU - a Gated Recurrent Unit.

The first operation is a bilinear warping that maps the elements of a two-dimensional matrix x_{ij} to y_{ij} according to the optical flow using the formula:

$$y_{\mathrm{ij}} = \sum_{m,n} x_{\mathrm{mn}} k(i + f^{y}_{\mathrm{ij}} - m, j + f^{x}_{\mathrm{ij}} - n) \tag{1}$$

where the optical flow vector at the pixel location (i, j) is given by $(f^{x}_{\mathrm{ij}}, f^{y}_{\mathrm{ij}})$, and $k(x, y)$ is the bilinear interpolation kernel:

$$k(x, y) = max(0, 1 - |x|)max(0, 1 - |y|) \tag{2}$$

We denote the bilinear operation with $\phi_{\mathrm{t-1,t}}(\bullet)$. This operation is employed to obtain a prediction from the previous color image $\phi_{\mathrm{t-1,t}}(I_{\mathrm{t-1}})$ or the previous semantic segmentation map $\phi_{\mathrm{t-1,t}}(h_{\mathrm{t-1}})$, which are further given as inputs to the next component to refine the current semantic segmentation.

The second module is a modified GRU with depthwise separable convolutions [26], which consider that the depth component (number of channels) and the spatial dimension (width and height) can be separated, thus the name separable. They are composed of two parts: a depthwise convolution - spatial convolution performed independently over each channel of the input, and pointwise convolution - a 1x1 convolution, projecting the previous outputs onto a new feature space. The key idea for this implementation stands behind the removal of non-linearity, thus making it a linear mapping function, which was proven to provide faster convergence, as well as a more accurate prediction. Each convolution change within the reset gate r_t, update gate z_t, and internal memory update on the current semantic segmentation x_{t} can be noticed in the following equations:

$$w_t = \phi_{\mathrm{t-1,t}}(h_{\mathrm{t-1}}) \tag{3}$$

$$r_t = 1 - tanh(|X^{1\times1}_{\mathrm{ir}} * (X_{\mathrm{ir}} \circledast (I_t - \phi_{\mathrm{t-1,t}}(I_{\mathrm{t-1}}))) + b_r|) \tag{4}$$

$$\tilde{h}_t = X^{1\times1}_{\mathrm{xh}} * (X_{\mathrm{xh}} \circledast x_t) + X^{1\times1}_{\mathrm{hh}} * (X_{\mathrm{hh}} \circledast (r_t \odot w_t)) \tag{5}$$

$$z_t = \sigma(X^{1\times1}_{\mathrm{zz}} * (X_{\mathrm{zz}} \circledast x_t) + X^{1\times1}_{\mathrm{hz}} * (X_{\mathrm{hz}} \circledast (r_t \odot w_t)) + b_z) \tag{6}$$

$$h_t = softmax(\lambda(1 - z_t) \odot r_t \odot w_t + z_t \odot \tilde{h}_t) \tag{7}$$

where $*$ denotes the convolution operation, $\circledast$ represents the depthwise convolution, and $\odot$ is the elementwise multiplication operator. W and b are the weights and biases which are learned by the network. The value of the constant is $\lambda = 2$ because $\tilde{h}_t$ is the sum between two images - the input semantic segmentation and the warped labeling map. The final output of the framework is the improved segmentation h_{t}.

We were inspired by the work presented in [27], where depthwise separable convolutions proved to significantly decrease the number of required FLOPs (floating-point operations) in the case of LSTM components. The authors reported that the new implementation takes only 12.1% of the computational cost of standard LSTM modules, therefore the same performance is expected for GRU. Thus we bring this change inside the $STGRU$ module, and in the next section, we present details of the experiments we carried.

IV. Experimental Results

To train and test the network, we use Mid-Air [28], which is a large synthetic dataset of realistic images recorded from a drone perspective. We use trajectories recorded in spring and fall, in three weather conditions - cloudy, sunny, and sunset. There are 11 classes used for training: sky, trees, dirt ground, ground vegetation, rocky ground, boulders, water plane, road, train track, road sign, and man-made objects. We divide the dataset into 80% training data and 20% for testing and validation purposes. To assess the performance of the framework, we carry out several experiments that sample the training space at various frame rates denoted with k.

A. Semantic Segmentation Results using $STGRU$ and Ground Truth Optical Flow

In Table I we present the results obtained when using the ground truth optical flow generated by PWC-Net. ERFNet(k) represents the semantic segmentation network, while GRU(k) represents the full framework trained using the previously obtained segmentation and the ground truth optical flow, both trained using every k-th label. We report the Intersection over Union (IoU) results, which is computed using the number of true positives (TP) - the model correctly predicts the class, false positives (FP) - an outcome where the model incorrectly predicts the positive class, true negatives (TN) - the model

TABLE I: IoU results on the Mid-Air test set, where GRU(k) represents the full system, trained with every k-th label, using the ground truh optical flow and the semantic segmentation from ERFNet.

Class	ERFNet(4)	GRU(4)	ERFNet(8)	GRU(8)	ERFNet(16)	GRU(16)	ERFNet(32)	GRU(32)	ERFNet(64)	GRU(64)
Sky	93.58	95.39	92.34	95.75	90.43	94.91	89.55	93.47	87.63	91.56
Trees	86.65	89.88	85.23	89.63	83.56	84.06	83.16	83.39	78.96	82.84
Dirt ground	75.48	81.13	70.73	75.48	68.20	72.23	67.52	73.70	65.83	67.16
Ground vegetation	83.16	85.15	79.61	80.39	77.91	84.38	76.22	82.10	73.54	74.40
Rocky ground	68.35	70.74	67.17	68.91	66.72	68.22	66.23	71.71	60.47	62.87
Boulders	67.56	71.66	65.74	68.84	65.30	69.52	65.19	69.98	58.26	60.82
Water plane	86.36	95.28	86.27	88.25	83.33	86.43	81.81	83.58	81.27	85.17
Road	89.38	93.03	86.22	87.72	83.54	87.52	81.33	82.83	79.61	81.12
Train track	81.48	87.90	73.92	74.46	71.80	75.78	70.59	74.54	68.53	69.82
Road sign	48.14	53.58	36.54	41.09	35.18	38.03	34.37	36.52	34.02	35.74
Man-made objects	57.71	65.35	55.41	58.98	54.73	63.60	53.70	56.87	50.84	51.66
Mean IoU	76.17	80.83	72.65	75.41	70.97	74.97	69.97	73.52	67.18	69.38

TABLE II: IoU results on the Mid-Air test set, where GRU(f,k) represents the full system, trained with every k-th label, f is using either the ground truth (GT) optical flow from PWC-Net or the one generated from the retrained VCN.

Class	ERFNet(8)	GRU(GT, 8)	GRU(VCN, 8)
Sky	92.34	92.50	92.67
Trees	85.23	89.45	89.75
Dirt ground	70.73	73.49	73.19
Ground vegetation	79.61	81.58	83.47
Rocky ground	67.17	68.85	72.35
Boulders	65.74	70.61	71.82
Water plane	86.27	88.47	88.82
Road	86.22	90.45	91.21
Train track	73.92	76.89	76.45
Road sign	36.54	39.02	40.83
Man-made objects	55.41	59.58	59.45
Mean IoU	72.65	75.54	76.36

TABLE III: Evaluation of runtime (ms), mean IoU and F1 scores for the three network modules: semantic segmentation, spatio-temporal component, and optical flow. The results are reported for k=4.

Task	Network	Runtime (ms)	mIoU	F1
Semantic Segmentation	ERFNet	7.5	76.17	-
	HRNet	13	80.13	-
Spatio-temporal propagation	LSTM	11	78.18	-
	GRU	8	80.27	-
	Our GRU	6	80.83	-
Optical Flow	PWC-Net	130	-	7.72
	VCN	88	-	6.3

correctly predicts the negative class, and false negatives (FN) - model incorrectly predicts the negative class, which give the following metric formula:

$$IoU = \frac{TP}{TP + FP + FN} \tag{8}$$

The performance of the network decreases with the sampling rate, as expected. However, we notice that using the $STGRU$ components results in an improvement of the segmentation results, no matter which percentage of the dataset we are using. The network's predictions improve for each class, since the system exploits both the spatial and temporal information needed when amounts of annotated data are missing from the training set.

Our proposed framework performs better when using a higher sampling rate than when using a lower sampling rate and a static segmentation network. For example, comparable $mIoU$ results can be noticed between $GRU(8)$ and $ERFNet(4)$ which means that by using the $STGRU$ module, one can obtain accurate results on sparse data.

B. Results of End-to-end Training

To further test the framework, we compare the results obtained by the end-to-end framework using the two optical flow networks, as reported in Table II. When optimizing both the semantic segmentation and the optical flow networks, this leads to an improvement of the segmentation results, meaning that an improved optical flow can aid in the task of scene understanding. We notice an improvement in the segmentation results for classes like ground vegetation, rocky ground, boulders, and road sign. Overall, the end-to-end network performs better by 4% as compared to the static segmentation, at the cost of an increased computational time.

Qualitative experimental results obtained on the test set are showcased in Fig. 2. We notice that the scenarios vary from forest to road areas, containing a variety of textures and day time conditions. The most difficult categories to segment were boulders and traffic signs, since these objects are less present in the training dataset.

C. Ablation Study

We carried out several experiments to test the runtime and accuracy of different network components, which we present in Table III. We use images of size 512 × 512, on an NVIDIA GeForce GTX 1080Ti GPU. For the static segmentation component, we took into consideration ERFNet and HRNet [29]. Even though the mean IoU is better for the second network, its runtime is higher since the model also more parameters.

Looking at the runtime for various RNN components, we notice that our implementation of GRU with depthwise separable convolutions performs better than the original implementation of LSTM and GRU, both in terms of runtime speed and segmentation accuracy.

To evaluate the optical flow networks' performance, we use the $F1$ metric, which can be computed as follows:

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{11}$$

We notice that VCN performs better than PWC-Net, both in terms of runtime and accuracy. The advantages VCN brings over other similar networks are faster convergence and improved generalization capabilities. Therefore it is more suited for an end-to-end training scenario.

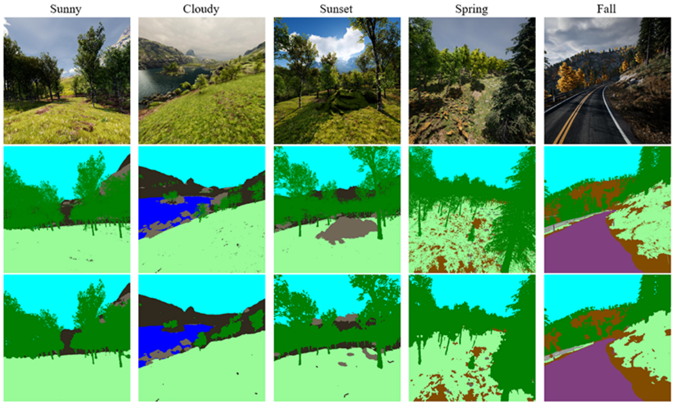

Fig. 2: Results of the framework on the test set, presented for the 5 different scenarios from Mid-Air.

V. Conclusions

In this paper, we presented a weakly supervised network which performs semantic segmentation on video sequences. The framework uses the recordings from the camera and the optical flow between frames to sequentially propagate and continuously improve the prediction, thus accounting for spatial and temporal constraints. We improve the GRU component by using depthwise separable convolutions, and test the system on a realistic synthetic dataset for a low-altitude drone flight scenario. Our work managed to successfully obtain pixel-level semantic segmentation for video recordings, and is more accurate compared to static networks when using lower percentages of training data.

In the future, we aim to enhance the framework by introducing a more performant semantic segmentation module and a less computationally expensive optical flow network, which will be trained and tested on real-world drone recordings. These improvements would lead to an increase in the segmentation accuracy, obtained in real-time, which needs fewer ground truth annotations.

Acknowledgment

This work was supported by the "SEPCA-Integrated Semantic Visual Perception and Control for Autonomous Systems" grant funded by the Romanian Ministry of Education and Research, code PN-III-P4-ID-PCCF-2016-0180.

References

[1] J. L. Morgan, S. E. Gergel, and N. C. Coops, "Aerial photography: A rapidly evolving tool for ecological management," *BioScience*, vol. 60, no. 1, pp. 47–59, 2010.

[2] J. C. Suárez, C. Ontiveros, S. Smith, and S. Snape, "Use of airborne LiDAR and aerial photography in the estimation of individual tree heights in forestry," *Computers & Geosciences*, vol. 31, no. 2, pp. 253–262, 2005.

[3] V. N. Nguyen, R. Jenssen, and D. Roverso, "Automatic autonomous vision-based power line inspection: A review of current status and the potential role of deep learning," *International Journal of Electrical Power & Energy Systems*, vol. 99, pp. 107–120, 2018.

[4] M. Erdelj, E. Natalizio, K. R. Chowdhury, and I. F. Akyildiz, "Help from the sky: Leveraging UAVs for disaster management," *IEEE Pervasive Computing*, vol. 16, no. 1, pp. 24–32, 2017.

[5] B.-C.-Z. Blaga and S. Nedevschi, "Semantic segmentation learning for autonomous uavs using simulators and real data," in *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2019, pp. 303–310.

[6] L. Schmarje, M. Santarossa, S.-M. Schröder, and R. Koch, "A survey on semi-, self-and unsupervised learning for image classification."

[7] B.-C.-Z. Blaga and S. Nedevschi, "Exploring deep learning solutions for scene perception from UAVs using simulators and real data," in *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2020, pp. 353–360.

[8] A. Giusti, J. Guzzi, D. C. Cireşan, F.-L. He, J. P. Rodríguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. Di Caro *et al.*, "A machine learning approach to visual perception of forest trails for mobile robots," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 661–667, 2015.

[9] M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Fathy, F. Huang, and R. Klette, "STFCN: spatio-temporal fully convolutional neural network for semantic segmentation of street scenes," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 493–509.

[10] Y. He, W.-C. Chiu, M. Keuper, and M. Fritz, "STD2P: RGBD semantic segmentation using spatio-temporal data-driven pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4837–4846.

[11] Z. Qiu, T. Yao, and T. Mei, "Learning deep spatio-temporal dependence for semantic video segmentation," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 939–949, 2017.

[12] J. Wu, Z. Wen, S. Zhao, and K. Huang, "Video semantic segmentation via feature propagation with holistic attention," *Pattern Recognition*, p. 107268, 2020.

[13] M. Yan, J. Wang, J. Li, K. Zhang, and Z. Yang, "Traffic scene semantic segmentation using self-attention mechanism and bi-directional GRU to correlate context," *Neurocomputing*, vol. 386, pp. 293–304, 2020.

[14] T. Zhang, G. Lin, J. Cai, T. Shen, C. Shen, and A. C. Kot, "Decoupled spatial neural attention for weakly supervised semantic segmentation," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2930–2941, 2019.

[15] K. Zhang, L. Wang, D. Liu, B. Liu, Q. Liu, and Z. Li, "Dual temporal memory network for efficient video object segmentation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1515–1523.

[16] C. Redondo-Cabrera, M. Baptista-Ríos, and R. J. López-Sastre, "Learning to exploit the prior network knowledge for weakly supervised semantic segmentation," *IEEE Transactions on Image Processing*, vol. 28, no. 7, pp. 3649–3661, 2019.

[17] D. Nilsson and C. Sminchisescu, "Semantic video segmentation by gated recurrent flow propagation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6819–6828.

[18] V. Lup and S. Nedevschi, "Video semantic segmentation leveraging dense optical flow," in *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE, 2020, pp. 369–376.

[19] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "ERFNet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[20] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] 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*, 2017, pp. 2462–2470.

[22] G. Yang and D. Ramanan, "Volumetric correspondence networks for optical flow." *NeurIPS*, vol. 5, p. 12, 2019.

[23] 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*, 2015, pp. 2758–2766.

[24] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4040–4048.

[25] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3061–3070.

[26] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[27] A. Pfeuffer and K. Dietmayer, "Separable convolutional LSTMs for faster video segmentation," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1072–1078.

[28] M. Fonder and M. Van Droogenbroeck, "Mid-Air: A multi-modal dataset for extremely low altitude drone flights," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[29] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.