

Optical Flow Augmented Semantic Segmentation Networks for Automated Driving

Hazem Rashed¹, Senthil Yogamani², Ahmad El-Sallab¹, Pavel Křížek³ and Mohamed El-Helw⁴

¹CDV AI Research, Cairo, Egypt

²Valeo Vision Systems, Ireland

³Valeo R&D DVS, Prague, Czech Republic

⁴Nile University, Cairo, Egypt

Keywords: Semantic Segmentation, Visual Perception, Dense Optical Flow, Automated Driving.

Abstract: Motion is a dominant cue in automated driving systems. Optical flow is typically computed to detect moving objects and to estimate depth using triangulation. In this paper, our motivation is to leverage the existing dense optical flow to improve the performance of semantic segmentation. To provide a systematic study, we construct four different architectures which use RGB only, flow only, RGBF concatenated and two-stream RGB + flow. We evaluate these networks on two automotive datasets namely Virtual KITTI and Cityscapes using the state-of-the-art flow estimator FlowNet v2. We also make use of the ground truth optical flow in Virtual KITTI to serve as an ideal estimator and a standard Farneback optical flow algorithm to study the effect of noise. Using the flow ground truth in Virtual KITTI, two-stream architecture achieves the best results with an improvement of 4% IoU. As expected, there is a large improvement for moving objects like trucks, vans and cars with 38%, 28% and 6% increase in IoU. FlowNet produces an improvement of 2.4% in average IoU with larger improvement in the moving objects corresponding to 26%, 11% and 5% in trucks, vans and cars. In Cityscapes, flow augmentation provided an improvement for moving objects like motorcycle and train with an increase of 17% and 7% in IoU.

1 INTRODUCTION

Semantic image segmentation has witnessed tremendous progress recently with deep learning. It provides dense pixel-wise labeling of the image which leads to scene understanding. Automated driving is one of the main application areas where it is commonly used (Horgan et al., 2015). The level of maturity in this domain has rapidly grown recently and the computational power of embedded systems have increased as well to enable commercial deployment. Currently, the main challenge is the cost of constructing large datasets as pixel-wise annotation is very labor intensive. It is also difficult to perform corner case mining as it is a unified model to detect all the objects in the scene. Thus there is a lot of research to reduce the sample complexity of segmentation networks by incorporating domain knowledge and other cues where-ever possible. In this work, we explore the usage of motion cues via dense optical flow to improve the accuracy.

Majority of semantic segmentation algorithms

mainly rely on appearance cues based on a single image and do not exploit motion cues from two consecutive images. In this paper, we address the usage of dense optical flow as a motion cue in semantic segmentation networks. In particular for automotive scenes, the scene is typically split into static infrastructure and set of independently moving objects. Motion cues which needs two consecutive frames could particularly improve the segmentation of moving objects.

The contributions of this work include:

- Construction of four CNN architectures to systematically study the effect of optical flow augmentation to semantic segmentation.
- Experimentation on two automotive datasets namely Virtual KITTI and Cityscapes.
- Ablation study on using different flow estimators and different flow representations.

The rest of the paper is organized as follows: Section 2 reviews the related work in segmentation, computation of flow and role of flow in semantic segmentation. Section 3 details the construction of

four architectures to systematically study the effect of augmenting flow to semantic segmentation networks. Section 4 discusses the experimental results in Virtual KITTI and Cityscapes. Finally, section 5 provides concluding remarks.

2 RELATED WORK

2.1 Semantic Segmentation

A detailed survey of semantic segmentation for automated driving is presented in (Siam et al., 2017a). We briefly summarize the relevant parts focused on CNN based methods which are split into mainly three subcategories. The first (Farabet et al., 2013) used patch-wise training to yield the final classification. In (Farabet et al., 2013) an image is fed into a Laplacian pyramid, each scale is forwarded through a 3-stage network to extract hierarchical features and patch-wise classification is used. The output is post processed with a graph based classical segmentation method. In (Grangier et al., 2009) a deep network was used for the final pixel-wise classification to alleviate any post processing needed. However, it still utilized patch-wise training.

The second subcategory (Long et al., 2015)(Noh et al., 2015)(Badrinarayanan et al., 2015) was focused on end-to-end learning of pixel-wise classification. It started with the work in (Long et al., 2015) that developed fully convolutional networks (FCN). The network learned heatmaps that was then upsampled with-in the network using deconvolution to get dense predictions. Unlike patch-wise training methods this method uses the full image to infer dense predictions. In (Noh et al., 2015) a deeper deconvolution network was developed, in which stacked deconvolution and unpooling layers are used. In Segnet (Badrinarayanan et al., 2015) a similar approach was used where an encoder-decoder architecture was deployed. The decoder network upsampled the feature maps by keeping the maxpooling indices from the corresponding encoder layer.

2.2 Optical Flow in Automated Driving Systems

Flow estimation is very critical for automated driving and it has been a standard module for more than ten years. Due to computational restriction, sparse optical flow was used and it is replaced by Dense Optical Flow (DOF) in recent times. As flow is already computed, it can be leveraged for semantic seg-

mentation. Motion estimation is a challenging problem because of the continuous camera motion along with the motion of independent objects. Moving objects are the most critical in terms of avoiding fatalities and enabling smooth maneuvering and braking of the car. Motion cues can also enable generic object detection as it is not possible to train for all possible object categories beforehand. Classical approaches in motion detection were focused on geometry based approaches (Torr, 1998)(Papazoglou and Ferrari, 2013)(Ochs et al., 2014)(Menze and Geiger, 2015)(Wehrwein and Szeliski, 2017). However, pure geometry based approaches have many limitations, motion parallax issue is one such example. A recent trend (Tokmakov et al., 2016)(Jain et al., 2017)(Drayer and Brox, 2016)(Vijayanarasimhan et al., 2017)(Fragkiadaki et al., 2015) for learning motion in videos has emerged. CNN based optical flow has produced state of the art results. FlowNet (Ilg et al., 2016) was a simple two stream structure which was trained on synthetic data.

There has been many attempts to combine appearance and motion cues for various tasks. Jain et al. presented a method for appearance and motion fusion in (Jain et al., 2017) for generic foreground object segmentation. Laura et al. (Sevilla-Lara et al., 2016) leverages semantic segmentation for customizing motion model for various objects. This has been commonly used in scene flow models. Junhwa et al. (Hur and Roth, 2016) uses optical flow to provide temporally consistent semantic segmentation via post processing. MODNet (Siam et al., 2017b) fuses optical flow and RGB images for moving object detection. FuseNet (Hazirbas et al., 2016) is the closest to the work in this paper where they augmented semantic segmentation networks with depth. They show that concatenation of RGBD slightly reduces mean IoU and two-stream network with cross links show an improvement of 3.65 % IoU in SUN RGB-D dataset. Motion is a complementary cue of color and its role is relatively less explored for semantic segmentation. Our motivation is to construct simple baseline architectures which can be built on top of current CNN based semantic segmentation. More sophisticated flow augmentation architectures were proposed in (Nilsson and Sminchisescu, 2016) and (Gadde et al., 2017), however they are computationally more intensive.

3 SEMANTIC SEGMENTATION MODELS

In this section, we discuss the details of the four different semantic segmentation networks used in this pa-

per. We construct two flow augmented architectures namely RGBF network (Figure 1 (c)) which does concatenation and two-stream RGB+F network (Figure 1 (d)). RGB only and flow only architectures serve as baselines for comparative analysis.

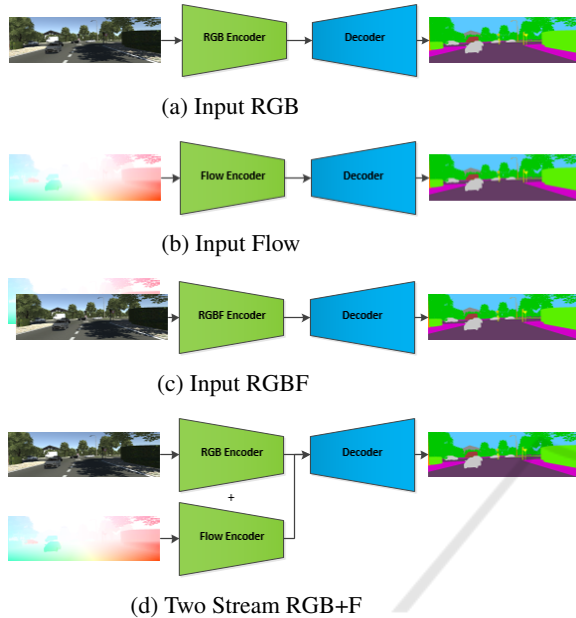


Figure 1: Four types of architectures constructed and tested in the paper. (a) and (b) are baselines using RGB and Flow only. (c) and (d) are flow augmented semantic segmentation architectures.

Table 1: Quantitative evaluation on Virtual KITTI data for our four segmentation networks.

Network Type	IoU	Precision	Recall	F-Score
RGB	66.47	78.23	75.6	73.7
F	42	63.92	55.28	55.76
RGBF (concat)	65.33	82.37	73.77	75.85
RGB + F (2-stream add)	70.52	83	76.4	78.9

3.1 One-stream Networks

An encoder-decoder architecture is used for performing semantic segmentation. Our network is based on FCN8s (Long et al., 2015) architecture, after which the remaining fully connected layers of the VGG16 are transformed to a fully convolutional network. The first 15 convolutional layers are used for feature extraction. The segmentation decoder follows the FCN architecture. 1x1 convolutional layer is used followed by three transposed convolution layers to perform up-sampling. Skip connections are utilized to extract high resolution features from the lower layers and then the extracted features are added to the partially upsampled results. The exact same one stream architecture is used for both RGB-only and flow-only

experiments. We use the cross-entropy loss function as shown below. where q denotes predictions and p denotes ground-truth. $C_{Dataset}$ is the set of classes for the used dataset.

3.2 RGBF Network

The input to this network is a 3D volume containing the original RGB image concatenated with the flow map. Several optical flow representations have been studied experimentally to find the best performing input, namely Color wheel representation in 3 channels, magnitude and direction in 2 channels, and magnitude only in 1 channel. It was found that an input of 4 channels containing RGB image concatenated with optical flow magnitude performs the best (Refer to Table 2 and 3). The flow map is normalized from 0 to 255 to have the same value range as the RGB. The first layer of the VGG is adapted to use the input of four channels and the corresponding weights are initialized randomly. The rest of the network utilizes the VGG pre-trained weights. In case of Virtual KITTI, we make use of flow ground truth as one of the inputs. This is done to simulate a near-perfect flow estimator which can be achieved by state-of-the-art CNN based algorithms and eliminate the variability due to estimation errors. In case of cityscapes, we make use of OpenCV Farneback function where the magnitude of the flow vectors are computed as $\sqrt{u^2 + v^2}$ where u and v are the horizontal and vertical flow vectors output from the function. The magnitude is then normalized to 0-255. In both the datasets, we also make use of the state-of-the-art flow estimator FlowNet v2 (Ilg et al., 2017).

3.3 Two Stream (RGB+F) Network

Inspired from (Simonyan and Zisserman, 2014)(Jain et al., 2017)(Siam et al., 2017b), we construct a two stream network which utilizes two parallel VGG16 encoders to extract appearance and flow features separately. The feature maps from both streams are fused together using summation junction producing encoded features of same dimensions. Then the same decoder described in the one-stream network is used to perform upsampling. Following the same approach as the one-stream network, skip connections are used to benefit from the high resolution feature maps. This network produces the best performance among the four as discussed in the experiments section. However, the two stream network is computationally more complex with more parameters relative to the RGBF network. The main difference is the fusion at the encoder stage rather than early fusion in RGBF network.

Table 2: Semantic Segmentation Results on Virtual Kitti (Mean IoU) for different DOF (GT) representations.

Type	Mean	Truck	Car	Van	Road	Sky	Vegetation	Building	Guardrail	TrafficSign	TrafficLight	Pole
RGBF (GT-Color Wheel)	59.88	41.7	84.44	40.74	93.76	93.6	66.3	49.43	52.18	61.21	49.61	21.52
RGBF (GT-Mag & Dir)	58.82	45.12	82.3	30.04	90.25	94.1	60.97	56.48	51.48	58.74	49.7	26.01
RGBF (GT-Mag only)	65.32	70.73	80.16	48.33	93.59	93.3	70.79	62.04	67.86	55.13	55.48	31.92
RGB+F (GT-3 layers Mag only)	67.88	35.7	91.02	24.78	96.47	94.06	88.72	74.4	84.5	69.48	68.95	34.28
RGB+F (GT-Color Wheel)	70.52	71.79	91.4	56.8	96.19	93.5	83.4	66.53	82.6	64.69	64.65	26.6

Table 3: Semantic Segmentation Results (Mean IoU) on CityScapes for different DOF (Farneback) representations.

Type	Mean	Bicycle	Person	Rider	Motorcycle	Bus	Car	Train	Building	Road	Truck	Sky	TrafficSign
RGBF (Mag only)	47.8	52.63	55.82	31.08	22.38	39.34	82.75	22.8	80.43	92.24	20.7	81.87	44.08
RGBF (Mag & Dir)	54.6	57.28	58.63	33.56	18.49	56.44	87.6	41.15	84.41	95.4	31.8	87.86	44.26
RGBF (Color Wheel)	57.2	61.47	62.18	35.13	22.68	54.87	87.45	36.69	86.28	95.94	40.2	90.07	51.64
RGB+F (3 layers Mag only)	62.1	65.15	65.44	32.59	33.19	63.07	89.48	43.6	87.88	96.17	57.2	91.48	55.76
RGB+F (Color Wheel)	62.56	63.65	66.3	39.65	47.22	66.24	89.63	51.02	87.13	96.4	36.11	90.64	60.68

Table 4: Semantic Segmentation Results (Mean IoU) on Virtual KITTI dataset.

Type	Mean	Truck	Car	Van	Road	Sky	Vegetation	Building	Guardrail	TrafficSign	TrafficLight	Pole
RGB	66.47	33.66	85.69	29.04	95.91	93.91	80.92	68.15	81.82	66.01	65.07	40.91
F (GT)	42	36.2	55.2	20.7	62.6	93.9	19.54	34	15.23	51.5	33.2	29.3
RGBF (GT)	65.328	70.74	80.2	48.34	93.6	93.3	70.79	62.05	67.86	55.14	55.48	31.9
RGB+F (GT)	70.52	71.79	91.4	56.8	96.19	93.5	83.4	66.53	82.6	64.69	64.65	26.6
F (FlowNet)	28.6	24.6	47.8	14.3	57.9	68	13.4	4.9	0.8	31.8	18.5	6.6
RGB+F (FlowNet)	68.84	60.05	90.87	40.54	96.05	91.73	84.54	68.52	82.43	65.2	63.54	26.54

Table 5: Semantic Segmentation Results (Mean IoU) on Cityscapes dataset.

Type	Mean	Bicycle	Person	Rider	Motorcycle	Bus	Car	Train	Building	Road	Truck	Sky	TrafficSign
RGB	62.47	63.52	67.93	40.49	29.96	62.13	89.16	44.19	87.86	96.22	48.54	89.79	59.88
F (Farneback)	34.7	34.48	37.9	12.7	7.39	31.4	74.3	11.35	72.77	91.2	19.42	79.6	11.4
RGBF (Farneback)	47.8	52.6	55.8	31.1	22.4	39.34	82.75	22.8	80.43	92.24	20.7	81.87	44.08
RGB+F (Farneback)	62.56	63.65	66.3	39.65	47.22	66.24	89.63	51.02	87.13	96.4	36.1	90.64	60.68
F (FlowNet)	36.8	32.9	50.9	26.8	5.12	25.99	75.29	15.1	65.16	90.75	25.46	50.16	29.14
RGBF (FlowNet)	52.3	54.9	58.9	34.8	26.1	53.7	83.6	40.7	79.4	94	28.1	79.4	45.5
RGB+F (FlowNet)	62.43	64.2	66.32	40.9	40.76	66.05	90.03	41.3	87.3	95.8	54.7	91.07	58.21

RGB+F network also has the advantage of being able to re-use pre-trained image encoders as it is decoupled from flow.

4 EXPERIMENTS

In this section, we present the datasets used, experimental setup and results.

4.1 Datasets

We had the following criteria for choosing the semantic segmentation datasets for our experiments. Firstly, the dataset has to be for automotive road scenes as our application is focused on automated driving. Secondly, we needed a mechanism to have the previous image for computational of optical flow. We chose two datasets namely Virtual KITTI (Gaidon et al., 2016) and Cityscapes (Cordts et al., 2016) datasets which satisfy the above criteria. Cityscapes is a popular automotive semantic segmentation dataset developed by Daimler. It comprises of 5000 images having full semantic segmentation annotation and 20,000 images with coarse annotations. We only use the fine annotations in our experiment and evaluated performance on the validation set that comprises of 500 images. We also chose Virtual KITTI dataset for

obtaining ground truth for flow which can be used as a proxy for a perfect flow estimator. It is a synthetic dataset that consists of 21,260 frames generated using Unity game engine in urban environment under different weather conditions. Virtual KITTI provides many annotations of which two are utilized in our approach namely dense flow and semantic segmentation. We split the dataset into 80% training data and 20% testing data. The semantic information is given for 14 classes including static and moving objects.

4.2 Experimental Setup

For all experiments, Adam optimizer is used with a learning rate of $1e^{-5}$. L2 regularization is used in the loss function with a factor of value of $5e^{-4}$ to avoid over-fitting the data. The encoder is initialized with VGG pre-trained weights on ImageNet. Dropout with probability 0.5 is utilized for 1x1 convolutional layers. Input image resolution of 375x1242 is used for Virtual KITTI and 1024x2048 is used for Cityscapes which is downsized to 512x1024 during training. Intersection over Union (IoU) is used as the performance metric for each class separately and an average IoU is computed over all the classes. In addition to IoU, precision, recall and F-score were calculated for Virtual KITTI dataset.

4.3 Experimental Results

We provide several tables of qualitative evaluation on the two datasets and discuss the impact of various classes due to flow augmentation. We also provide video links on outputs of the four architectures on test sequences in both datasets. In case of Virtual KITTI, we use the synthetic flow annotation as input. This is used to simulate a perfect flow estimator so that it can act as a best case baseline. Then it is compared with FlowNet v2 algorithm. In case of Cityscapes, we use a commonly used dense optical flow estimation algorithm namely Farneback's algorithm in OpenCV 3.1 with default parameters. We intentionally use this algorithm to understand the effects of relatively noisy flow estimations and compare it with FlowNet v2.

Table 1 shows the evaluation of our four architectures on Virtual KITTI dataset with different metrics. Flow augmentation consistently provides improvement in all four metrics. There is an improvement of 4% in IoU, 4.3% in Precision, 4.36% in Recall and 5.89% in F-score. Tables 2 and 3 shows a quantitative comparison of different DOF representations in RGBF and RGB+F architectures for both VKITTI and Cityscapes datasets. Results show that Color Wheel based representation in RGB+F architecture provides the best capability of capturing the motion cues and thus the best semantic segmentation accuracy. For all the following experiments, we make use of Color Wheel representation format for flow.

Table 4 shows detailed evaluation for each class separately. Although the overall improvement is incremental, there is a large improvement for certain classes, for example, trucks, van, car show an improvement of 38.14%, 27.8% and 5.8% respectively. Table 5 shows the evaluation of our proposed algorithm on Cityscapes dataset and shows a marginal improvement of 0.1% in overall IoU. However there is a significant improvement in moving object classes like motorcycle and train by 17% and 7% even using noisy flow maps. It is important to note here that the average IoU is dominated by road and sky pixels as there is no pixel frequency based normalization and the obtained improvement in moving object classes is still significant. Figure 2 and Figure 3 illustrates qualitative results of the four architectures on Virtual KITTI and Cityscapes respectively. Figure 2 (f) shows better detection of the van which has a uniform texture and flow cue has helped to detect it better. Figure 2 (h) shows that flownet provides good segmentation for the moving van, however fusion in Figure 2 (i) still needs to be improved. Figure 3 (f) and (h) illustrate better detection of the bicycle and the cyclist after

augmenting DOF. These examples visually verify the accuracy improvement obtained as shown in Table 4 and Table 5. Unexpectedly, Table 5 shows that Farneback and FlowNet v2 provides similar results, however FlowNet shows better results for some classes like Truck. One of the design goals in this work is to have a computationally efficient model, our most complex architecture namely RGB+F runs at 4 fps on a TitanX Maxwell architecture GPU. Autonomous driving systems need real-time performance and better run-time can be obtained by using a more efficient encoder instead of the standard encoder we used in our experiments.

Semantic segmentation results in test sequences of both datasets for all the four architectures along with ground truth and original images are shared publicly in YouTube. The Virtual KITTI results are available in ¹ and Cityscapes results are available in ². We included flow only network to conduct an ablation study to understand what performance flow alone can produce. To our surprise, it produces good results especially for road, vegetation, vehicles and pedestrians. We noticed that there is degradation of accuracy relative to RGB baseline whenever there is noisy flow. It would be interesting to incorporate a graceful degradation via a loose coupling of flow in the network. In spite of flow only network showing good performance, the joint network only shows incremental improvement. It is likely that the simple networks we constructed do not effectively combine flow and RGB as they are completely different modalities with different properties. It could also be possible that the RGB only network is implicitly learning geometric flow cues. In our future work, we plan to understand this better and construct better multi-modal architectures to effectively combine color and flow.

4.4 Future Work

The scope of this work is to construct simple architectures to demonstrate the benefit of flow augmentation to standard CNN based semantic segmentation networks. The improvement in accuracy obtained and visual verification of test sequences shows that there is still a lot of scope for improvement. Flow and color are different modalities and an explicit synergistic model would probably produce better results compared to learning their relationships from data. We summarize the list of future work below:

(1) Understand the effect of different encoders of varying complexity like Resnet-10, Resnet-101, etc.

¹<https://youtu.be/4M1IS2-2w5U>

²<https://youtu.be/VtFLpklatrQ>

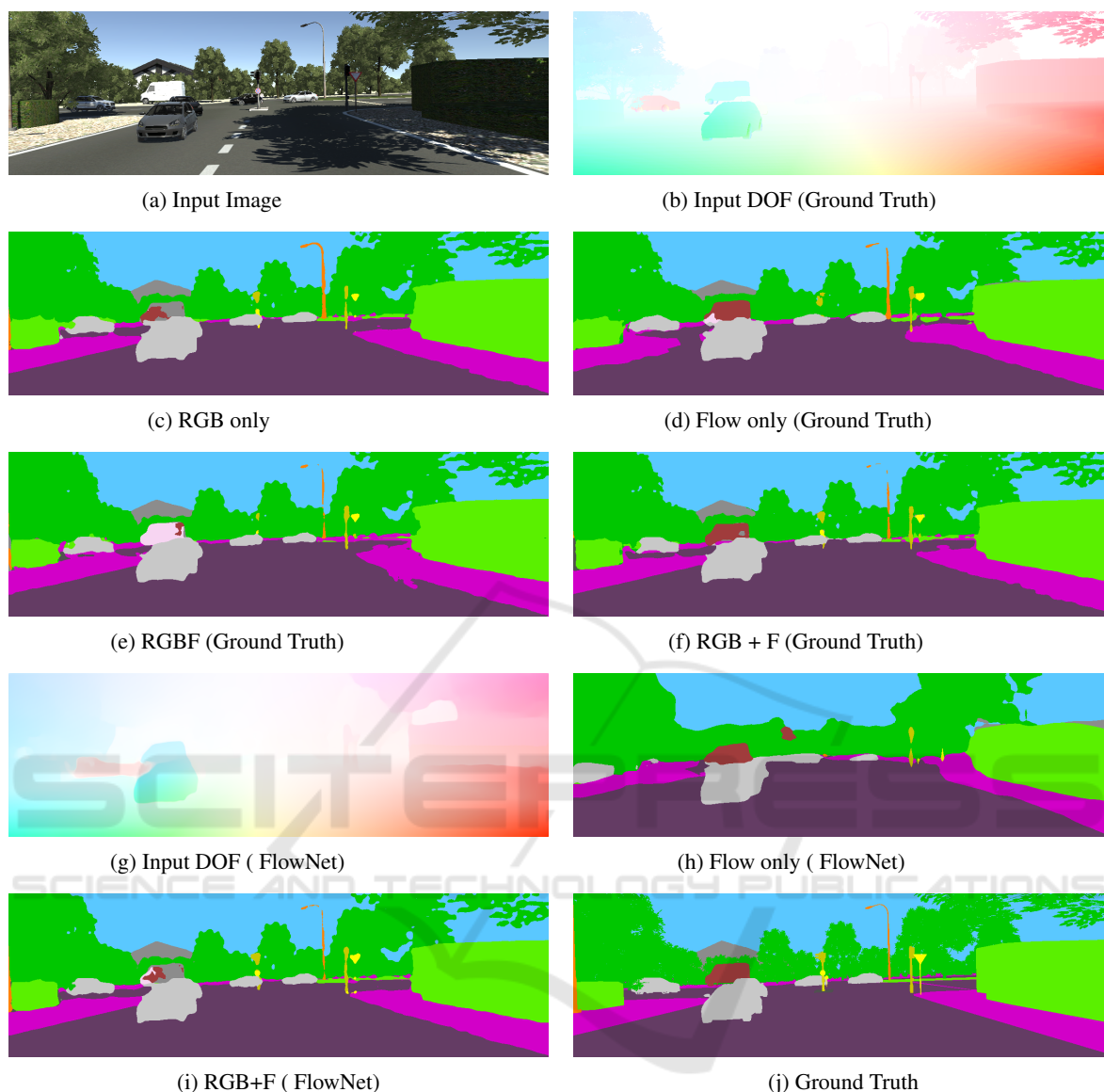


Figure 2: Qualitative comparison of semantic segmentation outputs from four architectures on Virtual KITTI dataset.

- (2) Evaluation of state-of-the-art CNN based flow estimators and joint multi-task learning.
- (3) Construction of better multi-modal fusion architectures.
- (4) Auxiliary loss induced flow feature learning in semantic segmentation architecture.
- (5) Incorporating sparsity invariant metrics to handle missing flow estimates.

5 CONCLUSION

In this paper, we explored the problem of leveraging flow in semantic segmentation networks. In applica-

tions like automated driving, flow is already computed in geometric vision pipeline and can be utilized. We constructed four variants of semantic segmentation networks which use RGB only, flow only, RGBF concatenated and two-stream RGB and flow. We evaluated these networks on two automotive datasets namely Virtual KITTI and Cityscapes. We provided class-wise accuracy scores and discussed them qualitatively. The simple flow augmentation architectures demonstrate a good improvement for moving object classes which are important for automated driving. We hope that this study encourages further research in construction of better flow-aware networks to fully utilize its complementary nature.

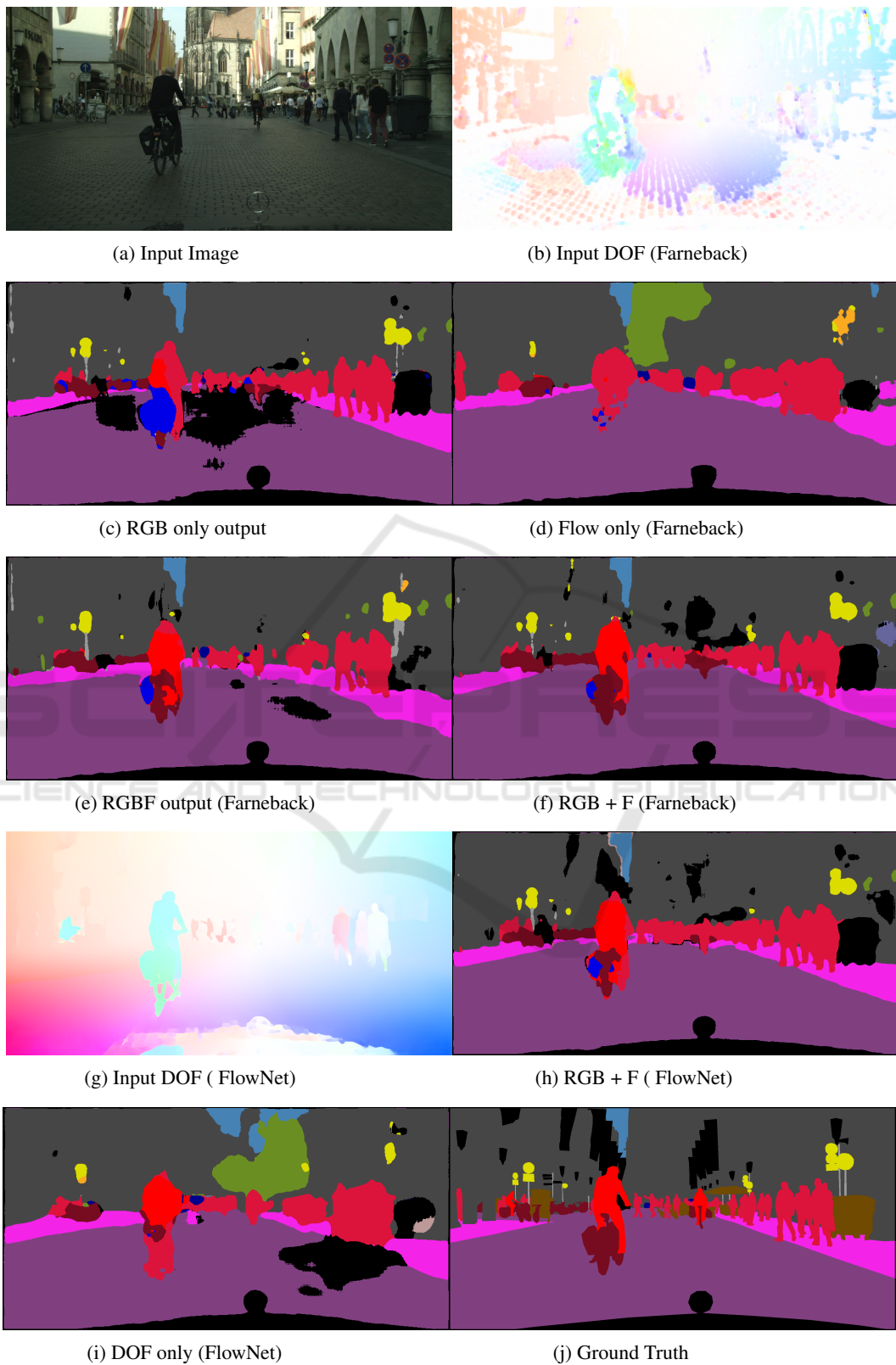


Figure 3: Qualitative comparison of semantic segmentation outputs from four architectures on Cityscapes dataset.

REFERENCES

- Badrinarayanan, V., Kendall, A., and Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*.
- Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *arXiv preprint arXiv:1604.01685*.
- Drayer, B. and Brox, T. (2016). Object detection, tracking, and motion segmentation for object-level video segmentation. *arXiv preprint arXiv:1608.03066*.
- Farabet, C., Couprie, C., Najman, L., and LeCun, Y. (2013). Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929.
- Fragkiadaki, K., Arbelaez, P., Felsen, P., and Malik, J. (2015). Learning to segment moving objects in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4083–4090.
- Gadde, R., Jampani, V., and Gehler, P. V. (2017). Semantic video cnns through representation warping. *CoRR, abs/1708.03088*, 8:9.
- Gaidon, A., Wang, Q., Cabon, Y., and Vig, E. (2016). Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*.
- Grangier, D., Bottou, L., and Collobert, R. (2009). Deep convolutional networks for scene parsing. In *ICML 2009 Deep Learning Workshop*, volume 3. Citeseer.
- Hazirbas, C., Ma, L., Domokos, C., and Cremers, D. (2016). Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Asian Conference on Computer Vision*, pages 213–228. Springer.
- Horgan, J., Hughes, C., McDonald, J., and Yogamani, S. (2015). Vision-based driver assistance systems: Survey, taxonomy and advances. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 2032–2039. IEEE.
- Hur, J. and Roth, S. (2016). Joint optical flow and temporally consistent semantic segmentation. In *European Conference on Computer Vision*, pages 163–177. Springer.
- Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2016). FlowNet 2.0: Evolution of optical flow estimation with deep networks. *arXiv preprint arXiv:1612.01925*.
- Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., and Brox, T. (2017). FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*, volume 2, page 6.
- Jain, S. D., Xiong, B., and Grauman, K. (2017). Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. *arXiv preprint arXiv:1701.05384*.
- Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.
- Menze, M. and Geiger, A. (2015). Object scene flow for autonomous vehicles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3061–3070.
- Nilsson, D. and Sminchisescu, C. (2016). Semantic video segmentation by gated recurrent flow propagation. *arXiv preprint arXiv:1612.08871*, 2.
- Noh, H., Hong, S., and Han, B. (2015). Learning deconvolutional network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528.
- Ochs, P., Malik, J., and Brox, T. (2014). Segmentation of moving objects by long term video analysis. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1187–1200.
- Papazoglou, A. and Ferrari, V. (2013). Fast object segmentation in unconstrained video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1777–1784.
- Sevilla-Lara, L., Sun, D., Jampani, V., and Black, M. J. (2016). Optical flow with semantic segmentation and localized layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3889–3898.
- Siam, M., Elkerdawy, S., Jagersand, M., and Yogamani, S. (2017a). Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges. *arXiv preprint arXiv:1707.02432*.
- Siam, M., Mahgoub, H., Zahran, M., Yogamani, S., Jagersand, M., and El-Sallab, A. (2017b). Modnet: Moving object detection network with motion and appearance for autonomous driving. *arXiv preprint arXiv:1709.04821*.
- Simonyan, K. and Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576.
- Tokmakov, P., Alahari, K., and Schmid, C. (2016). Learning motion patterns in videos. *arXiv preprint arXiv:1612.07217*.
- Torr, P. H. (1998). Geometric motion segmentation and model selection. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 356(1740):1321–1340.
- Vijayanarasimhan, S., Ricco, S., Schmid, C., Sukthankar, R., and Fragkiadaki, K. (2017). Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804*.
- Wehrwein, S. and Szeliski, R. (2017). Video segmentation with background motion models. In *BMVC*.