Three for one and one for three: Flow, Segmentation, and Surface Normals

Hoang-An Le
hoang-an.le@uva.nl
Anil S. Baslamisli
a.s.baslamisli@uva.nl
Thomas Mensink
thomas.mensink@uva.nl
Theo Gevers
th.gevers@uva.nl

Computer Vision Group,
Informatics Institute,
University of Amsterdam,
the Netherlands

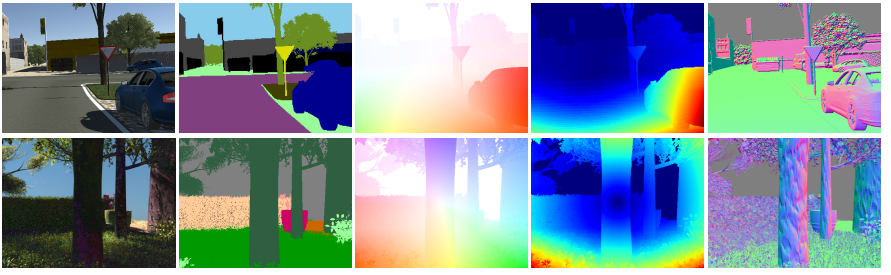
Abstract

Optical flow, semantic segmentation, and surface normals represent different information modalities, yet together they bring better cues for scene understanding problems. In this paper, we study the influence between the three modalities: how one impacts on the others and their efficiency in combination. We employ a modular approach using a convolutional refinement network which is trained supervised but isolated from RGB images to enforce joint modality features. To assist the training process, we create a large-scale synthetic outdoor dataset that supports dense annotation of semantic segmentation, optical flow, and surface normals. The experimental results show positive influence among the three modalities, especially for objects' boundaries, region consistency, and scene structures.

1 Introduction

Optical flow, semantic segmentation, and surface normals represent different aspects of objects in a scene i.e. object motion, category, and geometry. While they are often approached as single-task problems, their combinations are of importance for general scene understanding as humans also rarely perceive objects in a single modality. As different information sources provide different cues to understand the world, they could also become complementary to each other. For example, certain objects have specific motion patterns (flow and semantics), an object's geometry provides specific cues about its category (surface normals and semantics), and object's boundary curves provide cues about motion boundaries (flow and surface normals).

Scene-based optical flow estimation is a challenging problem because of complicated scene variations such as texture-less regions, large displacements, illumination changes, cast shadows, and specularities. As a result, optical flow estimation tends to perform poorly in homogeneous areas or around objects' boundaries. Another hindrance for many optical flow estimators is the common assumption of spatial homogeneity in the flow structure across



(a) (b) (c) (d) (e)

Figure 1: Visualization from VKITTI [18] (*top*) and Nature (*bottom*) dataset. From left to right: (a) RGB image, (b) semantic annotation, (c) color-coded optical flow, (d) corresponding flow magnitude; (e) surface normal.

an image [33]. That assumption poses difficulties as different objects have different motion patterns: objects closer to the viewer have stronger optical flow; independent objects have their own flow fields, while static objects follow the camera's motion (Figure 1). Thus, by modeling optical flow based on image segmentation, one could improve flow accuracy, especially at objects' boundaries [1, 33].

The goal of image segmentation is to partition an image into different parts that share common properties. In particular, semantic segmentation assigns to each image pixel the object category to which the pixels belong. It is a challenging task especially in videos, due to inherent video artifacts such as motion blur, frame-to-frame object-to-object occlusions and object deformations. As optical flow encodes temporal-visual information of image sequences, it is often exploited to relate scene changes over time [29, 41, 42]. Yet, optical flow, as an approximation to image motion [5], also encodes the 3D structure of a viewed scene. If the camera's translation is known beforehand, an optical flow image can be used to recover the scene depth [3]. Considering a moving camera, closer objects appear with stronger motion fields than distant ones, independent moving objects have prominent motion patterns compared to the background, and different object shapes generate different motion fields because of depth discontinuities. Therefore, temporal and structural information provided by optical flow can guide semantic segmentation by providing cues about scene ambiguities.

Surface normals, on the other hand, represent changes of depth, i.e. the orientation of object surfaces in 3D space. Thus, they are independent of illumination effects or object textures. That information is particular helpful for texture-less objects, regions of strong cast shadows, or scenes of less visibility. Additionally, object boundaries provide cues about both motion and semantic boundaries. Therefore, surface normal information is expected to assist the optical flow estimation process by providing various cues. Figure 1 illustrates the relationship of object boundaries depicted in segmentation and 3D structure cues in optical flow and surface normal images.

In this paper, we study the mutual interaction of optical flow, semantic segmentation, and surface normals and analyze their contribution to each other. We employ a modular approach and adapt a convolution based supervised refinement network to examine the efficiency of joint features from the different modalities.

Large-scale datasets with optical flow and surface normal ground truths are hard to obtain. Apart from semantic segmentation, it is not intuitive for humans to manually come up with pixel-wise annotations for optical flow and surface normals. Hence, only a small number of real-world datasets provide annotations for either optical flow or surface normals, but

not for both. Although one may construct surface normals from depth images, the process tends to produce unwanted artifacts. Thus, synthetic data is favorable. To that end, we construct a large-scale synthetic dataset of nature scenes such as gardens and parks, that provides ground-truth annotations for optical flow, surface normals and semantic segmentation.

In summary, our contributions are: (1) the connection among the three modalities (optical flow, semantic segmentation and surface normals), (2) adapting a convolutional based supervised refinement network to improve one of the three using the other two, (3) an experimental study to estimate all three in a joint fashion, (4) a large-scale scene-level synthetic images of nature scenes such as gardens and parks with ground-truth annotations for optical flow, surface normals and semantic segmentation.

2 Segmentation, Flow, and Surface Normals

2.1 Related Work

In this section, we review the work on optical flow, semantic segmentation, and surface normals and how they are mostly targeted as single tasks.

Optical flow is defined as the apparent motion field resulted from an intensity displacement in a time-ordered sequence of images. It is an approximation to image motion, because estimating optical flow is an ill-posed problem [5]. To model the displacement of image intensities that are caused solely by the objects' motion in the physical world, several priors are derived to constrain the problem. Two of the most exploited ones are the assumptions of brightness constancy and Lambertian surface reflectance [5, 16, 36]. Besides, [6] makes use of robust statistics to promote discontinuity-preservation. Many popular methods also apply coarse-to-fine strategies [7, 32, 38]. On the other hand, deep convolutional neural networks (CNNs) are dominating the field more recently. For instance, [37] applies a coarse-to-fine strategy with the help of a CNN framework. Then, Dosovitskiy *et al*. [13] proposes an end-to-end CNN called FlowNet, which is later improved by Ilg *et al*. [25] to perform state-of-the-art optical flow estimations.

Semantic Segmentation is vital for robot vision and scene understanding tasks as it provides pixel-wise annotations to scene properties. Traditional methods approach the problem by engineering hand-crafted features and perform pixel-wise classification with the help of a classifier [12, 34]. Other works try to group semantically similar pixels [10, 31]. Like most of the computer vision tasks, semantic segmentation also benefits from powerful CNN models. After the pioneering work of [30], many other deep learning based methods are proposed such as [8, 39].

Surface Normals provide information about an object's surface geometry. Traditional methods that infer 3D scene layout from single images rely on primitives detection such as oriented 3D surfaces [24] or volumetric primitives [20]. Their performance depends on the discriminative appearance of the primitives [17]. In the context of deep learning, Wang *et al*. [35] proposes a method to predict surface normals from a single color image by employing a scene understanding network architecture with physical constraints; Bansal *et al*. [2] predicts surface normals and use them as an intermediate representation for 3D volumetric objects for model retrieval. Eigen and Fergus [14] designs an architecture that can be used for predicting each of the three modalities, including depth, surface normals, and semantic segmentation, in a separating manner.

2.2 Cross-modality Influence

Three for Optical Flow

Image motion, although varies across image regions, is often treated in the same way by many optical flow methods. Semantic segmentation provides a way to partition an image into different groups of predefined semantic classes. Hence it provides optical flow with information of object boundaries, and helps to enforce motion consistency within similar object regions. Similar ideas employed by [1] with object instance-segmentation or [33] with 3 semantic classes (things, planes, and stuff) have shown to improve optical flow accuracy.

On the other hand, surface normals represent the orientation of objects' surfaces in 3D space. They contain geometry information that is invariant to scene lighting and objects' appearance, which is rendered useful for optical flow in case of intricate lighting such as cast shadows, texture-less regions, specularities, etc. Additionally, surface normals can be beneficial for optical flow at depth order reasoning, and may improve accuracy at occlusion boundaries.

Three for Semantic Segmentation

Optical flow is often exploited for its ability in relating scene changes along time-axis: He *et al*. [22] aggregates information from multiple views using optical flow to perform semantic segmentation for a single frame, while Zhu *et al*. [42] uses optical flow to propagate image features from keyframe images to nearby frames, speeding up the segmentation process. Several methods exploit motion information to segment images into foreground objects from a moving background, such as SegFlow [9] or FusionSeg [27]. However, they solely rely on the motion properties of objects and do not take the objects' identities into account.

In this paper, we leverage the notion of segmentation into a more semantic meaning of a scene, i.e. we do not limit the segmentation to just foreground/background [9] or coarse general classes (things, planes, stuff) [33], but rather adhere to the current segmentation problems in the literature, which on average, consist of 10-20 classes [11, 15, 18].

As image motion is the projection of 3D object motions, depth discontinuities correspond to motion boundaries, making optical flow an indication of scene depth. At the same time, surface normals represent the changes of depth and the alignment of object surfaces. Such information is particularly useful for semantic segmentation, as objects can be recognized not only from their appearances, but also from their shape and geometric characteristics. Thus, similar to methods that recognize objects from depth images, by associating each object with their motion type, or surface normals, it is feasible to recognize them from surrounding regions using geometry information signified from these modalities.

Three for Surface Normals

Similar to the case of optical flow, semantic segmentation provides object boundary information, which in many cases corresponds to depth disruptions, thus helps to enhance object boundary, and local coherence in surface normal prediction. Ladicky *et al*. [28] employs segment cues to enforce smooth surface normal results estimated with contextual information.

Optical flow represents motion information, yet also signifies geometry structure of a scene. As surface normals are identified by the rate of change in the location of objects, they are highly correlated. Thus, optical flow can provide useful cues to enhance surface normals, in terms of objects' inner structure as well as their boundaries.

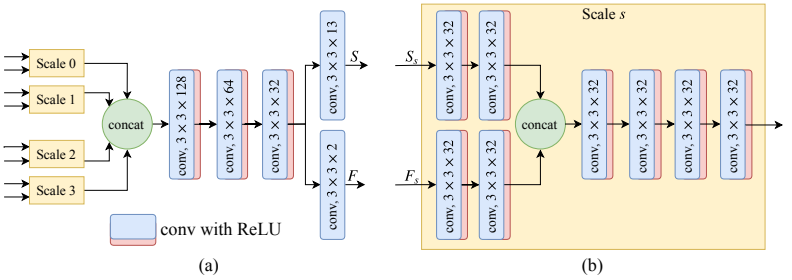


Figure 2: Joint refinement network for two modalities with *tight* feature coupling (*left*), inspired by [26]. The outputs of the modal-specific networks are integrated at different scales, using scale branch architecture (*right*), and up-sampled before concatenation.

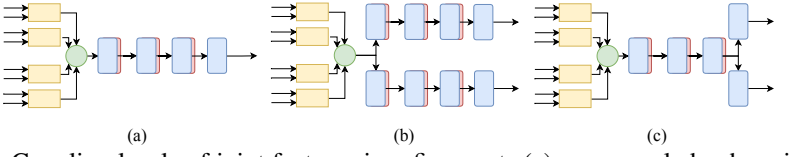


Figure 3: Coupling levels of joint features in refinement: (a) *zero* coupled, where joint features refine a single task; (b) *loosely* coupled, where joint features branch to refine each task separately; (c) *tightly* coupled, where joint features share a decoder to refine all tasks.

3 Method

To study the relationship between the three modalities, we follow a refinement strategy. That is based on the work by Jafari *et al*. who uses a similar network to analyze the relationship between semantic segmentation and depth [26]. In this paper, we adapt their network architecture to train a joint refinement network designed for semantic segmentation, surface normals and optical flow.

An overview of the architecture is shown in Figure 2a. As an example for joint refining optical flow and semantic segmentation, the refinement network takes in a preliminarily predicted segmentation and optical flow at different scales (Figure 2b) and couples them in a joint optimization process. The input to the branch scale s is composed of a segmentation image S_s and a flow image F_s, both sub-sampled to $\frac{1}{2^s}$ of the original size. The outputs of the scale branches are up-sampled, as the refinement network provides output at the original image size. We include a scale branch at the original size, and expand the network depth to increase its capacity to cope with different input modalities. When there are 3 modality inputs, the scale branch will have a third input, which is then concatenated to the other two.

We also leverage the study of cross-modality influence performed in [26], and the ability of the refinement network to learn a joint representation that benefits from both modalities. We keep the multi-scale encoder part fixed (up to the *concat* layer) and partially decouple the decoder. That allows the network to have different capacities in using the joint features learned from the encoder to refine different modalities. Specifically, we examine three different architectures that impose different coupling levels of joint features as illustrated in Figure 3. Namely, we study the ability of refinement when there is only one task required, hence *zero*-coupling (Figure 3a), when both 2 tasks are *loosely* coupled (Figure 3b), or *tightly* coupled (Figure 3c).

4 Experiments

4.1 Experimental Setup

Datasets

Nature. We created a synthetic dataset covering nature scenes like gardens and parks. The dataset features different vegetation types such as trees, bushes, flowering plants, and grass, being laid out in various terrain and landscape types. The models' textures and skies are used from real-world images to provide a realistic look of the scenes. For each construction, we put up random paths to run the cameras around and capture the scene from different positions and angles. The images are rendered with the physics-based Blender Cycles engine[1]. To obtain annotations, the rendering pipeline is modified to output RGB images, optical flow, surface normals, and objects' identities (semantic labels). The dataset consist of 300 images from 10 scenes, each with 5 different lighting conditions (clear, cloudy, overcast, sunset, and twilight), resulting in 15K images.

Virtual KITTI [18] (VKITTI) is a large-scale synthetic dataset following the setting of KITTI dataset [19] for autonomous driving problems. Each image frame comes with pixel-wise annotation of ground truth instance-object segmentation, optical flow, and depth information. To get surface normal ground truth, we convert ground truth depth images using the method described in [4]. As the depth images are produced by a simulation renderer, they are free from noise and uncertainties, producing less artifacts in the conversion results.

Baselines & Evaluation metrics

Each of the three modalities have their own baseline and evaluation metric.

Optical Flow. We use FlowNetC [13] as our baseline because of its balance in speed and accuracy [9, 41] (and is therefore preferable over the less accurate FlowNetS or the more expensive FlowNet2 [25]). We fine-tune the network for each dataset and report the results as baseline. Performance is evaluated by the average endpoint error, the lower the better.

Semantic Segmentation. We use the ResNet-101 architecture [21] as the baseline, and follow the practice of [9, 40] to add corresponding decoder layers so that the network can output full-resolution image. Performance is measured by mean intersection-over-union, shown in percentage, the higher the better.

Surface Normals. We follow the MarrRevisited [2] architecture to run and report their results on our datasets as baseline. Evaluation is based on the angular difference between predicted normals and ground truth [17]. The 3 error measurements *mean*, *median*, *rmse* show the difference (degrees) between predicted and ground truth normal vectors, thus lower is better. The 3 measurements *11.25*, *22.5*, *30* count the number of pixels within the indicated angle thresholds (in degrees); the results are shown in percentage, and higher is better.

4.2 Baseline & Oracle Experiments

The first set of experiments is to test the hypothesis that the modalities have a positive impact on each other and to establish the baselines. For each experiment, we train the aforementioned baseline networks, and have the output results passed into the refinement architecture (described in Section 3) together with ground truths of other modalities.

[1] https://www.blender.org/

Optical Flow

Figure 4a shows the baseline and refinement results for optical flow, using ground truth segmentation and surface normals. In general, both modalities help to improve optical flow. The refined results in Figure 4b appear more crispy, especially along the objects' boundaries.

Semantic Segmentation

As shown in Figure 5a, optical flow and surface normals also improve semantic segmentation over the baseline, since the outline of the objects obtained from these modalities are more informative. The geometry information provided by flow and normals helps to retain details in semantic segmentation; e.g. the lamppost, tree branches, and fence wires are well retained in Figure 5b. However, as the refinement module does not have access to the original image (raw RGB), geometry information alone cannot help much in correcting semantic errors that are present in the input (yellow regions in the second row).

Surface Normals

As shown in Figure 6a, using optical flow and semantic segmentation helps to imrpove surface normal prediction significantly. The refinement using flow seems to be better than using segmentation for the VKITTI case, whereas it is the other way around for the Nature case. This can be explained as estimating surface normals requires a network to understand the geometry information of the scene, which is easier for optical flow in Nature as all of the objects are static and thus optical flow field depends solely on the camera ego-motion, while it is not the case for VKITTI, segmentation has more advantage as objects' shapes are more uniform (e.g. houses, cars, roads' surface) than those in Nature (e.g. bushes, rocks, grass). The refined result using oracle flow produces sharper details, while the ones with oracle segmentation are more accurate (Figure 6b).

To conclude, different modalities, when being used in their most accurate form (GT), provide complementary cues to each other, thus improving the performance of other modalities: segmentation provides flows and normals with objects' identities and boundaries; optical flow provides segmentation and normals motion and geometry information; surface normals provide segmentation and flow with geometry and objects boundaries. In the following experiments, we examine the usefulness of different modalities when they are not perfect (given by an approximation scheme), and the interaction of more than one modalities.

4.3 Cross-Modality Influence

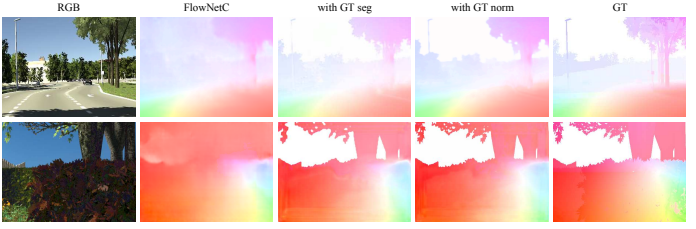
Refinement Coupling

In this experiment, we consider joint learning of semantic segmentation and optical flow on the VKITTI dataset [8]. We examine the different coupling levels during refinement: *zero* coupling, *loose* coupling, and *tight* coupling (see Figure 3). We also expand the *tight* coupling, into an end-to-end learning pipeline (denoted by *tight+*), where the segmentation and optical flow networks are fine-tuned together with the refinement module.

The results are shown in Table 1. From the results, we observe that using ground-truth or predicted segmentation to refine optical flow, the performance always improves. The difference between ground-truth and predicted segmentation is small compared to the difference between the baseline and the refined models. However, for segmentation, refining based on flow is only beneficial when the *zero* coupling is used. Likely, this is because the predicted flow does not contain accurate semantic cues to improve segmentation. Based on this experiment we use the *zero* coupling for the remaining experiments.

Dataset	FlowNetC [13]	with GT seg	with GT norm
VKITTI	2.68	2.37	**2.36**
Nature	16.19	14.09	**13.92**

(a) Average EPE of optical flow baseline and oracle refinement, lower is better.

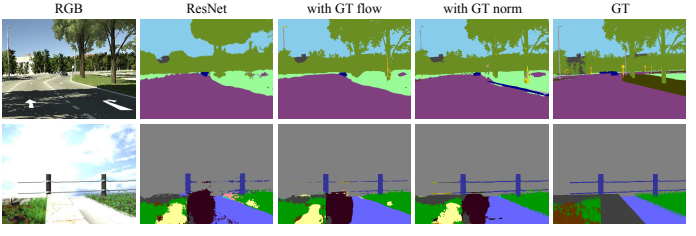


(b) Qualitative examples of optical flow on VKITTI (*top*) and Nature (*bottom*)

Figure 4: Results of optical flow oracle refinement: adding ground truth segmentation and surface normal improves the crispness of objects' boundaries and outperforms all the baselines.

Dataset	ResNet [9]	with GT flow	with GT norm
VKITTI	44.11	46.90	**50.0**
Nature	37.88	38.4	**41.6**

(a) Mean IOU (in %) of segmentation baseline and oracle refinement, higher is better.

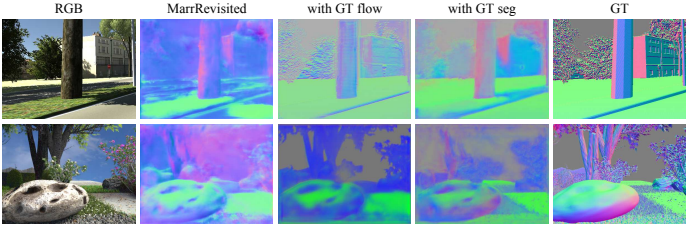


(b) Qualitative examples of segmentation on VKITTI (*top*) and Nature (*bottom*).

Figure 5: Results of segmentation oracle refinement: adding ground truth flow or surface normals captures more fine-detailed segmentations (*e.g.* tree leaves and fence wires) and improves performance over the baseline, since the shape of the objects obtained is informative.

		mean (↓)	median (↓)	rmse (↓)	11.25 (↑)	22.5 (↑)	30 (↑)
VKITTI	MarrRevisited [9]	48.13	48.34	57.44	17.39	19.81	27.44
	with GT flow	11.26	3.18	17.29	**62.17**	**76.27**	87.14
	with GT seg	**11.16**	**3.17**	**16.78**	60.70	75.67	**88.43**
Nature	MarrRevisited [9]	40.25	46.68	50.25	29.44	30.68	34.42
	with GT flow	**9.62**	**8.01**	**13.20**	**61.78**	**89.74**	**97.53**
	with GT seg	10.39	8.75	13.48	59.65	89.46	97.30

(a) Quantitative results of surface normals baseline and oracle refinement.



(b) Qualitative examples of surface normal on VKITTI (*top*) and Nature (*bottom*).

Figure 6: Results of surface normals oracle refinement: adding ground truth flow and segmentation removes the blur in the baseline and improve the results, indicating the interest of combining modalities.

Target	Baseline	GT	Predicted			
		zero	*zero*	*loose*	*tight*	*tight+*
Semantic segmentation (↑)	44.11	46.9	**44.78**	41.2	41.1	43.9
Optical flow (↓)	2.68	2.37	**2.40**	2.43	2.41	2.42

Table 1: Different levels of coupling of segmentation and optical flow on the VKITTI dataset. Performance is measured in Mean IOU (↑) and Average EPE (↓) respectively. Refinement can improve over the baseline, and *zero* coupling, when both modalities refine a single task, works best for both semantic segmentation and optical flow.

Flow from Segmentation and Normals

The refinement results for optical flow using predicted modalities are shown in Figure 7a. Because of inaccuracies in predicted normals, the refined results do not improve as much as with predicted segmentation, or even hurt in case of Nature dataset. In general, the two modalities help optical flow to obtain better delineation. Figure 7b shows an occlusion case where part of the car is occluded by a traffic sign, FlowNetC recognizes it but fails to obtain the correct shape of the occlusion, which can be recovered with surface normal information and improved using segmentation and surface normals.

Segmentation from Flow and Normals

Predicted optical flow and surface normals contain different inaccuracies. Thus, when used to refine the semantic segmentation, they confuse the semantic cues, and to some extent, have negative impact on the preliminary segmentation results. This explains the decreasing results and slight improvements in Figure 8a. Visual inspection on Figure 8b shows that refinement of flow and surface normals make the boundaries smoother, reducing the effect of incorrect areas. In combination, they capture more details and help to produce better segmentation.

Normals from Flow and Segmentation

Surface normal refinement results are provided in Figure 9a and illustrated in Figure 9b. The confusion in the sky and tree regions of the baseline estimation is removed when refined with different modalities. Information of objects' categories and boundaries provided by semantic segmentation helps retaining fine details in the results (e.g. pavement in VKITTI, fences in Nature). However, the inaccuracies of predicted flow leave some artifacts and makes the results less accurate (e.g. the sky, tree and fences regions). Surprisingly, refinement using predicted segmentation outperform refinement with ground truth in Figure 6a. This could be explained by the idea of knowledge distillation [23], the predicted segmentation with smoothed softmax scores, describes the knowledge of the network better than the hard one-hot encoding of the ground-truth. This allows the refinement module to learn better.

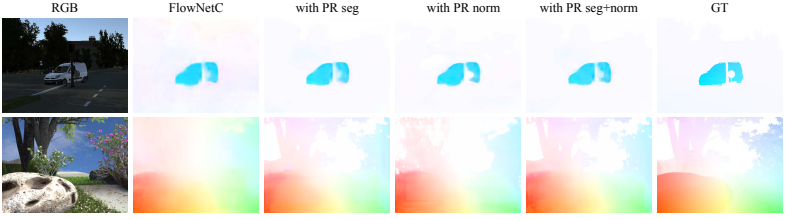
5 Conclusion and Future Work

We have analyzed the combination of three important modalities in computer vision, namely optical flow, semantic segmentation, and surface normals, and their impact on each other. Because each modality contains different type of information, in combination, they provide complementary cues to enhance each other. We approached the problem at a modular level where the inputs are kept fixed at the preliminary estimation. Future work will include end-to-end training of modalities to exploit raw image features.

Acknowledgements: This project was funded by the EU Horizon 2020 program No. 688007 (TrimBot2020). We would like to thank Leo Dorst for his helpful discussions and advice.

Method	FlowNetC [3]	with PR seg	with PR norm	with PR seg+norm
VKITTI	2.68	2.40	2.50	**2.39**
Nature	16.19	**14.16**	16.62	14.21

(a) Average EPE of optical flow based on prediction, lower is better.

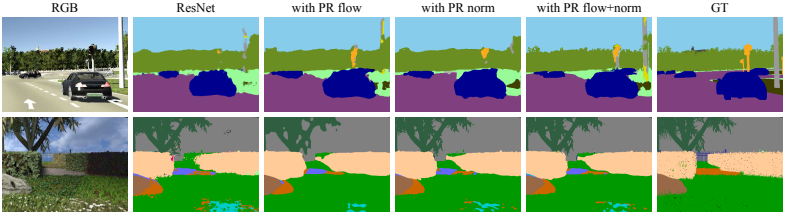


(b) Qualitative examples on VKITTI (*top*) and Nature (*bottom*).

Figure 7: Results of optical flow baseline and refinement based on prediction.

Method	ResNet [8]	with PR flow	with PR norm	with PR flow+norm
VKITTI	44.11	44.78	45.36	**47.55**
Nature	37.88	37.57	**38.83**	38.00

(a) Mean IOU (in %) of segmentation based on prediction, higher is better.

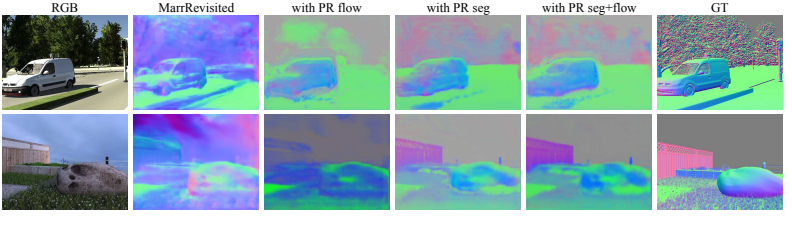


(b) Qualitative examples on VKITTI (*top*) and Nature (*bottom*).

Figure 8: Results of segmentation baseline and refinement based on prediction.

	Method	mean (↓)	median (↓)	rmse (↓)	11.25 (↑)	22.5 (↑)	30 (↑)
VKITTI	MarrRevisited [8]	48.13	48.34	57.44	17.39	19.81	27.44
	with PR flow	12.32	4.91	18.02	58.43	73.46	84.08
	with PR seg	**11.44**	**2.82**	**17.24**	**60.30**	74.02	86.58
	with PR flow+seg	11.58	3.45	17.28	60.04	**74.63**	**86.59**
Nature	MarrRevisited [8]	40.25	46.68	50.25	29.44	30.68	34.42
	with PR flow	11.39	9.80	14.38	55.36	86.88	96.66
	with PR seg	**9.09**	**7.08**	**12.56**	**65.26**	**91.11**	**97.70**
	with PR flow+seg	9.22	7.27	12.71	64.39	90.94	97.59

(a) Quantitative results of surface normal refinement based on prediction.



(b) Qualitative examples on VKITTI (*top*) and Nature (*bottom*).

Figure 9: Results of surface normals baseline and refinement based on prediction.

References

[1] Min Bai, Wenjie Luo, Kaustav Kundu, and Raquel Urtasun. Exploiting semantic information and deep matching for optical flow. In *Lecture Notes in Computer Science*, volume 9910 LNCS, pages 154–170, 2016.

[2] Aayush Bansal, Bryan Russell, and Abhinav Gupta. Marr Revisited: 2D-3D Alignment via Surface Normal Prediction. In *CVPR*, pages 5965—-5974, 2016.

[3] P. Baraldi, E. D. Micheli, and S. Uras. Motion and Depth from Optical Flow. In *Procedings of the Alvey Vision Conference 1989*, pages 35.1–35.4, 1989.

[4] Jonathan T. Barron and Jitendra Malik. Intrinsic Scene Properties from a Single RGB-D Image. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Intrinsic*, volume 2, pages 17–24, 2013.

[5] S S Beauchemin and J L Barron. The Computation of Optical Flow. *ACM Comput. Surv.*, 27(3):433–466, sep 1995.

[6] Michael. Black and P Anandan. Nonmetric calibration of wide-angle lenses and polycameras. *Computer Vision and Image Understanding*, 63(1):75–104, 1996.

[7] Thomas Brox and Jitendra Malik. Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):500–513, mar 2011.

[8] L.C. Chen, G Papandreou, I Kokkinos, K Murphy, and A.L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016.

[9] Jingchun Cheng, Yi-hsuan Tsai, Shengjin Wang, Ming-Hsuan Yang, and Shengjin Wang Ming-hsuan Yang. SegFlow : Joint Learning for Video Object Segmentation and Optical Flow. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[10] D Comaniciu and P Meer. Mean Shift: A Robust Approach Toward Feature Space Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5): 603–619, may 2002.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops*, volume 3, 2016.

[12] G Csurka and F Perronnin. An Efficient Approach to Semantic Segmentation. *International Journal of Computer Vision, 95(2)*, pages 198–212, 2011.

[13] Alexey Dosovitskiy, Philipp Fischery, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 11-18, pages 2758–2766, 2016.

[14] David Eigen and Rob Fergus. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2650–2658, 2015.

[15] M Everingham, L Van-Gool, C K I Williams, J Winn, and A Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88 (2):303–338, jun 2010.

[16] Denis Fortun, Patrick Bouthemy, Charles Kervrann, Denis Fortun, Patrick Bouthemy, and Charles Kervrann. Optical flow modeling and computation : a survey. *Computer Vision and Image Understanding*, 134:1–21, 2015.

[17] David F Fouhey, Abhinav Gupta, and Martial Hebert. Data-Driven 3D Primitives for Single Image Understanding. In *2013 IEEE International Conference on Computer Vision*, pages 3392–3399, 2013.

[18] A Gaidon, Q Wang, Y Cabon, and E Vig. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In *CVPR*, 2016.

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[20] Abhinav Gupta, Alexei A Efros, and Martial Hebert. Blocks World Revisited: Image Understanding Using Qualitative Geometry and Mechanics. In *Proceedings of the 11th European Conference on Computer Vision: Part IV*, ECCV'10, pages 482–496, Berlin, Heidelberg, 2010. Springer-Verlag.

[21] K He, X Zhang, S Ren, and J Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, jun 2016.

[22] Yang He, Wei-Chen Chiu, Margret Keuper, and Mario Fritz. STD2P: RGBD Semantic Segmentation Using Spatio-Temporal Data-Driven Pooling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[23] G. Hinton, O. Vinyals, and J. Dean. Distilling the Knowledge in a Neural Network. *ArXiv e-prints*, March 2015.

[24] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering Surface Layout from an Image. *Int. J. Comput. Vision*, 75(1):151–172, oct 2007.

[25] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[26] Omid Hosseini Jafari, Oliver Groth, Alexander Kirillov, Michael Ying Yang, and Carsten Rother. Analyzing modular CNN architectures for joint depth prediction and semantic segmentation. In *Proceedings - IEEE International Conference on Robotics and Automation*, pages 4620–4627, 2017.

[27] Suyog Jain, Bo Xiong, and Kristen Grauman. FusionSeg: Learning to combine motion and appearance for fully automatic segmention of generic objects in videos. *ICCV*, 2017.

[28] L'ubor Ladický, Bernhard Zeisl, and Marc Pollefeys. Discriminatively Trained Dense Surface Normal Estimation. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 468–484, Cham, 2014. Springer International Publishing.

[29] Byungjae Lee, Enkhbayar Erdenee, Songguo Jin, Mi Young Nam, Young Giu Jung, and Phill Kyu Rhee. Multi-class multi-object tracking using changing point detection. In *ECCV 2016 Workshops*, volume 9914 LNCS, pages 68–83, aug 2016.

[30] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 3431–3440, 2015.

[31] G Mori, X Ren, A. Efros, and J Malik. Recovering human body configurations: combining segmentation and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[32] J Revaud, P Weinzaepfel, Z Harchaoui, and C Schmid. EpicFlow: Edge-preserving interpolation of correspondences for optical flow. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1164–1172, jun 2015.

[33] Laura Sevilla-Lara, Deqing Sun, Varun Jampani, and Michael J Black. Optical Flow With Semantic Segmentation and Localized Layers. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2016.

[34] J Shotton, J Winn, C Rother, and A Criminisi. TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context. *International Journal of Computer Vision, 95(2)*, pages 2–23, 2009.

[35] Xiaolong Wang, David F Fouhey, and Abhinav Gupta. Designing Deep Networks for Surface Normal Estimation. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[36] Andreas Wedel and Daniel Cremers. *Optical Flow Estimation*, pages 5–34. Springer London, London, 2011.

[37] Philippe Weinzaepfel, Jerome Revaud, Zaid Harchaoui, and Cordelia Schmid. DeepFlow: Large Displacement Optical Flow with Deep Matching. In *Proceedings of the 2013 IEEE International Conference on Computer Vision*, ICCV '13, pages 1385–1392, Washington, DC, USA, 2013. IEEE Computer Society.

[38] L Xu, J. Jia, and Y Matsushita. Motion Detail Preserving Optical Flow Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1744–1757, 2012.

[39] F Yu and V Koltun. Multi-Scale Context Aggregation by Dilated Convolutions. In *International Conference on Learning Representations*, 2016.

[40] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid Scene Parsing Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 310–313, 2017.

[41] Xizhou Zhu, Yujie Wang, Jifeng Dai, Lu Yuan, and Yichen Wei. Flow-Guided Feature Aggregation for Video Object Detection. In *ICCV*, 2017.

[42] Xizhou Zhu, Yuwen Xiong, Jifeng Dai, Lu Yuan, and Yichen Wei. Deep Feature Flow for Video Recognition. In *CVPR*, 2017.