SMSnet: Semantic Motion Segmentation using Deep Convolutional Neural Networks

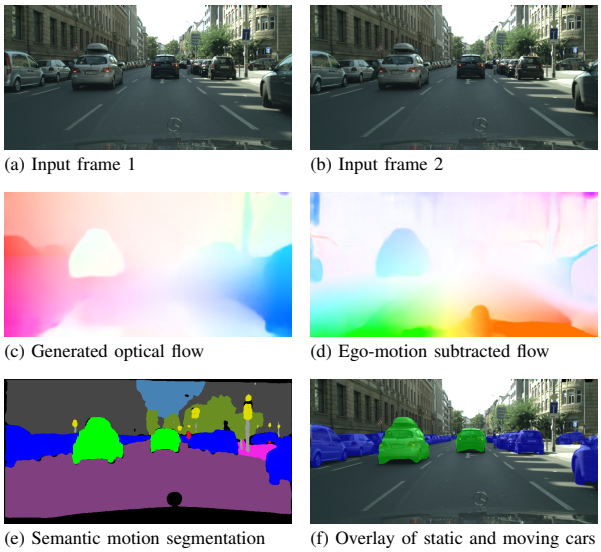
Johan Vertens* Abhinav Valada* Wolfram Burgard

***Abstract*— Interpreting the semantics and motion of objects are prerequisites for autonomous robots that enable them to reason and operate in dynamic real-world environments. Existing approaches that tackle the problem of semantic motion segmentation consist of long multistage pipelines and typically require several seconds to process each frame. In this paper, we present a novel convolutional neural network architecture that learns to predict both the object label and motion status of each pixel in an image. Given a pair of consecutive images, the network learns to fuse features from self-generated optical flow maps and semantic segmentation kernels to yield pixel-wise semantic motion labels. We also introduce the Cityscapes-Motion dataset which contains over 2,900 manually annotated semantic motion labels, which is the largest dataset of its kind so far. We demonstrate that our network outperforms existing approaches achieving state-of-the-art performance on the KITTI dataset, as well as in the more challenging Cityscapes-Motion dataset while being substantially faster than existing techniques.**

I. Introduction

The advancement in robotic technology and machine learning has led to the successful use of robots to accomplish tasks in various structured and semi-structured environments such as factory floors, domestic homes and offices. This recent success has now paved the way to tackle more complex tasks in challenging urban environments that contain many dynamic objects. Thus scene understanding plays a crucial role in ensuring the viability and safe operation in such scenarios. The ability to classify, segment and infer the state of motion of dynamic objects such as cars and pedestrians, allows robotic systems to increase their awareness, reason about behaviours and plan autonomous actions. Previous work [20], [7] has successfully shown the benefit of solving the problem of semantic motion segmentation jointly, as features learned for semantic labelling can help infer motion labels and vice versa. However, the multistage pipelines currently in use have long processing times deeming them impractical for real-world applications.

Deep convolutional neural network (DCNN) based approaches have significantly improved the state-of-art in both semantic segmentation [16] and motion estimation [10]. Yet, their applicability to the task of joint semantic motion segmentation has not been explored. There are several challenges that make this problem inherently hard including the ego-motion of the camera, lighting changes between consecutive frames, motion blur and varying pixel displacements due to motion with different velocities. Another major hindrance is the lack of a large enough dataset with ground truth semantic motion annotations that enable training of deep networks and allow for credible quantitative evaluations.



(a) Input frame 1 (b) Input frame 2 (c) Generated optical flow (d) Ego-motion subtracted flow (e) Semantic motion segmentation (f) Overlay of static and moving cars

Fig. 1. Illustration of semantic motion segmentation using our proposed SMSnet on the Cityscapes dataset. (e) shows the output from our network consisting of semantic and motion labels, while (f) shows the segmented moving cars in green and static cars in blue for better visualization.

In this work, we propose a composite deep convolutional neural network architecture that learns to predict both the semantic category and motion status of each pixel from a pair of consecutive monocular images. The composition of our SMSnet architecture can be deconstructed into three components: a section that learns motion features from generated optical flow maps, a parallel section that generates features for semantic segmentation, and a fusion section that combines both the motion and semantic features and further learns deep representations for pixel-wise semantic motion segmentation. For this work, we consider an urban driving scenario containing moving cars that appear in different scales. Therefore we utilize our previously proposed multi-scale ResNet skip layers [23] in the architecture to incorporate scale invariance. Training such a network requires several thousands of labelled consecutive image pairs. Currently the only publicly available semantic motion segmentation dataset contains 200 labelled images from the KITTI [9] benchmark, which is highly insufficient for training networks of this scale. To overcome this impediment, we release the Cityscapes-Motion dataset containing over 2,900 labelled images and a KITTI-Motion dataset with 255 labelled images, both with additional preceding frames. Additionally, we investigate the utility of combining these datasets to

*These authors contributed equally. All authors are with the Department of Computer Science, University of Freiburg, Germany. This work has been supported by Samsung Electronics Co., Ltd. under the GRO program.

measure the generalization capabilities to scenes of unseen cities. Utilizing efficient GPU implementations, our approach is several times faster than existing techniques and achieves state-of-the-art performance on multiple datasets.

II. Related Work

Semantic segmentation and motion segmentation are two fundamental problems in scene understanding that both have substantial amount of literature in their areas. Early convolutional neural network (CNN) based segmentation approaches involve small model capacities, multi-scale pyramid processing, saturating tanh non-linearities and post-processing such as superpixel computation, filtering and random field regularization. A recent breakthrough that does not require any pre or post processing was proposed by Long *et al.* [16], in which they extend a CNN designed for classification with learned deconvolution layers that are able to upsample low-resolution feature maps to higher resolution segmentation outputs. Since then, several improvements to this fully convolutional network (FCN) architecture have been proposed that improve the resolution of segmentation with additional refinement layers [18], alternative schemes for non-linear upsampling eliminating the need for learning to upsample [1], and deeper networks based on the residual learning framework that incorporates scale invariance [23].

There are numerous approaches that have been proposed for segmenting moving objects from stationary camera images [22], [8]. However, they cannot be directly applied to moving camera images, as the movement causes a dual motion appearance which consists of the background motion and the object motion. In general, methods that detect motion from freely moving cameras partition the image into coherent regions with homogenous motion. This process splits the image into background and moving clusters. These methods can be categorized into optical flow based and tracking based approaches. Optical flow based techniques [19], [21] check if the motion speed and direction of a region is consistent with its radially surrounding pattern. It is then classified as a moving object if the motion of this region deviates from this pattern. The disadvantage of these methods is that they are prone to occlusion, noise in the optical flow map and edge effects. In recent work [12], the authors derive a geometric model that relates 2D motion to a 3D motion field relative to the camera based on estimated depth and motion of vanishing points in the scene. Spectral clustering is then applied on the recovered 3D motion field to obtain the moving object segmentation. Although qualitative evaluations have been shown on the KITTI benchmark, no quantitative comparison has been reported. Tracking based techniques [3], [5], [14], [13] on the other hand, aim to detect and localize target objects in successive frames. Tracking of objects yields movement trajectories and by estimating the ego-motion of the camera, objects can be segmented from the background motion. These approaches typically have long processing pipelines resulting in high computation times and coarse segmentations.

There is only a handful of work that jointly estimates the semantic motion labels, mostly accounting to the last half a decade. Chen *et al.* [2] propose an approach that detects object-level motion from a moving camera using two consecutive image frames and provides 2D bounding boxes as the output. They design a robust context-aware motion descriptor that considers moving speed, as well as the direction of objects and combines them with an object classifier. The descriptor measures the inconsistency between local optical flow histograms of objects and their surroundings, giving a measure of the state of motion.

Dinesh *et al.* [20] propose an approach that generates motion likelihoods based on depth and optical flow estimations, while combining them with semantic and geometric constraints within a dense conditional random field (CRF). More recently, a multistep framework was proposed in [7], where first sparse image features in two consecutive stereo image pairs are extracted and matched. The matched feature points are then classified using RANSAC into inliers caused by the camera and outliers caused by moving objects. Following which, the outliers are clustered in a U-disparity map which provides the motion information of objects. Finally, a dense CRF is used to merge the motion information with the semantic segmentation provided by a FCN. A major disadvantage of these approaches is their long run-times that range from several seconds to even minutes, making them unusable for applications that require near real-time performance such as autonomous driving. In contrast to these existing multistage techniques, we propose an approach that is entirely based on convolutional neural networks, composing of a simpler but deeper structure while being more accurate and several orders faster than existing techniques.

III. Semantic Motion Segmentation

In this section, we first formulate the problem statement for semantic motion segmentation. We then describe our SMSnet architecture in detail and the training procedure that we employ. We represent the training set as $S = \{(X_{n-1}, X_n, Y_n), n = 1, \ldots, N\}$, where $X_n = \{x_j, j = 1, \ldots, |X_n|\}$ denotes the input frame and X_{n-1} denotes the preceding frame. The corresponding ground truth mask can be denoted as $Y_n = \{y_j, j = 1, \ldots, |X_n|\}, y_j \in \{0\} \cup \mathscr{C} \times \mathscr{M}$, where $\mathscr{C} = \{1, ..., C\}$ is the set of C semantic classes and each class can also take the label of static or moving $\mathscr{M} = \{m_1, m_2\}$. Where, m_1 denotes a static pixel and m_2 denotes a moving pixel. Let θ be the network parameters and $a = f(x_j; \theta)$ be the activation function. The goal of our network is to learn semantic motion features by minimizing the cross-entropy $(softmax)$ loss that can be computed as $\mathscr{L}(k) = -\frac{\exp(a_k)}{\sum_{l=1}^{C} \exp(a_l)}$. Using stochastic gradient descent, we then solve

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N \times |X_n|} \mathscr{L}(f(x^i; \theta), y^i) \tag{1}$$

A. Network Architecture

We propose a novel fully convolutional neural network architecture that represents the sought $f(x_j; \theta)$. Figure 2 depicts our SMSnet architecture for semantic motion segmentation. The network consists of three different streams:

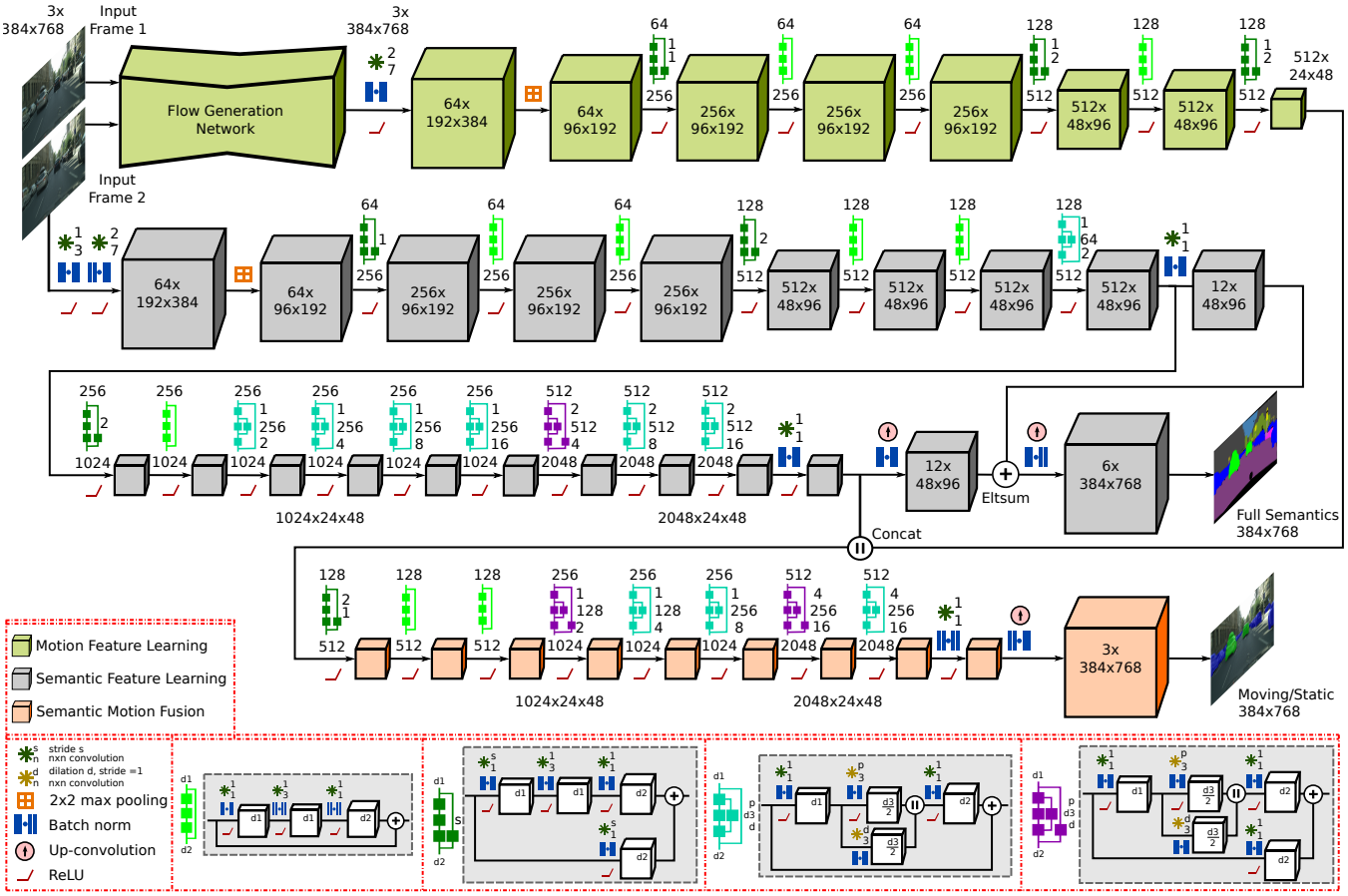


Fig. 2. Depiction of the proposed SMSnet architecture for semantic motion segmentation from two consecutive images as input. The stream shown in green learns deep motion features and in parallel the stream in gray learns semantic features, which are then both concatenated and further fused representations are learned in the stream depicted in orange. The legend for the network architecture is shown with a red outline.

Motion Feature Learning, *Semantic Feature Learning*, and *Semantic Motion Fusion*. The following sections describe each of these streams in detail.

a) Motion Feature Learning: This stream generates features that represent motion specific information. Successive frames (x_{j-1}, x_j) are first passed through a section of this stream that generates high quality optical flow maps $\hat{X}$. In this work, we embed the recently proposed deep convolutional architecture FlowNet2 [10] for this purpose. However, any network with the ability to generate optical flow maps can be embedded in place. The flow generation network yields the optical flow in the x and y direction and in addition we also compute the magnitude of the flow. This output tensor is of size $3 \times 384 \times 768$, which is the same dimensions as the input RGB images. Figure 1 (c) shows a generated optical flow image from this section, while the consecutive input frames are shown in Figure 1 (a) and (b).

Moving objects appear as motion patterns that differ in scale, geometry and magnitude. In order to enable the network to reason about object class and its borders, we further convolve and pool the optical flow features through multiple network blocks. These additional network blocks can be represented as a function $f_o(\hat{X}; \theta_o) \mid \theta_o \subset \theta$ of the optical flow maps yielding a feature map tensor of size $512 \times 24 \times 48$.

b) Semantic Feature Learning: The final output of our network is a combined label that denotes a semantic class $\mathscr{C}$ and the state of motion $\mathscr{M}$. While the stream described in the previous section yields information about the motion in the scene, the network still requires semantic features to learn the combined semantic motion segmentation. The semantic feature learning stream depicted in gray blocks in Figure 2 takes as input the image x_j and generates semantic features $f_s(x_j; \theta_s) \mid \theta_s \subset \theta$. The structure of this stream is similar to our previously proposed unimodal AdapNet [23] architecture for semantic segmentation. The architecture follows the design of a contractive segment that aggregates semantic information while decreasing the spatial dimensions of the feature maps and an expansive segment that upsamples the feature maps back to the full input resolution. The architecture incorporates many recent improvements including multiscale ResNet blocks that learn scale invariant deep features, skip connections that enable training of the deep architecture and dilated convolutions that enable the integration of information from different spatial scales. In our proposed SMSnet, the low resolution features from the last layer of the contractive segment are fused with the learned motion features in the Semantic Motion Fusion stream that follows. The expansive segment then in parallel yields the full semantic labels for the input frame x_j.

c) Semantic Motion Fusion: The final stream in the SMSnet architecture depicted using orange blocks in Figure 2, fuses the complementary motion and semantic features which are generated in the aforementioned streams of the network. The feature tensors from $f_o(\hat{X};\theta_o)$ and $f_s(x_j;\theta_s)$ are concatenated and further deep representations are learned through a series of additional layers. No further pooling is performed on these features and therefore a downsampling factor of 16 is maintained in comparison to the input x_j. Similar to the semantic feature learning stream, multiscale ResNet blocks from [23] that utilize dilated convolutions for aggregating information over different field of views are used in the layers that follow the concatenation segment. Finally, towards the end of this stream, we use deconvolution, also known as transposed convolution, for upsampling the low resolution feature maps from $2048 \times 24 \times 48$ back to the input resolution of $|\mathscr{C}| \times |\mathscr{M}| \times 384 \times 768$. This upsampled output has joint labels in $\mathscr{C} \times \mathscr{M}$ corresponding to a semantic class and a motion status: static or moving. Thus the final activation function of the SMSnet is given by:

$$f(x^i;\theta) = f_m(f_o(\hat{X};\theta_o), f_s(x_j;\theta_s);\theta_f) \mid \theta_o, \theta_s, \theta_f \subset \theta \quad (2)$$

B. Introducing Ego-Flow Suppression

Movement of the camera leads to ego-motion introducing additional optical flow magnitudes that are not induced by moving objects. This induced flow can cause ambiguities since objects can appear with high optical flow magnitudes although they are not moving. In order to circumvent this problem, we propose a further variant of the SMSnet that predicts the optical flow map $\hat{X}'$ which is purely caused by the ego-motion. We first estimate the backward camera translation T and the rotation matrix R from the position at the current frame x_j to the previous frame x_{j-1}. Using IMU and odometry data we can then estimate $\hat{X}'$ as:

$$\hat{X}' = KRK^{-1}X + K\frac{T}{z} \quad (3)$$

where, K is the intrinsic camera matrix, $X = (u, v, 1)^T$ is the homogenous coordinate of the pixel in image coordinates and z is the depth of the corresponding pixel in meters. Calculating the flow vector for every pixel coordinate yields the 2-dimensional optical flow image which purely represents the ego-motion. For estimating the depth z, we use the recently proposed DispNet [17] which is based on DCNNs and has fast inference times. We then subtract the ego-flow $\hat{X}'$ from the optical flow calculated by the embedded flow generation network $\hat{X}$ within the SMSnet architecture. This subtraction yields to suppression of the ego-flow while keeping the flow magnitudes evoked from other moving objects. An example of the optical flow with ego-flow suppression (EFS) is shown in Figure 1 (d). We present evaluations on both variants of our SMSnet, without and with EFS in Section V-A.

C. Training

We train our network on a system with an Intel Xeon E5 with 2.4 GHz and four NVIDIA TITAN X. We first train the Semantic Feature Learning stream in SMSnet that generates semantic features for all the C classes. Consecutively, we train the embedded flow generation network that produces the optical flow maps which are processed and generated in the SMSnet architecture. Finally, we train the entire SMSnet while keeping the weights of the semantic feature learning stream and the flow generation network fixed. We train the network with an initial learning rate $\lambda_0 = 10^{-7}$ and with the poly learning rate policy as, $\lambda_N = \lambda_0 \times \left(1 - \frac{N}{N_{\max}}\right)^c$, where λ_N is the current learning rate, N is the iteration number, $N_{\max}$ is the maximum number of iterations and c is the power. We train using stochastic gradient decent with a momentum of 0.99 and a mini-batch of 2 for $50,000$ iterations which takes about a day to complete.

IV. Dataset

One of the main requirements to train a neural network is a large dataset with ground truth annotations. Data augmentation can help expand datasets but for training a network from scratch and optimizing a network with millions of parameters, thousands of labelled images are required. While there are several large datasets for various scene understanding problems such as classification, segmentation and detection, for the task of semantic motion segmentation however, there only exists one public dataset [9] with 200 labelled images which is highly insufficient for training DCNNs. Obtaining ground truth for pixel-wise motion status is particularly hard as visible pixel displacement quickly decreases with increasing distance from the camera. In addition, any ego-motion can make the labelling an arduous task. To facilitate training of neural networks for semantic motion segmentation and to allow for credible quantitative evaluation, we create the following datasets and make them publicly available at `http://deepmotion.cs.uni-freiburg.de/`. Each of these datasets have pixel-wise semantic labels for 10 object classes and their motion status (static or moving). Annotations are provided for the following classes: *sky, building, road, sidewalk, cyclist, vegetation, pole, car, sign and pedestrian*.

KITTI-Motion: The KITTI benchmark itself does not provide any semantic or moving object annotations. Existing research on semantic motion segmentation has been benchmarked using the annotations for 200 images from the KITTI dataset provided by [9], however, there are no images with annotations that can be used for training learning based approaches. In order to train our neural network, we create a KITTI-Motion dataset consisting of 255 images taken from the KITTI Raw dataset and which do not intersect with the test set provided by [9]. The images are of resolution 1280×384 pixels and contain scenes of freeways, residential areas and inner-cities. We manually annotated the images with pixel-wise semantic class labels and moving object annotations for the category of cars. In addition, we combine two publicly available KITTI semantic segmentation datasets [6] and [24] for pretraining the semantic stream of our network, which yields a total of 253 images. These images also do not overlap with the test set [9] or the KITTI-Motion dataset that we introduced.

Cityscapes-Motion: The Cityscapes dataset [4] is a more recent dataset containing 2975 training images and 500 validation images. Semantic annotations are provided for 30 categories and images are of resolution 2048×1024 pixels. The Cityscapes dataset is highly challenging as it contains images from over 50 cities and different weather conditions, varying seasons and many dynamic objects. We manually annotated all the Cityscapes images with motion labels for the category of cars. We use this dataset in addition to KITTI-Motion for benchmarking the performance.

City-KITTI-Motion: As the KITTI-Motion dataset by itself is not sufficient to train deep networks and to facilitate comparison with other approaches that are evaluated on KITTI data, we merge the KITTI-Motion and Cityscapes-Motion training sets. Additionally, we merge the 200 image KITTI test set [9] with the 500 validation images from Cityscapes to compose a corresponding evaluation set. Combining them also helps the network learn more generalized feature representations. As we use an input resolution of 768×384 for our network, we downsample the Citiscapes-Motion images to this size. However, as the images in the KITTI-Motion dataset have wider resolution 1280×384, we slice each image into three partially overlapping images. In total the combined dataset yields 3734 training images and 1100 for validation. Furthermore, the dataset also contains 15 preceding frames for every annotated image and is thus perfectly suited for sequence based approaches.

In order to create additional training data, we randomly apply the following augmentations on the training images: rotation, translation, scaling, vignetting, cropping, flipping, color, brightness and contrast modulation. As the SMSnet takes two consecutive images as input, we augment the pair jointly with the same parameters.

V. Experimental Results

For the network implementation, we use the Caffe [11] deep learning library with cuDNN backend for acceleration. We quantify the performance using the standard Jaccard Index which is commonly known as average intersection-over-union (IoU) metric. It can be computed as $IoU = TP/(TP + FP + FN)$, where TP, FP and FN correspond to true positives, false positives and false negatives respectively.

A. Baseline Comparison

In order to compare the performance of our network with state-of-the-art techniques, we train our network on the combined City-KITTI-Motion dataset and benchmark its performance on the KITTI set from [9] on which the other approaches have reported their results. We compare the motion segmentation against three state-of-the-art techniques including geometric-based motion segmentation (GEO-M) [13], joint labelling of motion and superpixels based image segmentation (AHCRF+Motion) [14] and CRF-based semantic motion segmentation [20]. Table I summarizes the results of this experiment and shows the average IoU of the moving object, static object and background classes. Other approaches consider all the elements in the scene that are movable but not moving such as a stationary car and permanently static elements such as buildings to be under the same static class, which we denote as background in our evaluations. However, as it is more informative in the context of robotics to split these two cases into different categories, we consider the static class to only contain objects that are movable but are stationary at that time.

TABLE I
Comparison of motion segmentation performance with state-of-the-art approaches on the KITTI dataset.

Approach	IoU		
	Moving	Static	Background
GEO-M [13]	46.50	N/A	49.80
AHCRF+Motion [14]	60.20	N/A	75.80
CRF-M [20]	73.50	N/A	82.40
SMSnet 10-class	73.98	80.28	97.65
SMSnet 10-class with EFS	80.87	83.77	97.84
SMSnet 2-class	74.03	80.78	97.59
SMSnet 2-class with EFS	**84.69**	**84.50**	**98.01**

It can be seen that the method that jointly predicts the semantic class and motion (CRF-M) substantially outperforms approaches that perform only motion segmentation (GEO-M and AHCRF+Motion). This can be attributed to the fact that these approaches learn to correlate motion features with the learned semantic features which improves the overall motion segmentation accuracy. Intuitively, the approaches learn that there is a higher probability of a car moving than a building or a pole. Although Fan *et al.* [7] also propose an approach for semantic motion segmentation, the KITTI scene flow dataset that they evaluate on have inconsistent class labels which does not allow for meaningful comparison. Finally, we show the performance using variants of our proposed SMSnet architecture, specifically, with and without the subtraction of the optical flow induced by the ego-motion (ego-flow), as well as considering all the semantic classes in KITTI and considering only the semantic classes that are potentially moveable. All the SMSnet variants shown in Table I outperform the existing approaches, while our best performing models achieve the state-of-the-art performance of 84.69% for the moving classes, 84.50% for the static classes and 98.01% for the background class. It can be observed the subtraction of the ego-flow helps in improving the moving object segmentation.

Since we are interested in predicting both the motion status and the semantic label, we show the performance of semantic segmentation in comparison to recent neural network based approaches in Table II. As described in Section IV, the KITTI benchmark does not provide any official ground truth for semantic segmentation, therefore to train the semantic stream of our network, we combine the Cityscapes dataset with the KITTI semantic ground truth from [6] and [24] to obtain the most generalized training set. We then test the performance individually on the Cityscapes test set, as well as on the KITTI semantic motion test set that was also used in the motion segmentation comparison. For the experiments on the KITTI semantic motion test set, we observe that our SMSnet outperforms the other approaches for most of the semantic classes. Secondly, the KITTI semantic motion

TABLE II

COMPARISON OF SEMANTIC SEGMENTATION PERFORMANCE WITH STATE-OF-THE-ART APPROACHES ON THE KITTI AND CITYSCAPES DATASETS.

Test Set	Approach	Sky	Building	Road	Sidewalk	Cyclist	Vegetation	Pole	Car	Sign	Pedestrian
KITTI	FCN-8s [16]	77.35	74.24	74.41	**51.41**	35.79	78.80	15.99	76.20	35.97	40.87
	SegNet [1]	77.27	60.34	75.03	43.62	19.76	76.58	24.34	63.88	17.01	21.96
	ParseNet [15]	**81.26**	70.42	73.85	42.12	**41.04**	71.48	32.02	77.20	31.60	**47.49**
	SMSnet (ours)	78.39	**74.27**	**78.10**	46.11	26.85	**79.88**	**34.84**	**83.63**	**37.70**	42.88
Cityscapes	FCN-8s [16]	76.05	75.94	92.73	59.68	46.50	78.78	15.27	76.54	37.96	41.57
	SegNet [1]	69.93	59.87	83.25	43.35	27.25	68.83	19.23	60.80	23.81	23.14
	ParseNet [15]	77.58	76.23	92.76	60.04	47.96	79.68	22.66	76.85	40.99	44.54
	SMSnet (ours)	**85.43**	**81.08**	**94.50**	**66.89**	**49.26**	**84.85**	**37.92**	**82.40**	**47.48**	**46.47**

test set consists of images containing sidewalks with outgrown grass labelled as sidewalk as opposed to vegetation. Such examples are consistently labelled as vegetation in the Cityscapes dataset, consequently causing misclassification. Whereas, while testing on the Cityscapes test set, our proposed SMSnet substantially outperforms other networks in all the classes.

B. Influence of Range on Motion Segmentation Accuracy

In this section, we investigate the performance of motion segmentation using SMSnet to various ranges within which the moving objects might lie. One of the primary challenges is learning motion features of moving objects that are at far away distances from the camera, as the appearance of the object and the pixel displacement are both very small. To quantify this influence, we train models on examples that have moving objects within certain maximum distance from the camera and objects that lie beyond this distance are ignored for training. We then evaluate each of these models on test sets containing moving objects at varying distances. On the one hand, including training examples that are far away might enable learning of more multiscale features that can cover a wide variety of motion appearances, but on the other hand these highly difficult training examples can also confuse the training if the network is unable to learn features that can distinguish the state of distant objects. For this experiment, we train models on the City-KITTI-Motion dataset and also evaluate on the corresponding test set as we want the evaluation to generalize over both the Cityscapes and KITTI datatsets. The results of this experiment are shown in Figure 3. As hypothesized, the best accuracy is obtained for a maximum distance of 20 m and the accuracy gradually decreases with increasing maximum distance. The best trade-off is obtained for a maximum distance of 40 m. It can also be seen that the model trained with the maximum distance at infinity performs impressively well even for challenging moving object examples that are at far away distances.

C. Generality of the Network to Different Datasets

A large amount of good quality training data that encompasses the possible scenarios is essential for successful training of neural networks. Ideally, the training dataset should generalize to previously unseen scenes. In this section, we investigate the performance of models trained on various

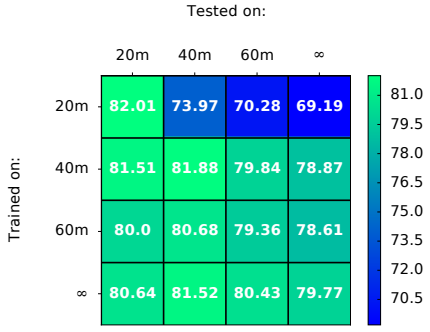


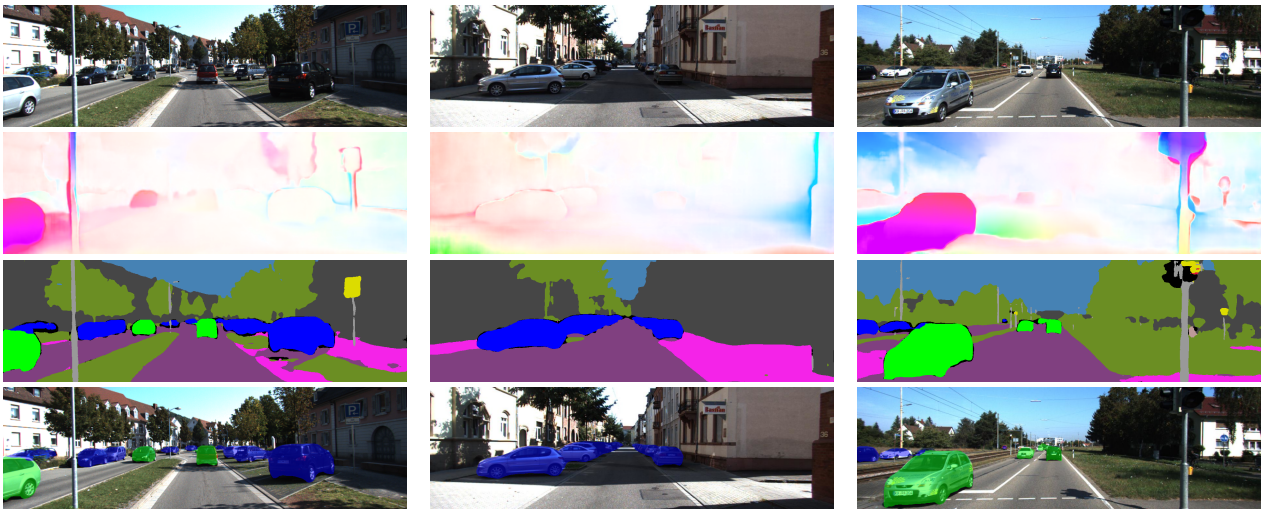
Fig. 3. Moving object segmentation performance of our proposed SMSnet while considering objects within different maximum ranges.

public datasets, including our newly proposed City-KITTI-Motion dataset and evaluate its efficacy on test sets from complementary datasets. Specifically, we train our network individually on KITTI, Cityscapes and the combined City-KITTI-Motion dataset and evaluate each of them on all their individual test sets. For this experiment, we use the 2-class model with EFS trained and evaluated with a maximum range of 40 m. In Table III we summarize the results from this experiment and show the mean IoU for the static and moving classes. It can be seen in Table III that the model trained on the combined City-KITTI dataset performs well on both the Cityscapes and KITTI test sets than the models trained on their individual counterparts. The models trained only the KITTI dataset or only on the Cityscapes dataset have a substantially lower performance when they are tested on the Cityscapes or KITTI test set respectively than the model trained on the City-KITTI-Motion dataset. This shows the utility of combining these datasets and the good generalization that it provides.

TABLE III

COMPARISON OF MODELS TRAINED - TESTED ON DIFFERENT DATASETS.

Trained On	Static		Moving	
	KITTI	Cityscapes	KITTI	Cityscapes
KITTI	78.51	52.05	70.32	34.84
Cityscapes	61.29	84.27	51.65	75.31
City-KITTI	86.10	84.03	87.02	72.78



(a) (b) (c)

Fig. 4. Qualitative semantic motion segmentation results on the KITTI dataset. In each column top to bottom: input frame 2, corresponding ego-motion subtracted flow, semantic motion segmentation output and segmented static and moving cars overlay. The network accurately segments and classifies moving objects of different scales and at varying velocities, in addition to pixel-wise classification of the entire scene.

D. Prediction Time Comparison

Fast prediction time is one of the most critical requirements for perception algorithms used in real-world robotics. Therefore, we designed SMSnet while keeping this critical factor in mind. To the best of our knowledge, there exists two alternate approaches that perform semantic motion segmentation to which we compare our inference time with in Table IV. Our proposed SMSnet takes 153 ms to predict a single frame which is significantly faster that the other two existing approaches. The SMSnet with ego-flow subtraction takes 313 ms which includes the disparity prediction and solving Equation 3. In contrast to existing approaches both variants of our proposed SMSnet enable interactive speeds which is a prerequisite for robotic applications.

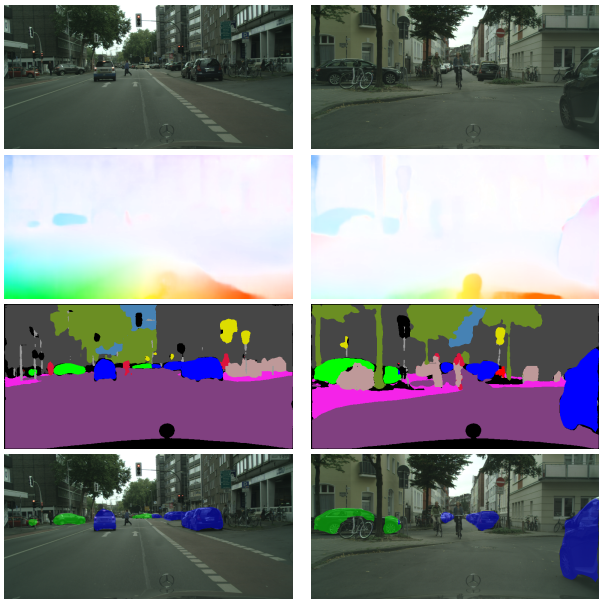
TABLE IV

COMPARISON OF PREDICTION TIME WITH STATE-OF-THE-ART.

Approach	Time
CRF-M [20]	240,000 ms
U-Disp-CRF-FCN [7]	1,060 ms
SMSnet (ours)	**153** ms
SMSnet with EFS (ours)	313 ms

E. Qualitative Evaluation

In this section, we show qualitative results on various datasets with our SMSnet trained on City-KITTI-Motion and critique its performance in diverse scenes. Figure 4 shows results on images from the KITTI test set. The segmented images are color coded according to the labels shown in Table II. Dark blue pixels indicate static cars and light green pixels indicate moving cars. Figure 4 (a) and (b) are scenes from residential areas which have cars moving with low velocities and Figure 4 (c) shows a scene on a highway which has cars moving at much higher velocities. These scenes also have objects of different scales and lighting conditions. We can see that the network accurately segments the scene and distinguishes between the static and moving cars even in these diverse situations. Figure 5 presents results on the Cityscapes test set which contains more complex scenes than the KITTI dataset. Figure 5 (a) shows a moving car over 80 m away and SMSnet succeeds in capturing this motion while precisely segmenting the object. Figure 5 (b) shows a scene with a moving car that is partially occluded by a tree, yet the entire car is captured in the segmentation. This demonstrates the ability of the SMSnet to handle diverse real-world scenarios.



(a) (b)

Fig. 5. Qualitative semantic motion segmentation results on the Cityscapes dataset. The network demonstrates robustness to complex scenes with many different dynamic objects, some that are even partially occluded.

F. Evaluation of Transferability and Platform Independence

In this section, we demonstrate the platform independence of our SMSnet model trained on the City-KITTI-Motion

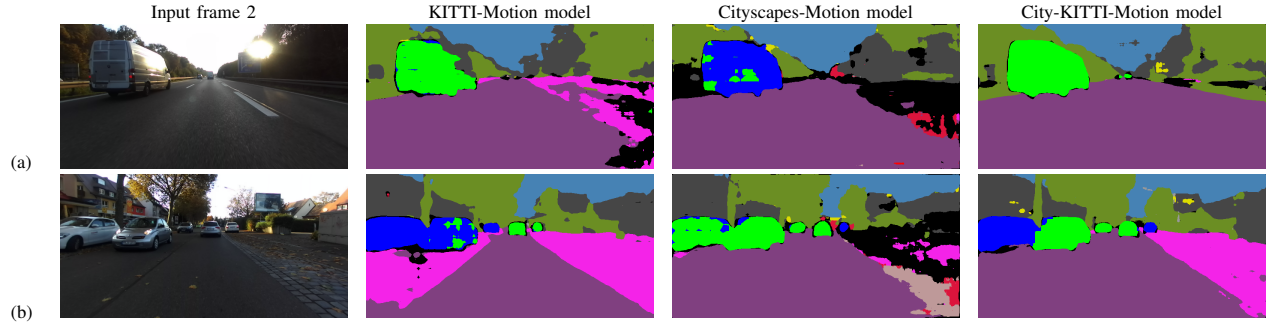


Fig. 6. Qualitative comparison of semantic motion segmentation models trained on various datasets and evaluated on real-world data from Freiburg. Note that the model trained on the City-KITTI-Motion dataset generalizes better to the previously unseen city than others. Our network robustly handles challenging conditions such as glare (a) and low lighting (b).

dataset by presenting qualitative evaluations on images captured using a different camera setup than those used in KITTI and Cityscapes datasets. We mounted a ZED stereo camera on the hood of a car and collected over 61,000 images of driving scenes in Freiburg, Germany. The recorded images comprise of adverse conditions such as low lighting, glare and motion blur, which pose a great challenge for semantic motion segmentation. Figure 6 shows a comparison of results using the SMSnet trained on KITTI-Motion, Cityscapes-Motion and the combined City-KITTI-Motion datasets. In Figure 6 (a), we see that the Cityscapes-Motion model misclassifies the car as static while it is moving and it has false positives on the sides of the image due to motion blur. The KITTI-Motion model on the other hand, segments the car as moving but fails to segment it as a whole, in addition to having numerous false positives. Figure 6 (b) shows a residential scene in low lighting. We see that the KITTI-Motion model misclassifies the moving car in the left as static and the Cityscapes-Motion model misclassies the static car as moving. Both these models also have a difficulty segmenting the sidewalk entirely. Overall, one can note that the model trained on City-Kitti-Motion performs substantially better in segmenting the static and moving classes as well as having negligible false positives, demonstrating the good generality of the learned kernels.

VI. Conclusion

In this paper, we presented a convolutional neural network that takes as input two input images and learns to predict both the semantic class label and motion status of each pixel in an image. We introduced two large first-of-a-kind datasets with ground-truth annotations that enable training of deep neural networks for semantic motion segmentation. We presented comprehensive quantitative evaluations and demonstrated that the performance of our network exceeds the state of the art, both in terms of accuracy and prediction time. We investigated the performance of motion segmentation to varying object distances and showed that our network performs well even for distant moving objects. We also presented extensive qualitative results that show the applicability to autonomous driving scenarios. Furthermore, we presented qualitative evaluations of various SMSnet models on real-world driving data from Freiburg that contain challenging perceptual conditions and showed that the model trained on our City-KITTI-Motion dataset generalized effectively to previously unseen conditions.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv: 1511.00561*, 2015.

[2] T. Chen and S. Lu, "Object-level motion detection from moving cameras," *TCSVT*, vol. PP, no. 99, pp. 1–1, 2016.

[3] W. Choi, C. Pantofaru, and S. Savarese, "A general framework for tracking multiple people from a moving camera," *PAMI*, 2013.

[4] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *CVPR*, 2016.

[5] B. Drayer and T. Brox, "Object detection, tracking, and motion segmentation for object-level video segmentation," *arxiv:1608.03066*, 2016.

[6] G. R. et al., "Vision-based offline-online perception paradigm for autonomous driving," in *WACV*, 2015.

[7] Q. Fan *et al.*, "Semantic motion segmentation for urban dynamic scene understanding," in *CASE*, 2016.

[8] P. Gao, X. Sun, and W. Wang, "Moving object detection based on kirsch operator combined with optical flow," in *IASP*, 2010.

[9] N. Haque, D. Reddy, and K. M. Krishna, "Kitti semantic ground truth," https://github.com/native93/KITTI-Semantic-Ground-Truth/, 2016.

[10] E. Ilg *et al.*, "Flownet 2.0: Evolution of optical flow estimation with deep networks," *arXiv:1612.01925*, Dec 2016.

[11] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[12] J. Y. Kao *et al.*, "Moving object segmentation using depth and optical flow in car driving sequences," in *ICIP*, 2016, pp. 11–15.

[13] A. Kundu *et al.*, "Moving object detection by multi-view geometric techniques from a single camera mounted robot," in *IROS*, 2009.

[14] T. H. Lin and C. C. Wang, "Deep learning of spatio-temporal features with geometric-based moving point detection for motion segmentation," in *ICRA*, May 2014, pp. 3058–3065.

[15] W. Liu, A. Rabinovich, and A. C. Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv: 1506.04579*, 2015.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[17] N.Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *CVPR*, 2016.

[18] G. Oliveira, A. Valada, C. Bollen, W. Burgard, and T. Brox, "Deep learning for human part discovery in images," in *ICRA*, 2016.

[19] M. P. Patel and S. K. Parmar, "Moving object detection with moving background using optic flow," in *ICRAIE*, 2014.

[20] N. D. Reddy, P. Singhal, and K. M. Krishna, "Semantic motion segmentation using dense CRF formulation," in *ICVGIP*, 2014.

[21] C. S. Royden and K. D. Moore, "Use of speed cues in the detection of moving objects by moving observers," *Vision Research*, vol. 59, pp. 17 – 24, 2012.

[22] P. Spagnolo, T. Orazio, M. Leo, and A. Distante, "Moving object segmentation by background subtraction and temporal analysis," *Image Vision Comput.*, vol. 24, no. 5, pp. 411–423, May 2006.

[23] A. Valada *et al.*, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *ICRA*, 2017.

[24] P. Xu *et al.*, "Multimodal information fusion for urban scene understanding," *Machine Vision and Applications*, pp. 331–349, 2016.