Joint Semantic and Motion Segmentation for Dynamic Scenes using Deep Convolutional Networks

Nazrul Haque, Dinesh Reddy and K. Madhava Krishna
International Institute of Information Technology, Hyderabad, India
nazrul.athar@research.iiit.ac.in, dinesh.andromeda@gmail.com, mkrishna@iiit.ac.in

Keywords: Monocular Semantic Motion Segmentation, Scene Understanding, Convolutional Neural Networks.

Abstract: Dynamic scene understanding is a challenging problem and motion segmentation plays a crucial role in solving it. Incorporating semantics and motion enhances the overall perception of the dynamic scene. For applications of outdoor robotic navigation, joint learning methods have not been extensively used for extracting spatio-temporal features or adding different priors into the formulation. The task becomes even more challenging without stereo information being incorporated. This paper proposes an approach to fuse semantic features and motion clues using CNNs, to address the problem of monocular semantic motion segmentation. We deduce semantic and motion labels by integrating optical flow as a constraint with semantic features into dilated convolution network. The pipeline consists of three main stages i.e Feature extraction, Feature amplification and Multi Scale Context Aggregation to fuse the semantics and flow features. Our joint formulation shows significant improvements in monocular motion segmentation over the state of the art methods on challenging KITTI tracking dataset.

1 INTRODUCTION

Visual understanding of dynamic scenes is a critical component of an autonomous outdoor navigation system. Interpreting a scene involves associating a *semantic concept*, also referred to as a *label* with each image pixel. These semantics can then be incorporated in a higher-level to reason about the image holistically. Traditional scene understanding approaches (Chen et al., 2014)(Athanasiadis et al., 2007) (Shotton et al., 2008) have focused on extracting pixel-level semantic labels, and have demonstrated superior performance in static scenes. Motion and semantics provide complementary cues about a dynamic scene, and can be used to generate a comprehensive understanding of the scene. Some recent approaches (Reddy et al., 2014) (Wedel et al., 2009) leverage stereo information to incorporate motion cues into the scene understanding framework.

We focus on the problem of obtaining semantic motion segmentation from monocular images. Recent success in scene understanding using convolutional neural networks, motivated us to extend existing models that perform semantic segmentation to incorporate motion cues. The success of deep neural network architectures can be attributed to the efficient learning and inference mechanisms employed. Learning involves determining a set of parameters using multiple iterations of stochastic gradient descent over randomly sampled *batches* of labeled images, and inference on a target image involves only a forward pass of the image through the network.

Figure 1: The eventual output of our Semantic Motion Segmentation approach. Semantic labels get prefixed with motion labels such as Moving Car and Stationary Pedestrian.(Best viewed in color).

Deep learning architectures used for scene understanding incorporate semantic labels for learning scene descriptions. We aim to generate richer descriptions by *prefixing* motion labels to semantics such as 'Moving Car' and 'Stationary Car', and do so in a joint framework. Currently, deep architectures

Haque, N., Reddy, D. and Krishna, K.
Joint Semantic and Motion Segmentation for Dynamic Scenes using Deep Convolutional Networks.
DOI: 10.5220/0006129200750085
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017) - Volume 5: VISAPP*, pages 75-85
ISBN: 978-989-758-226-4

model either motion (Fischer et al., 2015) or semantics (Long et al., 2015) in an exclusive manner. To the best of our knowledge, this is the first effort towards seamlessly integrating motion cues with deep architectures that are trained to predict only semantics. The proposed joint learning pipeline is efficient, and learning can be performed end-to-end. Fig. 1 shows a sample output of the proposed framework.

In settings where images are obtained from a monocular camera, motion detection has been tackled by taking into account the optical flow between two subsequent images, which tends to fail with large camera displacements. For outdoor robotic vision, the camera displacement is unavoidable. Although, this has been tackled in (Tourani and Krishna, 2016) where motion models are generated and merged using trajectory clustering into different motion affine subspaces. The moving object proposals generated from the prior model are sparse collection of points lying on the object, resulting into a sparse motion segmentation.(Fragkiadaki et al., 2015) exploit appearance similarity to capture parts of moving objects using two stream CNN with optical flow and rank spatio-temporal segments over a video sequence by mapping clustered trajectories to the pixel tubes. In contrast, our approach performs joint optimization for pixel wise motion and semantic labels, owing to the fact that they are interrelated. An intuitive example to demonstrate the relation is that the likelihood of a moving car or moving pedestrian is more than that of a moving tree or wall. To exploit the correlation, our pipeline proposes integration of semantic and motion cues in three stages, namely, Feature extraction, Feature amplification and multi-scale context aggregation. The proposed approach is shown to be effective for motion segmentation even with a moving camera, on outdoor scenes.

In summary, following are the key contributions of our work.

- We present an end-to-end convolutional neural network architecture that performs joint learning of motion and semantic labels, from monocular images.
- We provide a novel method for seamless integration of motion cues with networks trained for predicting semantic labels.
- We present results on several sequences of the challenging KITTI benchmark and achieve results superior to the state of the art.

The remainder of the paper is organized as follows. Section 3 presents the architecture and approach used for joint learning of motion and semantic labels. In section 4, we summarize the experiments carried out, dataset used and training procedure for our joint module. We also show evaluation and comparison of our approach in section 4.3.

2 RELATED WORK

Fair amount of literature has been done in the field of semantic and motion understanding of scene. Traditional approaches for semantic segmentation involve extracting features from a image and use different methods to classify each pixel. Multiple works have been used to train for semantic labels (Fields, 2001) (Reddy et al., 2014) (Russell et al., 2009) (Koltun, 2011). However, with the emerging era of Deep Learning, there has been a large amount of literature in the field of semantic segmentation which has shown large improvements compared to the previous baselines. Approaches using deep convolutional neural networks (LeCun et al., 1989) have shown to outperform most of the methods in all the basic problems of vision. The literature includes works by (Lin et al., 2015) (Liu et al., 2015)(Dai et al., 2015), where techniques such as bounding box, Deep net followed by CRF formulation and MRFs were put to use, achieving significant results. Further, (Long et al., 2015) adapted the VGG Net model (Simonyan and Zisserman, 2014b) to predict pixel-to-pixel semantic labels, with fusion at pool layers for output up-sampling. Yu and Koltun(Yu and Koltun, 2015) proposed an adaptation of VGG-16 architecture for systematic expansion of receptive fields using dilated convolutions for dense image segmentation, giving more accurate results than prior adaptations. The approach involves carrying over a global perspective without loss in resolution using repetitive deep convolutional layers.

Motion segmentation has been extensively addressed, particularly for outdoor robotic navigation. Most of the works use geometric constraints to attain significant accuracy. In the seminal work contributed by (Elhamifar and Vidal, 2009), trajectory points were modeled as sparse combination of evaluated trajectories. (Tourani and Krishna, 2016) used in frame shear constraints to generate and merge affine models, achieving state of art results in sparse motion segmentation using monocular camera. Recently many deep convolution nets have been used to learn motion labels (Rozantsev et al., 2014) (Fragkiadaki et al., 2015) (Tokmakov et al., 2016) for motion segmentation. Although they work very well, they suffer from unavailability of large datasets or rely on stereo information, therefore proving ineffective for monocular systems. (Fragkiadaki et al., 2015) presents state of art results in the detection of per frame moving ob-

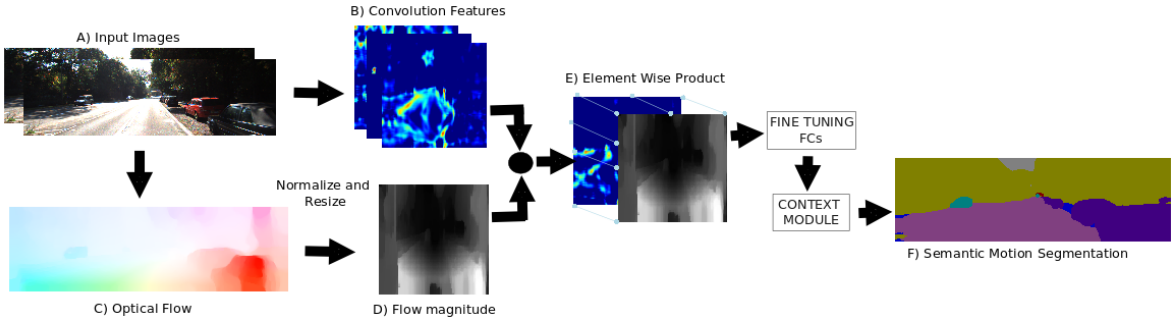

Figure 2: Illustration of the proposed approach. Images at t and t+1 are provided to the network(A). The dilated network undergoes fine tuning with addition of motion labels and the last Conv. features are extracted(B).Optical flow between the two frames(C) is scaled and resized to the size of feature maps(D). The dilated features are amplified using optical flow magnitude by element wise product(E). Further, convolution layers are freezed and fully connected layers are fine tuned. The augmented feature maps are further enhanced with end-to-end training with the Context Module, learning dependencies between object class and motion labels. The predictions obtained from the softmax layer are upsampled to give a joint label to each pixel(F).(Best viewed in color).

ject proposals. The work emphasizes segmentation on monocular uncalibrated video sequences by a ranking heuristics and regression using a two stream network with optical flow, followed by supervoxel projection.

Joint classification of semantic and motion labels is relatively new in the field, and much of the work has been carried out by (Reddy et al., 2014) using dense CRF joint formulation on stereo image sequences. This however would prove ineffective for monocular situations as it heavily relies on the depth information. We draw analogy from works (Fischer et al., 2015) (Simonyan and Zisserman, 2014a) (Karpathy et al., 2014) (Park et al., 2016) where two parallel streams of convolution neural networks are fused for action recognition in videos or generating optical flow. Due to unavailability of large scale datasets for semantic motion segmentation, training a neural network from scratch becomes unfeasible. However, we adapt the concept of feature amplification highlighted in (Park et al., 2016) to our problem in a joint formulation approach, resulting in an end-to-end model for semantic motion segmentation. We outperform state of art results for monocular motion segmentation using our joint model.

3 MONOCULAR SEMANTIC MOTION SEGMENTATION

In this section, we present our semantic motion segmentation framework. A joint formulation is proposed for the overall learning task and is composed of three main modules, viz. features from dilated convolutions, feature amplification, and multi-scale context aggregation. We also provide an illustration of our approach in Fig. 2.

3.1 Features from Dilated Convolutions

To obtain semantic features, we use a neural network architecture which employs dilated convolutions, specifically engineered for dense predictions. Originally proposed in (Yu and Koltun, 2015), a dilated convolution operator is a traditional convolution operator modified to apply a filter at different ranges using different dilated factors.

In relation to a discrete function $H : \mathbb{Z}^2 \rightarrow \mathbb{R}$ and $q : \Omega_s \rightarrow \mathbb{R}$, a discrete filter with size $(2s+1)^2$, where $\Omega_s = [-s,s]^2 \cap \mathbb{Z}^2$, the convolution operator is defined as:

$$(H *_d q)(c) = \sum_{r+dt=c} H(r)q(t) \qquad (1)$$

where, d is the dilation factor. Such an operator $*_d$ is referred to as d-dilated convolution. The operator can be intuitively understood as follows. Given a 1D signal f and a kernel q, with dilated convolution the kernel touches the signal at every d^{th} entry.

Expansion of receptive fields in existing pooled architectures leads to an ungainly increase in parameters to the same extent. The architecture proposed by Fisher et. al. (Yu and Koltun, 2015) is inspired from the fact that dilated convolutions sustain exponential expansion of the effective receptive field without loss in coverage area. While pooling architectures leads to loss in resolution, the dilated architecture enables initialization with the same parameters and producing higher resolution output.

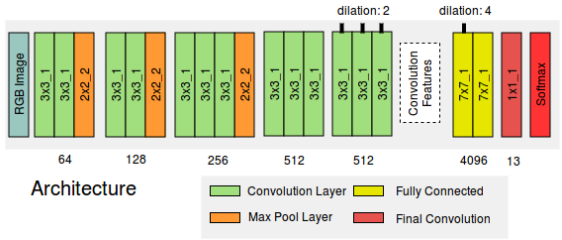

Figure 3: Network Architecture - $w \times h$_s: Layer with kernels of width w, height h, and stride s. The dilated factor in layers, if any, is shown on the top of each layer. Number of channels in the outputs from each layer is depicted below each layer. For instance, the fully connected has 4096 channels in its output block.

3.1.1 Network Architecture

Our network architecture is primarily adapted from the VGG-16 framework proposed by (Simonyan and Zisserman, 2014b), with modifications applied from the work by Fisher on dilated convolutions. The VGG-16 architecture incorporates a stack of convolutions, followed by three fully-connected layers. This was tailored for dense predictions by Long et al. (Long et al., 2015). The architecture proposed by Long includes two major shifts. First, the inner product layers are converted to convolutions. This overcomes the restriction on the size of the input image owing to the fact that the architecture does not contain any inner product layers. Second, an upsampling layer is introduced, which brings back the spatial resolution of the output through a learned operation. The upsampling operation is carried out at different intermediate layers and are fused to obtain dense predictions. This allows the architecture to predict finer details with global or high level information in place.

We adapt the fully convolutional network of Long. and integrate modifications proposed by (Yu and Koltun, 2015). Our network architecture is shown in Fig. 3. The last two pooling layers in the VGG-16 architecture (Simonyan and Zisserman, 2014b) were removed. Furthermore, for each of the removed layers, the following convolution layers are replaced with a dilation factor of 2. This enables the network to generate high resolution features with the same initialization parameters.

3.1.2 Network Initialization

We present a novel method for initialization of the ConvNet for obtaining convolution features. We use the model by (Yu and Koltun, 2015), pre-trained for semantic labels. For training with joint semantic and motion labels, we modify the final convolution layer and change the number of outputs to (C+M), where C is the number of semantic labels predicted by the dilated ConvNet and M is the number of motion labels. For instance, M can be 2, with two labels being moving car and moving Pedestrians in an outdoor scene. Furthermore, we copy the weights from the pretrained dilated ConvNet to the modified network for all layers except the final layer. For the final convolution layer, we copy the weights in the given fashion:

$$weights['final'][i] \rightarrow data[1:C,:,:,:] = weights_p['final'][i] \rightarrow data$$

where, $i \in \{0,1\}$, 1 for weights and 0 for bias, $weights_p$ is the pre-trained weights array of the dilated network for semantic features and $weights$ is the weights array of the modified network. Weights for M motion labels in the final convolution layer are initialized using Xavier initialization (Glorot and Bengio, 2010).

We propose that the initialization scheme works well for training with fairly small annotated datasets. The proposed initialization subjugates the limitation of unavailability of large scale annotated dataset to perform training for joint semantic and motion labels. The network is trained on our annotated dataset with the given initialization. Furthermore, the 'Convolution features'(see Fig.3) from the network are extracted for joint learning with flow features. Also, the joint labels obtained from the softmax layer forms our *Baseline* results for future comparisons.

3.2 Feature Amplification

We leverage optical flow for learning motion cues in an image. Conventionally, training two stream networks is found useful to the task where one is focused on learning semantic features using RGB image input, while the other is tasked for learning motion cues. The features from the two streams are fused at an intermediate layer for joint learning. However, unavailability of a large annotated dataset with joint semantic and motion labels is a major bottleneck for learning with two stream architectures. Akin to the ideas proposed in (Park et al., 2016), we present an approach for learning relationship between semantic and motion class of an object. The method proposed in (Park et al., 2016) is used primarily for action recognition tasks. Features from the last convolution layer in a Convolutional Network tasked for learning semantic features is amplified using optical flow magnitude to identify the moving parts in an image before the fully connected layers are evaluated. However, we extend the underlying idea for the task of semantic motion

segmentation. An intuitive reasoning behind such an adaptation is the similarity in recognition of motion cues and integration with semantic features in both the problems.

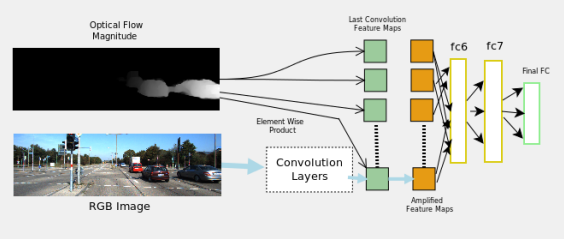

Figure 4: Fine tuning fully connected layers using amplified features as described in (Park et al., 2016). Amplified features are obtained by taking element wise product of scaled optical flow magnitude with the feature maps obtained from last convolution layer.

We propose a method to augment feature maps obtained from the last convolution layer of our dilated network (see Fig 3), for incorporating motion cues. Optical flow is generated between the consecutive frames at t and t+1. Next, we compute Euclidean norm of the flow vector and normalize the magnitudes in the range 1-2. With the flow information in hand, we quantize the scaled magnitudes and subsequently convert it to grayscale image. The image is further resized to the size of feature maps of the last convolution layer obtained from our spatial network. Given a 900x900 RGB image as input, our dilated network outputs feature maps of dimension 512x90x90. Hence, flow image is resized to 90x90 dimension. Thereafter, element wise product is performed between the flow image and each feature map in the stack. The intuition of scaling the magnitudes from 1 rather than 0 is to not zero out the feature values obtained from the spatial network, which is equally important. Further, we *freeze* the convolution layers of the network and fine tune the fully connected layers with the amplified feature maps as input to the fully connected layers. The amplification process is visualized in Fig. 4. The semantic features are enhanced with motion cues, as a consequence of feature amplification. We benefit with the amplification due to incorporated temporal consistency with optical flow and difference in flow magnitude between moving objects and it's surroundings. Also, object boundaries are retained due to amplification over baseline semantic motion features, thereby handling disorientation in optical flow boundaries. Label probabilities obtained after fine tuning from the softmax layer are up-sampled to obtain dense predictions with joint labels. Image predictions obtained forms our *Joint* results for evaluations.

3.3 Multi-Scale Context Aggregation

We use the context module introduced by Fisher (Yu and Koltun, 2015) for enhancement of the amplified features. The architecture was proposed as an extension to existing CNN architectures for overall increase in accuracy for dense predictions. The module improves upon the feature maps, by successive dilated convolutions, supporting exponential expansion of receptive field, without losing resolution. This is effectuated by continuous increase in dilation with increasing layer depth. The architecture consists of 7 convolution layers. The layers are dilated with factors - 1, 1, 2, 4, 8, 16 and 1, and each of these layers apply 3x3 convolutions with the specified dilation factors. The module aggregates contextual information at multiple scales and outputs feature maps of the same size as that of input by padding the intermediate layers.

The weights in this module are initialized with a form of identity initialization, commonly used for recurrent networks. In mathematical terms:

$$q^j(\mathbf{t},i) = 1_{\mathbf{t}=0} 1_{i=j} \tag{2}$$

where i and j are index of input and output feature maps respectively. The identity initialization of such a form, initiates filters which can relay the inputs to the next layer.

We learn the parameters for the context module with our amplified feature maps as input to the module. Fully connected and softmax layers from our network is appended to the module. We obtain joint label predictions from the softmax layer.

4 EXPERIMENTS AND EVALUATION

Experiments were carried out with pre-existing architectures, adapted to our problem. Concept of two stream architectures have been recently used in the field of action recognition, where spatial and temporal nets are combined at the fully connected layer. We tailor the architecture to our problem. Two VGG-16(Simonyan and Zisserman, 2014b) networks with image and optical flow as input to the respective networks were trained on our annotated dataset. The weights for both the streams were initialized with VGG models pre-trained for semantic segmentation task. We also inspect and implement Flow net (Fischer et al., 2015) to our problem, which has shown to outperform state of art in learning optical flow. The network was initialized with pretrained FlowNet-C weights and trained on our annotated dataset with inputs as image at t and t+1 respectively for the two

streams. However, both the formulations did not work well in combining motion cues with semantic information in hand. This is attributed to the failure of CNNs in learning and extracting useful features with smaller datasets. Collection of large scale scene datasets with joint semantic and motion labels is very expensive. In contrast, our joint learning approach reduces the burden of learning motion features from scratch with large labeled datasets and proves effective with fairly smaller annotated datasets.
In this section, we describe the details of ConvNet training and evaluations on KITTI tracking dataset.

4.1 Dataset

We have used renowned *KITTI* dataset (Geiger et al., 2012), for evaluation of our approach. The dataset contains over 40,000 images taken by a camera mounted on a driving car through European Roads. The driving sequences contain images from residential and urban scenes posing it as a challenging dataset. The dataset was chosen to showcase proficiency of our approach with multiple moving cars for outdoor scenes, which is uncommon in other datasets. 40 images were chosen from five sequences each, giving 200 images for training. Each of the images were manually annotated with 13 labels. To be specific, the labels given were *Building, Vegetation, Sky, Car, Sign, Road, Pedestrian, Fence, Pole, Sidewalk, Cyclist* and *Moving Car, Moving Pedestrian* for objects in motion. For testing, 60 images from KITTI tracking sequences were chosen as validation set and annotated with the given label spectrum. For validation set, we use challenging sequences with multiple moving cars and ensure no overlap between train and validation deck. We have used DeepFlow(Weinzaepfel et al., 2013) for dense optical flow computation, known for its state of art results for KITTI benchmark dataset. We plan to release the code, trained models and dataset with joint labels to encourage future work in the field.

4.2 Learning

In this section we describe the training procedure for our proposed approach. Our implementation is based on publicly available Caffe(Jia et al., 2014) framework. First, we describe the input to the data channel in the network. This applies to all modules in our proposed method. Input image resolution is 1242 x 375, obtained from KITTI tracking dataset. Images are padded using reflection padding and 900x900 random crops are sampled. It then undergoes randomized horizontal flipping. Further, each input batch contains crops from randomly selected images from the training dataset. This shapes the input to the module.

Training: Training is performed in three stages. At first, the network architecture (see Fig. 3) is fine tuned with motion labels added, to obtain convolution features and weights initialization for joint training with flow features. Learning rate and momentum was set to 10^{-4} and 0.9, respectively. Training was carried out for 10,000 iterations with batch size 1, using stochastic gradient descent. The dense predictions obtained from the module forms our *baseline* for further comparisons. We use these learned weights to train the joint model with augmented feature maps as input. Optical flow magnitude is computed between the frame at t and t+1. Flow image is padded and cropped to 900x900, with respect to the RGB crop. Furthermore, the convolution layers are *freezed* and the network is trained with the amplified feature maps as input to the fully connected layers. Training was carried out for 10,000 iterations. Other parameters stay the same.

Then, the context model is plugged into the architecture and end-to-end training is performed for 20,000 iterations with batch size 1. Learning rate and momentum is set to 10^{-5} and 0.99, respectively. We refer joint label predictions obtained from the softmax layer of this model as *Joint+Context*.

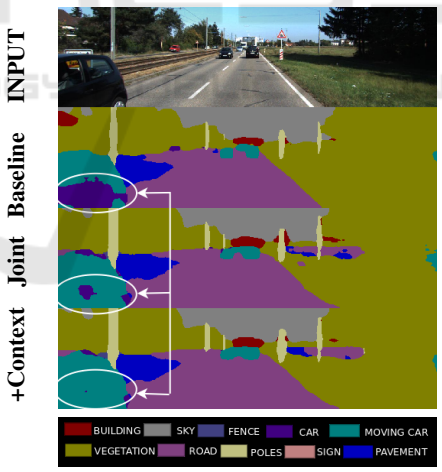

Figure 5: Figure outlining the labels from each stage of our end-to-end module. Image is taken from our KITTI tracking test dataset. The baseline predictions outputs wrong labeling to moving car patches and the motion labels of the car[Cyan] improve significantly using our joint model.(Best viewed in color).

4.3 Results

We evaluate the proposed approach on our manually annotated KITTI Tracking test dataset. The testing

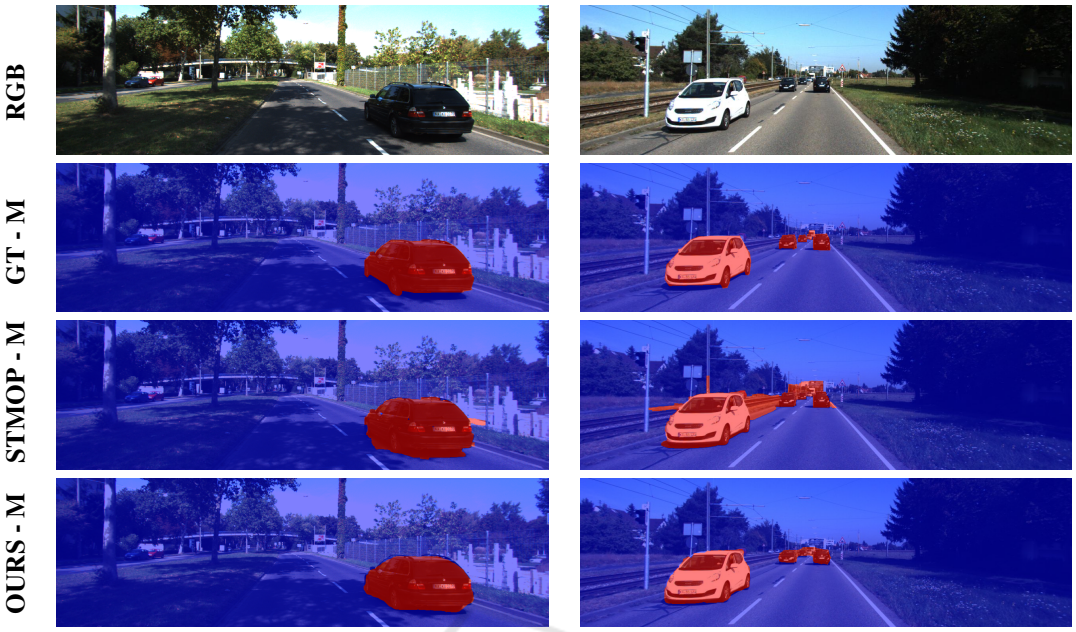

Figure 6: Qualitative evaluation - Motion segmentation on our KITTI test dataset. On the left, the images, consist of single moving car. On the contrary, we have multiple cars in the image, on the right. Blue pixels represent stationary and red pixels depict motion. We compare our approach with STMOP-M moving object proposals (Fragkiadaki et al., 2015). In the figure, GT - M is ground truth motion annotation, STMOP - M is output from (Fragkiadaki et al., 2015) and OURS - M is the motion segmentation obtained from the proposed approach. In contrast to STMOP-M where over-segmentation and False Positive cases are observed on the roads and fence, our proposed approach yields better segmentation and motion boundaries with cars in motion. (Best viewed in color).

images(see Sec. 4.1) chosen from different sequences pose challenging scenarios for motion segmentation with multiple moving objects. Also, there are prominent cases where moving cars lie in the camera subspace. To demonstrate qualitative results we take four sequences consisting of 116, 143, 309 and 46 images. Qualitative results are provided in Fig. 7, Fig. 8 and on complete sequences in the *supplementary video*. To the best of our knowledge, there are no available monocular joint semantic and motion baseline. Hence, we show independent semantic and motion evaluation with the existing state of art in the respective fields. For instance, for a pixel bearing joint label - 'Moving Car', we say 'Car' as the semantic label or object class and 'Moving' as the motion class of the pixel. Comparative evaluations are carried out for semantic segmentation and monocular motion segmentation. However, for joint semantic and motion labels, we demonstrate evaluations against manually annotated Ground Truth labels.

4.3.1 Qualitative Evaluation

In this section, we show our results with joint labels for different stages proposed in the paper, in comparison to Ground Truth. We also show qualitative assessment of motion segmentation in monocular settings with STMOP-M. (Fragkiadaki et al., 2015). In the Figures, 'Stationary Car' and 'Stationary Pedestrian' labels are abbreviated as 'Car' and 'Pedestrian' respectively, while the label is prefixed with 'Moving' in case of motion.

Table 1: Quantitative evaluation on our KITTI test tracking dataset. We compare PPV (Positive predicted value) from our approach with the state of the art sparse motion segmentation SHEAR-M(Tourani and Krishna, 2016) and STMOP-M(Fragkiadaki et al., 2015). We achieve 4.9% gain in the metric over the existing state of art.

Model	Stationary	Moving
STMOP-M	98.34	83.91
SHEAR-M	99.85	84.37
Ours (Joint+Context)	99.55	**89.28**

Motion: We show improvements over our baseline results in Fig. 5. Baseline results labels parts of moving car as stationary. However, with optical flow based feature amplification, pixels for cars in motion are rectified as moving. Further, via feature enhancement with Context Module, labels improve significantly. We attribute the improvements shown by using feature amplification, to the fact that temporal consistency has been incorporated using opti-

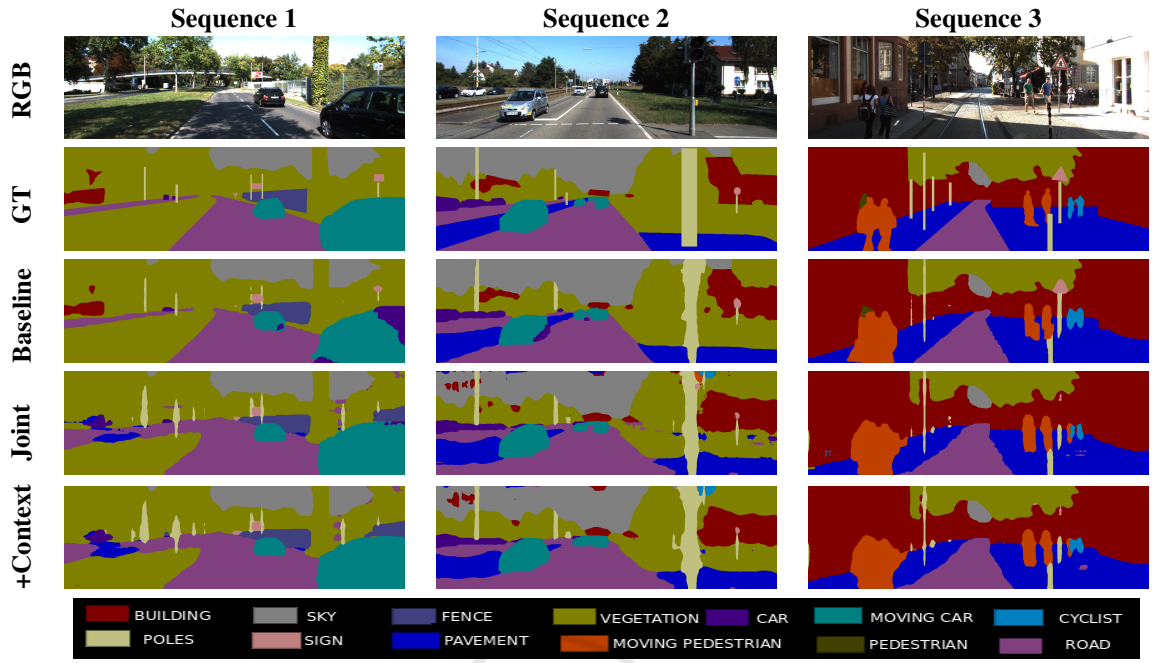

Figure 7: Qualitative evaluation of joint labels with Ground Truth annotations on our KITTI test dataset. Top to Bottom: (1) Input image from KITTI sequences (2) Ground Truth for semantic motion segmentation (3) Baseline predictions: joint labels using dilated convolution. (4) Joint Module: Results obtained after feature amplification with optical flow.(5) Context Module: joint predictions after feature enhancement with context module.(Best viewed in color).

cal flow into the baseline. Further, the proposed feature amplification has clear demarcation between the boundaries of the moving objects and stationary surroundings due to variance in flow vector magnitude, which is being incorporated into the final segmentation.

STMOP (Fragkiadaki et al., 2015) generates moving object proposals on video sequences. We use the code available and generate proposals on KITTI sequences. For fair comparison, we take the proposals with best supervoxel projection on the objects. We show our monocular motion segmentation results in comparison to Ground Truth and STMOP(Fragkiadaki et al., 2015) moving object proposals. In Fig. 6, consisting of images with single and multiple moving cars, STMOP-M leads to over segmentation, while our approach correctly segments the moving car, also removing extra segments of road and fence. In the above cases, STMOP fails in outdoor robotic scenarios essentially due to large camera motion and optical flow *bleeding*, while our approach uses semantic priors and benefits from motion and semantic correlation.

Joint Semantic and Motion: We also evaluate our approach with the Ground Truth semantic motion labeling. In the sequences demonstrated in Fig. 7, the Baseline results incorrectly labels patches of the moving car closer to the camera(in Sequence 1) as stationary(seen with *Violet* color). Similar observation is found in the baseline results of the moving car in Sequence 2. The patches are rectified as moving as a consequence of joint training with amplified features. This again reiterates the utility of joint learning and inference between motion and semantic cues. The improper patches on the moving cars in the sequences are further rectified by the context module using multi scale context aggregation. To perceive joint segmentation, other than car scenes, we consider a sequence(Sequence 3) from KITTI with Moving Pedestrians. The results for each stage are depicted in Fig. 7. Parts of moving pedestrians on the left are labeled as stationary in baseline results. The joint learning with context aggregation corrects the motion domain of pedestrians. Also, for pedestrians far away from the camera, false positive cases are observed from our approach in tiny patches due to inconsistency in optical flow magnitude of the pedestrian with large distance from the camera. Further, for consistency of our joint labels in challenging outdoor scenes, we show joint semantic motion results on both highway and city street scenes in Fig. 6.

Table 2: Quantitative analysis of motion label predictions with STMOP. Left: On our annotated Kitti(tracking sequence 4) test dataset - consisting of lone moving object. Right: On our annotated kitti images, consisting of multiple moving cars. We compare our results with (Fragkiadaki et al., 2015) moving object proposals.

Model	Stationary	Moving
(Fragkiadaki et al., 2015)	97.75	62.97
Baseline	**99.44**	76.36
Joint	99.35	81.94
Joint+Context	99.28	**83.69**

Model	Stationary	Moving
(Fragkiadaki et al., 2015)	97.63	44.53
Baseline	**99.05**	66.23
Joint	99.03	70.67
Joint+Context	98.97	**71.98**

Table 3: Quantitative evaluation of semantic label predictions from our proposed approach - Joint+Context (Ours - S) on our KITTI test dataset. We compare our method with DeepLab-LFOV(Chen et al., 2014) and Segnet(Badrinarayanan et al., 2015), known for semantic segmentation on outdoor driving scenes.

Method	Building	Vegetation	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Cyclist	mean IOU
Segnet	66.70	78.11	89.32	69.74	12.45	71.69	12.09	25.03	21.12	44.01	11.2	45.61
Deeplab	73.35	84.17	91.33	70.76	7.66	69.63	24.41	68.30	16.51	26.14	13.53	49.62
Ours-S	78.52	84.99	90.07	88.18	19.28	75.82	8.46	76.60	29.31	36.84	66.70	**59.53**

4.3.2 Quantitative Evaluation

In this section, we perform a quantitative assessment of both semantic and motion segmentation. We show evaluations with (Tourani and Krishna, 2016) and (Fragkiadaki et al., 2015). For semantic segmentation we compare our results with (Chen et al., 2015) and (Badrinarayanan et al., 2015), which have shown results for semantic segmentation on driving scenes.

Motion: For quantitative evaluation of motion segmentation, we compare our results with STMOP moving object proposals. Evaluation is staged by cross verification of each predicted pixel with corresponding ground truth motion label - stationary or moving. The evaluation is unfolded in two models. First, we compare our dense motion segmentation with STMOP moving object proposal. We use intersection over union as the evaluation metric for dense motion segmentation. The metric is defined as *TP/(TP+FP+FN)*, where TP denotes true positive, FP false positive and FN false negative. Table 2 summarizes our quantitative motion segmentation evaluation.

The assessment is done in two broad categories,i.e, on annotated sequences with lone moving object and sequences with multiple objects in motion. In the case with single moving car, we achieved 70.67% accuracy in detection of the moving car from our joint module, while STMOP yields 59.97% detection accuracy. The increase in accuracy is attributed to incorporated label and motion correlation. Further, using context aggregation, the context module yields further improvement in the efficiency. In case of multiple moving objects, STMOP yields 41.53% efficiency. The decrease in accuracy from STMOP is due to large camera motion observed in the scenes, while our joint module provides 70.67% success rate. The joint learning exploits the fact that the likelihood of a moving tree or pole is less compared to a moving car or moving person, resulting in substantial improvement in motion segmentation.

Another keynote observation would be a slight decrease in stationary accuracy over our baseline results. This is due to the fact that different objects can exhibit different optical flow depending on the depth from the camera, even though they share the same global motion. The decrease is although marginal as shown in Table 2. We also show motion segmentation evaluation with existing state of art in sparse monocular motion segmentation. The IOU metric used for dense motion segmentation is not known to be used in case of sparse evaluations. Therefore, for fair comparison with sparse segmentation, we use positive predictive value (PPV) or precision- (TP/TP+FP) as the evaluation metric. The results are summarized in Table 1. We gain 4.9% in motion label precision over the state of art SHEAR-M(Tourani and Krishna, 2016) on our test dataset.

Semantics: For quantitative evaluation of semantic image segmentation, we use per class Intersection over Union similar to the metric used for dense motion segmentation evaluation. This is done for 11 semantic labels on our KITTI test dataset. We perform quantitative semantic evaluation of our approach against Segnet(Badrinarayanan et al., 2015), which has shown results on outdoor driving scenes such as

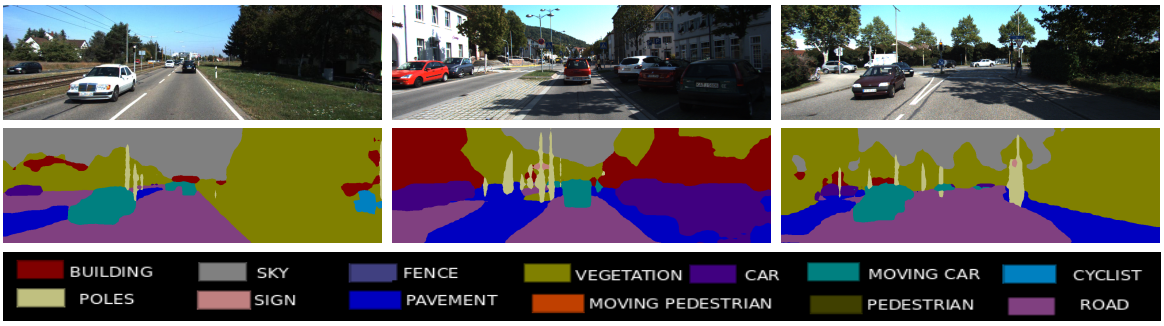

Figure 8: Joint Semantic and motion labeling obtained from the proposed approach on challenging urban scenes. Specifically, in the figure, from Left to Right: Highway scene, City Streets and a drive scene with relatively less traffic. The joint labels obtained in these settings depict robustness and consistency of our proposed approach.(Best viewed in color).

KITTI, and DeepLab-LFOV (Chen et al., 2015). For comparison with (Chen et al., 2015) we use the publicly available pre-trained model on PASCAL dataset and fine tune it on our KITTI training dataset. We run both the algorithms, Segnet and DeepLab, on our KITTI test dataset. The *semantic label* accuracy of the models on the test set is reported in Table 3. Our approach (Joint+Context) outperforms the other two architectures. This is due to the fact that dilated architecture produces higher resolution output crucial to dense prediction in comparison to the strided and pooled architectures in the former propositions.

5 CONCLUSIONS

In this paper, we have proposed a joint approach to predict semantic and motion labels using a monocular camera. We incorporate spatial and temporal information to learn object class and motion labels jointly. Evaluations show an increase in pixel wise motion segmentation accuracy without using stereo information. We learn pixel wise labels without the need for training temporal networks for motion cues, which has proved to be a pitfall with unavailability of large annotated datasets. To contribute and encourage future works on monocular semantic motion segmentation, we plan to release the annotated dataset and trained models.

We believe that the proposed work will be extended for pixel-wise labelling of individual moving objects. The end-to-end system can be used for better dynamic scene understanding in complex outdoor environments.

ACKNOWLEDGEMENTS

We would like to thank J. Krishna Murthy for providing insights into the formulation. We are also grateful to Parv Parkhiya and Aman Bansal for help with dataset annotation on KITTI Tracking benchmark.

REFERENCES

Athanasiadis, T., Mylonas, P., Avrithis, Y., and Kollias, S. (2007). Semantic image segmentation and object labeling. *IEEE transactions on circuits and systems for video technology*, 17(3):298–312.

Badrinarayanan, V., Handa, A., and Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. (2015). Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*.

Dai, J., He, K., and Sun, J. (2015). Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1635–1643.

Elhamifar, E. and Vidal, R. (2009). Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797. IEEE.

Fields, R. (2001). Probabilistic models for segmenting and labeling sequence data. In *ICML 2001*.

Fischer, P., Dosovitskiy, A., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., van der Smagt, P., Cremers, D., and Brox, T. (2015). Flownet: Learning optical

flow with convolutional networks. *arXiv preprint arXiv:1504.06852*.

Fragkiadaki, K., Arbeláez, P., Felsen, P., and Malik, J. (2015). Learning to segment moving objects in videos. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4083–4090. IEEE.

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Koltun, V. (2011). Efficient inference in fully connected crfs with gaussian edge potentials. *Adv. Neural Inf. Process. Syst*.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Lin, G., Shen, C., Reid, I., et al. (2015). Efficient piecewise training of deep structured models for semantic segmentation. *arXiv preprint arXiv:1504.01013*.

Liu, Z., Li, X., Luo, P., Loy, C.-C., and Tang, X. (2015). Semantic image segmentation via deep parsing network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1377–1385.

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.

Park, E., Han, X., Berg, T. L., and Berg, A. C. (2016). Combining multiple sources of knowledge in deep cnns for action recognition. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE.

Reddy, N. D., Singhal, P., and Krishna, K. M. (2014). Semantic motion segmentation using dense crf formulation. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, page 56. ACM.

Rozantsev, A., Lepetit, V., and Fua, P. (2014). Flying objects detection from a single moving camera. *arXiv preprint arXiv:1411.7715*.

Russell, C., Kohli, P., Torr, P. H., et al. (2009). Associative hierarchical crfs for object class image segmentation. In *2009 IEEE 12th International Conference on Computer Vision*, pages 739–746. IEEE.

Shotton, J., Johnson, M., and Cipolla, R. (2008). Semantic texton forests for image categorization and segmentation. In *Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Simonyan, K. and Zisserman, A. (2014a). Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems*, pages 568–576.

Simonyan, K. and Zisserman, A. (2014b). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tokmakov, P., Alahari, K., and Schmid, C. (2016). Weakly-supervised semantic segmentation using motion cues. *arXiv preprint arXiv:1603.07188*.

Tourani, S. and Krishna, K. M. (2016). Using in-frame shear constraints for monocular motion segmentation of rigid bodies. *Journal of Intelligent & Robotic Systems*, 82(2):237–255.

Wedel, A., Meißner, A., Rabe, C., Franke, U., and Cremers, D. (2009). Detection and segmentation of independently moving objects from dense scene flow. In *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 14–27. Springer.

Weinzaepfel, P., Revaud, J., Harchaoui, Z., and Schmid, C. (2013). Deepflow: Large displacement optical flow with deep matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1385–1392.

Yu, F. and Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.