2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Multi-Task Learning for Single Image Depth Estimation and Segmentation Based on Unsupervised Network

Yawen Lu, Michel Sarkis, and Guoyu Lu

Abstract— **Deep neural networks have significantly enhanced the performance of various computer vision tasks, including single image depth estimation and image segmentation. However, most existing approaches handle them in supervised manners and require a large number of ground truth labels that consume extensive human efforts and are not always available in real scenarios. In this paper, we propose a novel framework to estimate disparity maps and segment images simultaneously by jointly training an encoder-decoder-based interactive convolutional neural network (CNN) for single image depth estimation and a multiple class CNN for image segmentation. Learning the neural network for one task can be beneficial from simultaneously learning from another one under a multi-task learning framework. We show that our proposed model can learn per-pixel depth regression and segmentation from just a single image input. Extensive experiments on available public datasets, including KITTI, Cityscapes urban, and PASCAL-VOC demonstrate the effectiveness of our model compared with other state-of-the-art methods for both tasks.**

I. Introduction

Scene image depth estimation is a critical and challenging problem in the field of computer vision and robotics that can be applied into autonomous vehicles [14], 3D scene reconstruction [12] and augmented reality [16]. Based on multi-view geometry, the depth map can be estimated from temporal image sequences [34] and image pairs [32], where multiple images from different perspectives are required. With the development of deep neural networks, end-to-end learning frameworks [8] [25] have been applied to estimate the scene depth from a single image, which largely increases the estimation accuracy. The most significant limitation for these methods is that the ground truth labels are not always ready to be used in real-world scenarios.

Image segmentation has been initially explored by various clustering methods such as mean shift [5] and Markov random field [45] in an unsupervised learning manner. Recently, deep learning has enhanced the image segmentation performance by supervised learning methods, which requires extensive pixel-wise semantic ground truth labels. Segmenting the images based on unsupervised clustering is much more challenging task than supervised learning algorithms, as it is difficult to segment an image into an arbitrary number of meaningful regions without any prior knowledge, especially when the semantics exhibited in the images are complicated and composed of both background and foreground objects such as *Cityscapes* dataset [6].

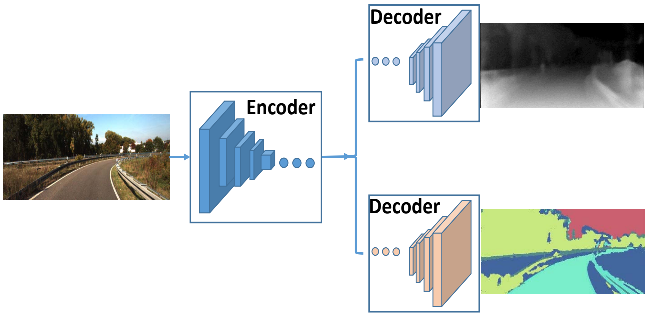

Fig. 1: Our method conducts the depth map estimation and segmentation from a single image during testing stage.

In this paper, we explore an unsupervised multi-task learning framework for simultaneous single image depth estimation and image segmentation. Different from monocular VO systems [39], [43], [47], [35], [44], [33] that all take unlabeled video sequence as input and stereo images as input [42], our network is for single (just one) image depth estimation and segmentation. Instead of applying ground truth labels for training [29], [23], we propose an unsupervised CNN framework consisting of loss constraints from both spatial and spectral perspectives to simultaneously train the neural network for each task. Given a single image, our model is able to segment it into different components and at the same time estimate a dense scene depth map as shown in Fig. 1. Our method can better understand and extract the common feature representations across these related tasks, leading to superior performance compared with existing methods learning each task separately. More specifically, the proposed model learns scene depth estimation and segmentation simultaneously under a multi-task unsupervised learning framework. Image segmentation can provide evidence to avoid gaps in the estimated scene depth. Meanwhile, image pixel clustering can also benefit from predicted single depth, as the pixel depths of the same object usually share a common pattern. We demonstrate that our method can improve both single image depth estimation and segmentation via our fully unsupervised end-to-end framework.

Our main contributions are listed as follows: 1) We introduce an interactive CNN model for image pairs with novel depth consistency loss and dual depth consistency constraint to improve the single depth estimation performance; 2) We utilize the proposed CNN to extract high-level feature vectors that can help to determine the pixel clusters. This is one of the few research works conducted on unsupervised deep neural network for segmentation tasks; 3) We propose a unified framework for simultaneously training image segmentation and depth estimation. Our work is one

Yawen Lu and Guoyu Lu are with Rochester Institute of Technology, NY, USA yl4280@rit.edu and luguoyu@cis.rit.edu. Michel Sarkis is with Qualcomm Technologies Inc., USA msarkis@qti.qualcomm.com. This work is funded by Qualcomm Technologies Inc.

978-1-7281-7395-5/20/$31.00 ©2020 IEEE

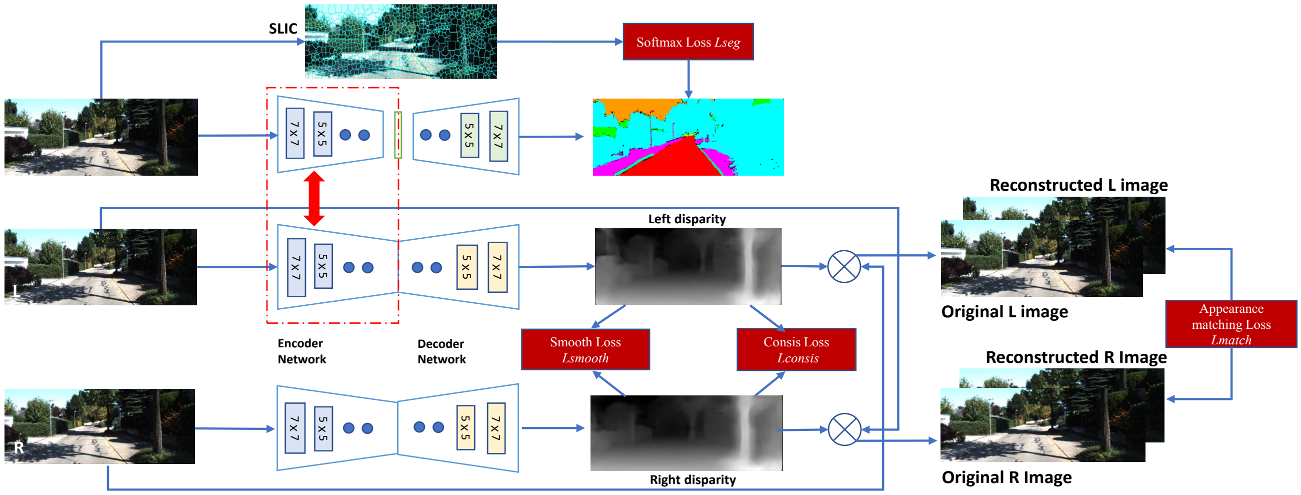

Fig. 2: Overview of the proposed learning framework. The proposed architecture consists of two tasks (single image depth estimation and segmentation) and five loss constraints from both spectral and spatial perspectives across the tasks. The shared encoder is connected to the respective decoder of each task to produce a pixel-wise depth map and segmentation.

of the first methods that conducts single depth estimation and segmentation tasks on challenging datasets without any prior knowledge or supervision.

II. Related Work

Single Image Depth Estimation based on deep neural network mainly relies on ground truth labels to train the model, which generates promising results. Eigen et al. [8] used two deep networks to perform a multi-scale deep network with a scale-invariant loss function for depth estimation. Liu et al. [24] [25] proposed to deal with the depth estimation problem based on deep CNN and Conditional Random Field (CRF) learning. A sequential network using CRF and CNN is then deployed for single depth estimation [41]. It fused complementary information from multiple CNN outputs by means of Continuous Conditional Random Fields (CCRFs). A fully convolutional residual architecture is proposed to predict the depth information given an RGB image in [22], which creates up-projection blocks as a more efficient scheme for upsampling and introduces a suitable reverse Huber loss term to optimize the network during their training process. Semi-supervised approaches are proposed in [21], [33] train CNN by using both supervised and unsupervised learning clues. Such models are first offline trained by a large number of training samples providing both images and their corresponding depth labels as the training input. The most significant limitation is that extensive labeled data is not always available in real-world scenarios. Unsupervised learning methods [31], [30], [46], [28], [10], [13] are applied to train the depth estimation neural network.

Image Segmentation based on classical unsupervised image segmentation methods usually rely on features such as color, intensity or textures to realize pixel-wise clustering. Various clustering algorithms have been applied to segment images, such as k-means [27], mean-shift [5], normalized Cuts [37] and Markov random field [45]. As deep neural networks have emerged as an efficient model for semantic segmentation, researchers start to deploy fully convolutional networks (FCN) to segment the images. By using fully connected layers, an image with an arbitrary size can be taken as input and produce a corresponding prediction. Based on FCN, a variety of research is conducted to perform semantic segmentation [19] [36] [2] [40] [33] [3] using training labels. Our method can use any training data without labels. In [19], an efficient algorithm for fully connected conditional random field model is utilized to refine the output segmentation map. In [36], a U-shape architecture named U-Net trained an end-to-end supervised neural network consisting of a contracting path to capture context and a symmetric expanding path to enable precise localization. However, most of these learning-based methods require thousands of annotated training samples, consuming a huge amount of annotation effort and cost.

Unlike the above methods, our objective of unsupervised simultaneous single image depth estimation and segmentation is much more challenging because it requires to conduct both tasks without acquiring any labeling information.

III. Joint Learning Framework

The proposed network aims to estimate the depth map and conduct image segmentation simultaneously given only one single image as input. Fig. 2 depicts the basic framework of our network for jointly learning these two tasks. The main idea is that segmentation helps to avoid the discontinuities and gaps in single depth estimation. Meanwhile, accurately predicted depth contributes to producing an accurate scheme for image pixel clustering accordingly. More specifically, we first closely correlate these two tasks by sharing weights in their encoder network when extracts essential geometry features from input images. By enforcing the estimated disparity to be consistent with the original disparity map in dual directions, we can build more accurate disparity map for depth inference. We also impose the smoothness term to prevent significant discontinuity. Rather than using a single encoder for different tasks [33], our network first applies a separate encoder for each task to explore their unique features, and then share them. In the decoder, we use the pyramid concept to output multi-scale depth maps, and each depth map is optimized by the disparity consistency and smoothness constraints. By deploying a fully connected neural network to guide and constrain SLIC algorithm in superpixel extraction, we can have a clustering algorithm to adaptively assign pixels to the correct cluster for complex scenes with multiple kinds of objects. By adding the segmentation context to the depth encoder, we can obtain more accurate feature representation than just focusing on depth task. The same for encoding depth context into segmentation task. We demonstrate that our method can improve the accuracy for scene segmentation and depth estimation simultaneously via using fully unsupervised end-to-end training scheme, which mitigates the training conditions and can be easily applied into other datasets.

A. Geometrical CNN for Depth Estimation

Inspired by [13], our interactive model also applies CNN to transfer the single image depth estimation as an image reconstruction problem, as shown in Fig. 2. For every single image from a stereo pair, our model constructs a CNN to predict its corresponding disparity map. Then the predicted output disparity can be used to reconstruct the other image in the stereo pair. During the training process, the reconstructed left image $\widetilde{I^l}$ can be obtained from the predicted right image disparity map $\widetilde{d^r}$ together with the right input image I^r. The disparity map is estimated from left image CNN. The reconstructed right image $\widetilde{I^r}$ can also be generated from the predicted left disparity map $\widetilde{d^l}$ from right image CNN and the left input image I^l. While we only use left image CNN to predict the testing image depth in the real application, we design a CNN separately for the left image and right image and connect these two CNNs by optimizing the shared loss function. The accurate learning of the right image CNN can help to optimize the left image CNN and vice versa. While training the depth estimation framework, we enforce the mutual consistency between left and right reconstructed images. To extract geometry features and learn a pixel-wise regression output, we apply an encoder-decoder structure. After extracting feature representations from the encoder network, the decoder further constrains the output disparity to produce the wrapped image $\widetilde{I^l}$ or $\widetilde{I^r}$ by moving pixels from the original input image along the epipolar line using bilinear sampling [17] that linearly searches the closest pixel in the surrounding positions to form the wrapped image $\widetilde{I^w}$. The relationship between the wrapped image $\widetilde{I^w}$, predicted disparity map $\widetilde{d}$ and original input image I is as follows:

$$\widetilde{I^w} = I(x + \widetilde{d}) \tag{1}$$

In the rectified training images, the original image at the horizontal position x will have a motion d in the wrapped image. And then the reconstructed image can be expressed using the wrapped image in the following equation:

$$\widetilde{I}(p) = \sum_{i\in(t,b),j\in(r,l)} w_{ij}\widetilde{I^w}(p_{ij}) \tag{2}$$

where w_{ij} is proximity weight value for the top (t), bottom (b), right (r) and left (l) four closest pixels around the pixel p in the wrapped image, whose values are inversely proportional to the projected point distance and sum up to 1. The weighted sum of the neighboring pixels is the value of the corresponding pixel p in the reconstructed image $\widetilde{I}$.

The final predicted disparity can be used to calculate the absolute depth map based on the relationship $\widetilde{D} = fB/\widetilde{d}$, where f is the focal length of the camera and B is the baseline distance between the camera pairs. To further optimize the disparity prediction, we constrain the output disparity maps predicted from left image CNN and right image CNN from both spectral and spatial perspectives (Section IV).

B. Clustering Model for Scene Segmentation

To facilitate improving the performance of unsupervised clustering segmentation, we combine convolutional blocks followed by a Fully Connected Network (FCN) [26] with Simple Linear Iterative Clustering (SLIC) [1] algorithm for superpixel extraction. We first apply convolutional blocks composed of 2D convolutional layers, ReLU activation layers, and batch normalization layers, to capture the global context of the scene. Then a fully connected layer classifies those high-level features into n clusters. More specifically, for each pixel, we assign one cluster label by choosing the label with the maximum probability from the total n clusters. As a result, the assigned number to each pixel would vary from 1 to n. We then apply the superpixel algorithm to group similar pixels in the scene into a unique segment. Here, for each superpixel block, pixels in the block may have different cluster labels. We conduct a majority vote to select the label with the largest number of pixels to be the cluster label for the whole superpixel, which means we select the most frequent cluster label C_{max}, $|C_{max}| \geq |C_n|$ for $C_n \in \{1, 2, 3...n\}$ to be the representation for this superpixel block. At the end of this procedure, orphaned pixels that do not maintain the same label with their near neighbors will be corrected and merged into their nearest component using the Connected Components Algorithm (CCA). We iterate these two steps M epochs and update the parameters using Adam optimizer to obtain the final prediction map. Unlike previous unsupervised clustering methods, our proposed method can adaptively determine the number of segments in multiple different scene images by utilizing the pixel-wise cluster labels predicted from the deep neural network in the first step. In addition, the spatial relationship of pixels in the same superpixel block is well preserved by the SLIC algorithm, which is further refined by CCA.

IV. Unsupervised Loss Constraints

In our framework, we train a network for jointly single image depth estimation and image segmentation. Our scheme optimizes the network based on multiple spatial and spectral constraints through a weighted sum of each loss term with an L2 regularization:

$$\begin{aligned} L = \lambda_1 L_{match} + \lambda_2 L_{smooth} + \lambda_3 L_{consis} \\ + \lambda_4 L_{seg} + \frac{\lambda}{2n}\sum ||w||^2 \end{aligned} \tag{3}$$

where L_{match}, L_{smooth}, L_{consis} and L_{seg} are appearance matching loss, disparity smoothness loss, consistency loss and segmentation loss respectively that constrain the single image depth estimation and segmentation, which will be explained in Section IV-A, IV-B, IV-C and IV-D.

To further improve the robustness of the combined objective function, we introduce a "generalized Charbonnier" loss factor [38] that increases the robustness to noises and illumination change [4] [20] for the reconstructed image matching and disparity smoothness. The generalized Charbonnier loss function is defined as: $\rho = (x^2 + c^2)^a$. Experiments have shown that with a penalty for a=0.45, the prediction performance is the highest and most stable in our tasks. Here c is set to be 0.001.

A. Appearance Matching Loss

Appearance matching loss is to enforce the reconstructed images to match with the original images. In our case, as we construct an interactive CNN to reconstruct the wrapped left and right images, we constrain reconstructed images to be consistent with the original images. By combining the Structural Similarity Index Metric (SSIM) structure [13] and the introduced Charbonnier loss factor, the appearance matching loss for the right and left images is defined as:

$$L^l_{match} = \frac{1}{N}\sum_{ij}\frac{a}{2}\tilde{p}(1 - SSIM(I^l_{ij}, \tilde{I}^l_{ij})) \\ +(1-a)\tilde{p}(||I^l_{ij} - \tilde{I}^l_{ij}||_1) \tag{4}$$

where $\|\cdot\|_1$ represents the L_1 norm operator which calculates the mean absolute value. I_{ij} refers to the image pixel at i th row and j th column in the original left and right input images, and $\tilde{I_{ij}}$ represents our reconstructed left and right image pixels. α is a constant parameter, and we set it to be 0.85. N is the number of pixels. The output value of SSIM operation ranges from 0 and 1, where 1 indicates an absolutely accurate matching. Similarly, we defined the right image appearance matching loss L^r_{match}. The overall appearance matching loss is given by:

$$L_{match} = w_{match}(L^l_{match} + L^r_{match}) \tag{5}$$

which combines both left and right matching loss functions in mutual directions.

B. Disparity Smoothness Loss

The disparity smoothness loss embeds the original image gradient change information to enforce the disparity maps to be smooth. On a flat region of the original image without substantial gradient change, the corresponding depth gradient change is amplified in the loss function, which requires the flat region to have a small depth gradient in the disparity map to make the loss small. The disparity smooth loss can be defined as the following:

$$L^l_{smooth} = \frac{1}{N}\sum_{ij}\tilde{p}(|\partial x D^l_{ij}|e^{-||\partial x I^l_{ij}||_1}) \\ +\tilde{p}(|\partial y D^l_{ij}|e^{-||\partial y I^l_{ij}||_1}) \tag{6}$$

where ∂x and ∂y are horizontal and vertical disparity gradient operators respectively. I_{ij} is i th row and j th column pixel in the original input image and D_{ij} is the corresponding pixel's predicted disparity. With the same loss constraint for the right image L^r_{smooth}, the overall disparity smoothness loss L_{smooth} is provided as $w_{smooth}(L^l_{smooth} + L^r_{smooth})$.

Similar to appearance matching loss, the overall disparity smoothness loss is also the combination of gradient loss for both left image disparity and right image disparity.

C. Disparity Consistency Loss

In order to maintain the coherence between the predicted left disparity map and right disparity map, we apply a left-to-right consistency constraint with the reverse Huber (berHu) penalty term [48]. The left-to-right consistency loss L^{lr}_{consis} and right-to-left consistency loss L^{rl}_{consis} are defined as below:

$$L^{lr}_{consis} = \left\{\begin{matrix} |d^l_{ij} - d^r_{ij+d^l_{ij}}| & |d^l_{ij} - d^r_{ij+d^l_{ij}}| \leq c, \\ \frac{(d^l_{ij} - d^r_{ij+d^l_{ij}}(p))^2 + c^2}{2c} & |d^l_{ij} - d^r_{ij+d^l_{ij}}| > c. \end{matrix}\right\} \tag{7}$$

where $c = \frac{1}{5}max(|d^{lr}_{ij} - d^{rl}_{ij+d^{lr}_{ij}}|)$. As the berHu loss in L1 norm $(d^{lr}_{ij} - d^{rl}_{ij+d^{lr}_{ij}})$ is in the range of $[-c, c]$ and the corresponding L2 norm is out of this range, empirically L_{consis} shows a good balance between the two norms in this given task. With the same disparity consistency loss from right to left image L^{rl}_{consis}, We sum up the left and right disparity map consistency constraints as the final disparity consistency loss L_{consis} as $w_{cosis}(L^{lr}_{consis} + L^{rl}_{consis})$.

D. Segmentation Loss

As mentioned in the Section III-B, the segmentation loss is to obtain a more accurate cluster label for each input pixel by constantly minimizing the Softmax Loss between the predicted label directly from FCN and the labels of the merged clusters from low-level superpixels to exploit the relationships of brightness, color and texture affinities between pixels. For each input image, the segmentation loss is defined as follows:

$$y_n = \frac{e^{y_i + log(M)}}{\sum_{k=i}^{c} e^{y_k + log(M)}} \tag{8}$$

$$L_{seg} = -\sum_i C_i log(y_i) \tag{9}$$

where y_i is the output label from the fully connected layer in the segmentation network, and C_i is the refined labels after superpixel clustering. To make our softmax function numerically stable, we normalize the values in the output labeling vector by multiplying the numerator and denominator with a constant M.

V. Experimental Results

A. Neural Network Configuration

Our unsupervised learning framework involves single image scene depth and segmentation branches, which facilitates the learning process to each other. The single depth

Methods	Training type	Error				Accuracy		
		Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Eigen Coarse [8]	Supervised	0.214	1.605	6.563	0.292	0.673	0.884	0.957
Eigen Fine [8]	Supervised	0.203	1.548	6.307	0.282	0.702	0.890	0.958
Deep CRF [25]	Supervised	0.202	1.614	6.523	0.275	0.678	0.895	0.965
SfMLearner [46]	Unsupervised	0.208	1.768	6.856	0.283	0.678	0.885	0.957
Vid2depth [28]	Unsupervised	0.163	1.240	6.220	0.250	0.762	0.916	0.968
Garg et.al. [10]	Unsupervised	0.152	1.226	5.849	0.246	0.784	0.921	0.967
MonoDepth [13]	Unsupervised	0.124	1.388	6.125	0.217	0.841	0.936	0.975
Ours	Unsupervised	**0.115**	**1.202**	**5.828**	**0.203**	**0.850**	**0.944**	**0.980**

TABLE I: Comparison of single depth estimation with other state-of-the-art methods. For training, all methods are trained on KITTI dataset for a fair comparison. We compare with both supervised and unsupervised methods either taking single images [8] [25] [10] [13] or monocular videos [46] [28] as input.

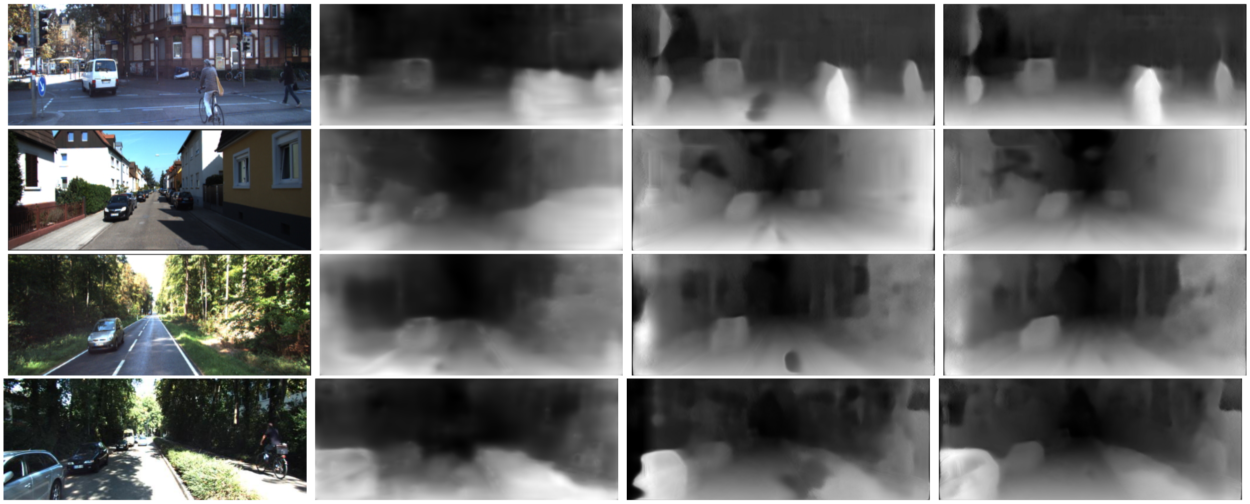

Fig. 3: Single depth estimation results on KITTI comparison between with other methods, SfMLearer [46] (second column), Monodepth [13] (third column) and our method (fourth column). Brighter color represents a closer pixel.

estimation involves a ResNet-18 [15] based encoder model consisting of 7 convolutional layers with Rectified Linear Units (ReLU) as the non-linear activation functions for all the convolutional layers. The size of the receptive fields in the network gradually decreases from 7×7 to 5×5 followed by 3×3 to extract features in finer regions. The decoder is made up of a series of deconvolution layers to recover the spatial features up to the input size after four-scale layers. The segmentation model sets the identical architecture and training process in the convolutional units to share weights with the depth encoder network component. The framework is implemented in PyTorch which is trained by NVIDIA 1080ti GPU for 100 epochs with a batch size of 8 using Adam optimizer [18] where β_1 equals to 0.9 and β_2 is 0.999. The learning rate is set with an initial rate of λ =2 $\times$ 10^{-4} and halved after the first 50 epochs till the end. We evaluate method on KITTI Stereo 2015 [11], Cityscapes Urban Scene [6] and PASCAL VOC2012 [9] datasets.

B. Comparison with State-of-the-art Methods

1) Depth Estimation Evaluation: For depth estimation, the performance of our method is evaluated on KITTI Eigen Split dataset [11], and compared with other state-of-the-art methods using either single image or monocular video in Table I. We use the same evaluation metrics as [8] [13], including the Absolute Relative difference (Abs Rel), Squared Relative difference (Sq Rel), Root Mean Square Error (RMSE), and RMSE log between the predicted depth map D_i and the ground truth depth map G_i. We also measure an accuracy metrics at three different thresholds a1, a2, and a3. Defining $\delta_i = max(\frac{G_i}{D_i}, \frac{D_i}{G_i})$, then a1 represents the maximum ratio between two depth values to be within $\delta < 1.25$. a2 controls the ratio of depths to be within $\delta < 1.25^2$ and a3 is the ratio within $\delta < 1.25^3$.

We can notice from Table I that our unsupervised learning method provides the best depth estimation accuracy and smallest errors compared with other state-of-the-art methods either taking a single image or monocular sequences as input, even compared with supervised learning methods. In general, the methods using video can benefit from the temporal information as they have multiple frames to estimate the depth. However, our method using just one single image can estimate more accurate depth map than the video-based estimation methods using many frames. Fig. 3 presents the estimated depth maps to compare with other methods visually. It can be seen that compared with [46], our method demonstrates higher accuracy and more clear outline. Compared with [13], our method has a superior performance in preventing the discontinuities and significant estimation errors for very close or infinite pixels like sky and ground. Our method also demonstrates better performance in predicting small and occluded objects such as trees and pedestrians that are not shown well in the compared methods.

2) Image Segmentation Evaluation: For semantic segmentation, the performance of our method is evaluated on

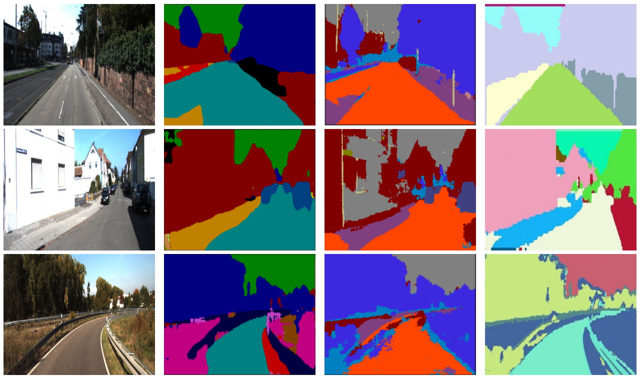

Fig. 4: Multi-class segmentation visual comparisons on KITTI. First column: Input RGB image; Second column: FCN_ResNet50 [15] [26]; Third column: Segnet [2]; Fourth column: Ours.

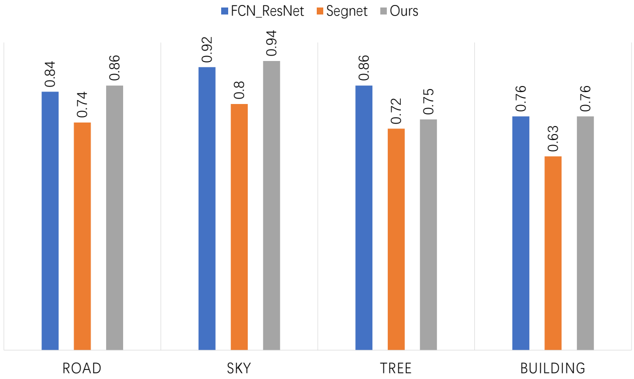

Fig. 5: Numerical comparison on the most typical scenes of KITTI.

KITTI dataset, PASCAL VOC2012 dataset, and Berkeley Segmentation dataset (BSDS500) [7]. Fig. 4 shows comparison results compared with SegNet[2] and FCN_ResNet50 [15] [26] on KITTI dataset, which demonstrates the segmentation precision of our method in segmenting complex urban scenes. We also conduct numerical analysis using metric segmentation covering (SC) as Fig. 5. We choose several most typical scenes to compare (road, sky, tree, building) segmentation effect. We can see that even though our method is unsupervised, it can still achieve higher or comparable segmentation accuracy compared with recent supervised deep learning segmentation methods. As two of the most popular supervised learning methods, SegNet and FCN_ResNet50 require extensive labeled training data to learn a segmentation deep neural network. Though our framework is unsupervised, our method can provide more accurate and clear segments, which demonstrates that our fully unsupervised method is comparable to supervised image segmentation methods. Fig. 6 shows segmentation results when our pre-trained model is directly applied to other datasets such as PASCAL VOC2012 dataset and BSDS500 without training on them. It can be observed that for these datasets that images mainly consist of background and a limited number of foreground objects, many meaningful segments can be obtained adaptively for different scenes by our method. Even if our model is not trained on these datasets, it can still produce a reasonable segmentation output map, which shows the robustness of our framework to unseen data.

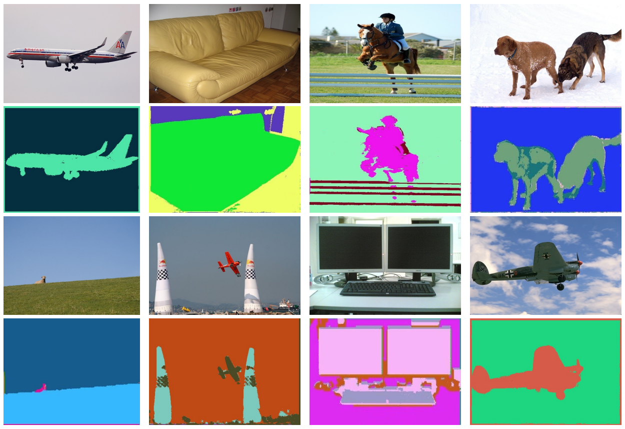

Fig. 6: Our multi-class segmentation prediction examples on PASCAL VOC 2012 and Berkeley Segmentation datasets. Note that our model is trained only on KITTI dataset.

Metric	Abs Rel	RMSE	IoU
Depth only	0.120	6.062	-
Segmentation only	-	-	47.8
Joint learning	0.115	5.828	51.4

TABLE II: Comparison of joint learning v.s. single task learning on KITTI dataset. We observe an improvement in performance when training with our joint framework and multi-task loss constraints.

C. *Framework Setting Analysis*

We analyze the benefit of jointly predicting depth and segmentation in our framework, then conduct a series of experiments to evaluate the effectiveness of our proposed architecture. Segmentation can be used to avoid small gaps in estimated depth. Meanwhile, the separation of foreground from the image can also prevent discontinuities and miss-prediction in sky and ground. On the other hand, accurate depth estimation can also help to capture the most salient boundaries to improve the image segmentation output.

As shown in Table II, our proposed network benefits each other on the two tasks compared with just deploying a single task. Segmentation task has a 7.5% improvement in Intersection over Union (IoU) for our predicted classes, and depth estimation task has a 4.2% and 3.8% decrease in Absolute Relative difference (Abs rel) and Root Mean Square Error (RMSE) respectively on the tests of KITTI, which indicates the performance of each task can be increased by learning from each other under our proposed network.

VI. Conclusion

In this paper, we propose a novel multi-task learning framework for fully unsupervised depth prediction and image segmentation from just one RGB image as input. To better leverage their relationships and common features, we designed convolutional modules with shared weights. The proposed framework can adaptively leverage on the common information of the two tasks, and encourage their interactive learning to benefit from each other. Comprehensive evaluations on benchmark datasets demonstrate the robustness of our network on jointly dealing with single image depth estimation and segmentation problems. Since our model is end-to-end and independent on any prior knowledge or labels of the scene, it can be easily applied to new scenarios.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[3] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018.

[4] Q. Chen and V. Koltun. Fast mrf optimization with application to depth reconstruction. In *CVPR*, pages 3914–3921, 2014.

[5] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):603–619, 2002.

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016.

[7] P. Dollár and C. L. Zitnick. Structured forests for fast edge detection. In *ICCV*, pages 1841–1848, 2013.

[8] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NIPS*, 2014.

[9] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[10] R. Garg, V. K. BG, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, pages 740–756. Springer, 2016.

[11] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361, 2012.

[12] A. Geiger, J. Ziegler, and C. Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 963–968, 2011.

[13] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017.

[14] A. Handa, T. Whelan, J. McDonald, and A. J. Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *ICRA*, pages 1524–1531, 2014.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[16] A. Holynski and J. Kopf. Fast depth densification for occlusion-aware augmented reality. In *SIGGRAPH Asia 2018 Technical Papers*, page 194, 2018.

[17] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. In *NIPS*, pages 2017–2025, 2015.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[19] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011.

[20] P. Krähenbühl and V. Koltun. Efficient nonlocal regularization for optical flow. In *ECCV*, pages 356–369. Springer, 2012.

[21] Y. Kuznietsov, J. Stückler, and B. Leibe. Semi-supervised deep learning for monocular depth map prediction. In *CVPR*, pages 6647–6655, 2017.

[22] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *3DV*, pages 239–248, 2016.

[23] X. Lin, D. Sánchez-Escobedo, J. R. Casas, and M. Pardàs. Depth estimation and semantic segmentation from a single rgb image using a hybrid convolutional neural network. *Sensors*, 19(8):1795, 2019.

[24] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *CVPR*, pages 5162–5170, 2015.

[25] F. Liu, C. Shen, G. Lin, and I. D. Reid. Learning depth from single monocular images using deep convolutional neural fields. *TPAMI*, 38(10):2024–2039, 2016.

[26] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[27] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[28] R. Mahjourian, M. Wicke, and A. Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, pages 5667–5675, 2018.

[29] A. Mousavian, H. Pirsiavash, and J. Košecká. Joint semantic segmentation and depth estimation with deep convolutional networks. In *3DV*, pages 611–619, 2016.

[30] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia. Towards real-time unsupervised monocular depth estimation on cpu. In *IROS*, pages 5848–5854, 2018.

[31] M. Poggi, F. Tosi, and S. Mattoccia. Learning monocular depth estimation with unsupervised trinocular assumptions. In *3DV*, pages 324–333, 2018.

[32] A. Rajagopalan, S. Chaudhuri, and U. Mudenagudi. Depth estimation and image restoration using defocused stereo pairs. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1521–1525, 2004.

[33] P. Z. Ramirez, M. Poggi, F. Tosi, S. Mattoccia, and L. Di Stefano. Geometry meets semantics for semi-supervised monocular depth estimation. In *ACCV*, pages 298–313, 2018.

[34] R. Ranftl, V. Vineet, Q. Chen, and V. Koltun. Dense monocular depth estimation in complex dynamic scenes. In *CVPR*, pages 4058–4066, 2016.

[35] A. Ranjan, V. Jampani, L. Balles, K. Kim, D. Sun, J. Wulff, and M. J. Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, pages 12240–12249, 2019.

[36] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[37] J. Shi and J. Malik. Normalized cuts and image segmentation. *Departmental Papers (CIS)*, page 107, 2000.

[38] D. Sun, S. Roth, and M. J. Black. Secrets of optical flow estimation and their principles. In *CVPR*, pages 2432–2439. IEEE, 2010.

[39] C. Wang, J. Miguel Buenaposada, R. Zhu, and S. Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, pages 2022–2030, 2018.

[40] X. Xia and B. Kulis. W-net: A deep model for fully unsupervised image segmentation. *arXiv preprint arXiv:1711.08506*, 2017.

[41] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *CVPR*, volume 1, 2017.

[42] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia. Every pixel counts: Unsupervised geometry learning with holistic 3d motion understanding. In *ECCV*, pages 0–0, 2018.

[43] Z. Yang, P. Wang, Y. Wang, W. Xu, and R. Nevatia. Lego: Learning edge with geometry all at once by watching videos. In *CVPR*, pages 225–234, 2018.

[44] Z. Yin and J. Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, pages 1983–1992, 2018.

[45] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.

[46] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017.

[47] Y. Zou, Z. Luo, and J.-B. Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 36–53, 2018.

[48] L. Zwald and S. Lambert-Lacroix. The berhu penalty and the grouped effect. *arXiv preprint arXiv:1207.6868*, 2012.