Disparity weighted loss for semantic segmentation of driving scenes

Abdelhak Loukkal, Yves Grandvalet, You Li

▶ To cite this version:

Abdelhak Loukkal, Yves Grandvalet, You Li. Disparity weighted loss for semantic segmentation of driving scenes. 22nd IEEE International Conference on Intelligent Transportation Systems (ITSC 2019), Oct 2019, Auckland, New Zealand. pp.3427-3432, 10.1109/ITSC.2019.8917171 . hal-02465013

HAL Id: hal-02465013
https://hal.science/hal-02465013

Submitted on 5 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/336926978

Disparity weighted loss for semantic segmentation of driving scenes

Conference Paper · October 2019

DOI: 10.1109/ITSC.2019.8917171

CITATIONS
0

READS
308

3 authors, including:

Abdelhak Loukkal
Université de Technologie de Compiègne
4 PUBLICATIONS **1** CITATION

SEE PROFILE

You Li
Renault
27 PUBLICATIONS **219** CITATIONS

SEE PROFILE

All content following this page was uploaded by You Li on 31 October 2019.

The user has requested enhancement of the downloaded file.

Disparity weighted loss for semantic segmentation of driving scenes

Abdelhak Loukkal[1,2], Yves Grandvalet[2], You Li[1]

Abstract— **Convolutional neural networks are the state of the art methods for semantic segmentation but their resource consumption hinders their usability for real-time mobile robotics applications. Recent works have focused on designing lightweight networks that require less resources, but their efficiency is accompanied with a drop in performance. In this work, we propose a pixel-wise weighting of the cross-entropy loss with the disparity map in order to give more importance to close objects during the optimisation procedure of the network. This weighting is applied to two lightweight networks, with different efficiency/performance trade-offs, that were designed for real-time autonomous driving. These networks are trained on CamVid and Cityscapes datasets and the disparity maps are obtained with an off-the-shelf unsupervised depth estimation network. Our method does not increase the number of parameters of the network nor imply any further manual labeling. This weighting is evaluated on both the regular mean intersection over union (mIoU) and a close-range mIoU. Compared to the standard weighting scheme, this new loss weighting improves the mIoU and the IoU of pertinent classes for autonomous driving especially at close range.**

I. INTRODUCTION

Semantic segmentation has been drawing a lot of attention from computer vision and autonomous driving communities for many years because in addition to detecting key elements in the scene, it adds semantic information to the global scene understanding problem. Convolutional neural networks (CNN) have achieved impressive semantic segmentation results and replaced the classic computer vision methods. However, state-of-the-art CNNs rely on a very important number of parameters and this comes with the price of a prohibitive resource consumption. This computational burden makes these networks less convenient for an autonomous car where multiple cameras are needed and resources and data bandwidth are limited. Recent works have focused on designing very efficient networks that can run with a sufficient frame rate on modern GPUs dedicated to autonomous driving. These networks have fewer parameters than their state-of-the-art counterparts and the direct consequence is a major drop in performance.

One of the issues in driving scene datasets is the imbalance between the labeled classes: critical classes like pedestrians or cyclists are under-represented compared to the sky or buildings. In order to compensate this imbalance, the loss function of the CNN is weighted according to the frequency of the classes, under-represented classes having the biggest weights. However, class imbalance is not the only issue. Finely segmenting an object located far from the ego vehicle does not seem to be a necessary asset for an autonomous pilot system. A coarse segmentation or a bounding in this situation should be sufficient , whereas having access to a fine semantic segmentation of close objects can be useful to have a better free space estimation.

In this work, a new weighting scheme, depicted in Figure 1, is proposed. Based on the assumption that objects that are close to the vehicle are more important than those located far away, this weighting scheme gives more weight to close objects in the training loss. This is accomplished by pixel-wise multiplying the cross-entropy loss by the disparity map for each training image. The disparity maps are precomputed for the whole dataset prior to training with an off-the-shelf unsupervised CNN. The efficacy of the weighting is evaluated with the mean Intersection over Union (mIoU) metric and a close range mIoU on two different networks and two datasets. A comparison with the frequency-based weighting shows that this weighting scheme improves the mIoU and the IoU for some important classes, especially in close range.

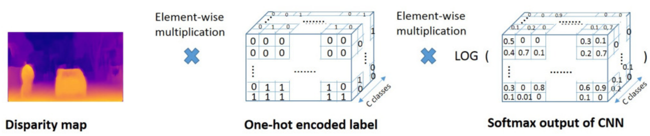


Fig. 1: Disparity weighting: Each pixel in the cross-entropy loss function is weighted by its value in the disparity map.

II. RELATED WORK

The success of deep CNNs for image classification [1] has encouraged researchers to explore the effectiveness of these networks for dense predictions tasks like semantic segmentation or depth estimation. The current state of the art approaches in semantic segmentation leverage the idea of fully convolutional neural networks [2] keeping the spatial information, that is eventually lost in regular CNNs, by avoiding fully connected layers. These networks come in two parts: the encoder extracts features from the input image and the decoder up-samples the encoded feature map to match the size of the input image. Skip connections between the input and output of convolution layers were introduced in [3] helping the gradient flow in deep architectures. Fully connected conditional random fields (CRF) [4] were the state of the art approach in semantic segmentation before the resurgence of neural networks. Used as a post processing step [5] or reformulated as a recurrent network [6], CRFs

[1]Renault S.A.S, 1 av. du Golf, 78288 Guyancourt, France.
[2]Sorbonne universités, Université de technologie de Compiègne, CNRS, Heudiasyc, UMR 7253, Compiègne, France.

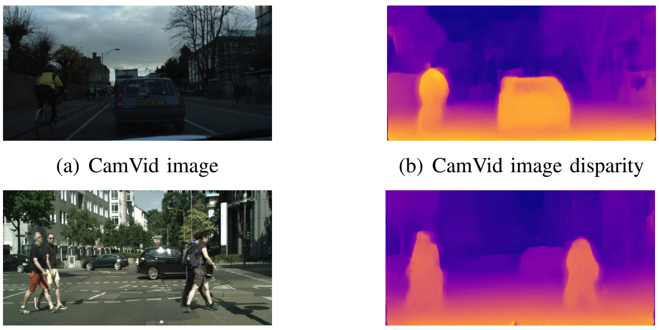
are still popular in the computer vision community. State of the art networks for pixel wise semantic segmentation take also advantage of spatial pyramid pooling [7] and dilated convolutions [5] to segment objects at different scales and enlarge the receptive fields, hence the context, of the convolution filters.

Another important aspect for CNNs, especially for mobile robotics, is their resource consumption. Adam Paszke et al, introduced ENET [8], an efficient CNN for semantic segmentation that is 18 times faster than [9] meanwhile it obtains better results on the mean Intersection over Union (mIoU) metric. Real-time computation is achieved with early downsampling that heavily reduces the size of the input. A consequence of the aggressive downsampling is the loss of information that incurs in much lower performance than state of the art networks. Introduced in [10], ERFNET is another efficient network based on the skip connections [3] and 1D convolutions to reduce resource usage. ERFNET is twice as slow as ENET but achieves much better mIoU.

Semantic segmentation is a very complex task for which satisfying results were obtained thanks to public datasets like KITTI [11], CamVid [12] and Cityscapes [13] which contains much more labeled images than the two previous ones. More recently, thanks to a new labeling pipeline, Apolloscapes [14], a huge dataset of more than 100k labeled images was released.

III. Disparity weighted cross-entropy loss

A. Disparity with an unsupervised network



(a) CamVid image (b) CamVid image disparity (c) Cityscapes image (d) Cityscapes image disparity

Fig. 2: Disparity estimation with an off-the-shelf unsupervised CNN

Disparity maps can be acquired with a stereo setup but the acquired maps are sparse and contain few measurements at far distance. Given enough depth labels, supervised CNNs [15] can also be used to estimate dense depth maps but require a large amount of labeled data.

Recent works [16] have leveraged epipolar geometry constraints to estimate disparity maps as an intermediate output in an image reconstruction network. It consists in taking the left image as input, estimating the disparity map and performing a reconstruction of the right image using the disparity and a bi-linear sampler. Two disparity maps are produced, left to right and right to left, and a left-right consistency term is added to the reconstruction loss function.

This CNN is not suitable for an application where a precise depth information is required but for this new weighting scheme, only a magnitude estimation is required. The disparity maps are precomputed for the whole dataset before training, thus not requiring any additional labelling efforts.

B. Loss weighting

The usual cross entropy loss is the following:

$$-\sum_{i=1}^{h}\sum_{j=1}^{w}\sum_{k=1}^{c} t_{ijk}\log(o_{ijk}), \quad (1)$$

where h, w are the dimensions of input images, c is the number of classes, and o and t are the (h, w, c) tensors corresponding respectively to the softmax output of the CNN and to the one-hot encoded segmentation labels.

To remedy class imbalance, the widely used technique is median class frequency balancing [17], that consists in weighting each pixel of class k by $\alpha_k = \text{median}(\{p_1, ..., p_c\})/p_k$ where p_k is the proportion of pixels of class k in the dataset. Another weighting, also based on the frequency of the classes in the dataset, is:

$$\alpha_k = 1/\log(\beta + p_k) \quad (2)$$

where β is a hyper-parameter [8]. This weighting is also used in [10]. The former weighting is referred to as FW and compared to disparity weighting in the experiments section. The disparity weighted cross-entropy is straightforward:

$$-\sum_{i=1}^{h}\sum_{j=1}^{w} d_{ij}\sum_{k=1}^{c} t_{ijk}\log(o_{ijk}) \quad (3)$$

where d is the (h, w) matrix of disparity map. Frequency based weightings put more weight on classes that are globally underrepresented in terms of surface. These weightings are not especially relevant for autonomous pilot systems, in the sense that a big truck located far from the vehicle may not be as important as a pedestrian close to the vehicle. Disparity weighting encourages the network to focus on the close range objects.

IV. Experiments

The disparity weighting is tested on the Pytorch implementations of ENET and ERFNET. These networks are trained using a single NVIDIA TITAN X with ADAM optimizer, momentum 0.9, a batch size of 4, initial learning rate of $5e^{-4}$ and weight decay of $2e^{-4}$. Pretrained weights on Cityscapes dataset are used for all networks.

These networks are trained on two semantic segmentation datasets, CamVid and Cityscapes. CamVid contains 367 training samples and 233 validation samples. Cityscapes contains 2975 training samples and 500 validation samples. For both datasets, the additional disparity labels are obtained with an unsupervised disparity CNN trained on Cityscapes. Tables I and II report the raw proportions of pixels belonging to each class.

TABLE I: Class proportions (in %) in the Cityscapes train set. Most represented classes are shown in bold.

road	sidewalk	building	wall	fence	pole	traffic_light	traffic_sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle	unlabeled
25.3	5.8	**25.3**	0.7	1.0	1.4	0.2	0.6	**17.6**	1.1	4.4	1.3	0.1	7.6	0.3	0.3	0.3	0.1	0.5	6.1

TABLE II: Class proportions (in %) in the CamVid train set. Most represented classes are shown in bold.

sky	building	pole	road	pavement	tree	traffic sign	fence	car	pedestrian	bicycle	unlabeled
21.1	**29.1**	1.2	**17.0**	4.4	**12.1**	1.5	1.4	7.0	0.8	0.3	4.2

The evaluation metric for each class is the intersection over union IoU defined as following:

$$IoU = \frac{TP}{TP + FP + FN} \ ,$$

where TP, FP and FN are respectively the true positive, false positive, and false negative pixel counts on the set of test images. The mIoU is the mean of the individual classes IoUs. A close-range IoU is introduced: it consists in filtering both the ground truth segmentation and the predicted segmentation with regard to the disparity. All pixels with a disparity value inferior to a defined threshold are ignored. The unsupervised depth estimation network outputs disparity values that are normalised by the image width: the depth values on each pixel can be retrieved by rescaling acccording to the image width. To obtain the value of the depth in meter, we use this relation relation between B, f and d respectively the baseline, the focal length of the camera and the disparity: $Depth = \frac{B*f}{d}$. We arbitrarily consider a depth of 30 meters as being our close range limit and use the corresponding disparity value as our threshold to filter the pixels. For each dataset and each efficient network, three weighting schemes are compared: No Weighting, Frequency Weighting and Disparity Weighting respectively referred to as NW, FW and DW. Here, frequency weighting corresponds to the weighting introduced in [8], see equ. 2, with $\beta = 1.02$.

The values in the result tables are averaged on 3 runs for each configuration.

A. Results on CamVid

Results for CamVid are presented in Table III.

a) ENET: ENET network with DW outperforms its FW counterpart on the mIoU by 1.1% but more importantly on important classes like pedestrians, cyclists, vehicles and traffic signs by respectively 6.4%, 2.7%, 0.4% and 3% . The network trained without any weighting has the worst mIoU and the worst IoU on most of the important classes. Another important effect of the disparity weighting is the decrease in the sky IoU: this class has always the smallest disparity value and has in consequence the smallest weight in the learning loss function.

b) ERFNET: The same effects are observed with ERFNET DW topping the FW by 1.5% on the mIoU and respectively 4.4%, 1.4%, 1.3% and 1.4% on the IoUs of pedestrians, cyclists, vehicles and traffic signs. We also observe the decrease on the sky IoU by a lesser margin.

TABLE III: Results on CamVid validation set. Best results are shown in bold. All networks are trained on 150 epochs.

	ENET			ERFNET		
	NW	FW	DW (ours)	NW	FW	DW (ours)
mIoU	51.6	52.6	**53.7**	66.6	68.1	**69.6**
Pedestrians	30.2	29.0	**35.4**	55.4	56.4	**60.8**
Cyclist	4.8	20.0	**22.7**	52.8	59.2	**60.6**
Vehicle	65.1	67.2	**67.6**	77.8	79.4	**80.7**
Traffic sign	25.6	25.1	**28.1**	44.6	44.7	**46.1**
Road	89.5	**89.7**	89.5	92.9	**94.2**	93.3
Fence	27.7	28.2	**31.9**	42.0	46.7	**51.4**
Pole	19.1	18.9	**19.6**	35.4	**38.4**	38.0
Building	**76.5**	73.6	75.7	82.3	81.7	**84.2**
Sky	89.1	**89.4**	82.5	**91.8**	91.4	89.4
Pavement	71.8	71.4	**72.2**	80.4	81.9	**82.0**
Tree	**67.8**	66.6	65.5	75.9	76.2	**77.1**

B. Results on Cityscapes

Results for Cityscapes dataset are presented in Table IV. Both ENET and ERFNET networks with disparity weighting improve on the mIoU by respectively 0.5% and 0.3%. However, the disparity weighting does not improve the IoU of all important classes.

a) ENET: We observe the effect of the disparity weighting on the road, sidewalk, bicycle and pedestrians IoUs with respective improvements of 1.7%, 4.7%, 3.5% and 4.2% compared to FW. However FW outperforms DW on important classes like rider, car, bus, train and motorcycle.

b) ERFNET: DW obtains the best IoUs for road, pedestrians, trains, bicycle, and cars with respective improvements of 0.5%, 1% , 2.9% , 0.5% and 0.4%. Overall, ERFNET achieves better results than ENET when weighted with the disparity map.

c) Close range evaluation: Given that DW puts more weight on closer objects, these two networks are evaluated with the close range IoU to verify their emphasis on close objects, see Table V. ENET DW obtains the best IoU on most of the important classes with the exception of bus, train and motorcycle. ERFNET obtains also the best IoU results except for motorcycle and traffic light. In close range, both networks perform better with our DW than with other weighting schemes on the regular IoU.

Figures 3, 4, and 5 show the evolution of the IoU with respect to the depth threshold for some important classes. The presented results are for ERFNET trained on Cityscapes

TABLE IV: Results on Cityscapes validation set. Best results are shown in bold. All networks are trained on 150 epochs.

	ENET			ERFNET		
	NW	FW	DW (ours)	NW	FW	DW (ours)
mIoU	50.4	51.2	**51.7**	66.6	70.5	**70.8**
Road	93.7	93.1	**94.8**	96.5	96.6	**97.1**
Sidewalk	68.7	67.1	**71.8**	80.3	80.6	**82.3**
Building	84.6	83.4	**84.7**	89.9	90.1	**90.6**
Wall	**26.3**	24.0	24.7	44.4	51.5	**52.8**
Fence	29.5	30.8	**33.3**	50.2	54.4	**55.2**
Pole	**39.2**	38.0	**39.2**	57.5	59.1	**59.2**
Traffic light	20.7	**23.9**	20.5	57.1	**60.4**	59.3
Traffic sign	**38.5**	33.8	37.3	68.1	**71.4**	71.2
Vegetation	81.1	85.5	**86.8**	90.7	91.0	**91.1**
Terrain	42.7	37.6	**44.1**	58.4	60.4	**61.9**
Sky	**87.1**	86.2	86.6	**93.3**	91.8	93.2
Pedestrians	56.3	53.6	**57.8**	72.5	75.1	**76.1**
Rider	21.7	**27.3**	25.4	46.9	**53.3**	53.1
Car	86.0	**86.5**	86.4	91.5	92.5	**92.9**
Bus	45.8	**54.2**	51.1	57.1	**74.5**	74.3
Train	23.0	**29.2**	27.8	46.5	58.0	**60.9**
Motorcycle	10.0	**21.6**	13.8	35.5	**44.3**	43.5
Bicycle	53.5	51.5	**55.0**	66.2	69.7	**70.2**

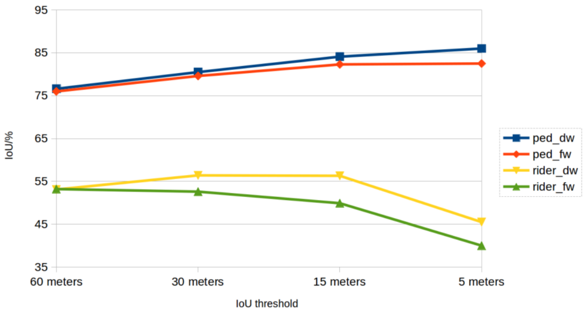


Fig. 3: IoU of pedestrians and riders classes with different depth thresholds. All pixels with a depth value superior to the threshold are ignored in the IoU computation.

for both frequency and disparity weightings(best of the 3 runs for each for each weighting). In this experiment, all pixels with a depth value above the specified thresholds are ignored in the IoU computation. The depth values are rough estimations as the unsupervised network is not precise. In these figures, we first observe that for some classes IoUs improve as the depth threshold decreases and for other classes IoUs increase then decrease when getting too close to the vehicle. This can be explained by the fact that some classes are not very well represented in close range during training: road or cars are represented at all ranges but this is not the case for the bus class for example. We also clearly observe the effect of disparity weighting when getting closer to the ego vehicle. The IoUs with a 60 meters threshold are equivalent for both weightings or slightly better for frequency weighting but when the threshold is at 30 meters or less, disparity weighting has consistently a better IoU (by a significant margin for bus class, 20%).

TABLE V: Close-range results on Cityscapes validation set. Best results are shown in bold. All networks are trained on 150 epochs.

	ENET			ERFNET		
	NW	FW	DW (ours)	NW	FW	DW (ours)
mIoU CR	52.2	52.4	**53.9**	67.6	71.2	**72.5**
Road	95.3	94.1	**96.3**	98.8	98.8	**99.0**
Sidewalk	72.7	70.8	**76.4**	81	81.3	**83.1**
Fence	36.7	39.3	**42.3**	50.1	**54.8**	52.5
Pole	48.7	47.3	**49.5**	66.6	67.2	**68.3**
Traffic light	25.4	23.8	**26.8**	67.0	**69.2**	68.9
Traffic sign	49.1	40.0	**50.1**	75.9	**78.2**	**78.2**
Pedestrians	59.7	59.7	**61.7**	76.6	79.0	**80.3**
Rider	28.9	33.0	**33.2**	48.6	53.6	**54.2**
Car	90.3	89.9	**91.3**	93.7	94.2	**94.7**
Bus	49.0	**52.3**	51.9	61.1	75.3	**77.7**
Train	**25.9**	24.9	19.2	50.9	63.3	**71.6**
Motorcycle	9.7	**26**	11.7	37.8	**47.1**	46.5
Bicycle	61.0	57.6	**63.7**	70.1	73.5	**74.2**

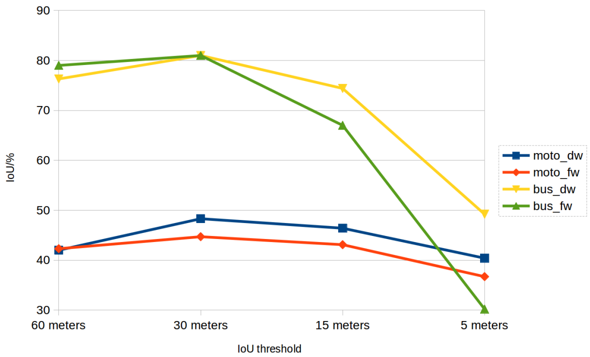


Fig. 4: IoU of motorcycle and bus classes with different depth thresholds. All pixels with a depth value superior to the threshold are ignored in the IoU computation. .

C. *Qualitative results*

Figure 6 presents different situations where the close range effect of the disparity weighting is observed. On these figures we observe that DW produces better road segmentation in close range which can avoid dangerous situations like in the sixth image where a portion of sidewalk is detected in the middle of the road. We also observe that DW ensures better pedestrian delineation and sometimes detects objects that are not detected by FW: on the first, the second and the sixth image the rider in front of the vehicle is detected by FW but not the bicycle he rides. On image 3 and 5 we observe that DW assigns the correct class to the truck and the bus but FW mistakes some part of these objects with other similar classes. A better segmentation in close range avoids

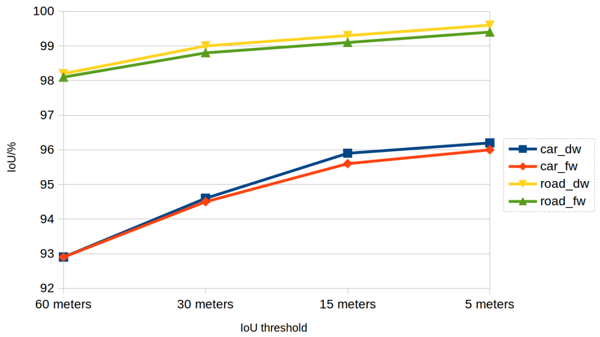


Fig. 5: IoU of road and car classes with different depth thresholds. All pixels with a depth value superior to the threshold are ignored in the IoU computation.

dangerous situations where the vehicle could brake because of a false alarm or hit an obstacle that was not detected.

D. Discussion

Experiments show improving but varying results depending on the dataset, the range of the metric and the chosen network. A first explanation of these variations is given in Tables I and II. The proportion of pixels corresponding to the class sky in CamVid dataset is roughly 5 times greater than in Cityscapes, which can explain why disparity weighting works better on the former. A large part of the images in CamVid corresponds to the sky and this part is ignored in the loss function which helps the network focus on the other more important classes. The performance increase is not very important because these efficient networks naturally segment better closer objects because close results in big objects. Nevertheless, the disparity weighting still has interesting properties and clearly observable effects on close range segmentation.

V. Conclusion

In this paper, a new loss weighting scheme is introduced for semantic segmentation of driving scenes. This weighting consist in multiplying each pixel in the training samples by its disparity value. The disparity maps for each sample are precomputed prior to training with an off-the-shelf unsupervised CNN. Intersection over union metric is improved on CamVid and Cityscapes dataset with better results on the former. This is partly explained by the proportion of classes in both datasets, CamVid sky class being over represented it is ignored because of its very low disparity, hence putting more emphasis on other classes. Further investigation on Cityscapes dataset show even more improvements when evaluating with a close range IoU. Without any additional labelling effort nor computation burden, disparity weighting improves semantic segmentation performance. A similar approach with dense optical flow could be tested with an unsupervised flow estimation network. Instead of putting the emphasis on close objects, it could be done on moving objects which would also make sense for autonomous driving.

Acknowledgement

This work has been conducted within SIVALAB, a joint research laboratory between Renault S.A.S and Heudiasyc laboratory, which targets issues pertaining to integrity questions in autonomous driving.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, pages 1106–1114, 2012.

[2] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[4] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pages 109–117, 2011.

[5] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[6] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6230–6239, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *13th European Conference on Computer Vision (ECCV)*, pages 346–361, 2014.

[8] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv preprint arXiv:1606.02147*, 2016.

[9] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017.

[10] Eduardo Romera, Jose M. Alvarez, Luis Miguel Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Trans. Intelligent Transportation Systems*, 19(1):263–272, 2018.

[11] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *I. J. Robotics Res.*, 32(11):1231–1237, 2013.

[12] Gabriel J. Brostow, Julien Fauqueur, and Roberto Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognit. Lett.*, 30(2):88–97, 2009.

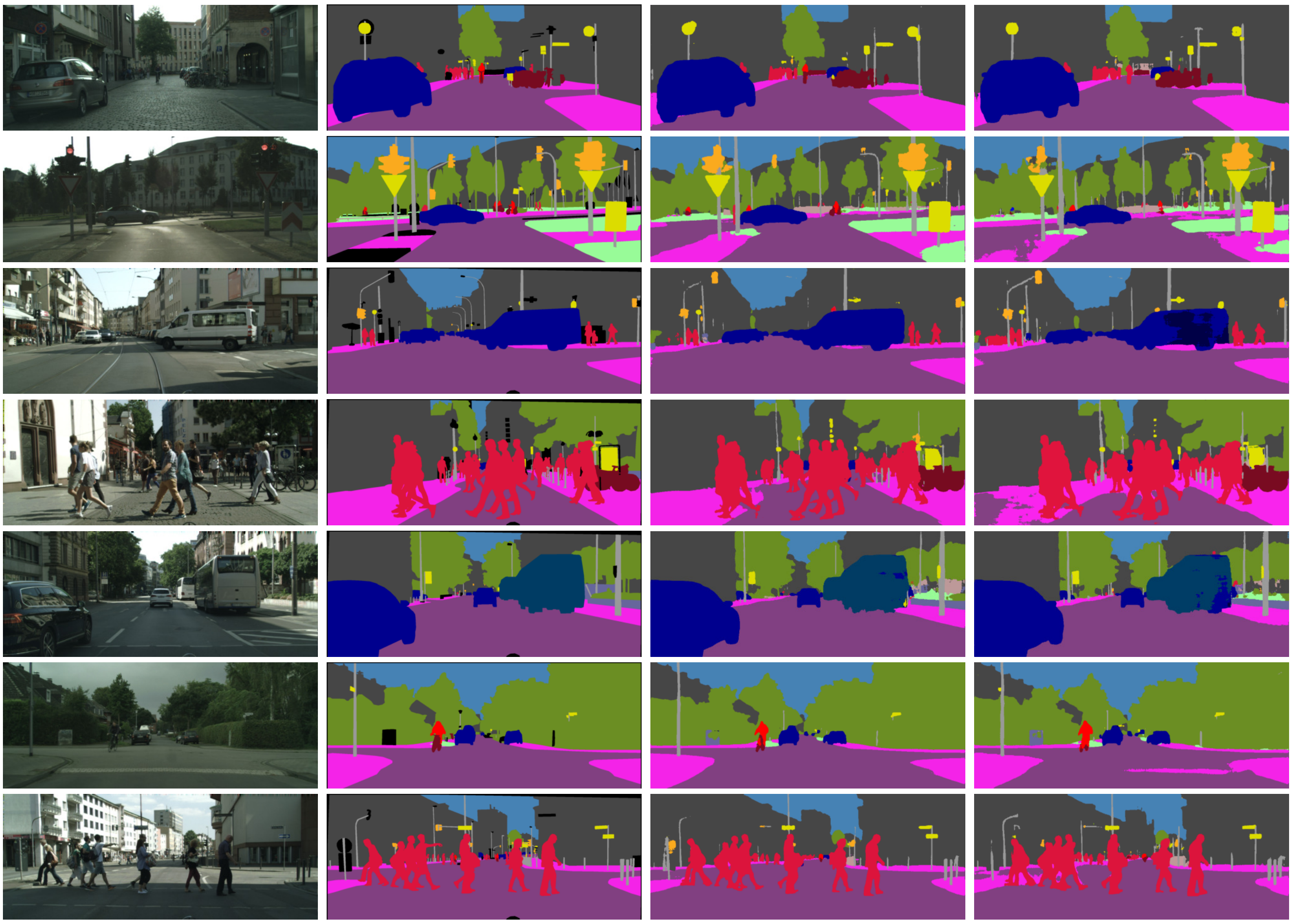
[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[14] Xinyu Huang, Xinjing Cheng, Qichuan Geng, Binbin Cao, Dingfu Zhou, Peng Wang, Yuanqing Lin, and Ruigang Yang. The ApolloScape dataset for autonomous driving. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 954–960, 2018.

[15] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *Fourth International Conference on 3D Vision (3DV)*, pages 239–248, 2016.

[16] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6602–6611, 2017.

[17] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *International Conference on Computer Vision (ICCV)*, pages 2650–2658, 2015.



(a) RGB image (b) Ground truth (c) DW (ours) (d) FW

Fig. 6: Qualitative results on Cityscapes with ERFNET: we observe that DW behaves better in close range

View publication stats