Dense Relation Network: Learning Consistent and Context-Aware Representation for Semantic Image Segmentation

Conference Paper · October 2018

DOI: 10.1109/ICIP.2018.8451830

CITATIONS
48

READS
1,365

9 authors, including:

Yueqing Zhuang
Peking University
5 PUBLICATIONS **205** CITATIONS
SEE PROFILE

Fan Yang
Peking University
25 PUBLICATIONS **390** CITATIONS
SEE PROFILE

Li Tao
The University of Tokyo
18 PUBLICATIONS **356** CITATIONS
SEE PROFILE

Cong Ma
Peking University
10 PUBLICATIONS **223** CITATIONS
SEE PROFILE

DENSE RELATION NETWORK: LEARNING CONSISTENT AND CONTEXT-AWARE REPRESENTATION FOR SEMANTIC IMAGE SEGMENTATION

Yueqing Zhuang, Fan Yang, Li Tao, Cong Ma, Ziwei Zhang, Yuan Li, Huizhu Jia, Xiaodong Xie, Wen Gao*

National Engineering Laboratory for Video Technology, Peking University, Beijing 100871, China

ABSTRACT

Semantic image segmentation, which aims at assigning pixel-wise category, is one of challenging image understanding problems. Global context plays an important role on local pixel-wise category assignment. To make the best of global context, in this paper, we propose dense relation network (DRN) and context-restricted loss (CRL) to aggregate global and local information. DRN uses Recurrent Neural Network (RNN) with different skip lengths in spatial directions to get context-aware representations while CRL helps aggregate them to learn consistency. Compared with previous methods, our proposed method takes full advantage of hierarchical contextual representations to produce high-quality results. Extensive experiments demonstrate that our method achieves significant state-of-the-art performances on Cityscapes and Pascal Context benchmarks, with mean-IoU of 82.8% and 49.0% respectively.

***Index Terms*—** Image Semantic Segmentation, Context-Aware Representation, Context-Restricted Loss.

1. INTRODUCTION

Semantic image segmentation is a fundamental computer vision problem whose goal is to assign category in pixel level. This topic attracts broad interest for applications such as automatic driving, remote sensing and medical image processing which need accurate boundaries of objects.

To solve this problem, in previous decades, traditional methods depend on pixel-level hand-crafted features [1] combined with a classifier [2]. Driven by powerful deep neural network in classification [3], pixel-level tasks like semantic image segmentation have achieved great success by replacing fully-connected layers with convolution layers in classifier which enables network to generate image [4]. Currently, state-of-the-art segmentation frameworks are mainly based on fully convolutional network (FCN) [4], which can be roughly divided into two parts, feature extraction and classification.

*means corresponding author(Email:hzjia@pku.edu.cn). This work is partially supported by the National Key Research and Development Program of China under contract No. 2016YFB0401904, Major National Scientific Instrument and Equipment Development Project of China under contract No. 2013YQ030967, National Science Foundation of China under contract No. 61602011 and NVIDIA NVAIL program

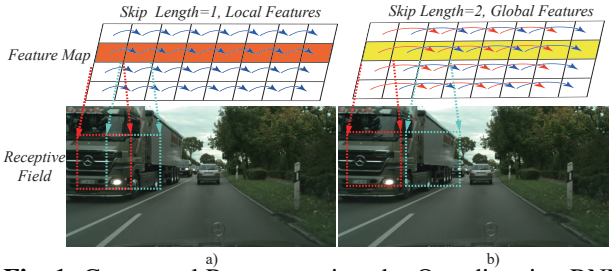

Fig. 1: Contextual Representations by One-direction RNN in Dense Relation Module. **a)** F_1 is produced from previous feature map by a 1-skip GRU; **b)** F_2 whose size of receptive field is 2 times than F_1 by 2-skip GRU;

Extracting a better feature representation is important to distinguish pixels. A popular way to design discriminative features is based on multi-scale technology which can be roughly summarized as two categories, image-level and network-level technologies. Image-level technology [5] uses images of different scale to extract features where low-resolution image contains the global contextual representation and high-resolution image incorporates the local representation. Network-level technology [6] uses characteristics of neural networks in which low layers attach importance to detail information while high layers capture global information. Besides, scale-aware operation [7, 8] in network is used to extract multi-scale features.

For classification, different from traditional classification methods such as SVM and Adaboost, structural classifications like CRF (conditional random field) [9] consider relatedness of pixels to refine results. The method in [10] refines networks by end-to-end modeling, which integrates CRF into convolutional network. Moreover, work [11] uses neural network to estimate CRF. Methods mentioned above are established in the factor of pair-wise consistency and relatedness.

However, these FCN-based methods [4, 5, 6, 7, 8] suffers from the lack of suitable capacity for utilizing the global contextual information due to the shrunken receptive field of CNN [12]. Structural classifications [9, 10, 11, 13] aim at solving pair-wise relation between pixels yet not extract suitable features to consider global context.

To handle this problem, we propose dense relation network which uses RNN with different skip lengths in spatial directions to enlarge receptive field and aggregate contextual

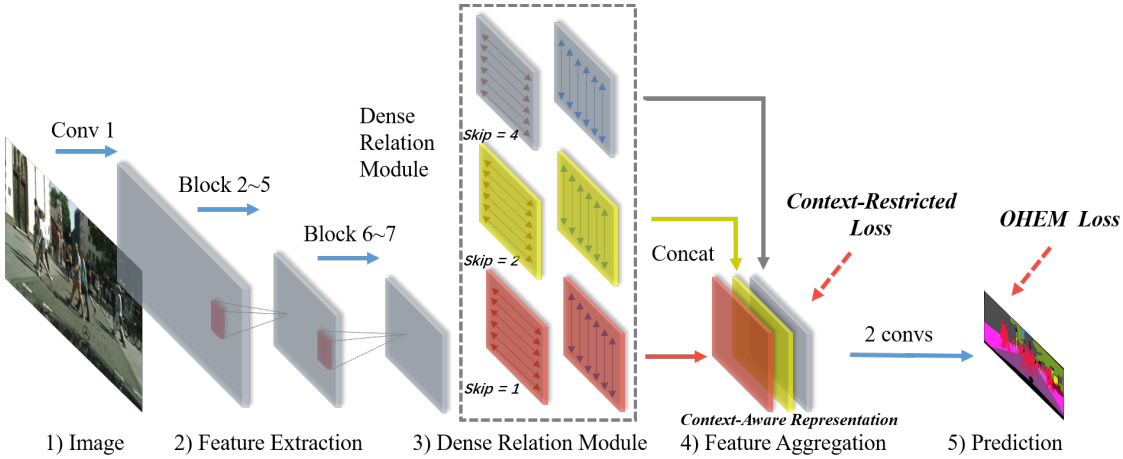

Fig. 2: Visualization of DRN, which is composed by feature extraction subnet, dense relation module, feature aggregation part.

information of different scale (Fig. 1). Meanwhile, contextual representations of the same label are more alike, Context-Restricted Loss (CRL) is proposed to constrict the consistency of contextual representations assigned to the same label. In this way, our approach achieves state-of-the-art performances on Cityscapes dataset [14] and Pascal Context dataset [15], with 82.8% and 49.0% in terms of mean-IoU respectively.

2. METHODS

CNN has limited receptive field which will result in insufficient utilization of contextual information [12]. To solve this problem, we make the best use of RNN to aggregate context. The overall framework of our approach is shown in Fig. 2, which can be divided into three parts, including feature extraction, dense relation module and classification subnet. Moreover, Online Hard Example Mining (OHEM) is used to deal with unbalanced examples while Context-Restricted Loss (CRL) is proposed to handle contextual consistency.

2.1. Dense Relation Module

In a CNN, the size of receptive field can roughly indicates how much contextual information we use. The empirical receptive field (ERF) of a convolutional network is limited though theoretical receptive field (TRF) is larger than the input image [12]. To fuse global context, global average pooling is a good choice to learn global description [16, 17]. However, this strategy is insufficient for complex environment, directly forming a single representation for contextual environment may lose the spatial relation and cause ambiguity. Fusing contexts of different sub-regions [7, 8, 18] is a way to get more powerful representations.

To resist the shrunken receptive field of CNN (the size of ERF is smaller than TRF) and make full use of contextual information, we utilize Recurrent Neural Network (RNN) to aggregate global contexts. Our proposed dense relation module uses four-direction RNNs to learn suitable contextual information (Fig. 2) so that the receptive field is larger than CNN. In order to aggregate different-scale contexts, skip length in RNN is set as Fig. 1. Meanwhile, it's more important to consider context near the pixel rather than context far away from the pixel (which cannot be ignored as well). To evaluate the importance of different-scale representations, the output dimension of RNN (channel in feature map) is chosen by their importance. Features of different scales are concatenated into final features to be evaluated by a classifier, which composes of two convolutions.

Therefore, our proposed network is formulated as below:

$$\begin{aligned} fo_{m,n,s} &= \overrightarrow{\mathcal{C}}(fi_{m,n}, fi_{m+\triangle s,n+\triangle s}) \\ fo_{m,n} &= \mathcal{F}(fi_{m,n,1}, fi_{m,n,2}, fi_{m,n,4}) \end{aligned} \tag{1}$$

where we restrict the channel of contextual features as belows:

$$|fi_{m,n,1}| = 2|fi_{m,n,2}| = 4|fi_{m,n,4}| \tag{2}$$

where $\overrightarrow{\mathcal{C}}$ is a formation of Recurrent Neural Network (RNN), m, n is the spatial coordinates in a feature map. $\mathcal{F}(\cdot)$ is an aggregation function, in which the channel of global and local representations is restricted by Equation (2). Our dense relation module is composed by three scale submodules and four paralleled GRUs in each one (Fig. 2). We choose GRU [19] because it converges faster and can learn suitable receptive field than vanilla RNN.

2.2. Network Architecture

Our proposed Dense Relation Network (DRN) is illustrated in Fig. 2. Given an input image, we use ResNet38 [20] with dilated strategy [21] to extract features. The spatial size of network output is $1/8$ of the input image. On the top of the feature map, we use dense relation module to gather hierarchical global contexts. We set skip lengths of DRN $1, 2, 4$ to gather contextual information of different scales. At last we concatenate contextual representations of different dimensionality, which are followed by two convolution layers to generate final prediction map.

2.3. Loss Function

Unbalanced samples in sematic image segmentation datasets cause the preference on common categories that appears frequently and less improvement on the hard objective at training stage. In order to solve this problem, we adopt Online Hard Example Mining [22] from [23] as below:

$$\mathcal{L}_{ohem} = \frac{1}{\sum_i^N \sum_j^K \mathcal{I}\{y_i = j \; and \; p_{ij} < t\}} * \sum_i^N \sum_j^K \mathcal{I}\{y_i = j \; and \; p_{ij} < t\} log p_{ij} \tag{3}$$

where K is the number of category c_j in label space. Suppose that we flatten an image into a one-dimensional pixel array and there are N pixels we shall predict. i is the mark number identifying the pixel. p_{ij} is the probability of the $pixel_i$ assigned to the category c_j. y_i is the target label of $pixel_i$. $\mathcal{I}(\cdot)$ is indicator function whose value is set to 1 if condition is satisfied and is set to 0 when condition fails. As $\mathcal{L}_{ohem}$ discards high-confidence loss according to threshold t, network would pay more attention on hard example at the training stage.

Pixel-wise labeling depends on contextual information because pixels make up objects. Pixels assigned to the same label should have more consistency in high-level contextual representations. Therefore, inspired by Center Loss [24], we compose context-restricted loss for hierarchical contextual representations as below:

$$\mathcal{L}_{crl} = \sum_{s=1,2,4} \sum_{m=1}^{W} \sum_{n=1}^{H} \frac{\|f_{mn,s} - C_{k_{mn},s}\|^2}{2N_{k_{mn}}} \tag{4}$$

where $f_{mn,s}$ is hierarchical features out of DRN. k_{mn} is the category of pixel, and $N_{k_{mn}}$ is total number of category k_{mn}. $C_{k,s}$ is the feature center of hierarchical environment for the scale s and the category k. With these settings, contextual representations for the category k_{mn} are constricted.

To update contextual representations for different categories at each iteration, we formulate update function as follows:

$$C_{\hat{k},\hat{s}}^{t+1} = C_{\hat{k},\hat{s}}^{t} + \eta \sum_{m=1}^{W} \sum_{n=1}^{H} \frac{f_{mn,\hat{s}} \mathcal{I}\{\; k_{mn} = \hat{k}\}}{N_{k_{mn}}} \tag{5}$$

For a specific category $\hat{k}$ and contextual $\hat{s}$, contextual representation $C_{\hat{k},\hat{s}}$ would be updated according to updating rates η. $\mathcal{I}(\cdot)$ is indicator function.

We define our loss function as the sum of L_{ohem} and L_{crl}, which are weighted by λ. The loss function for DRN is as below:

$$\mathcal{L} = \mathcal{L}_{ohem} + \lambda \mathcal{L}_{crl} \tag{6}$$

3. EXPERIMENTS

3.1. Datasets and Evaluation Protocol

We evaluate the performance of our DRN with CRL on two widely used semantic image segmentation datasets, Cityscapes [14] and Pascal Context [15]. Cityscapes contains 5000 high quality pixel-level finely annotated images (2975, 500, 1525 images for training, validation and testing) and 20k coarsely annotated images, whose pixels can be classified into 19 classes *(eg. car, bus, person, rider)* and 7 categories *(eg. flat, object, construction)*. For comparison, we test our methods on testing set over cityscapes benchmark sever. Pascal Context consists of 4998 images for training and another 5105 images for validation, whose pixels either belong to background category or 59 semantic categories *(eg. bag, food, sign, ceiling, ground, and snow)*. As test set of Pascal Context is not available, we directly test our result on the validation set as [20]. For ablative studies, we use fine training data of Cityscapes, and evaluate each part of our proposed method on the fine validation set.

We report metrics as [4], which contains: *1) Acc.:* pixel accuracy, which is the percentage of correctly labeled pixels. *2) mAcc:* mean value of class-wise pixel accuracies. *3) mIoU:* mean IoU score, which is the mean value of class-wise intersection-over-union scores. Among these, mIoU is the most important metric which evaluates effectiveness of method.

3.2. Implementation Details

We use MXNet [25] framework for DRN implementation. The network is shown as Fig. 2, the number of filter in three submodule of DRN is 512, 256, 128 according to Equation (2). Batch size in training stages is set to 8. The initial learning rate is set to 5×10^{-4} for first half epochs, and decreases linearly to 5×10^{-6} for last half epochs. For data arguments, we randomly flip, resize images ranging from 0.55 to 1.3 and randomly crop it to (512, 544). Moreover, λ and η are set to 10^{-4} and 5×10^{-2} respectively. Inspired by [26], which uses *Alternating Training* to train RCNN together with RPN, we train our final model in alternating ways. Firstly train with fine data, then fine and coarse data, finally fine data with $60, 30, 15$ epoches respectively. Meanwhile, we use [27] to replace interpolation in order to get accurate boundary when testing. All of our results is done without post-processing like CRF [9].

3.3. DRN Evaluation

3.3.1. Ablative Studies

DRN helps to learn powerful hierarchical features to classify each pixels. To evaluate DRN, we conduct experiments with several setting, including different settings of skip length and importance dimension reduction in dense relation module. The results are tested on the validation set with single-scale input. As listed in Table 1, DRN with single-skip length works worse than DRN with multi-skips length, which means the multi-scale representations is superior. With experimental results, we find that different dimension reduction in GRUs of

different skip lengths would bring about significant improvement in terms of mIoU, exceeding network without dimension reduction by 0.56% mIoU. In summary, our proposed DRN yields 79.69%/86.03% in terms of mIoU and mAcc, outperforming the baseline by 2.30%/3.87%.

Table 1: Investigation of Dense Relation Module on Cityscapes validation. '1','2','4' mean skip length in three submodule over Dense Relation Module. 'I' means dimensions reduce at different importance.

Network	mIoU	mAcc	Acc
Our ResNet38-based FCN	77.39	82.16	95.27
Our DRN(Skip-111)	78.76	85.07	96.21
Our DRN(Skip-222)	78.98	85.15	96.23
Our DRN(Skip-444)	78.76	84.60	96.17
Our DRN(Skip-124)	79.13	85.33	96.23
Our DRN(Skip-124-I)	**79.69**	**86.03**	**96.25**

The introduced CRL helps to embed context information while not influencing learning process in the main master. We use λ to control the relative weights of OHEM and CRL. We experiment with setting CRL weight λ between 10^{-9} to 10^{-1} and show the results in Fig. 3. The baseline is Dense Relation Network without CRL. The ablation experimental result is expected, small λ would not influence learning proceeding for master branch. Meanwhile, λ will damage learning process for master branch if weight λ is too big. Only suitable λ would take effect at the training stage. The weight $\lambda = 10^{-4}$ yields the best performance, which outperforms the baseline with an improvement of $0.39\%/0.39\%$ (mIoU/mAcc).

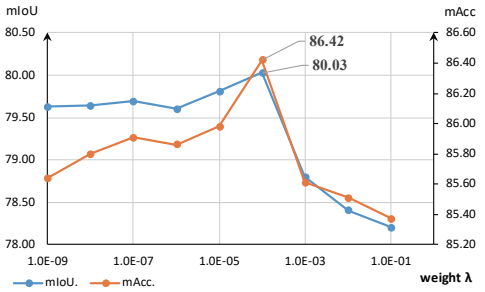

Fig. 3: Quantitative analysis of weight λ.

3.3.2. *Comparison with state-of-the-art results*

1) Cityscapes: Table 2 shows our proposed method is superior than previous methods. Our methods yields 82.8% mean-IoU over the benchmark[1], which outperforms previous state-of-the-art results by 1.6%, 0.6% in terms of mIoU class and category. Fig. 4 shows our visual comparison with previous state-of-the-art method, in which our method not only can learn consistency among pixels but also extract suitable contextual representations.

2) Pascal Context: Table 3 demonstrates that our method gets performance 49.0% in term of mIoU, which outperforms previous state-of-the-art result by 0.9%.

[1]https://www.cityscapes-dataset.com/benchmarks/

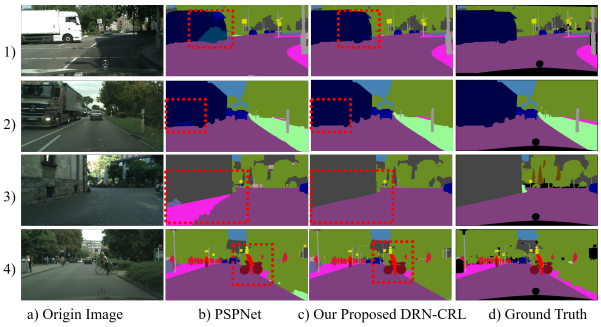

Fig. 4: Visualization comparisons with previous state-of-the-art method [7]. Red bounding box indicates our superiority. 1) and 2) demonstrate our DRN can predict semantic image as a whole. 3) shows that DRN predicts unitary plane. 4) indicates contextual representations plays a important role on distinguishment between rider and pedestrian.

Table 2: Results on Cityscapes testing set, iIoU and iAcc are instance-level intersection-over-union metrics respectively. ‡ means training using both fine and coarse data.

Method	IoU class	iIoU class	IoU category	iIoU category
CRF-RNN [10]	62.5	34.4	82.7	66.0
FCN [4]	65.3	41.7	85.7	70.1
DPN [13]	66.8	39.1	86.0	69.1
LRR [5]	69.7	48.0	88.2	74.7
DeepLabv2_CRF [28]	70.4	42.6	86.4	67.7
Piecewise [11]	71.6	51.7	87.3	74.1
Global-Local-Refinement [29]	77.3	53.4	90.0	76.8
TuSimple [30]	77.6	53.6	90.1	75.2
SAC_multiple [31]	78.1	55.2	90.6	78.3
PSPNet [7]	78.4	**56.7**	90.6	78.6
Our DRN-CRL	**79.9**	56.1	**91.1**	**79.4**
Segmodel [32] ‡	79.2	56.4	90.4	77.0
TuSimple_Coarse [30] ‡	80.1	56.9	90.7	77.8
Netwarp [33] ‡	80.5	59.5	91.0	79.8
ResNet38 [20] ‡	80.6	57.8	91.0	79.1
PSPNet [7] ‡	81.2	59.6	91.2	79.2
Our DRN-CRL ‡	**82.8**	**61.1**	**91.8**	**80.7**

Table 3: Results on Pascal Context [15] validation set.

Method	mIoU(%)	mAcc(%)	Acc(%)
FCN-8s [4]	35.1	46.5	65.9
BoxSup [34]	40.5	-	-
Context [35]	53.9	43.3	71.5
VeryDeep [21]	44.5	54.8	72.9
DeepLab_v2 [8]	45.7	-	-
ResNet38 [20]	48.1	58.1	75.0
Our DRN-CRL	**49.0**	**59.6**	**75.5**

4. CONCLUSION

In this paper, we propose an effective DRN for semantic image segmentation. Dense Relation Module aggregates multi-scale features with dimension reduction at different importance to provide hierarchical contextual information. Extensive experiments suggest that our proposed CRL will help learn consistent representations. Significant improvement and the state-of-the-art results on the Cityscapes dataset (82.8% mIoU) and Pascal Context (49.0% mIoU) demonstrate the superiority of the proposed DRN-CRL.

5. REFERENCES

[1] Caseiro Rui, Jorge Batista, and Cristian Sminchisescu, "Semantic segmentation with second-order pooling," in *European Conference on Computer Vision*, 2012, pp. 430–443.

[2] Philipp Krähenbühl and Vladlen Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in neural information processing systems*, 2011, pp. 109–117.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.

[5] Golnaz Ghiasi and Charless C Fowlkes, "Laplacian pyramid reconstruction and refinement for semantic segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 519–534.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[7] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia, "Pyramid scene parsing network," *CoRR*, vol. abs/1612.01105, 2016.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *CoRR*, vol. abs/1606.00915, 2016.

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *CoRR*, vol. abs/1412.7062, 2014.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr, "Conditional random fields as recurrent neural networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 1529–1537.

[11] Guosheng Lin, Chunhua Shen, Anton van den Hengel, and Ian Reid, "Efficient piecewise training of deep structured models for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba, "Object detectors emerge in deep scene cnns," *CoRR*, vol. abs/1412.6856, 2014.

[13] Ziwei Liu, Xiaoxiao Li, Ping Luo, Chen-Change Loy, and Xiaoou Tang, "Semantic image segmentation via deep parsing network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1377–1385.

[14] 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," *CoRR*, vol. abs/1604.01685, 2016.

[15] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille, "The role of context for object detection and semantic segmentation in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[16] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014.

[17] Wei Liu, Andrew Rabinovich, and Alexander C. Berg, "Parsenet: Looking wider to see better," *CoRR*, vol. abs/1506.04579, 2015.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision*. Springer, 2014, pp. 346–361.

[19] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014.

[20] Zifeng Wu, Chunhua Shen, and Anton van den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *CoRR*, vol. abs/1611.10080, 2016.

[21] Fisher Yu and Vladlen Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[22] Zifeng Wu, Chunhua Shen, and Anton van den Hengel, "High-performance semantic segmentation using very deep fully convolutional networks," *CoRR*, vol. abs/1604.04339, 2016.

[23] Abhinav Shrivastava, Abhinav Gupta, and Ross B. Girshick, "Training region-based object detectors with online hard example mining," *CoRR*, vol. abs/1604.03540, 2016.

[24] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao, *A Discriminative Feature Learning Approach for Deep Face Recognition*, Springer International Publishing, 2016.

[25] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," *CoRR*, vol. abs/1512.01274, 2015.

[26] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[27] Zifeng Wu, Chunhua Shen, and Anton van den Hengel, "High-performance semantic segmentation using very deep fully convolutional networks," *CoRR*, vol. abs/1604.04339, 2016.

[28] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *CoRR*, vol. abs/1412.7062, 2014.

[29] Min Lin Jintao Li Shuicheng Yan Rui Zhang, Sheng Tang, "Global-residual and local-boundary refinement networks for rectifying scene parsing predictions," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 3427–3433.

[30] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell, "Understanding convolution for semantic segmentation," *arXiv preprint arXiv:1702.08502*, 2017.

[31] Rui Zhang, Sheng Tang, Yongdong Zhang, Jintao Li, and Shuicheng Yan, "Scale-adaptive convolutions for scene parsing," in *Proc. 26th Int. Conf. Comput. Vis.*, 2017, pp. 2031–2039.

[32] Falong Shen, Rui Gan, Shuicheng Yan, and Gang Zeng, "Semantic segmentation via structured patch prediction, context crf and guidance crf," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1953–1961.

[33] Raghudeep Gadde, Varun Jampani, and Peter V Gehler, "Semantic video cnns through representation warping," *CoRR, abs/1708.03088*, 2017.

[34] Jifeng Dai, Kaiming He, and Jian Sun, "Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation," *CoRR*, vol. abs/1503.01640, 2015.

[35] Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid, "Exploring context with deep structured models for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.