SINGLE IMAGE DEPTH PREDICTION USING SUPER-COLUMN SUPER-PIXEL FEATURES

Xufeng Guo, Kien Nguyen, Simon Denman, Clinton Fookes, Sridha Sridharan

Image and Video Laboratory, Queensland University of Technology (QUT), Brisbane, Australia
Email: {felix.guo, k.nguyenthanh, s.denman, c.fookes, s.sridharan}@qut.edu.au

ABSTRACT

Depth prediction from a single monocular image is a challenging yet valuable task, as often a depth sensor is not available. The state-of-the-art approach [1] combines a deep fully convolutional network (DFCN) with a conditional random field (CRF), allowing the CRF to correct and smooth the depth values estimated by the DFCN according to efficient contextual modeling. However, using the output of the DFCN as unary input for CRF is limited by using only the last layer of the DFCN. The middle layers of the DFCN have been shown to carry useful information for other scene understanding tasks, which may help to improve the prediction quality. This paper proposes a novel super-column super-pixel (SCSP) feature that is the combination of multiple layers of the DFCN after a super-pixel pooling process. The proposed approach based on the SCSP features reduces the root mean square (rms) error of the prediction by more than 16% in NYUv2 dataset.

Index Terms— depth, fully covolutional, super pixel, super column

1. INTRODUCTION

Predicting depth from a single image is an important task in the computer vision as it provides depth information for an image when such a sensor is not available [1]. Without reliable cues such as motion, depth prediction from a single image is a tough task. Prior research [2, 3] generally calculated the relationship between an image feature and the depth; however, using of hand-crafted features has some limitations such as the images needing to be aligned horizontally [2], and the error rate in the estimated depth image is high [3].

Recent approaches combine deep fully convolutional neural networks (DFCN) with a conditional random fields (CRF) [1,4,5] for depth prediction, since DFCNs can produce reliable unary depth prediction from local features, and the CRF can capture the contextual information to smooth and interpolate the depth map, removing local errors and improving prediction. The system in [1] propose a novel super pixel pooling (sp-pooling) method that not only takes advantage of super-pixels to perform a better up-sampling from the DFCN output to the original size depth output, but also reduces the training by performing sp-pooling after forwarding the whole image to the DFCN, rather than forwarding the super pixel patches to the DFCN individually. This system [1] uses only features taken from the last layer of the DFCN however, the last layer of DFCN is too coarse and there is valuable information in other intermediate layers of the DFCN, which can potentially be useful for scene understanding tasks [6, 7]. In particular the high level information in latter layers assists in boundary detection tasks [6, 7], and 'the combination of semantic information from a deep coarse layer with appearance information from a shallow layer provides large gains in detection and object segmentation [8, 9].

Motivated by this observation, we propose SCSP features that combine information from multiple layers of the DFCN and processes them with the super-pixel pooling process. In contrast to the conventional super-pixel features [1], the proposed SCSP feature helps to explore and embed hidden information from the middle layers of the DFCN, such that the proposed feature captures detail at multiple scales, providing a richer context from which to estimate depth.

The main contributions of this paper are: 1) we propose the SCSP feature that makes use of important information hidden in the intermediate layers of DFCN; 2) coupled with the super-pixel pooling method of [1], the super column stacking better estimates object boundaries than the traditional interpolation up-sampling in [6] and [7]; 3) we perform evaluations over a number of combinations of layers from the DFCN and show under which conditions improved performance can be achieved.

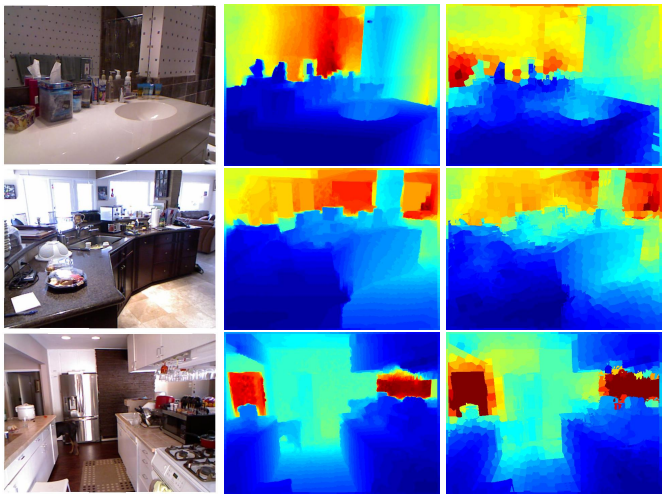
2. PRIOR WORK

Deep fully convolutional neural networks have been applied to solve problems with arbitrary image inputs and dense outputs. The related image recognition and enhancing problems include but are not limited to image super resolution [10], image random noise removal [11], semantic object pixel-level labelling [12], semantic image segmentation [8], and object detection or recognition [13, 14]. Compared to traditional CNNs, DFCNs have the advantage of allowing any input image size, preserving spatial information, and fast computational speed. Since DFCNs usually down-sample the image during the networks forward pass, however, researchers have to apply approximation methods to upsample the network output to the same size as the input, such as interpolation, which does not preserve the object boundaries.

Super pixel methods segment an image into multiple homogeneous 'blobs', and each 'blob' is considered to be a super pixel, within which, the colour and texture of the pixels are the same. As such, when super-pixel segmentation is applied to an image, typically a super pixel will be part of one single object or surface in the image. Therefore when super pixels are used up-sampling a DFCN output, object boundaries can be retained [1, 4, 5, 15].

For tasks that produce dense output it is helpful to use multiple levels of object scale and abstraction [16]. Some architectures merge parallel trained networks with different resolutions for the tasks of human post estimation [17], or scene labelling [4]. Others use a super column method to merge different layers from the same network for tasks such as boundary detection [6, 7], object segmentation [8, 9] and pedestrian detection [16].

As CRF is capable of modelling the spatial relationship between visual elements. It has recently been combined with CNNs which are good at extracting local features. A common strategy is to use the CNN to calculate unary probability, and use the CRF to calculate pairwise relationships. Tasks that apply CRFs and CNNs jointly include: depth learning from a single image [1–3], scene 3-D structure learning from a single image [18], scene labelling [4], human post estimation [17] and structured regression [19].



(a) Input (b) GroudTruth (c) Proposed System

Fig. 1: Examples of a): input images; b): ground truth depth, in which a warmer colour indicates a further distance from the camera, and a cooler colour indicates a closer distance from the camera; c): proposed system output. The output of the proposed system retains much of the object detail. All images CC-BY-SA NYU and QUT

3. PROPOSED APPROACH

3.1. System Overview

The pre-trained network in [1] has five major modules: 1) a deep fully convolutional network module (denoted 'DFCN' in Fig. 3), 2) a super-pixel pooling module ('sp-pooling'), 3) 3-layer fully connected network module ('3fc'), 4) super pixel pair-wise relationship information generator module ('pw-gen'), and 5) a CRF module ('CRF').

The first module 'DFCN' consists of 15 fully convolutional layers shown in Fig. 2 which is the combination of all VGG-16 convolutional layers trained from ImageNet [14], and two newly added convolutional layers [1]. To simplify the training process, we keep the whole 'DFCN' module unchanged. For convenience, we denote the intermediate output of the 15 convolutional layers from the 'DFCN' module with one hex digit: i.e. 1, 2, 3, ... E, F.

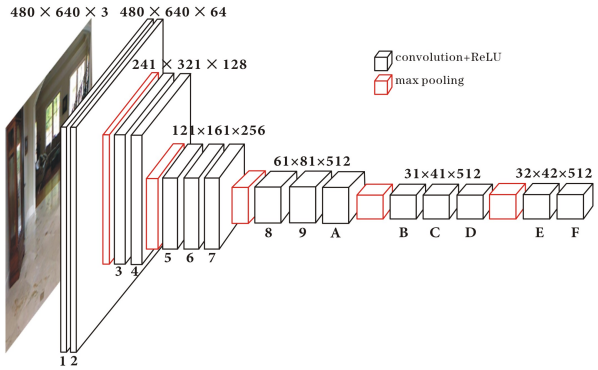


Fig. 2: Detail of the 15-layer DFCN module. The DFCN is the combination of all VGG-16 convolutional layers trained from ImageNet [14] and two newly added convolutional layers [1]. Hex labels for the layers are shown in the figure.

The baseline approach estimates depth by using the output of the DFCN to obtain the unary probability, ignoring valuable information from other layers of the DFCN. In our proposed system we extract SCSP feature from multiple layers of the DFCN. We denote the SCSP feature as consisting of multiple layer indices (1-F) in descending order. For instance, the SCSP feature 'SCSP-F2' is the combination of layers '2' and 'F'; and the baseline system [1] which uses only the last layer is denoted as 'SCSP-F'. We replace the '3fc' module with 4-layers fully connected layers ('4fc'), in which the input layer is the SCSP feature, with the additional complexity required to handle the higher dimensional input compared to the baseline.

Since the super pixel segmentation and pairwise relationship information between neighbouring super pixels is purely based on the input image, the modules 'sp-pooling', 'pw-gen' and 'CRF' are also unchanged. The proposed system is shown in Fig. 4, in which, only the dashed-line area is different from that in Fig. 3.

In the proposed system, each of the selected convolutional layers is individually processed by the super pixel pooling module and the corresponding $N_{sp} \times N_{dim}$ individual super pixel (SP) features are generated (N_{sp} is the number of super pixels in the image and N_{dim} is the number of filters in the selected intermediate convolutional layer output). The individual features are then concatenated to form a super column super pixel (SCSP) feature

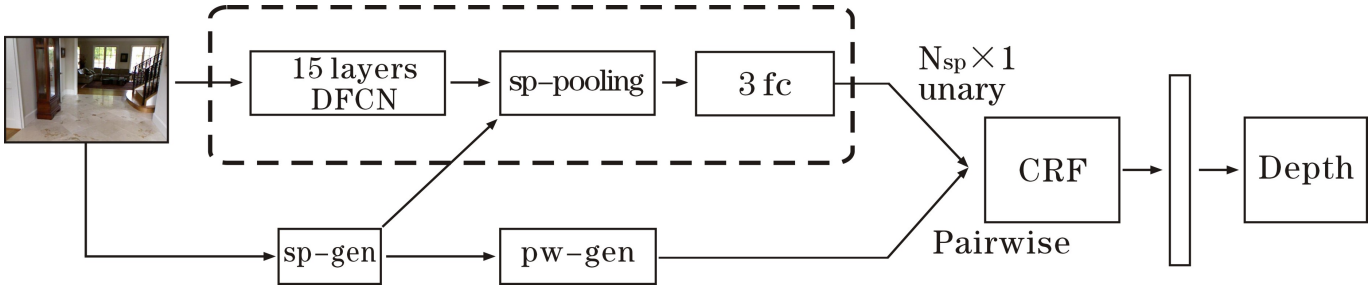


Fig. 3: Original System Framework. There are five modules: 1) a deep fully convolutional network module ('DFCN'), 2) a super-pixel pooling module ('sp-pooling'), 3) 3-layer fully connected network module ('3fc'), 4) super pixel pair-wise relationship information generator module ('pw-gen'), and 5) a CRF module ('CRF').

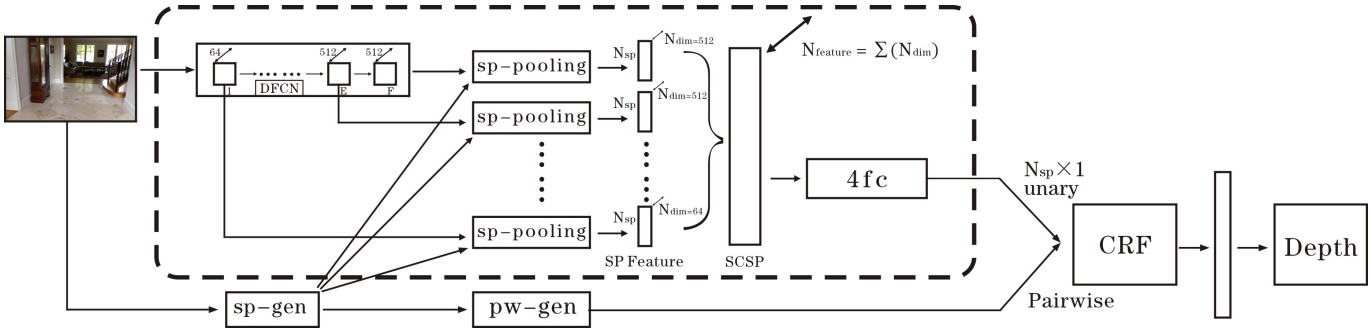


Fig. 4: Proposed System Framework. The dashed-line area differs from the baseline system. Rather than forwarding only the output of the DFCN to '3fc', we use the combination of multiple intermediate layers output in the DFCN. We apply sp-pooling and concatenate the selected layer outputs and forward it to the newly trained '4fc' network. The '4fc' is four-layer fully connected regression network that produces unary depth prediction from the SCSP input.

of size $N_{sp} \times N_{feature}$, where $N_{feature}$ is the sum of the number of filters of all selected convolutional layer outputs (see Fig. 4). The SCSP feature is fed the '4fc' network, and generates the unary depth output with dimension of $N_{sp} \times 1$. Each value in the output represents the predicted depth of a super pixel. The system then forwards the unary output and pairwise information to the same CRF in Fig. 3 to yield the final super pixel depth prediction.

3.2. Model Training and Layer Selection

The new '4fc' network is a four-layer fully connected regression network that takes an input of $1 \times N_{feature}$ SCSP feature and outputs one floating point value. The training data is a $N_{Sample} \times N_{feature}$ matrix, where $N_{Sample} = N_{train} \times N_{sp}$, such that each training image contributes N_{sp} samples. To train, we use mean square error (mse) as the loss function, and batch size is set to 2048. We set the maximum number of epochs of 400 with early an stopping scheme so that the training stops when the validation loss begins increasing.

4. EVALUATION

4.1. Experimental Setup

The experimental data contains 1449 indoor photos provided by New York University (the NYUv2 dataset), as used in the baseline system [1]. Depth ground truth images are generated by a Kinect. We follow the same training-testing splits that is used in the baseline [1]. We follow the evaluation protocol and evaluation metrics of [1], using average relative error (rel), root mean squared error (rms), and average log_{10} error (log10). Interested readers are referred to [1] for further details.

4.2. Experimental Results

We investigate how the combination of different convolution outputs can be used to improve performance. In general, when earlier layers are combined with the last layer, the performance improves, however, blindly concatenating all layers does not necessarily reduce the error rate. We test different layers combinations, and list the combinations with the lowest error rates in Table. 1 and Fig. 5.

We can see that when the '4fc' module input is only one layer ('SCSP-1' to 'SCSP-F'), the error rate is higher than the baseline, although 'SCSP-E' 'SCSP-F' give similar results to the baseline system. Unsurprisingly, earlier layers on their own perform poorly, with only the later layers, which can be seen as very rich features being the end result of a long cascade of convolutional filters, performing well. However, although single layers such as '1', '4' or 'B' achieve unsatisfactory results on their own, when they are combined with later layers they improve performance, for instance 'SCSP-FEB41' which reduces the rms by 0.123 (about

Table 1: The error metrics for different combination of layers.

rel	log10	rms	including layers
0.214	0.087	0.764	Baseline
0.368	0.151	1.316	SCSP-1
0.292	0.124	1.112	SCSP-4
0.220	0.097	0.869	SCSP-B
0.194	0.086	0.777	SCSP-E
0.215	0.086	0.773	SCSP-F
0.175	0.075	0.691	SCSP-FE
0.214	0.090	0.821	SCSP-FED
0.199	0.080	0.696	SCSP-FEB
0.209	0.083	0.721	SCSP-FE2
0.186	0.075	0.655	SCSP-FE1
0.162	0.070	0.653	SCSP-FEC1
0.159	**0.068**	**0.636**	**SCSP-FEB1**
0.185	0.073	0.637	SCSP-FEB4
0.170	0.071	0.641	SCSP-FEB41

16%); indicating that the features from earlier layers provide valuable complementary information for depth prediction.

However not all layers aid performance, as can be seen by the performance of 'SCSP-FE2' and 'SCSP-FED' where the addition of layer '2' or 'D' has reduced performance. We find that in general, layers that increase system error are usually directly before the a 'max pooling' layer in the DFCN (i.e. layer '2' and 'D'). In contrast layers that come after a max pooling layer, such as layers 'B' and 'E', in contrast are more likely to provide useful complementary information to the system. We argue this is due to the max-pooling removing unnecessary features, leaving a richer and more compact representation, that simplifies subsequent learning.

Fig. 6 shows some example results. We can clearly see that by including earlier layers, the depth image produced by our system provides more detail than the baseline. For example in case B and case C, the predicted depth provides even more item detail than the ground truth. Our approach also avoid errors in calculated depths for the window (case A) and wall (case D). In cases E and F our system produces a higher error compared to the baseline; although we note that in case E, even though the proposed depth image obtains a worse error score, regions such as the edge of the door are better described than that in the baseline.

5. CONCLUSION

We have shown that for the task of depth estimation from a single image, the proposed SCSP feature reduces the error metrics of rel, log10 and rms by 0.055 (25.7%), 0.019 (21.8%), and 0.128(16.7%) respectively; and the combination of multiple layers in the DFCN preserves more object detail in the predicted depth image compared to the using only a single layer. Rather than using interpolation for up-sampling and layer stacking, the use of super-pixel pooling ensures the SCSP feature of each super pixel describes the depth of only one small region. The experimental results highlight that while blindly stacking all layers from DFCN is not ideal, selecting multiple layers that follow a max pooling operations provides a positive boost to performance. We believe that our super column stacking method as well as super column layer selection could help improve the performance of other systems that also utilise super columns.

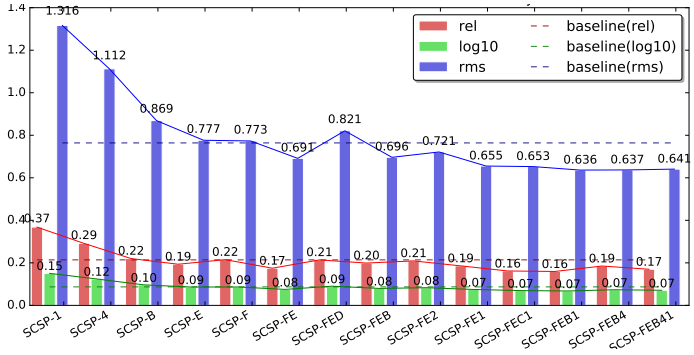
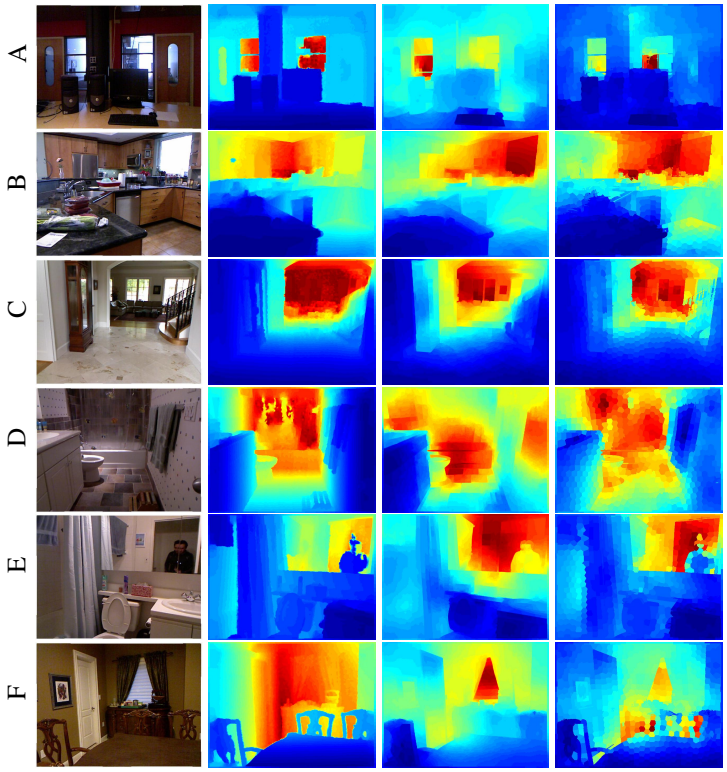


Fig. 5: The error metrics for different combination of layers.



(a) Input (b) GroudTruth (c) Baseline (d) Proposed

Fig. 6: Examples of a):input images; b): ground truth depth, in which a warmer colour indicates a further distance from the camera, and a cooler colour indicates a closer distance from the camera; c): baseline system [1] ; d): Propose approach, with layers 'FEB41'. All images CC-BY-SA NYU and QUT

6. REFERENCES

[1] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2024–2039, Oct 2016.

[2] Ashutosh Saxena, Sung H Chung, and Andrew Y Ng, "Learning depth from single monocular images," in *NIPS*, 2005, vol. 18, pp. 1–8.

[3] Miaomiao Liu, Mathieu Salzmann, and Xuming He, "Discrete-continuous depth estimation from a single image," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[4] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, Aug 2013.

[5] Dan Ciresan, Alessandro Giusti, Luca M. Gambardella, and Juergen Schmidhuber, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds., pp. 2843–2851. Curran Associates, Inc., 2012.

[6] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani, "High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision," *CoRR*, vol. abs/1504.06201, 2015.

[7] Gedas Bertasius, Jianbo Shi, and Lorenzo Torresani, "Semantic segmentation with boundary neural fields," *CoRR*, vol. abs/1511.02674, 2015.

[8] Jonathan Long, Evan Shelhamer, and Trevor Darrell, "Fully convolutional networks for semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[9] Bharath Hariharan, Pablo Arbelaez, Ross Girshick, and Jitendra Malik, "Hypercolumns for object segmentation and fine-grained localization," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[10] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199.

[11] David Eigen, Dilip Krishnan, and Rob Fergus, "Restoring an image taken through a window covered with dirt or rain," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[12] Pedro HO Pinheiro and Ronan Collobert, "Recurrent convolutional neural networks for scene labeling.," in *ICML*, 2014, pp. 82–90.

[13] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," *CoRR*, vol. abs/1312.6229, 2013.

[14] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[15] Yaroslav Ganin and Victor Lempitsky, *N4 -Fields: Neural Network Nearest Neighbor Fields for Image Transforms*, pp. 536–551, Springer International Publishing, Cham, 2015.

[16] Pierre Sermanet, Koray Kavukcuoglu, Soumith Chintala, and Yann LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3626–3633.

[17] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in neural information processing systems*, 2014, pp. 1799–1807.

[18] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824–840, May 2009.

[19] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency, "Continuous conditional neural fields for structured regression," in *European Conference on Computer Vision*. Springer, 2014, pp. 593–608.