DOI: 10.3217/978-3-85125-528-7-06

# Semantic Labeling Enhanced by a Spatial Context Prior

Daniel Steininger, Csaba Beleznai

Austrian Institute of Technology, Austria
*Daniel.Steininger.fl@ait.ac.at*
*Csaba.Beleznai@ait.ac.at*

***Abstract***

*Our observed visual world exhibits a structure, which implies that scene objects and their surroundings are not randomly arranged relative to each other but typically appear in a spatially correlated manner. Thus, the structural correlation can be exploited to make the visual recognition task predictable to a certain extent. Modeling relations between categories is, however, non-trivial, since categories are often represented at different granularities across distinct datasets. In this paper, we merge fine-level semantic descriptions into basic semantic classes which allows the generation of spatial contextual priors from a wide range of datasets. In this way, a contextual model is derived with the objective to employ the learned contextual prior to enhance visual recognition via improved semantic labeling. The prior is captured explicitly by computing occurrence and co-occurrence probabilities of specific semantic classes and class pairs from a diverse set of annotated datasets. We show improved semantic labeling accuracy by incorporating the contextual priors into the label inference process, which is evaluated and discussed on the Daimler Urban Segmentation 2014 dataset.*

## 1. Introduction

Semantic segmentation of digital images links two core computer vision challenges: visual object recognition and segmentation. In recent years, great improvement in accuracies to both task domains has been demonstrated, mainly due to a transition from learned hand-crafted representations towards representations distributed within hierarchies and embedded into compositional schemes, enabling a rich generalization for a large number of object classes.

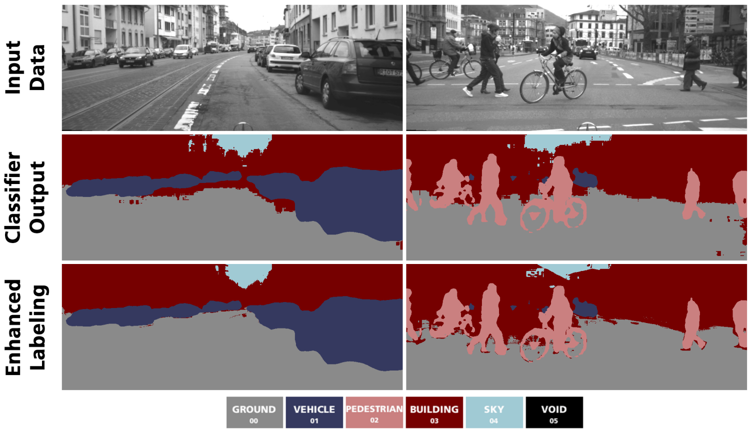

**Figure 1. Semantic Labeling enhanced by a Spatial Context Prior and Conditional Random Field.**

As representations and learning schemes have grown capable of accommodating the sheer variability in the data, this progress is also imposing new requirements on the employed datasets. Current learned models are often optimized for specific datasets they have been trained on, and their capture modalities are restricted by their implicit design. Real world scenarios are highly diverse, therefore, a single dataset solely represents a small fraction of all possible visual appearances. Although datasets have become more elaborate and diverse lately [17], class coverage, balancing and variability are still relevant issues to be tackled. Motivated by the diversity in the characteristics of prevailing datasets, in terms of number and granularity of annotated classes and scene-specific view attributes, we propose to capture the spatial relationship between various semantically labeled regions across several datasets. We demonstrate that the modeled spatial prior can enhance recognition accuracies leading to state-of-the-art results, as illustrated in Figure 1.

## 2. Related Work

Spatial context is an important type of information in the human cognitive process [12] when recognizing objects, especially in the presence of a cluttered background. Certain objects predominantly co-occur in the real world. Thus analyzing vast amounts of visual data can result in meaningful contextual statistics which can be used to robustify visual object recognition [5].

Pixel-wise semantic labeling is a relatively novel domain since large-scale object recognition with shared informative representations is a prerequisite for this task. Starting with manually selected low-level features, discriminatively trained Random Forests or Boosting have been used to perform classification patch-wise [16] or to additionally incorporate local structural information within the analysis patch [7]. Based on recent advances in deep learning, several frameworks [13, 18] have demonstrated significant improvements in the accuracy of per-pixel class estimates. Recently, multi-scale deep architectures have been proposed in order to represent local and global context by employing multiple input images at different resolutions [2], or combining feature maps from different layers of the convolutional architecture [6]. Both techniques aim to combine fine detail representations with relational information established at a coarse resolution level in order to generate accurate segment boundaries between labeled regions. The immense representational power of deep convolutional architectures captures rich details of the object classes to be represented and yields segmentation frameworks which surpass learned hand-crafted representations. Capturing spatial context within convolutional architectures, however, is linked with complexities in terms of training (augmented parameter space) and increased computational expense due to the computation of multiple scale-specific features.

Our proposed approach employs a previously learned spatial prior model as an additional step to switch class labels at locations where per-pixel estimates are ambiguous. We term our model as the *Explicit Priors* model. Per-pixel ambiguity is quantified from class posterior probabilities at the given pixel by examining the distance between first and second rank probabilities. Our method, while limited in representing spatial context at a wide range of spatial scales and orientations, yields a remarkable improvement at a negligible increase of computational complexity.

## 3. Methodology and Experimental Setup

The proposed approach for combining learned information from multiple datasets and thereby enhancing existing classifiers is based on the concept of *Explicit Priors*. By aggregating statistical data on the level of individual pixels and capturing spatial context, we generate additional cues for training and classification, while remaining independent of the underlying machine learning algorithm. The method and its integration throughout the entire processing pipeline is described in this chapter and demonstrated on the Daimler Urban Segmentation 2014 dataset [14].

The dataset consists of image sequences captured by a camera mounted on a moving car. The images are provided without color information at a resolution of 1024x440 px, with every 10th frame of the sequences being annotated with pixel-wise segmentations. For a reasonable comparison, only the test sequences, as specified by the evaluation protocol, are considered. The dataset is supplemented with precomputed disparity maps and additional information, like time-stamps, vehicle speed and yaw rate. The ground truth distinguishes between two foreground (*Vehicle* and *Pedestrian*) and three background classes (*Ground*, *Sky* and *Building*). Within the test data 36.3% of all pixels are defined as *Void*. The frequency of occurrence of the labeled pixels is 54.1% for *Ground*, 14.8% for *Vehicle*, 4.6% for *Pedestrian*, 2.4% for *Sky* and 24.0% for *Building*, resulting in a background ratio of 80.6%.

### 3.1. Training

**Dataset Analysis** As a preliminary step for the training and classification process, an appropriate choice of input data with regard to the intended application scenario is a decisive aspect. For this purpose, a statistical analysis of multiple datasets was conducted according to the concept of *Explicit Priors*. The resulting data ranges from basic statistics, such as label frequency and the ratio of background to foreground classes, to more sophisticated aspects concerning occurrence distribution and spatial context. For each application scenario, this dataset analysis can be used to select a subset of additional cues for identifying appropriate datasets. For the demonstrated task, for instance, the most useful information was provided by the concept of *Location Bins*. By dividing the image dimensions into a coarse grid and capturing the spatial distribution of each class across the resulting cells over the entire dataset, probabilities for the occurrence of certain labels with regard to their location can be derived. The resulting representation provides clearly arranged patterns closely related to certain characteristics of the dataset, such as the method of image acquisition. In the case of Vehicles, for instance, the analysis clearly showed that images taken with a hand-held camera are mostly centered on the these objects, while for the datasets using a camera mounted on a car they are most often found in the lower half of the image. Comparing these statistics for candidate training datasets to the intended application scenario facilitates the evaluation of their compatibility.

Other available statistical measures proved to add less distinct cues for the given task, such as the analysis of co-occurrence, which provides a measure of probability for each combination of labels to appear in the same image. Since the application scenario only includes five labels arranged within a consecutive image sequence, the resulting correlation matrix did not show significant peaks. However, an adapted version in the form of *Local Label Neighborhood* (*LLN*), which limits the co-occurance measure to label transitions, was successfully applied, as described in detail in Section 3.2.

Based on the aggregated information of class frequency and *Location Bins*, the CamVid dataset [1] could be identified as an appropriate choice for training background classes, since it offers a background ratio of 80.9%, as well as a fitting spatial arrangement of class probabilities. The foreground classes, on the other hand, are trained on the PascalContext dataset [11], in particular the version including 33 categories, which contains 46% foreground pixels.

**Classifier Setup** Based on the selected datasets, two classifiers are applied to cover the background and foreground classes separately. The former classifier uses the pre-trained model pascal-fcn8s-tvgdag provided by Zheng et al. [18], which is evaluated on the foreground classes of the PascalContext33 dataset. The background classifier was trained using TextonBoost [8] on randomly sampled images of the CamVid dataset. For this purpose, feature descriptors based on filter banks, location and gradient orientations were applied for training a total of 950 Textons, which represents a compromise between computational complexity and accuracy. Since the test dataset consists of gray-scale images, the learning input is restricted to the intensity channel.

**Label Aggregation and Mapping** The main obstacle in aggregating multiple datasets during the training stage results from variations in the denomination of object classes. Furthermore, since in many cases not all labels of the training datasets are required for classifying the test images, and multiple labels of one dataset can relate to a single label of another, a generalized mapping strategy is a prerequisite for combining label information. For this purpose, an automatic method for label clustering was developed based on version 3.0 of the Wordnet database [10]. This knowledge representation was trained exclusively on lexical data and is capable of providing a similarity measure among semantic descriptions. Based on this, labels of the training dataset can be assigned to the final denominations by applying a threshold and giving preference to classes with higher similarity.

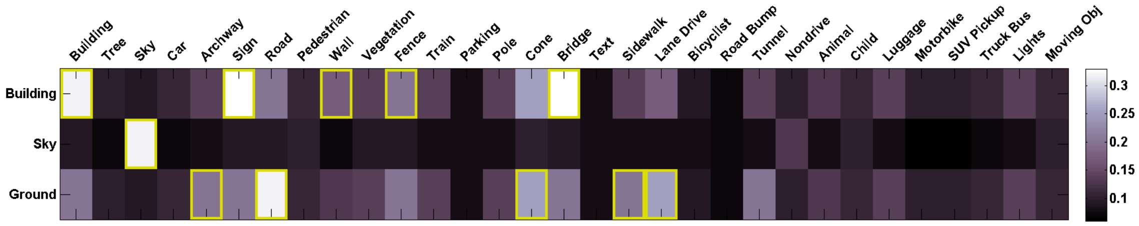

**Figure 2. Label Mapping of CamVid (Columns) to Daimler (Rows) dataset based on Wordnet similiarity (selected labels are marked in yellow color).**

In the case of the CamVid dataset this process resulted in a selection of eleven labels, as visualized in Figure 2, while the remaining ones are not required for the application task and therefore suppressed. The selected labels were assigned to the background classes *Building*, *Sky* and *Ground* of the final dataset based on the corresponding similarity. Analogously, the two foreground objects *Pedestrian* and *Vehicle* are assigned the PascalContext labels of *Pedestrian*, *Bicyclist*, *Child* and *Moving Object*, as well as *Car*, *Motorbike*, *SUV Pickup* and *Truck*, respectively.

### 3.2. Classification

The foreground and background classifiers are applied to each input image of the test set resulting in two complementary segmentations, which are further refined by applying the label mapping method described in Section 3.1. This step results in both images being segmented into the labels required by the test dataset. In order to further improve the segmentation quality of background classes, the two highest ranked labels of each pixel are retained, as well as the probability distance between them. This information is required for enhancing the results with *Local Label Neighborhood* priors and further refinement by inference based on a Conditional Random Field (CRF).

**Local Label Neighborhood** The concept of *Local Label Neighborhood* is based on statistically learning conditional probabilities of transitions between specific labels in vertical and horizontal direction. Each annotated pixel within the selected training images is evaluated to capture this prior based on spatial context. For the given task, this results in a measure of probability for each background class to be found on a specific side of either of the two foreground classes. The probabilities extracted from the CamVid dataset using this method are weighted by the frequency of occurrence for each background class and aggregated into the final labels, as visualized in Table 1. The learned a-priori knowledge is used to resolve ambiguous classifications.

| | *Ground* | *Sky* | *Building* |
|---|---|---|---|
| *LLN* Vehicle ↑ | 0.026 | 0.005 | 0.195 |
| *LLN* Pedestrian ↑ | 0.025 | 0.002 | 0.222 |
| *LLN* Vehicle ↓ | 0.383 | 0.000 | 0.004 |
| *LLN* Pedestrian ↓ | 0.086 | 0.001 | 0.081 |

**Table 1. *Local Label Neighborhood* learned on CamVid dataset.**

The resulting statistics show the probability of encountering each background class above or below a label transition from each foreground class. For instance, the *Vehicle* prior in upward direction indicates a significant chance of detecting *Buildings* above the class and the prior in downward direction increases the probability of detecting *Ground* below it. Using this information, areas between foreground classes and image borders in vertical direction are marked as candidates for the corresponding background label based on the probability indicated by the prior. If the candidate labels correspond to the second-ranked label for a pixel and the probability distance to the current class is sufficiently low, the second rank is recovered and replaces the first.

**Conditional Random Field** In order to further increase segmentation accuracy, especially in areas of label transitions, a framework [8] for inference based on CRF is applied with empirically determined parameters. As an input, the existing intermediate background segmentation is integrated in the form of a unary potential with a globally defined confidence of 80%. Additionally, two pairwise potentials, based on label compatibility and intensity information within a defined radius, are added, the latter weighted four times higher than the former. After conducting the inference process in five iterations, the eventual segmentation is combined with the foreground classes.

## 4. Experiments and Discussion

Semantic Labeling was performed on the test sequences of the Daimler Urban Segmentation 2014 dataset with and without the learned spatial context prior. In order to compare the results to previously published methods, the Intersection-over-Union (*IoU*) metric is used according to the official Pascal VOC definition [3],

$$IoU_{l_i} = \frac{TP_i}{TP_i + FP_i + FN_i}, \tag{1}$$

where $L = \{l_1, ..., l_k\}$ is a set of labels and $TP_i$, $FP_i$ and $FN_i$ are the true positive, false positive and false negative detections corresponding to label $l_i$, is used. The detailed results are shown in Table 2. Additionally, we show the average IoU over all classes, as well as a separate average value for the dynamic classes *Vehicle* and *Pedestrian.* The global per-pixel accuracy (PPA) represents the ratio of correctly classified pixels to the total number of annotated pixels in the test dataset. Each column shows the results for the baseline method and its enhancement with the proposed *LLN* and CRF, which are compared to state-of-the-art methods. The best-performing results are displayed in bold numbers.

|  | *Ground* | *Vehicle* | *Pedestrian* | *Sky* | *Building* | *Avg* | *Avg$_{dyn}$* | *PPA* |
|---|---|---|---|---|---|---|---|---|
| Stixmantics [14] | 93.8 | 78.8 | 66.0 | 75.4 | 89.2 | 80.6 | 72.4 | 92.8 |
| ALE [14] | 94.9 | 76.0 | 73.1 | **95.5** | 90.6 | 86.0 | 74.5 | **94.5** |
| Darwin pw. [4] | 95.7 | 68.7 | 21.2 | 94.2 | 87.6 | 73.5 | 44.9 | - |
| PN-RCPN [15] | **96.7** | 79.4 | 68.4 | 91.4 | 86.3 | 84.5 | 73.8 | **94.5** |
| Layered Ip. [9] | 96.4 | 83.3 | 71.1 | 89.5 | **91.2** | **86.3** | 77.2 | - |
| BL | 92.9 | **85.5** | **75.4** | 54.2 | 80.1 | 77.6 | **80.5** | 92.8 |
| BL \| LLN | 92.9 |  |  | 54.2 | 81.3 | 77.9 |  | 93.0 |
| BL \| LLN \| CRF | 94.8 |  |  | 74.1 | 85.1 | 83.0 |  | **94.5** |

**Table 2. Intersection-over-Union measures and Per-Pixel Accuracy (BL: baseline method, LLN: Local Label Neighborhood, CRF: Conditional Random Field).**

Compared to recently published approaches, the proposed method leads to an improved segmentation of dynamic classes by 3.3%. The concept of Label Aggregation applied to a pre-trained model proves to be an appropriate choice for both labels. The classification of background classes, on the other hand, is quite competitive for the *Ground* class with a distance of 1.9% to the leading method, while being slightly inferior to the others concerning *Building* and *Sky*. However, these results are still promising, considering several influencing factors. Firstly, the proposed method is presently based exclusively on intensity information, while the other algorithms, except [15], incorporate additional cues such as depth and motion data. However, this limitation can still be partially compensated by the application of *LLN* and CRF. While *LLN* leads to an increase 0.3% concerning the average IoU, CRF contributes an additional 5.1%. For the PPA, improvements of 0.2% and an additional 1.5% can be achieved. An example of the overall results is provided in Figure 3.

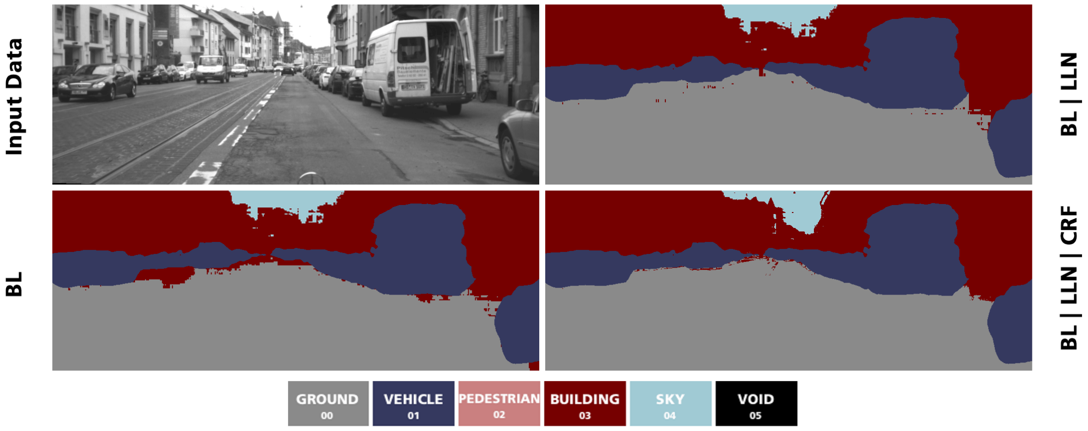

**Figure 3. Improvement of segmentation quality of background classes (BL: baseline method, LLN: Local Label Neighborhood, CRF: Conditional Random Field).**

Please note that the lowest accuracy corresponds to the *Sky* class, which has a frequency of occurrence of solely 2.4% in the testing dataset. Therefore, its influence on the PPA is almost negligible, which leads to an accuracy equal to the currently best results.

More detailed insights can be retrieved by analyzing precision and recall measures for each class, as displayed in Table 3. Both values present highly promising results for the *Ground* class, which is the most frequent background class. The remaining two background classes show higher interdependency. While *Building* offers a high recall but lower precision value, *Sky* shows the opposite characteristics, which indicates that the *Building* class tends to inaccurate over-segmentation into *Sky*

|  | *Ground* | *Vehicle* | *Pedestrian* | *Sky* | *Building* | *Avg* |
|---|---|---|---|---|---|---|
| BL | 97.8 \| 95.0 | **98.5** \| **86.7** | **97.1** \| **77.2** | 72.9 \| 67.9 | 82.0 \| 97.1 | 89.6 \| 84.8 |
| BL \| LLN | 97.2 \| 95.5 |  |  | 72.9 \| 67.9 | 83.5 \| 96.9 | 89.8 \| 84.8 |
| BL \| LLN \| CRF | **97.7** \| **96.9** |  |  | **89.1** \| **81.5** | **86.3** \| **98.4** | **93.7** \| **88.1** |

**Table 3. Precision (left) and recall (right) of each label class.**

regions. Concerning the influence of *LLN*, the *Building* class reaches an increase in precision of 1.5% combined with an insignificant decrease of recall. Simultaneously, the optimization leads to a decrease in precision for the *Ground* class, while increasing its recall. It can be concluded that the method successfully recovers misclassified *Ground* pixels originally labeled as *Building*. CRF further increases the average precision and recall by an additional 3.9% and 3.3%, respectively.

## 5. Conclusions

This paper introduces a concept to capture spatial context between labeled regions for diverse datasets annotated at different semantic granularity, referred to as *Explicit Priors*, which was successfully applied to enhance the entire training and classification process of semantic segmentation demonstrated on the Daimler Urban Segmentation 2014 dataset. The approach provides a generalized way to select an appropriate subset of multiple training datasets and to efficiently combine their labels to fit a given application scenario. The segmentation quality of foreground classes is comparable to, and in terms of certain measures even surpasses, state-of-the-art methods. The results for the background classes proved to be competitive as well. Their relatively high precision, combined with lower recall correspond to a classification accuracy of certain labels slightly inferior to currently leading methods. Further improvements concerning background labeling were achieved by applying priors based on *Local Label Neighborhood* as well as inference using CRF. In order to exploit additional potentials, the next step would be to integrate complimentary modalities, such as depth and motion cues.

## Acknowledgments

This work is supported by the research initiative 'Mobile Vision' with funding from the Austrian Federal Ministry of Science, Research and Economy and the Austrian Institute of Technology.

## References

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

[2] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2650–2658, 2015.

[3] Mark Everingham, S.M. Ali Eslami, Luc Van Gool, Christopher K.I. Williams, John Winn, and Andrew Zisserman. The Pascal Visual Object Classes Challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[4] Stephen Gould. Darwin: A framework for machine learning and computer vision research and development. *The Journal of Machine Learning Research*, 13(1):3533–3537, 2012.

[5] Michelle R. Greene, Christopher Baldassano, Andre Esteva, Diane M. Beck, and Li Fei-Fei. Visual scenes are categorized by function. *Journal of Experimental Psychology: General*, 145(1):82, 2016.

[6] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 447–456, 2015.

[7] Peter Kontschieder, S. Rota Bulò, Horst Bischof, and Marcello Pelillo. Structured class-labels in random forests for semantic image labelling. In *Proceedings of IEEE International Conference on Computer Vision*, pages 2190–2197, 2011.

[8] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, 2011.

[9] Ming-Yu Liu, Shuoxin Lin, Srikumar Ramalingam, and Oncel Tuzel. Layered interpretation of street view images. In *Proceedings of Robotics: Science and Systems*, Rome, Italy, July 2015.

[10] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[11] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[12] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007.

[13] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1742–1750, 2015.

[14] Timo Scharwächter, Markus Enzweiler, Uwe Franke, and Stefan Roth. Stixmantics: A medium-level model for real-time semantic scene understanding. In *Proceedings of the European Conference on Computer Vision*, pages 533–548. Springer, 2014.

[15] Abhishek Sharma, Oncel Tuzel, and David W. Jacobs. Deep hierarchical parsing for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 530–538, 2015.

[16] Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[17] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

[18] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H.S. Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1529–1537, 2015.