HDPA: HIERARCHICAL DEEP PROBABILITY ANALYSIS FOR SCENE PARSING

Yuan Yuan[1], *Zhiyu Jiang*[1,2], *and Qi Wang*[3,*],

[1]Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, Shaanxi, P. R. China.
[2]University of Chinese Academy of Sciences, 19A Yuquanlu, Beijing 100049, P. R. China.
[3]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.
yuany@opt.ac.cn, jiangzhiyu@opt.cn, crabwq@nwpu.edu.cn

ABSTRACT

Scene parsing is an important task in computer vision and many issues still need to be solved. One problem is about the non-unified framework for predicting *things* and *stuff* and the other one refers to the inadequate description of contextual information. In this paper, we address these issues by proposing a *Hierarchical Deep Probability Analysis*(HDPA) method which particularly exploits the power of probabilistic graphical model and deep convolutional neural network on pixel-level scene parsing. To be specific, an input image is initially segmented and represented through a CNN framework under Gaussian pyramid. Then the graphical models are built under each scale and the labels are ultimately predicted by structural analysis. Three contributions are claimed: unified framework for scene labeling, hierarchical probabilistic graphical modeling and adequate contextual information consideration. Experiments on three benchmarks show that the proposed method outperforms the state-of-the-arts in scene parsing.

Index Terms— Computer vision, scene parsing, semantic segmentation, probabilistic graphical model, CRF

1. INTRODUCTION

Scene parsing has been widely investigated for its important role in computer vision. Many challenging tasks such as image or video captioning, autonomous navigation, and traffic scene analysis[1, 2, 3] have proven to benefit from scene parsing. In some literature, scene parsing is known as semantic segmentation, semantic annotation, image parsing, and Full Scene Labeling (FSL). Concretely, scene parsing tends to label every pixel in the image with the category of things or stuff it belongs to. After scene parsing, every element of objects is segmented and tagged.

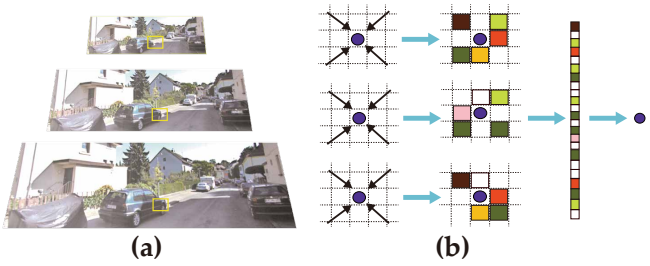

Fig. 1. Highlights of the proposed method. Given a fixed-size segmented region under multi-scale images, different levels of contextual information can be provided. (b) shows the strategy of considering different range of contextual information.

Two issues of primary importance in the context of scene parsing should be clearly described. One issue is about two items: *thing* and *stuff*. *Thing* is defined as an object with a specific size and shape[1], such as the car in Fig. 1. Instead, *stuff* tends to be a homogeneous or repetitive pattern, with no specific spatial extent/shape[1], such as the sky or the road in Fig. 1. Like traditional computer vision systems, a perfect scene parsing model should be unified which can cooperatively segment all the things and stuff in the image. The other issue is about contextual information, which provides important cues for scene parsing. Context refers to the semantic correlation between one object and its neighboring objects. For example, a car is more likely to appear on road and it is not likely to be surrounded by sky in Fig. 1(a). In some cases, contextual information is the most significant cue when the object is ambiguously represented in feature space.

Previous methods have made significant progress addressing the mentioned issues over the past few years. Probabilistic Graphical Models(PGM) have been widely investigated[4, 5] to enhance the scene parsing tasks for its ability of multivariate joint probability distribution representation[6]. Par-

*Qi Wang is the corresponding author of this paper. This work is supported by the National Basic Research Program of China (Youth 973 Program) (Grant No. 2013CB336500), the State Key Program of National Natural Science of China (Grant No. 60632018, 61232010), and the National Science Foundation of China(Grandt No. 61379094).

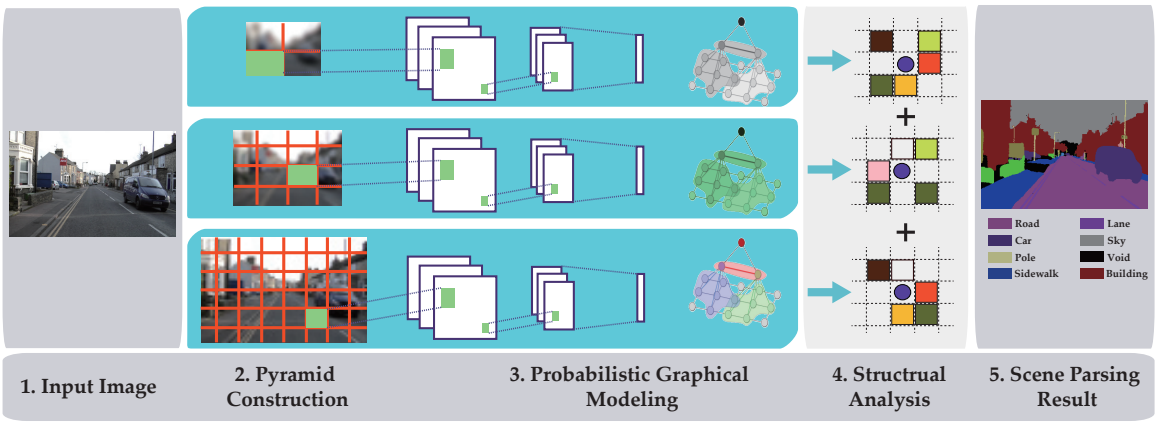

Fig. 2. HDPA pipeline. The input image is initially segmented under Gaussian pyramid with specific block numbers and each block is represented through a convolutional neural network. Subsequently, a probabilistic graphical model is utilized under each pyramid scale and the final pixel labels are obtained by structural analysis.

ticularly, Markov Random Fields(MRFs) and its variant Conditional Random Fields(CRFs) have witnessed great success in this field[7, 8]. Moreover, they have became one of the most successful models in computer vision. For scene parsing, CRFs formulate the label assignment task as a conditional probability inference problem, with the ability of combining local appearance information and smoothness priors[9]. Recently, Deep Neural Network(DNN) approaches such as Convolutional Neural Networks(CNNs) have been successfully equipped in high level tasks such as object detection and object recognition[10, 11]. This phenomenon motivates researchers to explore CNN for scene parsing. The superiority of CNNs is fully data-driven, therefore being more accurate in representing training samples and able to find feature patterns. While hand-crafted feature is domain inspired and tends to be more suitable for a specific task. Furthermore, a number of recent DNN based frameworks including Fully Convolutional Networks(FCN)[12] and Recurrent Neural Network(RNN)[13] have witnessed significant performance boost by end-to-end strategy and memory characteristic.

Although significant progress has been made when adapting CRFs and CNNs for scene parsing, some challenges still exist. Firstly, CRFs by themselves are not able to capture large input contexts which are essential for detecting larger object classes such as road[14, 15]. Meanwhile, traditional pixel-level scene parsing models, especially CRF based methods, are time-consuming. Secondly, CNNs are more likely to be feature extractor and lack smoothness constraints that encourage label agreement between similar pixels[16]. Starting from these drawbacks, this paper proposes a scene parsing method named *Hierarchical Deep Probability Analysis*(HDPA). Our formulation jointly considers the strengths of Deep Neural Network and probabilistic graphical models in a unified framework. More specifically, the input image is initially segmented under Gaussian pyramid with specific block numbers and each block is represented through a convolutional neural network. Subsequently, a probabilistic graphical model is utilized under each pyramid scale and the final pixel labels are obtained by probability analysis. The main contributions of this research are listed as follows.

(1) *Unified framework for scene parsing.* Traditional methods tend to separately label things and stuff, such as labeling things through detection method and labeling stuff through segmentation method, which would weaken the neighboring constraints. In this work, a unified framework is proposed which simultaneously considers the neighboring things and stuff.

(2) *Hierarchical probabilistic graphical modeling.* The proposed probabilistic graphical model is defined over a set of patches. Compared to the previous works, this strategy jointly takes different levels of the scene into consideration and most scene-level relationships can be captured.

(3) *Adequate contextual information consideration.* One of the main challenges for scene parsing is how to take wide range of contextual information into consideration. As is known, probabilistic graphical models, such as CRFs, are not able to capture large input contexts which are essential for labeling larger object classes such as road. To solve this problem, the proposed probabilistic graphical models are built on patches under Gaussian pyramid. Smaller number of squared patches are defined in smaller scale. This strategy can efficiently take wide contexts into consideration.

The rest of this paper is organized as follows. The proposed HDPA model is elaborated in Section 2. Experimental results are presented in Section 3 and Section 4 concludes the paper.

2. HIERARCHICAL DEEP PROBABILITY ANALYSIS FOR SCENE PARSING

In this section, scene parsing problem is elaborated formulated. We present the detailed procedure of our *Hierarchical Deep Probability Analysis* (HDPA) method and the pipeline is depicted in Fig. 2.

2.1. Multi-Scale Modeling

Given an input image $\mathbf{I}$, a multiscale pyramid of images $\mathbf{X}_s, \forall s \in \mathcal{S} = \{1, \ldots, S\}$ is constructed. The multiscale pyramid is based on Gaussian pyramid for its linear property which would not pollute the original image, and is typically pre-processed, so that local neighborhoods have zero mean and unit standard deviation. As is shown in Fig. 2, each scale of the Gaussian pyramid is segmented into patches. For the scale image with larger Gaussian kernel size, smaller number of patches are segmented. The reason can be explained as follows: for a certain model, it's a tradeoff between the ability of recognizing detailed information and owning a wide perceptual field. Based on this, detailed texture information is eliminated for the image processed by large Gaussian kernel and wide perceptual field is naturally captured for the proposed model.

CNN model, which is widely used for its data-driven ability, is utilized to represent the high-level semantic information of each patch. Given a classical convolutional network with parameters θ_s, the multiscale network is obtained by instantiating one network per scale s, and sharing all parameters across scales: $\theta_s = \theta_0, \forall s \in \mathcal{S}$. Considering the complexity of the CNN models, images under different scales share the same CNN models with the same parameters. In this paper, a pre-trained VGG model [17] is utilized to initialize the CNN model and the fine-tune strategy is also considered. For the ith patch x_i of $\boldsymbol{X}_s$, where $i \in \{1, ..., N\}$ indicates the number of patches in current scale, the CNN feature can be written as $\boldsymbol{f}_i$.

For simplicity, the scale parameter s is eliminated and $\mathbf{X}$ represents a segmented patch of $\mathbf{X}_s$ in the following part.

2.2. Conditional Random Fields

A brief overview of Conditional Random Fields(CRFs) for pixel-wise labeling and the notations used in this work are introduced. CRF is a discriminative undirected probabilistic graphical model, a sort of Markov Random Field. The chief advantage of CRF lies in the fact that it models the conditional distribution $P(\boldsymbol{Y}|\boldsymbol{X})$ rather than the joint distribution $P(\boldsymbol{Y}, \boldsymbol{X})$. The major difference between CRF with some other existing methods is that it is a global model that considers all residues as a whole rather than focuses merely on a local window around the tag to be labeled. In the inference, the states of all tags are predicted simultaneously in a way that maximizes the overall likelihood.

Let $G = (\mathcal{V}, \mathcal{E})$ be the associated undirected graph of an CRF. Nodes $\mathcal{V}$ represent random variables $\boldsymbol{X} = \{x_1, \ldots, x_N\}, \boldsymbol{Y} = \{y_1, \ldots, y_N\}$ and edges $\mathcal{E}$ represents conditional dependencies. Typically, random variable y_i represents the label assigned to pixel i, x_i represents the feature of pixel i and edges represent the relationship between neighboring pixels. Accordingly, the random variables $\boldsymbol{Y}$ ranges over possible pixel label space $\mathcal{L} = \{1, \ldots, l\}$ and $\boldsymbol{X}$ ranges over input image size N.

According to the Markov property, the conditional probability of $\boldsymbol{Y}$, given observations $\boldsymbol{X}$, can be expressed as

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{Y})} \prod_i \phi(\boldsymbol{y}_i|\boldsymbol{X}) \prod_c \phi(\boldsymbol{y}_c|\boldsymbol{X}), \quad (1)$$

where the first product in Eq.1 is over all individual variables, while the second is over the set of *cliques* c in the graph. From the aspect of Gibbs distribution, Eq.1 can be rewritten as

$$P(\boldsymbol{Y}|\boldsymbol{X}) = \frac{1}{Z(\boldsymbol{Y})} e^{-E(\boldsymbol{Y}|\boldsymbol{X})}, \quad (2)$$

where $Z(\boldsymbol{Y}) = \sum_{\boldsymbol{X}} e^{-E(\boldsymbol{Y}|\boldsymbol{X})}$ is a normalization term called *partition function* and $E(\boldsymbol{Y}|\boldsymbol{X})$ is the Gibbs *energy function* of labeling $\boldsymbol{Y} \in \mathcal{L}^N$. For notational convenience, the conditioning $\boldsymbol{X}$ is omitted in the rest of this paper and the Gibbs energy of fully connected pairwise CRF model can be written as

$$E(\boldsymbol{Y}) = \sum_i \psi_u(y_i) + \sum_{i<j} \psi_p(y_i, y_j), \quad (3)$$

where $i, j \in \{1, \ldots, N\}$. The unary energy components $\psi_u(y_i)$ measure the distribution over the label assignment y_i given image features. The pairwise energy components $\psi_p(y_i, y_j)$ measure the cost of assigning labels y_i, y_j to pixels i, j simultaneously. In this work, unary energies are obtained from a pre-trained CNN, which roughly predicting labels without considering the smoothness of the label assignments. The pairwise energies provide an image data-dependent smoothing term that encourages assigning similar label to pixels with similar position. As was done in [18], we model pairwise function as

$$\psi_p(y_i, y_j) = \mu(y_i, y_j) \sum_{m=1}^{M} \omega^{(m)} k^{(m)}(\boldsymbol{f}_i, \boldsymbol{f}_j), \quad (4)$$

where $\mu(y_i, y_j) = 1$ if $y_i \neq y_j$ and $\mu(y_i, y_j) = 0$ if $y_i = y_j$. Each $k^{(m)}$ is a Gaussian kernel depends on pixel feature $\boldsymbol{f}$ and $\omega^{(m)}$ is weighted parameters. The CRF energy defined in Eq. 3 is minimized using truncated EM in [18].

2.3. Structural Analysis

Based on the CRF models described above, the conditional distributions $P(\boldsymbol{Y}_s|\boldsymbol{X}_s), \forall s \in \mathcal{S} = \{1, \ldots, S\})$ are obtained and the predicted labels of current scale s can be written as

$$\boldsymbol{Y}_s = \arg \max_{\boldsymbol{Y}_s \in \{1,\ldots,l\}^N} P(\boldsymbol{Y}_s|\boldsymbol{X}_s), s \in \{1, ..., S\}, \quad (5)$$

where N is the number of patches in current scale.

How to take full advantage of the inferred results is a challenging problem. Traditional methods tend to utilize voting strategy although significant contextual information is eliminated. For example, the CRF model trained by the patches in small scale tends to label the testing patch as large things or stuff, such as road and sky. Moreover, the inferred labels between different scales also show strong correlations which can be regarded as the contextual information between scales. For example, given a pixel which is labeled as road in small scale model and labeled as car in large scale model, then we can strongly believe that the pixel belongs to car region. On the contrary, if a pixel is labeled as road region in small scale model and labeled as pedestrian class in large scale model, then we will be confused about labeling the pixel for the pedestrian is impossible to be located in sky region. One solution to this problem is to calculate the joint conditional probability$P(\boldsymbol{Y}|\boldsymbol{X}) = P(\boldsymbol{Y}|\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_S)$. However, the equation can not be easily calculated for the absence of essential information.

In order to solve this problem, a sparse based model is built to calculate the final scene parsing results considering the prediction results of neighboring pixels. For each scale, the predicted labels $\boldsymbol{Y}_s$ are firstly resized to the original image size with the nearest interpolation method. Subsequently, for a certain pixel x_i, its labels and neighboring pixels inferred from the total S scales can be written as $\boldsymbol{a}_i \in \mathcal{L}^{k\times 1}$ and $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_N]^T$, where k is the total number of its neighbours. The mathematical equation can be written as

$$\boldsymbol{w}^* = \arg\min \|\boldsymbol{A}\boldsymbol{w} - \boldsymbol{Y}\|_2^2, s.t. \|\boldsymbol{w}\|_1 \leq \varepsilon, \qquad (6)$$

where ε indicates the residual error and this problem can be solved by Lasso[19]. The final labels can be obtained by

$$\boldsymbol{y}_i^* = \arg\min_{\boldsymbol{y}_i \in \mathcal{L}} \|\boldsymbol{a}_i \boldsymbol{w}^* - \boldsymbol{y}_i\|_2, i \in [1, ..., N]. \qquad (7)$$

3. EXPERIMENTS

We evaluate our approach over three published datasets. All results are obtained using the same parameters across the different algorithms and datasets.

3.1. Datasets

A number of scene parsing datasets are available and three datasets [20, 21, 22] are chosen for their challenging properties.

- **CamVid** is a road scene understanding dataset with 468 training images and 233 testing images of day and dusk scenes [20]. The challenge is to segment 11 classes such as road, buildings, cars, pedestrians, signs, sidewalk, etc.
- **Stanford-Background** dataset [21] contains 715 images of outdoor scenes with two separate label sets: semantic and geometric. We conduct our experiments for predicting the semantic category only. The semantic classes include seven background classes and a generic foreground class.
- **KITTI** dataset [22] is a large publicly available road scene dataset and some images were extracted and manually annotated for scene parsing. For comparison, the labeled images in [23] are utilized as experimental data which contains 142 images. Moreover, 11 semantic classes, such as buildings and road, are severely imbalanced distributed.

3.2. Performance Analysis

In order to evaluate the performance of the HDPA approach for scene parsing, two evaluation criterions are considered, one is pixel accuracy which indicates the percentage of pixels correctly labeled. The other one is per class accuracy which defined as the average of semantic category accuracies. Experimental results are evaluated by qualitative and quantitative measures. Typical scene parsing results of the three datasets are presented in Fig. 3. From the Table 1, it is obvious that the proposed method is robust in defining the scene labels.

For a more objective comparison, a more detailed analysis on the three datasets is presented as follows.

CamVid: The frames are sampled from two daytime and one dusk sequences and the first block of Table 1 shows the performance of the proposed method compare with state-of-the-arts. We can observe the positive impact of the proposed HDPA model in this work. For example, the appearance model [24] and the local labeling method [25] perform worse in the dust sequences for their low-level feature representation. On the contrary, our work exploits the power of CNN model and Gaussian pyramid strategy, adequate contextual information is utilized to improve the performance of HDPA. In addition, the CRF method [26] performs well when considering the class accuracy criteria. Our method takes advantage of the CRF model and takes different levels of the scene into consideration which leads to higher pixel accuracy.

Stanford-Background: Experiments on this dataset are conducted over 5-fold validation. Concretely, 572 images are served as training examples and the other 143 images are utilized to test the performance of the proposed HDPA method each time. The second block of Table 1 shows the superiority of our method. For example, Recursive Neural Network model [27] and Recurrent Neural Network model [13] can efficiently take the contextual constraints into account on the structure of the models. On this foundation, our work explores the power of contexts from two directions. The first one is focusing on the local contexts and the probabilistic graphical model is built. The other one exploits the strengths

Table 1. Quantitative scene parsing results, including pixel accuracy and class accuracy(%). The bold numbers represent the best scores.

Dataset	Approach	Pixel Acc.	Class Acc.
CamVid	SFM+Appearance [24]	69.1	53.0
	Boosting [26]	76.4	59.8
	Structured Random Forests [28]	72.5	51.4
	Local Label Descriptors [29]	73.6	36.3
	Boosting+pairwise CRF [26]	79.8	**59.9**
	Local Labeling+MRF [25]	77.6	43.8
	HDPA(ours)	**81.1**	49.9
Stanford	Stacked Labeling [30]	76.9	66.2
	Recursive Neural Networks [27]	78.1	N/A
	Recurrent Neural Networks [13]	80.2	69.9
	Hierarchical Features [5]	81.4	**76.0**
	WAKNN+MRF [31]	74.1	62.2
	HDPA(ours)	**81.7**	70.6
KITTI	Temporal Semantic Segmentation [23]	51.2	**61.6**
	Semantic Segmentation Retrieva [23]	47.1	58.0
	HDPA(ours)	**79.8**	45.84

of pyramid model by taking hierarchical inferred labels into solving a sparse problem. Experiments on pixel accuracy verified the contributions of this work.

KITTI: This sequence is captured with wide angle and was sampled from videos under a certain frequency. Moreover, the semantic label is imbalanced distributed and the long-tail phenomenon is obvious. Addressing these difficulties, temporal constraint is considered in [23] and high class accuracy verified the effectiveness of the temporal information. On the contrary, temporal context information does not take into account in our method temporarily and competitive results on the pixel criterion also show the superiority of the proposed method.

Although significant results have been reached when take the pixel accuracy as the evaluation criterion, the proposed HDPA method has shown its weakness on the aspect of class accuracy. On one hand, a unified framework which simultaneously considers the neighboring things and stuff can strengthen the object correlations. Furthermore, hierarchical CRF models based on Gaussian pyramid can take different level of object parsing into considerations which leads to adequate contextual for scene parsing. On the other hand, the proposed method is based on sampling certain number of patches from Gaussian pyramid and this strategy would ignore the small-sized semantic class. For example, for the KITTI dataset, the number of pixels defines as pole label[23] is very small and nearly zero number of pixels were correctly though our method. This phenomenon can explain the low class accuracy of the proposed HDPA method. Generally, high pixel accuracy can effectively reduce the negative effect of the low class accuracy on some aspects.

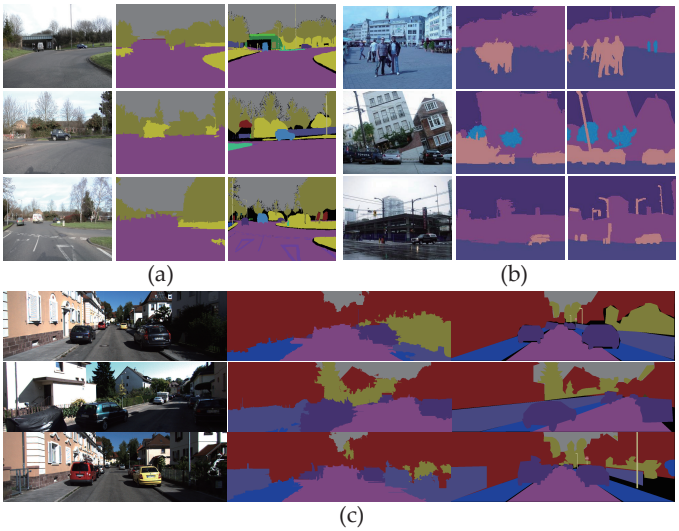

Fig. 3. Qualitative scene parsing results. CamVid results are shown in (a) and (b) is Stanford-Background results. KITTI results are demonstrated in (c). For each datset results, the first column indicates the input images, the second column shows the scene parsing result based on the proposed HDPA model and the last column provides the groundtruth.

4. CONCLUSION

In this paper, we present a hierarchical deep probability analysis model for pixel-level scene parsing. Specifically, the input image is firstly segmented under Gaussian pyramid with certain block numbers and each block is represented through a convolutional neural network. Subsequently, a probabilistic graphical model is utilized under each pyramid scale and the final pixel labels are obtained by probability analysis. The proposed method particularly exploits the power of probabilistic graphical model and deep convolutional neural network, and three contributions are claimed: unified framework for scene parsing, hierarchical probabilistic graphical modeling and adequate contextual information consideration. The superiority of the proposed method is verified on three benchmark datasets and the experimental results show that it outperforms the other competitors.

Although noticeable results have shown the effectiveness of the proposed method, some limitations still exist. For example, the average accuracy for small semantic object is low. How to increase the accuracy of small target is still a challenging problem.

5. REFERENCES

[1] D. Forsyth, J. Malik, M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler, "Finding pictures of objects in large collections of images," in *Proc. International Workshop on Object Representation in Computer Vision*, 1996, pp. 335–360.

[2] Q. Wang, M. Chen, and X. Li, "Quantifying and detecting collective motion by manifold learning," in *Proc. AAAI Conference on Artificial Intelligence*, 2017, pp. 4292–4298.

[3] Y. Yuan, Z. Jiang, and Q. Wang, "Video-based road detection via online structural learning," *Neurocomputing*, vol. 168, pp. 336–347, 2015.

[4] V. Lempitsky, A. Vedaldi, and A. Zisserman, "A pylon model for semantic segmentation," *Advances in Neural Information Processing Systems*, pp. 1485–1493, 2011.

[5] 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. 1915–29, 2013.

[6] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.

[7] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proc. European Conference on Computer Vision*, 2006, pp. 1–15.

[8] M. Cheng, S. Zheng, W. Lin, V. Vineet, P. Sturgess, N. Crook, N. Mitra, and P. Torr, "Imagespirit: Verbal guided image parsing," *ACM Transactions on Graph*, vol. 34, no. 1, pp. 1–11, 2014.

[9] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2–23, 2009.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. European Conference on Computer Vision*, 2014, pp. 346–361.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

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

[13] P. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling.," in *Proc. International Conference on Machine Learning*, 2014, pp. 82–90.

[14] Z. Jiang, Q. Wang, and Y. Yuan, "Adaptive road detection towards multiscale-multilevel probabilistic analysis," in *Proc. IEEE China Summit International Conference on Signal and Information Processing*, 2014, pp. 698–702.

[15] Q. Wang, J. Fang, and Y. Yuan, "Adaptive road detection via context-aware label transfer," *Neurocomputing*, vol. 158, pp. 174–183, 2015.

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

[17] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. British Machine Vision Conference*, 2014.

[18] J. Domke, "Learning graphical model parameters with approximate marginal inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2454–2467, 2013.

[19] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 58, no. 1, pp. 267–288, 1996.

[20] G. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.

[21] S. Gould, R. Fulton, and D. Koller, "Decomposing a scene into geometric and semantically consistent regions," in *Proc. IEEE International Conference on Computer Vision*, 2009, pp. 1–8.

[22] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[23] G. Ros, S. Ramos, M. Granados, A. Bakhtiary, D. Vazquez, and A. Lopez, "Vision-based offline-online perception paradigm for autonomous driving," in *Proc. IEEE Winter Conference on Applications of Computer Vision*, 2015, pp. 231–238.

[24] G. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla, "Segmentation and recognition using structure from motion point clouds," in *Proc. European Conference on Computer Vision*, 2008, pp. 44–57.

[25] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *Proc. European Conference on Computer Vision*, 2010, pp. 352–365.

[26] P. Sturgess, K. Alahari, L. Ladicky, and P. Torr, "Combining appearance and structure from motion features for road scene understanding," in *Proc. British Machine Vision Conference*, 2009.

[27] R. Socher, C. Lin, A. Ng, and C. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *Proc. International Conference on Machine Learning*, 2011, pp. 129–136.

[28] P. Kontschieder, S. Bulo, H. Bischof, and M. Pelillo, "Structured class-labels in random forests for semantic image labelling," in *Proc. International Conference on Computer Vision*, 2011, pp. 2190–2197.

[29] Y. Yang, Z. Li, L. Zhang, C. Murphy, J. Ver Hoeve, and H. Jiang, "Local label descriptor for example based semantic image labeling," in *Proc. European Conference on Computer Vision*, 2012, pp. 361–375.

[30] D. Munoz, J. Bagnell, and M. Hebert, "Stacked hierarchical labeling," in *Proc. European Conference on Computer Vision*, 2010, pp. 57–70.

[31] G. Singh and J. Kosecka, "Nonparametric scene parsing with adaptive feature relevance and semantic context," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3151–3157.