EDGE-AWARE INTEGRATION MODEL FOR SEMANTIC LABELING OF RARE CLASSES

Liangjiang Yu[1] *and Guoliang Fan*[1,2]
[1]School of Electrical and Computer Engineering, Oklahoma State University, USA
[2]School of Automation and Information Engineering, Xi'an University of Technology, China

ABSTRACT

Rare class objects in natural images often convey semantically important information than background for scene understanding, but they are often overlooked during image parsing due to their low occurrence frequency and limited spatial coverage. In this work, we present a superpixel-based and rare class-oriented scene labeling framework (sRCSL), which seamlessly incorporates edge features into an integration of global and local CNN models. A dual-mode coarse-to-fine superpixel representation is developed for accurate yet efficient labeling, where coarse and fine superpixels are applied to background and rare classes respectively. Furthermore, saliency detection is incorporated by the combination of probabilistic belief maps from local and global inference. Experimental results demonstrate promising performance of the proposed framework on the SIFTflow dataset both qualitatively and quantitatively for semantic labeling, especially for rare classes.

Index Terms— Semantic labeling, CNN, Label transfer, Superpixel, Rare classes.

1. INTRODUCTION

Scene labeling aims at assigning a semantic label to each pixel in a given image, resulting in semantic classification and segmentation simultaneously [1, 2] for comprehensive scene understanding. In most natural scene images, rare class objects could be unintentionally neglected to achieve higher overall labeling accuracy [3] due to their quantitative and spatial limitation. However rare classes (e.g., human, boat, vehicle, etc.) are usually important to visual understanding of a natural scene. Some approaches were proposed to ensure sufficient learning for rare class objects, most of which try to encourage a balanced data distribution among all classes. This is achieved by either manually controlling or combining various sampling methods [3–5], or assigning different weights to rare classes during the training [6]. Some other methods focus on merging multiple model components to enhance rare class learning [2, 7]. On the other hand, superpixel-based CNN has been successfully applied for edge learning, including feature masking for semantic segmentation [8], and a mapping from intensity to color contrast and distribution sequence through superpixels for CNN training [9].

This work is supported in part by the National Science Foundation (NSF) under grant NRI-1427345 (USA) and the Shaanxi Hundred Talent Program (China). Corresponding author: guoliang.fan@okstate.edu.

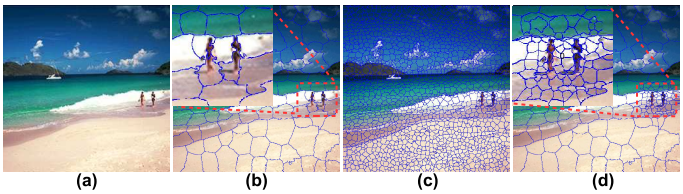

Fig. 1. Superpixel-based segmentation strategies. (a) A given observation; (b) coarse superpixel segmentation where rare classes are overlooked due to small spatial coverage; (c) finer superpixel segmentation -unnecessary for background; (d) the dual-mode coarse-to-fine (*c2f*) superpixel representation.

In this work, we propose a new superpixel-based and rare class oriented scene labeling (sRCSL) framework that integrates global-local CNNs and label transfer. Moreover, it supports dual-mode coarse-to-fine (*c2f*) scene labeling (as shown in Fig.1) for effective and efficient rare class recognition and segmentation. Specifically, we have three contributions:

- Edge information is incorporated during CNN learning/inference and label transfer to support perceptually meaningful segmentation of rare classes.
- The *c2f* superpixel representation balances the labeling accuracy of small rare class objects and the computational complexity in the large background areas.
- Rare class labeling and saliency detection are integrated probabilistically through the combination of multi-class likelihoods and foreground belief.

2. RELATED WORK

Deep convolutional neural networks (CNN) [10] have successfully advanced scene labeling research by learning effective and discriminative high-level representations [4, 11, 12]. Such networks are able to mitigate the limited system robustness caused by low-level human engineered features. On the other hand, nonparametric methods [13–20] involving image retrieval and label transfer have also achieved promising scene labeling performance. Furthermore, the label transfer method was incorporated into the CNN framework in [21],

where the local ambiguities from CNN models are alleviated by using global scene semantics, and scene-relevant class dependencies and priors are transfered by matching learned CNN features without additional training effort.

To cope with rare class objects with unbalanced distribution, model fusion [2, 7] and retrieval set expansion [3, 21] have been developed to find more discriminative representation among classes. In [22], two techniques are combined efficiently to extenuate the insufficient rare class training. One is the scene assisted rare class retrieval, which selectively add rare classes based on the global scene to reduce class ambiguities rather than random sampling (irrelevant classes may appear). While the second one is a rare class-balanced CNN with the focus on rare class objects for re-inference near potential rare classes. However, since edge information is not involved during the patch-based training, shapes of rare class objects may not be well preserved after scene labeling. In [9], salient object detection is formulated as a binary labeling problem using superpixel-based CNN. This network is learned from superpixel-wise contrast features between foreground objects and background areas hierarchically. Inspired by the fact that rare class objects are usually salient locally, we propose an edge-aware integration framework for rare class labeling, which provides not only multi-class local and global beliefs, but also a foreground belief map through superpixel-based CNN learning and inference for rare classes.

3. PROPOSED METHODS

We first present the proposed superpixel-based and rare class-oriented scene labeling framework (sRCSL as shown in Fig. 2) with a dual-mode coarse-to-fine (*c2f*) superpixel segmentation and inference flow. Then we introduce superpixel-based CNN for salient object detection, which leads to our proposed edge aware CNN learning and inference framework for an integrated belief map, which not only provides multi-class likelihood, but also a probabilistic foreground belief for semantically meaningful rare class segmentation and labeling.

3.1. Problem Formulation

Given an image $\mathbf{x}$ represented by superpixels [23], let s_i denote a superpixel and y_j the corresponding label ($j = 1, 2, \cdots, L$) from L classes. The proposed sRCSL framework incorporates two belief maps from local to global understanding of the scene,

$$y(s_i) = \arg \min_{j=1,\cdots,L} \big(E(s_i, y_j) + H(s_i, y_j)\big), \tag{1}$$

where $E(s_i, y_j)$ is an integration of class likelihood,

$$E(s_i, y_j) = -(P_L(s_i, y_j) + P_G(s_i, y_j)), \tag{2}$$

where $P_L(s_i, y_i)$ is the local belief computed using a traditional patch-based CNN denoted by *pCNN* [21, 24]. The global belief $P_G(s_i, y_i)$ is computed by label transfer using statistical features from the scene-assisted retrieval set [22]. $H(s_i, y_j)$ is the integrated belief of rare class balanced CNN and superpixel-based CNN, denoted by *rCNN* and *sCNN* [9].

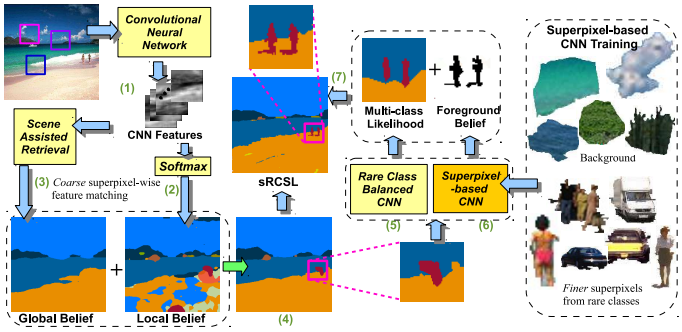

Fig. 2. The computational flow in the proposed sRCSL framework. (1) Traditional CNN based on patches; (2) local belief computation; (3) global belief by transferring labels; (4) rare class object localization by combining global and local beliefs; (5) refined local belief by rare class balanced CNN; (6) foreground belief by superpixel-based CNN; (7) Integration of multi-class likelihoods and the binary foreground belief.

Unlike using superpixels for post-processing in [22], we develop a dual-mode coarse-to-fine (*c2f*) superpixel representation to incorporate edge information during CNN learning and inference. First, superpixel-wise feature matching from coarse segmentation is implemented during global label transfer, and then the local belief is achieved by normalized class-likelihood within each superpixel. Combining these two beliefs gives the class likelihood $E(s_i, y_j)$. Second, we apply finer superpixels around potential rare classes obtained by $E(s_i, y_j)$, which is integrated with $H(s_i, y_j)$ by combining superpixel-wise *rCNN* and *sCNN* belief maps.

3.2. Superpixel-based CNN

We adopt a superpixel-based CNN framework to capture higher-level discriminative contrast features using hierarchical edge information [9] from finer superpixel segmentation. There are two motivations. First, small image patterns in the traditional CNN are not optimized to produce densely labeled maps; second, most rare class objects are locally salient due to the contrast with background. Specifically, given an image blob $\mathbf{x}$ of a salient object, it could be segmented as a group of N superpixels $\mathbf{S} = \{s_1, \cdots, s_x, \cdots, s_N\}$, each s_x is represented by a color uniqueness sequence $Q_x^C = \{q_1^c, \cdots, q_j^c \cdots, q_M^c\}$ and also a color distribution sequence $Q_x^D = \{q_1^d, \cdots, q_j^d \cdots, q_M^d\}$, each with size M, where $M \leq N$. q_j^c and q_j^d are defined as

$$q_j^c = \tau(s_j) \cdot |C(s_x) - C(s_j)| \cdot w(Z(s_x), Z(s_j)), \tag{3}$$

$$q_j^d = \tau(s_j) \cdot |Z(s_x) - Z(s_j)| \cdot w(C(s_x), C(s_j)), \tag{4}$$

where $\tau(s_j)$ counts the number of pixels in s_j, $C(s_x)$ and $Z(s_x)$ are the mean color and position within s_x. $w(Z(s_x),$

Algorithm 1 Pseudo code of the proposed sRCSL algorithm.

for each training image $\mathbf{x}_n$ **do**
  Compute Q^C and Q^d as (3) and (4) for rare classes
end for
Train *pCNN* and *rCNN*
Train *sCNN* using Q^C and Q^D by finely segmented superpixels
for each query image $\mathbf{x}_t$ **do**
  Coarse segmentation of $\mathbf{x}_t$ as N superpixels $\mathbf{S} = \{s_1, \cdots, s_N\}$
  for each s_i **do**
    Compute $E(s_i, y_j)$ using *pCNN* as (2)
    if $\zeta(s_i) = 1$ **then**
      finer segmentation of s_i as $\mathbf{S_i}' = \{s'_{i,1}, \cdots, s'_{i,M}\}$
      for each $s'_{i,k}$ **do**
        Compute $E(s'_{i,k}, y_j)$ using *pCNN* as (2)
        Compute $H(s'_{i,k}, y_j)$ using *rCNN* and *sCNN* as (7)
        Compute $y(s'_{i,k})$ according to (1)
      end for
    else
      Compute $y(s_i)$ as (1)
    end if
  end for
end for

$Z(s_j))$ weights the distance between s_x and s_j, while $w(C(s_x), C(s_j))$ gives the color similarity,

$$w(Z(s_x), Z(s_j)) = \exp(-\frac{1}{2\sigma_z^2}\|Z(s_x) - Z(s_j)\|^2), \quad (5)$$

$$w(C(s_x), C(s_j)) = \exp(-\frac{1}{2\sigma_c^2}\|C(s_x) - C(s_j)\|^2), \quad (6)$$

where σ_z^2 and σ_c^2 are two parameters to control the sensitivity of evaluating location and color similarity. These two color sequences provides a mapping from 2D intensity to a 1D feature space that captures color distribution spatially and chromatically, constrained by edge information. We apply these two features to rare class objects, and then feed them into *sCNN* for foreground belief computation.

3.3. Edge-aware Belief Integration

$H(s_i, y_j)$ given in (1) evaluates not only the class likelihood of $y_j, j = \{1, \cdots, L\}$ around rare class pixel s_i, but also foreground belief achieved by *sCNN*,

$$H(s_i, y_j) = -\big(P_r(s_i, y_j) + P_u(s_i, y_j)\big) \cdot \zeta(s_i), \quad (7)$$

where $\zeta(\cdot) = 1$ if s_i is within a rare class region determined by $E(s_i, y_j)$ in (2). $P_r(s_i, y_j)$ is the normalized belief obtained by passing all pixels from s_i into *rCNN* [22]. $P_u(s_i, y_j)$ is the superpixel-based foreground belief, which is achieved by *sCNN* trained using color uniqueness and distribution obtained by (3) and (4) for rare classes using finer segmentation. Therefore, the final sRCSL labeling $\mathbf{y}$ can be computed as $\mathbf{y} = \cup_{i=1:K}(y_i)$, where K is the total number of superpixels. Note that in order to build $\mathbf{y}$, we evaluate $E(s_i, y_j)$ in (1) twice for potential rare class areas (where the finer segmentation is applied). The first one is for coarse segmentation and the second one for finer segmentation. The full sRCSL framework is summarized in Algorithm 1.

4. EXPERIMENTS

The SIFTflow dataset [13] that was widely used for scene labeling [3,5,21] is tested for rare class labeling. There are 2688 images (256×256) captured from 8 natural scenes that have been manually labeled with 33 classes. The dataset is split into 2488 training and 200 testing images. First, we evaluate the overall *pixel accuracy* (percentage of correctly labeled pixels) and the *class accuracy* averaged overall all classes. Second, both accuracies are evaluated for rare classes only.

4.1. Overall Labeling Accuracy

We compare sRCSL with a set of recent ones in Table 4.1 in terms of overall labeling accuracy. It is shown that sRCSL achieves 9.5% and 1.7% better class accuracy than the baseline CNN algorithm [21] and a recent RCSL-Seg framework where superpixels are used for post-processing [22]. It also outperforms [3] where scene labeling is based on statistical inference on superpixels, showing the effectiveness of the *c2f* superpixel-based superpixel strategy.

Table 1. Overall performance on the SIFTflow dataset

	Pixel (%)	Class (%)
Multiscale ConvNet [4]	67.9	45.9
[4]+cover(balanced)	72.3	50.8
[4]+cover(natural)	78.5	29.6
CNN, Pinheiro et al. [11]	76.5	30.0
RCNN, Pinheiro et al. [11]	77.7	29.8
Superparsing [19]	76.9	29.4
Eigen et al. [17]	77.1	32.5
Tighe et al. [20]	78.6	39.2
Singh et al. [18]	79.2	33.8
Gould et al. [25]	78.4	25.7
Integration model (metric) [21]	80.1	39.7
Integration model (metric) [5]	81.2	45.5
Yang et al. [3]	79.8	48.7
RCSL [22]	80.8	41.2
RCSL-seg [22]	81.6	47.5
sRCSL	**82.3**	**49.2**

4.2. Rare Class Labeling Accuracy

We report pixel and class accuracies of rare class objects only ($< 5\%$ frequency in the training dataset [3]) in Table 4.2. Again, sRCSL outperforms several recent methods. Some qualitative results are shown in Fig. 3 where sRCSL is able to achieve semantically meaningful segmentation and labeling. For example in the 4^{th} column of the 4^{th} and 5^{th} row, silhouettes of human and vehicles from sRCSL are better preserved compared with other methods, and there is less ambiguity between rare classes and background. In the 1^{st} row, vehicles tend to have more detailed boundaries, and in the 3^{rd} row, we are able to clearly label and segment almost all of the windows. We are even able to detect some human in the last image, which is not correctly labeled in the ground truth.

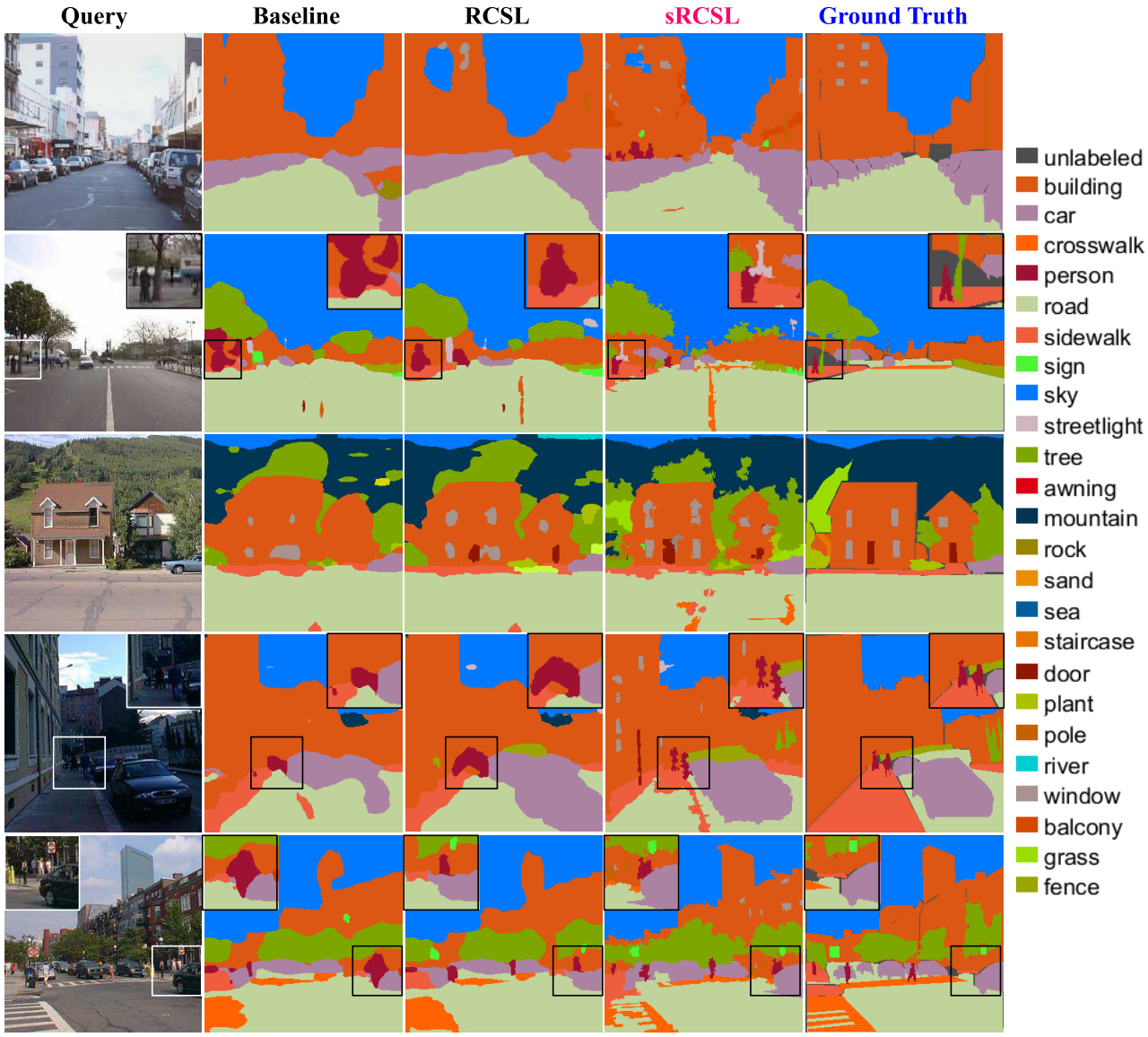

Fig. 3. Some qualitative examples of labeling results. From the first to the last column: input images; the baseline integration model [21]; RCSL (without post-processing) [22]; the proposed sRCSL algorithm with *c2f* superpixels; the ground truth.

Table 2. Performance of rare classes on the SIFTflow dataset

	Pixel (%)	Class (%)
Tighe et al. [20]	48.8	29.9
Yang et al. [3]	59.4	41.9
Shuai et al. [21]	-	30.7
Shuai et al. [5]	-	37.6
RCSL [22]	61.3	39.2
RCSL-Seg [22]	62.6	42.3
sRCSL	**64.5**	**43.1**

5. CONCLUSIONS

We have proposed a new scene labeling algorithm called sRCSL that is especially focused on rare class objects. The key idea is the incorporation of edge information in the CNN-based scene labeling framework. The dual-mode *c2f* superpixel-based CNN inference module is incorporated to carefully infer and label rare class objects while sustaining computational efficiency in background regions. Experimental results show promising performance of sRCSL in terms of overall classification accuracy, especially for rare classes. In the future, we plan to incorporate superpixels directly for edge-aware multi-class learning and inference.

6. REFERENCES

[1] L. Li, R. Socher, and F. Li, "Towards total scene understanding: Classification, annotation and segmentation in an automatic framework," in *Proc. CVPR*, June 2009, pp. 2036–2043.

[2] M. George, "Image parsing with a wide range of classes and scene-level context," in *Proc. CVPR*, June 2015, pp. 3622–3630.

[3] J. Yang, B. Price, S. Cohen, and M. H. Yang, "Context driven scene parsing with attention to rare classes," in *Proc. CVPR*, June 2014, pp. 3294–3301.

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

[5] B. Shuai, Z. Zuo, G. Wang, and B. Wang, "Scene parsing with integration of parametric and non-parametric models," *IEEE Trans. Image Processing*, vol. 25, no. 5, pp. 2379–2391, May 2016.

[6] B. Shuai, Z. Zuo, B. Wang, and G. Wang, "Dag-recurrent neural networks for scene labeling," in *Proc. CVPR*, June 2016, pp. 3620–3629.

[7] H. Caesar, J. Uijlings, and V. Ferrari, "Joint calibration for semantic segmentation," in *Proc. BMVC*. September 2015, pp. 29.1–29.13, BMVA Press.

[8] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. CVPR*, June 2016, pp. 3150–3158.

[9] S. He, R. W. H. Lau, W. Liu, Z. Huang, and Q. Yang, "Supercnn: A superpixelwise convolutional neural network for salient object detection," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 330–344, 2015.

[10] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[11] P. H. O. Pinheiro and R. Collobert, "Recurrent convolutional neural networks for scene labeling," in *Proc. ICML*, 2014, pp. 82–90.

[12] M. Liang, X. Hu, and B. Zhang, "Convolutional neural networks with intra-layer recurrent connections for scene labeling," in *Advances in NIPS 28*, pp. 937–945. Curran Associates, Inc., 2015.

[13] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing: Label transfer via dense scene alignment," in *Proc. CVPR*, June 2009, pp. 1972–1979.

[14] C. Liu, J. Yuen, and A. Torralba, "Nonparametric scene parsing via label transfer," *IEEE Trans. PAMI*, vol. 33, no. 12, pp. 2368–2382, Dec 2011.

[15] S. Gould and Y. Zhang, "Patchmatchgraph: Building a graph of dense patch correspondences for label transfer," in *Proc. ECCV*, 2012, pp. 439–452.

[16] F. Tung and J. J. Little, "Collageparsing: Nonparametric scene parsing by adaptive overlapping windows," in *Proc. ECCV*, 2014, pp. 511–525.

[17] D. Eigen and R. Fergus, "Nonparametric image parsing using adaptive neighbor sets," in *Proc. CVPR*, June 2012, pp. 2799–2806.

[18] G. Singh and J. Kosecka, "Nonparametric scene parsing with adaptive feature relevance and semantic context," in *Proc. CVPR*, June 2013, pp. 3151–3157.

[19] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with superpixels," in *Proc. ECCV*, 2010, pp. 352–365.

[20] J. Tighe and S. Lazebnik, "Finding things: Image parsing with regions and per-exemplar detectors," in *Proc. CVPR*, June 2013, pp. 3001–3008.

[21] B. Shuai, G. Wang, Z. Zuo, B. Wang, and L. Zhao, "Integrating parametric and non-parametric models for scene labeling," in *Proc. CVPR*, June 2015, pp. 4249–4258.

[22] L. Yu and G. Fan, "Rare class oriented scene labeling using cnn incorporated label transfer," in *Advances in Visual Computing: Proc. ISVC*, 2016, pp. 309–320.

[23] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. PAMI*, vol. 34, no. 11, pp. 2274–2282, Nov 2012.

[24] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," in *Proceeding of the ACM Int. Conf. on Multimedia*, 2015.

[25] S. Gould, J. Zhao, X. He, and Y. Zhang, "Superpixel graph label transfer with learned distance metric," in *Proc. ECCV*, 2014, pp. 632–647.