Semantic Segmentation for Real-World Data by Jointly Exploiting Supervised and Transferrable Knowledge

Li-Hsien Lu
s103062588@m103.nthu.edu.tw

Chiou-Ting Hsu
cthsu@cs.nthu.edu.tw

Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan

Abstract

This paper addresses two major challenges in semantic segmentation for real-world data. First, with ever-increasing semantic labels, we need a more pragmatic approach other than existing fully-supervised methods. Second, semantic segmentation for very small or rarely-appeared objects are still very challenging for existing methods. In this paper, we propose to (1) fully utilize the predicted label information from an existing supervised model and to (2) infer newly generated labels via label transfer from a real-world dataset. We propose a "content-adaptive" and "label-aware" MRF framework to jointly exploiting both the supervised and label-transferrable knowledge. The proposed method needs no off-line training and can easily adapt to real-world data. Experimental results on SIFT Flow and LMSun datasets demonstrate the effectiveness of the proposed method, and show promising performance over state-of-the-art methods under the real-world scenario.

1 Introduction

The goal of semantic segmentation is to assign a category label to each pixel in an image. Existing parametric methods [1, 2, 3, 4, 5, 6] infer the pixel-level labels based on fully-supervised algorithms. Earlier methods [7, 8] usually train a multi-category classifier with a densely annotated image dataset, then assign one category label to each region (e.g., sliding window, superpixel, object proposal), and finally derive pixel-level labels by including contextual information.

With the rapid advance of deep learning, Convolutional Neural Networks (CNN) have been largely included in many semantic segmentation approaches [9, 10, 11]. However, the idea of end-to-end dense segmentation was not considered tractable until the prominent approach: fully convolutional network (FCN) [1] was proposed. In [1], the fully connected layers in CNN models are replaced with convolutional layers; thus, given an arbitrary-sized image, the network extracts and combines multi-resolutional layer responses to predict the labelling result with the same spatial dimension as the query image. Based on the architecture of FCN, a number of methods [3, 4, 5] are proposed to further improve the labelling performance. In [3], the authors proposed to refine labelling results via a fully-connected Conditional Random Fields (CRF). In [4], instead of using bilinear interpolation to upsample the outputs of FCN, the authors proposed to learn a deconvolution network on

doi:10.5244/C.110.###

Pages 84.1-84.12
DOI: https://dx.doi.org/10.5244/C.30.84

top of the convolutional layers of FCN. In [5], an object regularization for FCN is proposed to refine the mislabelled objects.

Although existing parametric methods have achieved remarkable performance on datasets with a moderate size of category labels, these methods are not easy to adapt to real-world data. As new labels as well as new data are generated increasingly, it is impossible to predefine a large number of categories including every possible label. Moreover, even if we retrain a parametric model once new labels are given, we will need a large number of newly-annotated dataset for training. Therefore, nonparametric methods [11, 12, 13, 14, 15, 16, 17] have been proposed for semantic segmentation by transferring labels from similar images. Given a query image and an annotated data set, nonparametric methods first conduct a visual search to retrieve a set of reference images and then transfer labels from the reference images to the query image. Under this training-free scenario, nonparametric methods are more efficient and adaptive to real-world data. Nevertheless, because no off-line training stage is involved in nonparametric methods, their performance is inherently inferior to that of parametric methods.

In this paper, we focus on semantic segmentation problem for dynamically increasing real-world data. The scenario we assume in this paper is that, given one parametric model (off-line trained with a fixed set of labels) and one annotated real-world dataset (with a large label set), how we should take advantage of both to predict the pixel-level labels for any query image. Since parametric models (e.g., FCN [1]) have been shown to achieve promising results for known labels, we should leverage their well-learned knowledge through referring to the up-to-date dataset. Although the idea of combining parametric and nonparametric methods seems intuitive, we have to deal with challenges in both methods to ensure a high-quality result. On one hand, FCN-based parametric methods perform well only in certain spatial scales, because they involve fixed size receptive fields but consider little contextual information across the whole image. Therefore, the labelling result for objects larger or smaller than the receptive field are usually fragmentary [5] or miss-labelled. On the other hand, the performance of nonparametric methods is generally inferior to that of parametric methods and highly depends on the semantic quality of the reference images. Moreover, both methods usually fail to label objects of significantly smaller size (e.g., moon, streetlight, traffic sign), because the number of annotated pixels given for these labels (called “rare labels”) is negligibly smaller than that of other larger objects. We therefore need a good strategy to maximize the benefit from both methods and also tackle their different challenges.

We propose a Markov Random Field (MRF) framework to combine (1) the knowledge of known labels learned from an existing FCN-based model and (2) the knowledge of all possible labels transferred from a set of retrieved images. Because there involves no off-line training stage, the proposed method is efficient and adaptive to real-world data. We use two datasets: SIFT Flow [12] and LMSun [13], as the real-world data to evaluate the proposed method. Experiments demonstrate that our method achieves promising results even when there exist unknown labels to the parametric model.

2 Proposed method

Figure 1 shows the overview of our proposed method. Under the proposed scenario, we assume that there exists a parametric model trained with the label set C_{fs}, and a real-world dataset annotated with the label set C_r, where $C_{fs} \neq C_r$. Because the label set C_r contains

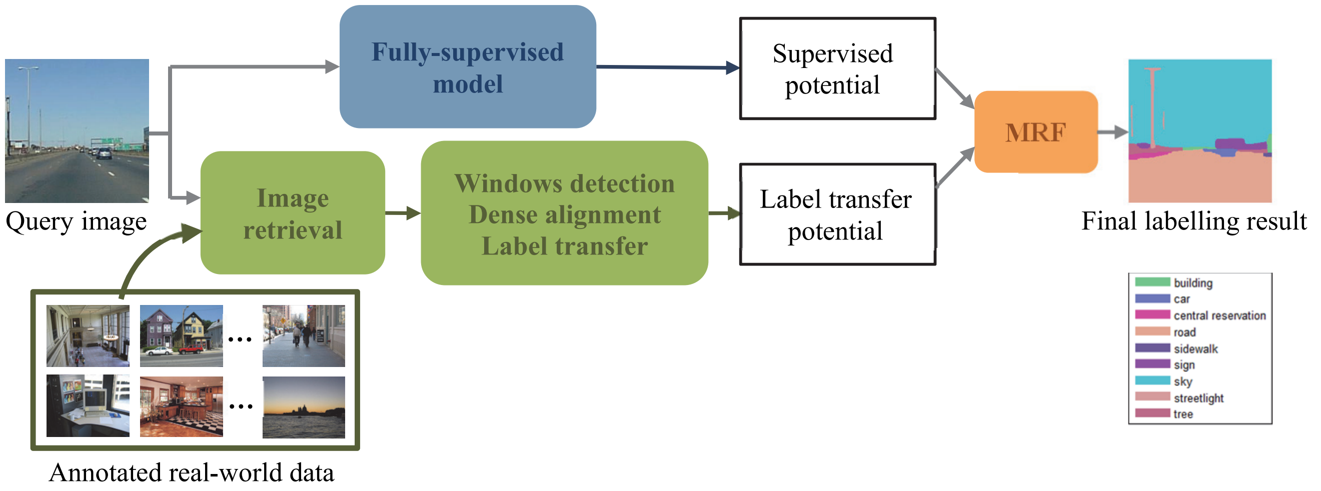


Figure 1: Overview of the proposed method.

new labels which are not included in C_{fs}, the parametric model alone is unable to predict labels $c \in C_r \backslash C_{fs}$.

Given a query image $I \in R^{m \times n}$, we model the semantic segmentation problem using Markov Random Field (MRF) inference. We formulate the MRF energy function over the field of labels $\boldsymbol{c} = \{c(\boldsymbol{p}), \boldsymbol{p} \in I\}$ by:

$$E(\boldsymbol{c}) = -\sum_{\boldsymbol{p} \in I} \left[(1 - \alpha(I)) \cdot \psi_{fs}(c, \boldsymbol{p}) + \alpha(I) \beta(c, I) \cdot \psi_{trans}(c, \boldsymbol{p}) \right] \\ + \lambda \sum_{(\boldsymbol{p}, \boldsymbol{q}) \in \varepsilon} \theta(c(\boldsymbol{p}), c(\boldsymbol{q})), \quad (1)$$

where $c(\boldsymbol{p})$ is the label of the pixel $\boldsymbol{p}$, ε defines the set of adjacent pixels, and λ is a smoothing constant. The term $\psi_{fs}(c, \boldsymbol{p})$ indicates the supervised potential derived by the parametric model; and the term $\psi_{trans}(c, \boldsymbol{p})$ indicates the label transfer potential obtained by the nonparametric model. The pairwise potential term $\theta(\cdot,\cdot)$ is similarly defined as in [13] according to the probabilities of label co-occurrence in the real-world dataset. The parameter $\alpha(I)$ is adaptive to different query image I so as to dynamically combine the two potentials. The term $\beta(c, I)$ is a label-aware parameter for balancing the priority of rare labels.

We introduce the supervised potential $\psi_{fs}(c, \boldsymbol{p})$, the label transfer potential $\psi_{trans}(c, \boldsymbol{p})$, and the MRF framework in Sections 2.1 – 2.3, respectively.

2.1 Supervised potential

In this paper, we use FCN [1] as the off-line trained fully-supervised model. Given a query image I, we derive $|C_{fs}|$ score maps $\mathbf{M}_c \in R^{m \times n}, \forall c \in C_{fs}$ from FCN, and define the supervised potential in terms of $\mathbf{M}_c$ by:

$$\psi_{fs}(c, \boldsymbol{p}) = \begin{cases} \frac{\mathbf{M}_c(\boldsymbol{p}) - m_{min}}{m_{max} - m_{min}}, & \text{for } c \in C_{fs} \\ 0, & \text{otherwise} \end{cases}, \quad (2)$$

where $m_{min} = \min_{\forall c \in C_{fs}; \forall \boldsymbol{p} \in I} (\mathbf{M}_c(\boldsymbol{p}))$ and $m_{max} = \max_{\forall c \in C_{fs}; \forall \boldsymbol{p} \in I} (\mathbf{M}_c(\boldsymbol{p}))$.

2.2 Label transfer potential

As mentioned before, the real-world data contains new labels $c \in C_r \backslash C_{fs}$ which are completely unknown to FCN. In addition, FCN is not adaptive to different object sizes because of its fixed size receptive field. We therefore modify the existing nonparametric methods [11, 14] and conduct label transfer in terms of content-adaptive windows.

Given a query image I, we first retrieve the top K similar images from the real-world dataset using either hand-crafted features or pre-trained CNN features [18]. Next, we follow the similar idea in [11, 14] to transfer labels through windows. However, unlike [11, 14], we adopt Faster R-CNN [19] to detect a set of category-independent windows and derive their feature representation simultaneously. For each query window w, we search the top κ matched windows from the K retrieved images using the feature representation derived by Faster R-CNN. Nevertheless, because the Faster R-CNN model released by [19] was trained with Pascal VOC dataset [20], the model works well when detecting "thing" windows (e.g., car) but can hardly recognize "stuff" windows (e.g., sky). To have a comprehensive window detector, we further collect a set of annotated "stuff" windows of different sizes to fine-tune the network. With the fine-tuned Faster R-CNN, our window detectors better characterizes various content across multiple scales in the real-world data.

Finally, we adopt the dense alignment method [11] to align the spatial layout of the reference window in accordance with its matched query window. With the SIFT flow fields [12] between the matched window pairs, we define the label transfer potential by:

$$\psi_{trans}(c,\mathbf{p}) = \begin{cases} \sum_{w \in W} \sum_{w_i \in W_{rs}} \delta[L(\widetilde{w}_i(\mathbf{p}+\boldsymbol{f}_i),\widehat{\mathbf{p}}) = c]\phi_{size}(w_i)\phi_{idf}(c), & \text{for } c \in C_r \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where W denotes the set of windows detected by the fine-tuned Faster R-CNN from I, W_{rs} denote the set of detected windows from the K retrieved images, w_i is a matched window of w, $\widetilde{w}_i$ is the resized w_i, $\widehat{\mathbf{p}}$ indicates the window-centric coordinates of $\mathbf{p}$ in w, $\boldsymbol{f}_i$ is the SIFT flow vector [12] from w to $\widetilde{w}_i$, and $L(\cdot,\cdot)$ [14] transfers the label from $\widetilde{w}_i$ to w.

Thus, in Eq. (3), $L(\widetilde{w}_i(\mathbf{p}+\boldsymbol{f}_i),\widehat{\mathbf{p}})$ assigns the label of the densely aligned window $\widetilde{w}_i$ to w; the term $\phi_{size}(w_i)$ is used to penalize large windows [11]:

$$\phi_{size}(w_i) = \frac{1}{N(w_i)}, \quad (4)$$

where $N(w_i)$ is the number of pixels in w_i. $\phi_{idf}(c)$ is similarly defined as in [14] to reflect the rareness of the label c in the retrieval set:

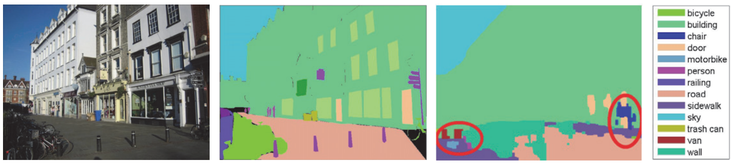
$$\phi_{idf}(\mathrm{c}) = \frac{1}{N(\mathrm{c})^{\gamma}}, \quad (5)$$

where $N(\mathrm{c})$ denotes the number of pixels of the label c in the retrieval set, and γ is a constant and is set as $\gamma = 0.38$ in our experiments.

2.3 MRF framework

We next explain how we adaptively combine the two potentials into the proposed MRF framework. Note that, even though there exist unknown labels to the FCN model, the model is unaware of this fact and will still assign one label to each pixel. We therefore have inconsistent labelling estimations from the two potentials and need to determine which one should be more reliable.

We first define the set of labels estimated by the supervised and label transfer potentials as $\hat{C}_{fs}(I) = \left\{c_{fs} \in C_{fs} \middle| c_{fs}(\mathbf{p}) = \mathrm{argmin}_{\forall c \in C_{fs}} \psi_{fs}(c,\mathbf{p}), \forall \mathbf{p} \in I\right\}$ and $\hat{C}_r(I) =$



(a) (b) (c)

Figure 2: Examples of noisy labelling results with false-positive new labels (marked with red circles) estimated via the label transfer model. (a) Query image; (b) ground truth; and (c) the labelling result by the label transfer model.

$\{c_r \in C_r | c_r(\boldsymbol{p}) = \text{argmin}_{\forall c \in C_r} \psi_{trans}(c,\boldsymbol{p}), \forall \boldsymbol{p} \in I\}$, respectively. The estimated label sets $\hat{C}_{fs}(I)$ and $\hat{C}_r(I)$ depend on the content of the query image I and usually contain only part of the original label sets, i.e., $\hat{C}_{fs}(I) \subseteq C_{fs}$ and $\hat{C}_r(I) \subseteq C_r$. Because the fully-supervised model generally performs well on known labels, whereas the label transfer model is more reliable for new labels, we propose to determine the content-adaptive parameter $\alpha(I)$ in Equation (1) by:

$$\alpha(I) = \frac{|\hat{C}_r(I) - \hat{C}_{fs}(I)|}{|\hat{C}_r(I) \cup \hat{C}_{fs}(I)|}, \tag{6}$$

where $|\hat{C}_r(I) - \hat{C}_{fs}(I)|$ denotes the number of labels estimated by the label transfer model but are not estimated by the fully-supervised model, and $|\hat{C}_r(I) \cup \hat{C}_{fs}(I)|$ denotes the number of labels estimated in either model. With Equation (6), we will rely more on $\psi_{fs}(c,\boldsymbol{p})$ if the two label sets $\hat{C}_{fs}(I)$ and $\hat{C}_r(I)$ are consistent and will rely more on $\psi_{trans}(c,\boldsymbol{p})$ if otherwise.

However, because the labelling result by the label transfer model is not 100% accurate, the difference set $(\hat{C}_r(I) - \hat{C}_{fs}(I))$ may contain some false-positive new labels. An example is given in Figure 2, where some "new" labels are wrongly estimated. In Equation (6), this noisy estimation will mislead Equation (1) to heavily rely on $\psi_{trans}(c,\boldsymbol{p})$. Therefore, instead of calculating the number of labels in the difference set $(\hat{C}_r(I) - \hat{C}_{fs}(I))$, we involve the number of pixels belonging to this difference set into consideration and modify Equation (6) into:

$$\alpha(I) = \frac{ln(\#pixels(\hat{C}_r(I) - \hat{C}_{fs}(I)))}{ln(m \times n)}, \tag{7}$$

where $\#pixels(\hat{C}_r(I) - \hat{C}_{fs}(I))$ denotes the number of pixels in I which are estimated as new labels.

In Equation (1), the term $\beta(c,I)$ concerns the "rare label" issue. We define the "rareness" of a label c in inverse proportion to the number of pixels estimated as the label c, and assign a larger weight to the labels with less than 20% pixels in the whole image according to the Pareto principle [15]:

$$\beta(c) = \begin{cases} 2 - \dfrac{\#pixels(c)}{\max\limits_{\forall c \in \hat{C}_r(I)} \#pixels(c)}, & \text{if } \left(c \in \hat{C}_r(I)\right) \text{ and } \left(\dfrac{\#pixels(c)}{m \times n} \leq 0.2\right) \\ 1, & \text{otherwise} \end{cases}. \tag{8}$$

Note that, in Equation (1), we do not apply β to the supervised potential $\psi_{fs}(c,\boldsymbol{p})$. Because FCN-based parametric models tend to over-segment the objects larger than receptive fields, we may mistakenly assign the fragmentary labels as rare labels by Equation (8). Therefore, we only apply β to the label transfer potential.

Finally, we use the alpha-beta swap algorithm [21] to minimize the MRF energy function defined in Equation (1) and obtain the final labelling result.

3 Experiments

3.1 Datasets and settings

We use two datasets: SIFT Flow [12] and LMSun [13], as the real-world datasets to verify our method. SIFT Flow dataset contains 2,488 training images and 200 testing images, which are all outdoor scenes; all the images are of size 256 × 256 with 33 labels. LMSun is a large-scale dataset contains 45,176 training images and 500 testing images of indoor and outdoor scenes; the size of images ranges from 256 × 256 to 800 × 600 with 232 labels. We retrieve reference images from the 2,488 and 45,176 training images in SIFT Flow and LMSun, respectively; and use their corresponding testing images as the query. We set the number of retrieved images $K = 40$, the number of matched windows $\kappa = 5$, and smoothing constant $\lambda = 0.02$ for SIFT Flow; $K = 120$, $\kappa = 15$ and $\lambda = 0.01$ for LMSun.

When testing on SIFT Flow dataset, we use *FCN-8s-pascal* and *FCN-8s-pascal-context* as the parametric model in two independent experiments. The SIFT Flow dataset is composed of outdoor scenes with $|C_r| = 33$ labels, whereas the parametric model *FCN-8s-pascal* and *FCN-8s-pascal-context* are trained on Pascal VOC [20] and Pascal-Context datasets [22] with $|C_{fs}| = 21$ and 59 labels, respectively. Because images in Pascal VOC dataset are centred objects, *FCN-8s-pascal* performs well on labelling “things” (e.g., car, boat) but covers only 7 labels (i.e., $|C_{fs} \cap C_r| = 7$) out of the 33 real-world labels. The Pascal-Context dataset contains scene annotations; thus *FCN-8-pascal-context* is more consistent with the real-world data with $|C_{fs} \cap C_r| = 19$ overlapped labels.

When using LMSun as the real-world dataset, we adopt *FCN-16s-siftflow* and *FCN-32s-pascal-context* as the parametric model. The model *FCN-16s-siftflow* is trained on SIFT Flow dataset. Thus, both *FCN-16s-siftflow* and *FCN-32s-pascal-context* models are capable of annotating scene content of LMSun. However, because LMSun is a large-scale dataset with $|C_r| = 232$ labels, the two parametric models cover only 32 and 55 overlapped labels. The big challenge here is to infer the large unknown label set by the proposed method.

3.2 Evaluation of the proposed method

In Table 1, we show the per-pixel and per-class accuracy under various settings to investigate the effectiveness of our method. Per-pixel accuracy measures the percentage of correctly labelled pixels; and per-class accuracy measures the averaged per-pixel accuracy of all the labels. There is usually a trade-off between the two measurements, because a higher per-pixel accuracy shows the method performs well on most common labels, whereas a higher per-class accuracy reflects how well the method perform across all the labels (including rare labels).

From Table 1, the overall performance of using parametric model alone is rather poor, because many labels in the testing datasets are unknown to the off-line trained FCN model. Especially, *FCN-8s-pascal* achieves only 0.9% per-pixel and 4.6% per-class accuracies on SIFT Flow. There are two major reasons: only very few labels in the overlapped label set ($C_{fs} \cap C_r$) appear in the testing images; and the frequently-appeared labels (e.g., sky and sea) in the testing images are labelled as “background” in Pascal VOC. One notable exception is the case of *FCN-32s-pascal-context* on LMSun, whose 61.5% per-pixel accuracy is as good as state-of-the-art parametric methods. The possible reason is that,

Table 1: Labelling accuracy of different settings on SIFT Flow and LMSun.

Method	# labels	α	β	Per-pixel acc. (%)	Per-class acc. (%)
SIFT Flow dataset $\lvert C_r \rvert = 33$					
FCN-8s-pascal	$\lvert C_{fs} \rvert = 21$ $\lvert C_{fs} \cap C_r \rvert = 7$	-	-	0.9	4.6
Proposed method *FCN-8s-pascal*		0.5		31.4	33.5
		Adaptive Eq. (6)		78.5	48.9
		Adaptive Eq. (7)		**82.0**	48.9
		Adaptive Eq. (7)	✓	81.7	**50.0**
FCN-8s-pascal-context	$\lvert C_{fs} \rvert = 59$ $\lvert C_{fs} \cap C_r \rvert = 19$	-	-	71.0	29.1
Proposed method (*FCN-8s-pascal-context*)		0.5		77.6	36.9
		Adaptive Eq. (6)		76.0	37.4
		Adaptive Eq. (7)		**80.4**	46.7
		Adaptive Eq. (7)	✓	79.6	**47.6**
LMSun dataset $\lvert C_r \rvert = 232$					
FCN-16s-siftflow	$\lvert C_{fs} \rvert = 33$ $\lvert C_{fs} \cap C_r \rvert = 32$	-	-	53.1	5.2
Proposed method (*FCN-16s-siftflow*)		0.75		63.0	13.7
		Adaptive Eq. (6)		65.1	14.2
		Adaptive Eq. (7)		**65.3**	16.2
		Adaptive Eq. (7)	✓	64.5	**16.4**
FCN-32s-pascal-context	$\lvert C_{fs} \rvert = 59$ $\lvert C_{fs} \cap C_r \rvert = 55$	-	-	61.5	11.1
Proposed method (*FCN-32s-pascal-context*)		0.75		**66.8**	14.1
		Adaptive Eq. (6)		61.6	11.1
		Adaptive Eq. (7)		66.2	16.3
		Adaptive Eq. (7)	✓	65.4	**16.5**

since the PASCAL-Context [22] originally contains 59 most frequently-appeared labels, some of the common labels in the overlapped label set achieves high accuracy and dominates the per-pixel accuracy.

The proposed method, when using the fixed value for α (0.5 on SIFT Flow and 0.75 on LMSun) in Equation (1), improves the performance significantly by transferring labels from similar images in the real-word dataset. When we adopt the content-adaptive α, both Equation (6) and Equation (7) outperform the case of a fixed value. Moreover, the content-adaptive parameter defined by Equation (7) outperforms the other cases, because we adjust α by considering the number of pixels in the newly detected labels to compensate the presence of false-positive labels. Furthermore, we show that including the label-aware parameter β indeed addresses the rare label issue and improves the per-class accuracy.

Figure 3 and Figure 4 show some qualitative results. These results demonstrate how the proposed method corrects the fragmentary labels of FCN by incorporating the label transfer model. For example, in the first row of Figure 3(c), some of the "road" and "sky" pixels are mislabelled as "sea" and "mountain" by FCN; these mislabelled pixels are successfully corrected by the proposed method. In Figure 4, we use the last two examples to demonstrate that the rare labels "window" (in the third row) and "books" (in the last row) can be correctly predicted with the label-aware parameter β.

3.3 Comparison with existing methods

Table 2 shows the quantitative comparison of our method with existing methods testing on LMSun and SIFT Flow. Note that, all the parametric methods listed in Table 2 are trained

Table 2: Comparison with existing methods on SIFT Flow and LMSun.

Method		Per-pixel acc. (%)	Per-class acc. (%)
SIFT Flow dataset			
Nonparametric method	C. H. Ma et al. [11]	78.3	46.1
	F. Tung et al. [14]	77.1	41.1
	F. Tung et al. [24]	79.9	49.3
Parametric + nonparametric method	B. Shuai et al. [10]	80.1	39.7
	M. George [16]	81.7	50.1
Proposed method ($C_{fs} \neq C_r$)	*FCN-8s-pascal*	81.7	50.0
	FCN-8s-pascal-context	79.6	47.6
Parametric method	A. Sharma et al. [2]	75.5	**52.8**
	J. Long et al. [1]	**85.6**	50.1
Proposed method ($C_{fs} = C_r$)	*FCN-16s-siftflow*	85.2	52.0
LMSun dataset			
Nonparametric method	F. Tung et al. [24]	60.8	**19.3**
Parametric + nonparametric method	M. George [16]	61.2	16.0
	J. Yang et al. [15]	60.6	18.0
Proposed method	*FCN-16s-siftflow*	64.5	16.4
	FCN-32s-pascal-context	**65.4**	16.5

under fully supervision of the whole label sets (i.e., without unknown labels) and are expected to perform the best among all the methods. As to the three "parametric + nonparametric" methods [10, 15, 16], they combine parametric and nonparametric models using different strategies. [16] includes an ensemble learning to train three boosted decision tree (BDT) models; [15] conducts an additional off-line training process for rare-class examplers; and [10] combines a nonparametric energy term into the proposedparametric model. Because the off-line training process in [10, 15, 16] has involved all the labels in the testing dataset, there is no unknown labels to these methods. Instead, our proposed method involves only a subset of labels in FCN and has to infer all the unknown labels via the proposed MRF framework.

On SIFT Flow dataset, our method (using *FCN-8s-pascal*) is compatible with [16] even though the parametric model *FCN-8s-pascal* has 26 unknown labels and the label transfer model uses a smaller retrieval set with $K = 40$ ($K = 64$ in [16]). In order to fairly compare with parametric methods [1, 2], we conduct an additional experiment by assuming that all the real-world labels are known to the parametric model (i.e., $C_{fs} = C_r$). Under this fully-supervised scenario (i.e., using *FCN-16s-siftflow* as the parametric model), our method outperforms [2] about 9.7% per-pixel accuracy and outperforms [1] about 2% per-class accuracy. (The accuracy of [1] in Table 2 is conducted by the model released on Model Zoo website [23]). The result verifies that the proposed method effectively address the fragmentary and mislabel issue as well as the rare label concern. On the challenging LMSun dataset, even though the parametric models in our method are aware of only 32 and 55 labels out of the whole 232 labels, our method outperforms [15] and [16] in terms of per-pixel accuracy. Our method also outperforms nonparametric method [24], i.e. the journal version of [14], in terms of per-pixel accuracy.

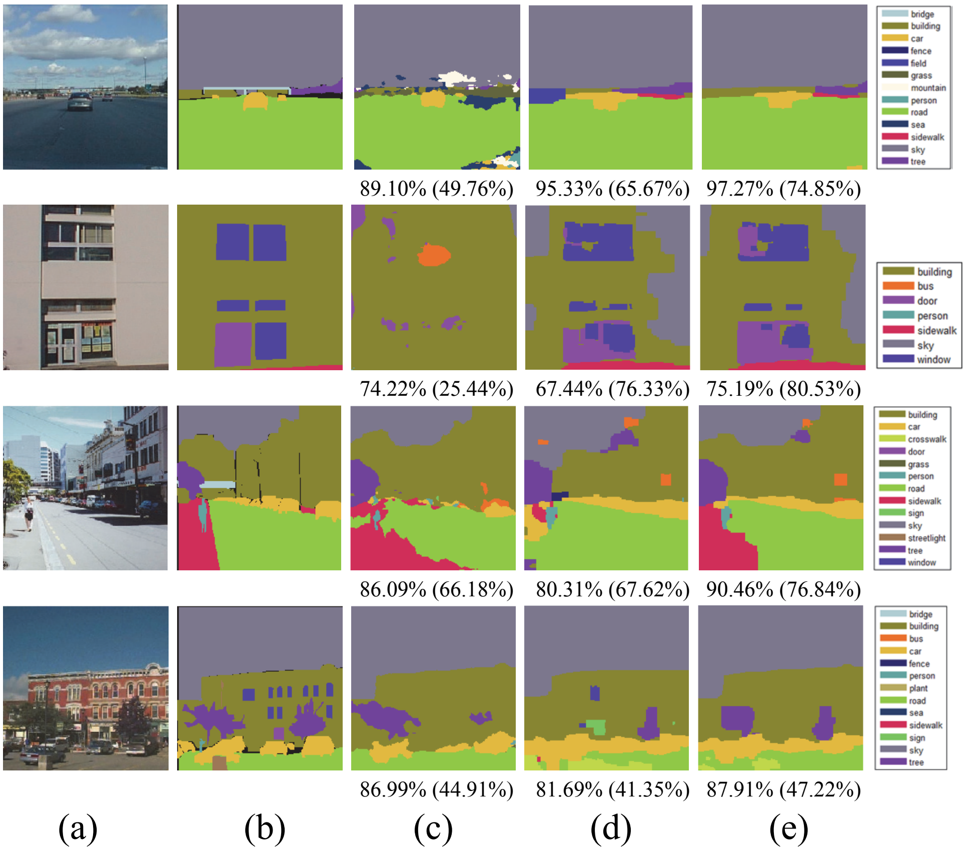


Figure 3: Examples of labelling results on SIFT Flow. The numbers given below each image are per-pixel and per-class accuracy (in brackets), respectively. (a) Query image; (b) ground truth; (c) results by *FCN-8s-pascal-context*; (d) results by the label transfer model; and (e) results of the proposed method.

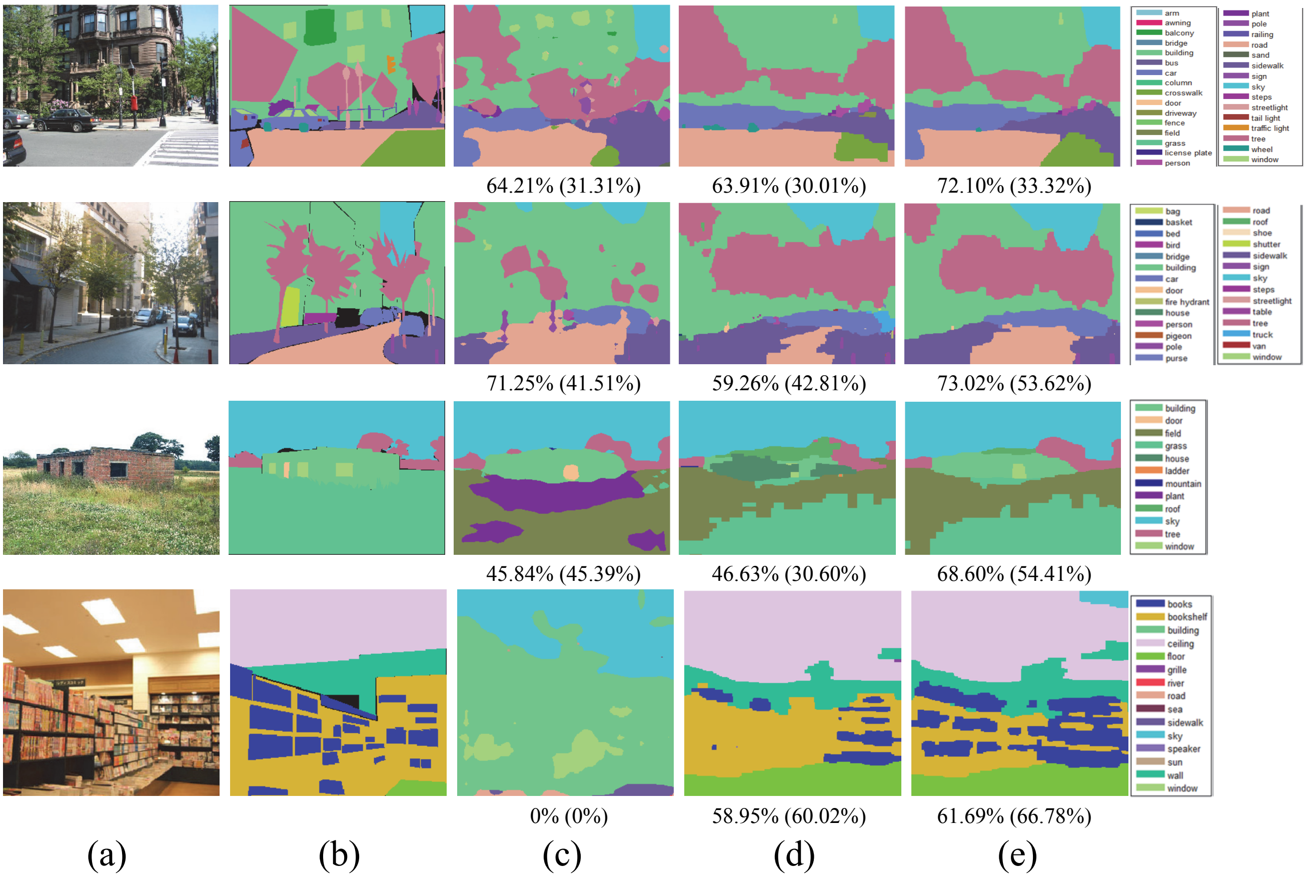


Figure 4: Examples of labelling results on LMSun. (a) Query image; (b) ground truth; (c) results by *FCN-16s-siftflow*; (d) results by the label transfer model; and (e) results of the proposed method.

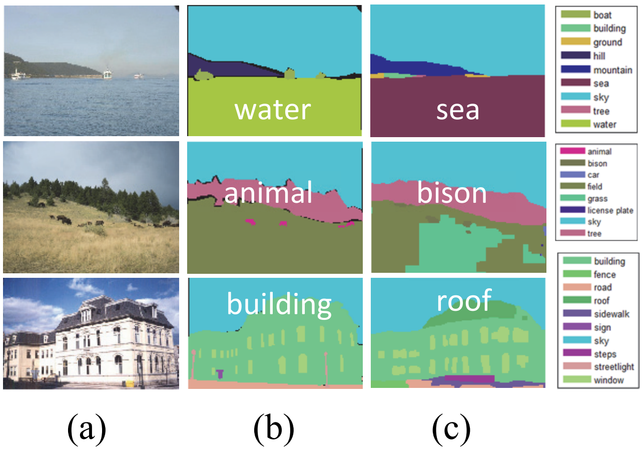


Figure 5: Examples of parent-child relationship between labels. (a) Query image; (b) ground truth; and (c) results of the proposed method.

Table 3: Efficiency evaluation on SIFT Flow and LMSun.

Dataset	Supervised potential (sec.)	Label transfer potential (sec.)	MRF framework (sec.)	Total (min.)
SIFT Flow	5.07	120.33	4.58	~ 2.2
LMSun	6.85	484.98	53.87	~ 9.1

4 Discussion

Figure 5 shows some examples where a category label is semantically divided into two or more sub-category labels. For example, “sea”, “bison” and “roof” predicted by our method actually belong to sub-category labels of “water”, “animal”, and “building” annotated in ground truth, respectively. However, the three cases in Figure 5 are treated as incorrect, because this parent-child label relationship is not yet included in the performance measurement. We believe further study on the semantic hierarchy may largely improve the accuracy of semantic segmentation.

Table 3 shows the execution time of our method. We conduct the supervised potential term (via FCN [1]), image global feature (via AlexNet), and windows detection (via Faster R-CNN) on a single GeForce GTX 780 Ti GPU. The label transfer potential and MRF framework are conducted on a CPU without code optimization. Inference on LMSun is more time-consuming than on SIFT Flow because it involves increased number of windows per image, larger window retrieval set size κ**,** and larger number of labels.

5 Conclusion

In this paper, we address the semantic segmentation problem for large scale real-world data. Instead of retraining a fully-supervised model, we propose to combine the partial information learned from FCN with the label information transferred from a real-world dataset. To adapt to dynamic content in real-world data, we design an MRF framework to adaptively combine the supervised potential and the label transfer potential. Moreover, we include a label-aware parameter to balance the priority of rare labels. Experimental results demonstrate that our method achieves better per-pixel accuracy and comparable per-class accuracy with state-of-the-art methods on the large-scale LMSun dataset, and achieves comparable overall performance with nonparametric methods on the SIFT Flow dataset.

References

[1] J. Long, E. Shelhamer and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[2] A. Sharma, O. Tuzel, and D. W. Jacobs. Deep hierarchical parsing for semantic segmentation. In *CVPR*, 2015.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015.

[4] H. Noh, S. Hong and B. Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015.

[5] X. Qi, J. Shi, S. Liu, R. Liao and J. Jia. Semantic segmentation with object clique potentials. In *ICCV*, 2015.

[6] G. Lin, C. Shen, A. van den Hengel and I. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *CVPR*, 2016.

[7] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR*, 2013.

[8] J. Tighe, M. Niethammer and S. Lazebnik. Scene parsing with object instances and occlusion ordering. In *CVPR*, 2014.

[9] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Trans. PAMI*, 35(8): 1915-1929, Aug. 2013.

[10] B. Shuai, G. Wang, Z. Zuo, B. Wang, and L. Zhao. Integrating parametric and non-parametric models for scene labeling. In *CVPR*, 2015.

[11] C. H. Ma, C. T. Hsu and B. Huet. Nonparametric scene parsing with deep convolutional features and dense alignment. In *ICIP*, 2015.

[12] C. Liu, J. Yuen and A. Torralba. Nonparametric scene parsing via label transfer. *IEEE Trans. PAMI*, 32(12): 2368-2382, Dec. 2011.

[13] J. Tighe and S. Lazebnik. Superparsing: Scalable nonparametric image parsing with superpixels. *IJCV*, 101(2): 329-349, Jan. 2013.

[14] F. Tung and J. J. Little. CollageParsing: Nonparametric scene parsing by adaptive overlapping windows. In *ECCV*, 2014.

[15] J. Yang, B. Price, S. Cohen and M. Yang. Context driven scene parsing with attention to rare classes. In *CVPR*, 2014.

[16] M. George. Image parsing with a wide range of classes and scene-level context. In *CVPR*, 2015.

[17] M. Rubinstein, C. Liu and W. T. Freeman. Joint inference in weakly-annotated image datasets via dense correspondence. *IJCV*, early access, 2016.

[18] A. Krizhevsky, I. Sutskever and G. Hinton. ImageNet classication with deep convolutional neural networks. In *NIPS*, 2012.

[19] S. Ren, K. He, R. Girshick and J. Sun. Faster R-CNN: Toward real-time object detection with region proposal networks. In *NIPS*, 2015.

[20] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88(2): 303-338, June 2010.

[21] Y. Boykov, O. Veksler and R. Zabin. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11): 1222-1239, Nov. 2001.

[22] R. Mottaghi, X. Chen, X. Liu, N. G. Cho, S. W. Lee, S. Fidler, R. Urtasun and Al Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

[23] Model Zoo. https://github.com/shelhamer/fcn.berkeleyvision.org

[24] F. Tung and J. J. Little. Scene parsing by nonparametric label transfer of content-adaptive windows. *CVIU*, 143: 191-200, Feb. 2016.