Integrating Parametric and Non-parametric Models For Scene Labeling

Bing Shuai, Gang Wang, Zhen Zuo, Bing Wang, Lifan Zhao
School of Electrical and Electronic Engineering, Nanyang Technological University.
50 Nanyang Avenue, Singapore, 639798.
{bshuai001,wanggang,zzuo1,wang0775,zhao0145}@ntu.edu.sg

Abstract

We adopt Convolutional Neural Networks (CNN) as our parametric model to learn discriminative features and classifier for local patch classification. As visual similar pixels are indistinguishable from limited context, we eliminate such ambiguity by introducing a global scene constraint. We estimate the global potential in a non-parametric framework. Furthermore, a large margin based CNN metric learning is proposed for better global potential estimation. The final pixel class prediction is performed by integrating local and global beliefs. Even without any post-processing, we achieve state-of-the-art on SiftFlow and competitive results on Stanford Background benchmark.

1. Introduction

Scene labeling builds a bridge towards better scene understanding. The goal is to relate one semantic class (road, water, sea, etc) to each pixel. Generally, “thing” pixels (car, person, etc) in real world images can be quite different due to their scale, illumination and pose variation, meanwhile “stuff” pixels are very similar (road, sea, etc) in a local close-up view. These issues pose scene labeling one of the most challenging problems in computer vision.

The recent advance of Convolutional Neural Networks (CNN) [13, 15] has dazzled computer vision community due to its outstanding performance ranging from large scale object recognition, detection [2, 13, 20, 32, 35] to pose estimation [27, 28]. It has also been demonstrated that this network is able to learn compact, discriminative and high-level features [34]. Recently, Farabet [4] and Pinheiro [21] has applied CNNs to scene labeling. In this scenario, CNNs are used to model the class likelihood of pixels from local context by parameterizing it in terms of features and classifiers. They can produce satisfactory labeling results by learning strong features and classifiers to discriminate visually dissimilar pixels.

However, CNNs struggle in visually similar pixels due to their limited context view. As shown in Figure 1, the sand pixels are highly confused with road and sidewalk pixels in a local view. Generally, previous works have addressed this issue from two perspectives:

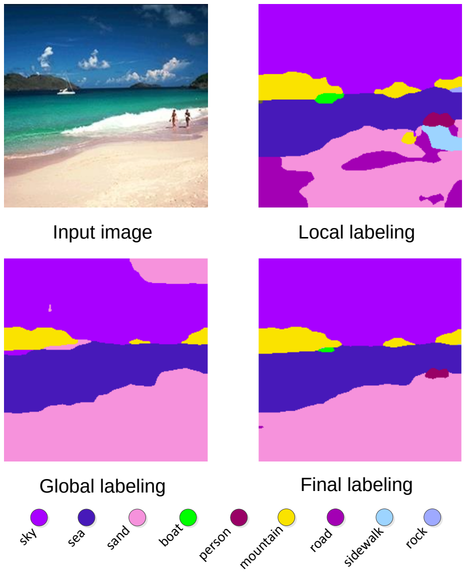

Figure 1. Motivation of our method: the parametric model can distinguish visually different pixels very well, but get confused for pixels that are visually similar in local context. However, the local features can be disambiguated from global scene semantics. A more consistent labeling result can be achieved by integrating their beliefs. The figure is best viewed in color.

- Augmenting the scale of context to represent pixels: [4] considers multi-scale context input, [21] naively increases the size of context input in a recurrent convolutional neural network. These methods somehow mitigate the local ambiguity, however they may have an negative effect for small objects and may also degrade the efficiency of the system.
- Building a graphical model to capture dependencies among pixels [9, 11, 23, 36]. However, the parametric

graphical model is usually hard and inefficient to optimize when the higher order potentials are involved, and the lower-order potentials suffer from low representation power.

Here, we propose to utilize global scene semantics to eliminate ambiguity of local context, for example, the confusion between 'road' and 'sand' pixels in Figure 1 can be easily removed if the "coast" scene is revealed. The global scene constraint is achieved by adding a global potential to the energy function. However, it's infeasible to model the global potential parametrically due to the extraordinarily huge labeling space, therefore we model it by transferring the class dependencies and priors from its similar exemplars. We further decouple the global potential to the aggregation of global beliefs over pixels. The final labeling result is obtained by integrating the local and global beliefs.

Furthermore, to make the estimation of global belief more accurate, a large margin based metric learning is introduced. We justify our method on the popular Stanford Background [6] and SiftFlow [16] benchmarks. Our integration model is able to achieve state-of-the-art result on SiftFlow and very competitive results on Stanford Background dataset. The contributions of this paper are summarized as follows:

1. We use global scene semantics to remove the ambiguity of local context by transferring class dependencies and priors from similar exemplars.
2. We demonstrate that our Convolutional Neural Network (CNN) can achieve significantly better results than other reported CNNs when they are fed with same local context.
3. We introduce a CNN metric learning approach, and show that the learned features and metrics are beneficial in our non-parametric global belief estimation.

For the rest of the paper, section 2 will review related works and compare theirs with ours and the formulation for our integration model is presented in Section 3. The details for each module of our approach are presented in section 4, and section 5 demonstrates the experimental evaluation of our method. Section 6 concludes and prospects our paper.

2. Related Work

Scene labeling (also termed as scene parsing, semantic segmentation) has attracted more and more attention these years. It serves as a bridge towards deeper scene understanding. Among all the related interesting works achieved so far, the directions how researchers approach this problem can be roughly grouped to three categories.

The first direction exploits extracting better unary features for classifying pixels/superpixels. This module is usually ignored by most researchers until recently when machine learning techniques are commonly used to learn discriminative features for various computer vision tasks [1, 4, 21, 30, 31, 38, 37]. Previously, low-level and mid-level hand-engineered features are designed to capture different image statistics, which usually lack discriminative power and suffer from high dimensionality, thus limiting the complexity of the full system. Recently, [4] bypassed this issue by feeding a convolutional neural network multi-scale raw data, and they have presented very interesting results on real-world image scene labeling datasets. Furthermore, [21] adopted a recurrent CNN to process the large size raw data. [1] learned a more compact random forest by substituting the random split function with a stronger Neural Network. They disambiguate the local context confusions via simply augmenting input context. In contrast, our CNN only takes limited context as input, thus forcing the network to learn strong features for local classification. The local context ambiguity is further eliminated by introducing global scene constraints.

Another line of works focuses on dependency modeling by formulating it as a structure learning problem. [23] formulated the unary and pairwise features in a 2nd-order sparse CRF graphical model. Later on, [22, 36] built a fully connected graph to enforce higher order labeling coherence. [11, 12, 18] modeled the higher order relations by considering patch/superpixel as a clique. [9] defined a multi-scale CRF that captures different contextual relationships ranging from local to global granularity. Our work is related to this batch of works, but approaches from a different angle. Their potentials are usually modeled parametrically, therefore extensive efforts are needed for learning these parameters. Our global potential doesn't require additional training effort and can be estimated very efficiently in a non-parametric framework.

Recently, non-parametric label transfer method [3, 16, 24, 25, 29, 33]has gained popularity due to its outstanding performance and scalability to increasing data. Their usual practice is to estimate the unit class likelihood from its nearest neighbors. The unit can be defined in pixel or superpixel level. Global scene information is firstly used to remove irrelevant images. A MRF is later employed to ensure neighborhood labeling coherence. The pioneering label transfer work [16] transformed RGB image to sift image, which was used to seek correspondences in pixel unit. Then, an energy function was defined over pixel correspondences, and by minimizing it can they obtain the labeling result. The Superparsing system [25] achieved a better result by performing label transfer over superpixel unit. [3] learned adaptive weights for each low-level features, which resulted in better nearest neighbor search. Gould[7, 8] built a graph for dense patch or superpixel to achieve label transfer. we adopt this framework to estimate our global potential. Compared with their hand-engineered features, we used the learned CNN features that are more compact and discriminative. We also believe that our features can further benefit their work in terms of accuracy and efficiency.

3. Formulation

The image labeling task is usually formulated as a clique based discrete energy minimization problem:

$$E(X,Y) = \sum_{c \in \mathcal{C}} \Phi(X, Y_c) \tag{1}$$

where $X = \{X_1, X_2, \ldots, X_N\}$ is the observed image and X_i corresponds to the ith pixel; $Y = \{Y_1, Y_2, \ldots, Y_N\}$, $Y \in \{1, 2, \ldots |L|\}^N$ denotes a labeling configuration; $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_M\}$ define the clique set, and $\Phi(X, Y_c)$ is the potential function for label assignment Y_c over clique c; Energy term associated to clique size 1 ($|\mathcal{C}_j| = 1$) is usually referred to unary energy term, which considers beliefs from local appearance cues. A more sensible and coherent labeling result can be achieved by adding pairwise or higher-order energy term, which are defined over clique size to be 2 or larger($|\mathcal{C}_j| \geq 2$).

Here in this paper, we consider clique size to be 1 and N. Therefore, our energy function can be written as:

$$E(X,Y) = \sum_{i \in X} \Phi_I(X_i, Y_j) + \Phi_G(X, Y) \tag{2}$$

where $\Phi_I(X_i, Y_j) = -P_I(X_i, Y_j)$ is the unary potential function defined as negative likelihood of pixel X_i being labeled as Y_j [1]; $\Phi_G(X, Y)$ is the global potential of image X taking labeling configuration Y. Since it's infeasible to model the huge labeling state of Y parametrically ($|L|^N$) , we adopt a similar non-parametric approach like [18] to model the global potential, which is defined as:

$$\Phi_G(X,Y) = -\sum_{i \in X} P_G^{\mathcal{S}(X)}(X_i, Y_j) \tag{3}$$

where $\mathcal{S}(X)$ is the similar exemplars of image X and $P_G^{\mathcal{S}(x)}(X_i, Y_j)$ is global class likelihood of X_i taking label Y_j. Even though the global potential is aggregated from independent beliefs, their dependencies have been implicitly modeled and transferred by $\mathcal{S}(X)$. For example, in Figure 1, the global exemplars define a "coast" scene, in which road and sidewalk pixels are invalid and sand pixels are more likely to appear in the bottom regions. By rewriting the energy functions, it gives us the following form:

$$E(X,Y) = -\sum_{i \in X} (P_I(X_i, Y_j) + P_G^{\mathcal{S}(X)}(X_i, Y_j)) \tag{4}$$

Therefore, the energy function can be interpreted as an integration of beliefs from two sources: (1), Local belief: $P_I(X_i, Y_j)$ measures the belief for local context centering on pixel X_i; (2), Global belief: $P_G^{\mathcal{S}(X)}(X_i, Y_j)$ denotes the belief for X_i from global scene view. Since the likelihood estimations of pixels is independent with each other, the inference can be done in a pixel-wise manner: $Y = \bigcup_{i=1:N} Y_i$, $Y_i = argmin_{1,\ldots,|L|} E(X_i, Y_j)$.

4. Approach

The framework of our method is depicted in Figure 2. The truncated CNN works as a feature extraction module. The local features are fed into two branches: (1), Local belief: they are independently classified based on the parametric CNN model; (2), Global belief: they are aggregated to generate the global feature, which are used to retrieve similar exemplars; the global belief is estimated based on them. Finally, the labeling result is generated by integrating local and global beliefs. We elaborate each module in the following sections.

4.1. Local Belief From Parametric CNN

The estimation of local belief $P_I(X_i, Y_j)$ is achieved by training a Convolutional Neural Network (CNN). The softmax layer outputs the class likelihood for pixel X_i. The structure of our CNN is demonstrated in Figure 2. Compared with other reported CNNs [4, 21], ours have three differences: (1), a location channel is appended to the RGB image, which denotes the normalized distance of each pixel to the image center. By doing this, the output pixel features are expected to carry spatial information. (2), our network is fed with small-size patch, rather than multi-scale or large-size contextual patches. (3), instead of using sigmoid or tanh activation function, we use the ReLU: $y(x) = max(0, x)$, which have been shown to converge faster during training in large scale object recognition task [13].The experiments will demonstrate that our CNN is able to achieve significantly better results than others when they are fed with same scale local context, while more efficient in terms of training and testing.

4.2. Global Belief From Non-parametric Estimation

The highly supervised trained CNN is capable of generating satisfactory results for the pixels with good local contextual support. However the predictions are quite random for pixels that are visually indistinguishable in local context. For example, waveless sea pixels can be confused with road pixels, thus an incorrect labeling configuration in which a sea region is surrounded by road pixels may occur.

These confusing pixels can be distinguished if neighborhood context is revealed to them. Previously, Farabet[4] fed the network with multi-scale patches to yield richer contextual aware local features, and likewise Pinheiro[21] took

[1] The negative log-likelihood can also be used when the global belief is not skewed towards frequent classes. The likelihood potential can be regarded as a non-linear transformation of log-likelihood to ameliorate the very small global belief, based on which MAP of Equation 4 is more robust.

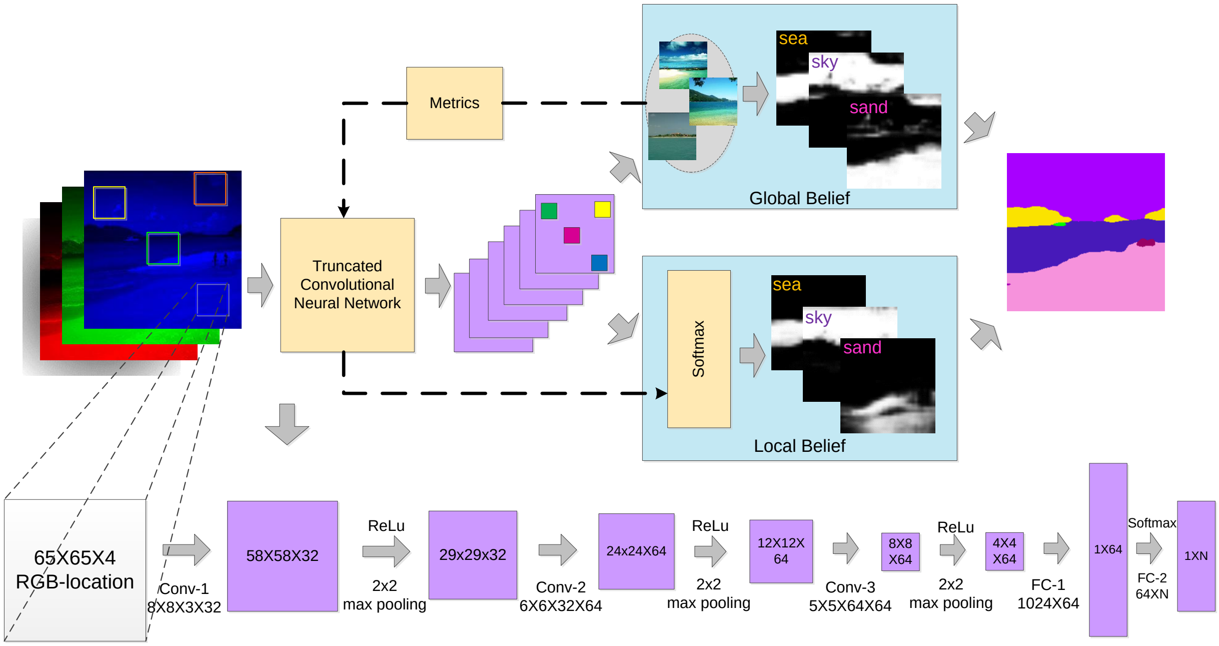

Figure 2. Framework of our approach: The final labeling is obtained by integrating local and global beliefs. The modules painted in yellow represents parametric part (CNN and Metrics).

network input as larger patch. In this paper, the disambiguation of local context is achieved by gleaning cues from pixels of the whole image. More specifically, a pixel is considered under global context: the class likelihood of the pixel should satisfy the scene layout and semantics of the image.

Given an image I, the corresponding CNN feature tensor [2] $F \in \mathbb{R}^{H\times W\times M}$ is obtained by passing I to the truncated CNN. The global feature H is aggregated from F, which is achieved by an average pooling operator *pool* [13]. Suppose an image is decomposed to regions [3] $\mathcal{R} = \{R^{(1)}, R^{(2)}, \ldots, R^{(N)}\}$. The region feature is pooled from the constituent pixel features: $H(R^{(i)}) = pool(F^i), \forall i \in R^{(i)}$. The global image representation is thus defined as concatenation of region features $H = [H(R^{(1)}), H(R^{(2)}), \ldots, H(R^{(N)})]$. As expected, this global image feature H not only conveys discriminative scene semantics but also encodes scene layout information.

The class likelihood of pixels (global belief) should match their scene semantics and layout, which is implicitly defined by its nearest exemplars $\mathcal{S}(X)$ in the global feature H space. Concretely, the global belief is transferred from the statistics of pixel features in $\mathcal{S}(X)$, and it is calculated in a weighted K-NN manner:

$$P_G^{\mathcal{S}(X)}(X_i, Y_j) = \frac{\sum_k \phi(X_i, X_k)\delta(Y(X_k) = Y_j)}{\sum_k \phi(X_i, X_k)} \tag{5}$$

where X_k is the k-th nearest neighbor of X_i among all the pixel features in $\mathcal{S}(X)$, $Y(X_k)$ is the ground truth label for pixel X_k; $\delta(Y(X_k), Y_j)$ is an indicator function; $\phi(X_i, X_k)$ measures the similarity between X_i and X_k, which is defined over spatial and feature space:

$$\phi(X_i, X_j) = exp(-\alpha||x_i - x_j||)exp(-\gamma||z_i - z_j||) \tag{6}$$

where $x_i = F(X_i)$ denotes the CNN pixel feature for X_i, z_i is the normalized coordinate along image height axis and α, γ controls the belief exponential falloff. Our non-parametric global belief estimation is reminiscent of popular label transfer works[3, 16, 24, 25, 29], two differences need to be highlighted:

- Instead of adopting hand-engineered low-level local and global features, we use more discriminative and compact features learned from CNN for label transfer.
- Our non-parametric model works as global scene constraints for local pixel features. Generally, small size retrieval images are sufficient to define the scene semantic and layout. However, previous works have to seek large retrieval set to cover all the possible semantic classes.

[2] The output of FC-1 layer in our CNN (Figure 2). $1 \leq h \leq H, 1 \leq w \leq W$ denotes the site location of the pixel feature, whose dimension is given by M.

[3] In our experiments, the image is divided into rectangular regions in a 2-layer spatial pyramid fashion [14].

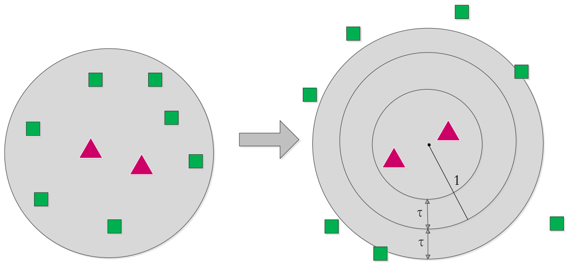

Figure 3. Illustration of our large margin idea. Before learning metric, the nearest neighbors of the testing feature (rectangle) are dominated by imposter classes (triangle) due to highly imbalance data distribution. After metric mapping, the imposters will stay far away from it, thus contributing little to the belief estimation.

4.3. CNN Metric Learning For Better Global Belief Estimation

As shown in Equation 5, the estimation of global belief $P_G(X_i, Y_j)$ is highly dependent on the distance metric between two pixel features. However, our features are learned by optimizing pixel/patch classification accuracy, while do not take distance metric into consideration. Therefore we propose to learn a large-margin based metric to mitigate the inaccurate class likelihood estimation for rare classes (Figure 3). In details, the Mahalanobis metric $M = W^T W$ is learned by minimizing the loss function, which is formally written as:

$$\begin{aligned} L &= \frac{\lambda}{2}||W||^2 + \frac{1}{2N}\sum_{i,j} g(x_i, x_j) \\ g(x_i, x_j) &= max(0, 1 - \ell_{i,j}(\tau - ||Wx_i - Wx_j||^2)) \\ x_i &= F(X_i) \end{aligned} \tag{7}$$

where $\ell_{i,j}$ indicates whether two features have the same semantic label or not, and $\ell_{i,j} = 1$ if X_i and X_j are from the same class, or $\ell_{i,j} = -1$ otherwise; $\tau(> 1)$ is the margin and λ controls the effect of regularization; $F(X_i)$ is the feature representation for X_i and N is the number of features. The objective function would enforce the pixel features from the same semantic class to be close and stay within the ball with radius $1 - \tau$, and enforce data from different classes to be far away from each other by at least $1 + \tau$ (Figure 3).

Instead of simply learning a metric based on the extracted CNN features, we further replace the softmax layer with our metric learning layer, so that the feature extraction parameters can also be adapted. We replace the softmax layer of previous CNN (CNN-softmax) with a fully connected layer parameterized by W (or more layers to learn nonlinear metrics [10]) and fix the biases to be zero, which serves as a Mahalanobis metric ($M = W^T W$). We call the new network CNN-metric. These two networks do not share any parameters except that the feature extraction parameters of CNN-metric are initiated from the corresponding layers of CNN-softmax. The errors are back propagated through the chain rule, and $\frac{\partial L}{\partial W}$ $\frac{\partial L}{\partial x_i}$ for the last layer are given in Equation 8.

```
Data: Images: $X^{(1)}, \ldots, X^{(N)}$;
Ground Truth: $Y^{(1)}, \ldots, Y^{(N)}$;
Result: CNN parameters $F$, metric $W$;
train CNN-Softmax;
while $iter \leq MAXITER$ do
    for $i$ = $1 \ldots N$ do
        $\mathcal{S}(X^{(i)})$ = NearestNeighbors($X_i$);
        $\mathcal{T}_i$ = RandomPixels($\mathcal{S}(X^{(i)})$);
        Fine tune CNN-Metric based on $\mathcal{T}_i$;
    end
end
```

Algorithm 1: CNN Training and fine tuning

$$\begin{aligned} \frac{\partial L}{\partial W} &= \lambda W + \frac{1}{N}\sum_{i,j} \zeta_{ij} \\ \zeta_{ij} &= g^{'}(c)\ell(i,j)(Wx_i - Wx_j)(x_i - x_j)^T \\ \frac{\partial L}{\partial x_i} &= \frac{1}{N}\sum_{i,j} g^{'}(c)(W^T\ell(i,j)(Wx_i - Wx_j)) \\ c &= 1 - \ell(i,j)(\tau - ||Wx_i - Wx_j||^2) \\ x_i &= F(X_i) \\ g^{'}(c) &= \begin{cases} 0, & c <= 0 \\ 1, & c > 0 \end{cases} \end{aligned} \tag{8}$$

Due to the large intra-class variations of pixel features(e.g. green and yellow tree pixels can be quite different), we only require that features from the same subcategory to be close. Since pixel features from the same semantic class in the nearest exemplars are usually quite similar, we implicitly divide original class space to finer-grained subclasses by only adapting their distance metrics. Stochastic gradient descent is adopted to fine tune the network based on the pixel features in the nearest exemplars. The whole training algorithm is summarized in Algorithm 1.

5. Experiments

5.1. Evaluation Datasets

We evaluate our approach on two benchmarks:

- Stanford-background [6]: It has 715 images from urban and natural scene composed of 8 semantic classes. Each image has around 320 × 240 pixels. We follow the training/testing protocol by randomly using 80% images as training, and the rest for testing. The results are reported under 5-fold cross validation.

	Stanford	SiftFlow
singlescale convnet (46×46) [4]	66.0%(56.5%)	-
Recurrent CNN (67×67) [21]	76.2%(67.2%)	65.5%(20.8%)
Ours (45×45)†	77.1% (68.0%)	73.5%(35.3%)
Ours (65×65)	**79.1%** (**70.1%**)	**75.1%** (**38.2%**)

† The network has the same structure as 65×65 CNN, except that the spatial dimensions for the first convolutional filter is 6×6.

Table 1. Performance comparison for different CNNs. The number following the networks indicates the size of input context. The percentage given outside and inside of parenthesis denotes pixel accuracy and class accuracy respectively.

- Sift Flow [16]: It has 2688 images generally captured from 8 typical natural scenes. Every image has 256 × 256 pixels, which belongs to one of 33 semantic classes. We use the training/testing (2488/200 images) split provided by [16] to conduct our experiments.

5.2. CNN Local Labeling Result

Stochastic Gradient Descent is adopted to train our CNN. For every epoch, we randomly sample 2×10^5 pixels from training pixel pools (approximate 80 million for Stanford background and Sift Flow). We start with learning rate of 0.001, and decrease by 10 times after 20 epoches. The momentum is set to 0.9. We take the batch size as 100 for each weight update iteration and the reported results are based on the model learned in 35 epoches. Each image is preprocessed by first subtracting the mean and then performing contrast normalization by dividing its variance. Unlike other CNNs that take days or even weeks for training [4] , our CNN only takes $3 \sim 4$ hours on a modern Telsa K40 GPU. The efficiency in training comes from the following aspects: (1), We use batch size 100, rather than mini-batch size 1; (2), we only randomly sample a small portion of patches to train the model for each epoch, other than considering all of them; we find that increasing the sampled patches from 2×10^5 to 5×10^5 during each epoch has a marginal effect. (3), we use ReLU function that has demonstrated faster convergence in large scale object recognition task [13].

Furthermore, a 'hybrid' sampling method is developed to mitigate the imbalance data distribution. A frequency threshold η is considered ($\eta = 0.01$ in our experiments). We require that all the class frequencies to be above η while still respecting their natural frequency distribution. Therefore, we firstly sample data naturally, and then augment data for classes whose frequencies are below η to make it reach the frequency threshold η.

Table 1 presents the result comparison for different CNNs. Our network can achieve significantly better results than other reported CNNs when they are fed with similar local context input.

	Dim	K=1	K=5	K=10
GIST[19]	512D	74.0%	70.7%	68.3%
SIFT-SPM[14]	2100D	76.5%	71.3%	69.1%
Our feature	**320D**	**90.5%**	**84.3%**	**82.1%**
GT	165D	94.0%	91.0%	89.5%

Table 2. Average genuine matching percentage in their K-nearest neighbors for different global features. GT is the semantic feature pooled from ground truth label map.

5.3. Evaluation of Discriminative Power for Global Features

The discriminative power of pixel CNN features has been presented in last section. Here, we demonstrate that the pooling operation for pixel features is able to generate semantic consistent global features. Specifically, The average genuine matching percentage in their K nearest neighbors is calculated: $p = \frac{\sum_i^N \sum_k^K \delta(I_i, NN(i,k))}{NK}$, where N is the number of test images, $NN(i,k)$ stands for the k-th nearest neighbor for image I_i, and $\delta(i,j)$ outputs value 1 if i and j are a genuine match, or 0 otherwise. A genuine matching pair means that they belong to the identical semantic class. We test the features in SiftFlow benchmark, as it provides scene labels for each image.

Four global features are compared in our experiment: GIST [19] is a global summary of scene images that captures scene structure and layout; SIFT-SPM (GT) [14] is pooled from low-level local SIFT [17] (ground truth label map [16]) in a 3(2)-layer spatial pyramid. They are are commonly used in scene classification and non-parametric label transfer framework. GT is the idea global semantic feature. Euclidean distance is used to retrieve nearest neighbors for non-histogram features (GIST, Ours), and histogram intersection distance is applied for the rest histogram features (SIFT-SPM and GT).

The quantitative genuine matching percentages for different global features are shown in Table 2. Our global feature pooled from CNN pixel features performs significantly better than its hand-engineered counterparts. The imperfect performance of GT features implies that different scenes can have very similar building blocks, for example, 'inside city' and 'street' scenes are dominated by sky and building pixels. We believe that the quality of nearest neighbors directly determines the correctness of global belief. Therefore, our global feature is expected to benefit other label transfer works. Figure 4 presents two nearest exemplars for each test image.

5.4. Non-parametric Global Labeling Result

We adopt non-overlapping patches as label transfer unit in the non-parametric model. In details, as our CNN has three subsampling layers (Figure 2), the dimension of the output feature map F is $\frac{1}{8}$ of original image size: one

[4] The author of [5] reports the training/testing time of different CNNs by personal communication to the original authors.

	Stanford	Sift Flow
SuperParsing[25]	77.5% (-)	76.9% (29.4%)
Liu[16]	-	74.8 % (-)
Gould[7]	73.9% (63.2%)	-
Ours	79.0% (69.0%)	78.0% (33.5%)
Ours+ metric tuning	**80.2**% (**69.9**%)	**78.2**% (**35.8**%)

Table 3. Performance for different label transfer approaches.

feature in F corresponds to a 8×8 image patch. We only consider estimating class likelihood for each feature in F. In this sense, the non-overlapping 8×8 patches work as the label transfer unit. Knowing that small-size object pixels make negligible contribution to the global scene semantics, we introduce an auxiliary transfer set to ameliorate the biased global potential. Specifically, the global belief estimation is based on the features from its similar exemplars and the auxiliary transfer set.[5]

For the parameters involved in the non-parametric model, they are quite robust. Referring to Equation 5, we set $|\mathcal{S}(X)|$ (size of nearest exemplar images) and K (size of nearest pixel/patch neighbors) to be 5 and 200 respectively. Both α and γ in Equation 6 are set to 5. To fine-tune CNN metric, 20 global nearest exemplars are retrieved for each training image and then 200 patches are randomly sampled among them to be a training batch. λ and τ in Equation 7 makes marginal difference to the performance, and they are fixed to 0.2 and 1.5 respectively. The learning rate is fixed to 10^{-5} and the reported results are obtained under the models learned in 25 epoches .

Table 3 clearly demonstrates that our method is able to attain very promising results that are comparable or even better than most complicated label transfer counterparts. We attribute the performance superiority to the highly discriminative CNN features that we adopt in the non-parametric framework. Moreover, as evidenced by Table 3, the learned metric is capable of further improving the quality of global belief, thus boosting the labeling accuracy. Some qualitative labeling results are presented in Figure 4.

5.5. Final Labeling Result

Our final labeling result is obtained by integrating local and global beliefs. The quantitative results are presented in Table 4. In comparison with other CNNs that are fed with richer context input, our integration model is able to yield significantly better results that are comparable to state-of-the-art. Some qualitative results are presented in Figure 4.

	Stanford	Sift Flow
Multiscale convnet[4]†	78.8% (**72.4**%)	-
Multiscale convnet[4]‡	-	67.9% (**45.9**%)
CNN (133×133)[21]	79.4% (69.5%)	76.5% (30.0%)
RCNN (133×133)[21]	80.2% (69.9%)	77.7% (29.8%)
[4]†+ CRF	**81.4**% (**76.0**%)	78.5% (29.4%)
[4]‡+ CRF	-	72.3% (**50.8**%)
Ours Final(65×65)	80.3% (70.9%)	**79.8**% (38.3%)
Ours Final(65×65, metric)	**81.2**% (71.3%)	**80.1**% (39.7%)
Gatta[5]	-	78.7% (32.1%)
Gould[8]	79.3% (69.4%)	78.4% (25.7%)
Tighe[26]	-	78.6 % (39.2%)
Singh[24]	-	79.2% (33.8%)

† Natural Sampling
‡ Class sampling

Table 4. Performance comparison with state-of-the-art.

	sky	tree	road	grass	water	building	mountain	object
CNN	90.3%	76.8%	90.1%	81.6%	61.0%	77.3%	17.4%	66.4%
Ours Final	**92.6%**	**78.7%**	**92.0%**	**84.5%**	**62.0%**	**80.1**%	13.4%	**67.0%**

Table 5. Per-class accuracy comparison for Stanford dataset.

As evidenced by Table 4, our integration model is capable of significantly boosting the qualitative results (global pixel accuracy) of CNN local labeling by introducing global scene constraint: 2.1% and 5% global pixel accuracy improvement for Stanford Background and Sift Flow benchmark respectively. As images in Sift Flow present obvious and distinct scene semantics, it's natural that confusions from local context be disambiguated. Moreover, our global belief can still remove some labeling errors for images in Stanford Background that exhibit huge variance.

We further investigate the effect of our global constraint on per-class accuracy. The detailed per-class accuracy for SiftFlow dataset is presented in Table 6. Our integration model boosts the accuracy significantly for frequent classes, while slightly washes away some rare "object" classes. In more details, the global potential is more helpful for classes which are more stable in positions, and large-size classes are preferred because the target classes to be included in the nearest exemplars. Two strategies are adopted to alleviate this issue: (1), As formulated in Equation 6, we only transfer spatial information on y-axis in a soft manner, thus the global prior is weakly position dependent. (2), we introduce an auxiliary set to mitigate the global belief bias towards frequent classes. In the end, our final integration model is able to improve the average class accuracy. As shown in Table 5, the same trend is also observed in Stanford Background dataset.

6. Conclusion

In this paper, we have shown that our parametric CNN model can generate significantly improved labeling results compared with other CNNs when they fed with limited context input. However, as the predictions for visual similar pixels are poor, we proposed to use global scene context

[5] The auxiliary set consists of features from classes whose frequency is lower than 0.01. The label transfer set for each image is augmented with infrequent classes from auxiliary transfer set until their numbers reach 100.

	awning	balcony	bird	boat	bridge	building	bus	car	crosswalk	door	fence	field	grass	mountain	person	plant	pole
Frequency	0.11	0.03	≈0	0.06	0.04	20.2	0.03	1.6	0.05	0.26	0.24	3.64	1.22	12.4	0.23	1.33	0.04
CNN	10.8	40.0	13.2	2.0	6.7	83.4	10.5	54.3	69.4	32.1	34.0	33.1	57.2	72.0	32.9	19.2	2.6
Ours Final	6.6	**40.4**	7.6	**2.3**	5.8	**89.5**	8.3	52.6	53.0	29.6	**37.1**	**38.8**	**71.2**	**78.2**	31.7	6.4	0

	river	road	rock	sand	sea	sidewalk	sign	sky	staircase	streetlight	sun	tree	window	Global	Average
Frequency	1.37	6.93	0.85	1.41	5.6	0.89	0.11	27.1	0.18	0.02	0.01	12.6	1.07	-	-
CNN	25.7	79.4	14.9	32.8	68.5	23.9	28.3	94.5	28.6	4.9	79.0	78.4	15.3	75.1	38.2
Ours Final	**45.5**	**85.7**	7.5	**37.3**	**77.1**	**46.6**	**31.4**	**95.7**	**32.1**	0.6	74.6	**84.7**	14.0	**80.1**	**39.7**

Table 6. Per-class accuracy comparison for SiftFlow dataset. All the numbers are scaled to percent range. The statistics for class frequency is obtained in test images and the frequent classes are highlighted in red.

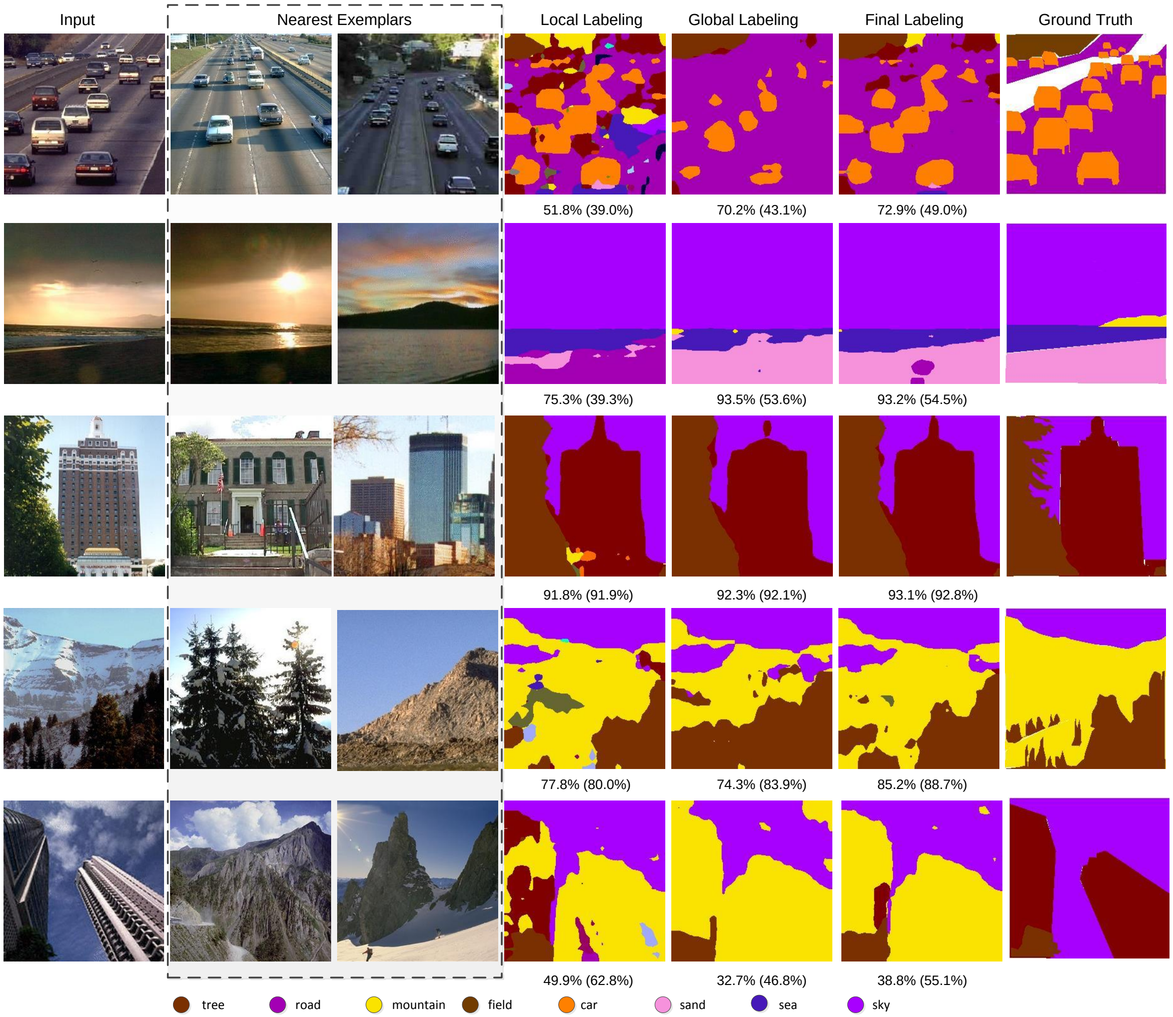

Figure 4. Qualitative result comparisons. The numbers given outside and inside of brackets represent pixel accuracy and average class accuracy respectively. The images shown in the dashed frame are two most similar exemplars. The last row shows an example, where the ambiguity of local features cannot be eliminated when their aggregated global features fail to reveal the true scene semantics.

to eliminate the local ambiguity. The global belief was estimated in a non-parametric framework, which transferred label dependencies from similar exemplars. Our final labeling is achieved by integrating local and global beliefs, and it achieved very competitive results on two real world scene labeling benchmarks. We may explore how to better apply discriminative CNN features to the current successful non-parametric models in the future.

Acknowledgements: The authors would like to thank Xingxing Wang, Rahul Rama Varior for fruitful discussions. The research is supported by Singapore Ministry of Education (MOE) Tier 2 ARC28/14, and Singapore A*STAR Science and Engineering Research Council PSF1321202099. We also would like to thank Nvidia for their generous donation of GPU.

References

[1] S. R. Bulo and P. Kontschieder. Neural decision forests for semantic image labelling. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014.

[2] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of The 31st International Conference on Machine Learning*, pages 647–655, 2014.

[3] D. Eigen and R. Fergus. Nonparametric image parsing using adaptive neighbor sets. In *Computer vision and pattern recognition (CVPR), 2012 IEEE Conference on*, pages 2799–2806. IEEE, 2012.

[4] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1915–1929, 2013.

[5] C. Gatta, A. Romero, and J. van de Veijer. Unrolling loopy top-down semantic feedback in convolutional deep networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014.

[6] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1–8. IEEE, 2009.

[7] S. Gould and Y. Zhang. Patchmatchgraph: Building a graph of dense patch correspondences for label transfer. In *Computer Vision–ECCV 2012*, pages 439–452. Springer, 2012.

[8] S. Gould, J. Zhao, X. He, and Y. Zhang. Superpixel graph label transfer with learned distance metric. In *Computer Vision–ECCV 2014*, pages 632–647. Springer, 2014.

[9] X. He, R. S. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on*, volume 2, pages II–695. IEEE, 2004.

[10] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1875–1882. IEEE, 2014.

[11] P. Kohli, P. H. Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009.

[12] P. Kontschieder, S. R. Bulo, H. Bischof, and M. Pelillo. Structured class-labels in random forests for semantic image labelling. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2190–2197. IEEE, 2011.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[15] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[16] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1972–1979. IEEE, 2009.

[17] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[18] P. Márquez-Neila, P. Kohli, C. Rother, and L. Baumela. Non-parametric higher-order random fields for image segmentation. In *Computer Vision–ECCV 2014*, pages 269–284. Springer, 2014.

[19] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[20] M. Oquab, L. Bottou, I. Laptev, J. Sivic, et al. Learning and transferring mid-level image representations using convolutional neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014.

[21] P. Pinheiro and R. Collobert. Recurrent convolutional neural networks for scene labeling. In *Proceedings of The 31st International Conference on Machine Learning*, pages 82–90, 2014.

[22] A. Roy and S. Todorovic. Scene labeling using beam search under mutex constraints. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014.

[23] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Computer Vision–ECCV 2006*, pages 1–15. Springer, 2006.

[24] G. Singh and J. Kosecka. Nonparametric scene parsing with adaptive feature relevance and semantic context. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3151–3157. IEEE, 2013.

[25] J. Tighe and S. Lazebnik. Superparsing: scalable nonparametric image parsing with superpixels. In *Computer Vision–ECCV 2010*, pages 352–365. Springer, 2010.

[26] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3001–3008. IEEE, 2013.

[27] J. Tompson, A. Jain, Y. Lecun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.

[28] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. *arXiv preprint arXiv:1312.4659*, 2013.

[29] F. Tung and J. J. Little. Collageparsing: Nonparametric scene parsing by adaptive overlapping windows. In *Computer Vision–ECCV 2014*, pages 511–525. Springer, 2014.

[30] A. Wang, J. Lu, G. Wang, J. Cai, and T.-J. Cham. Multimodal unsupervised feature learning for rgb-d scene labeling. In *Computer Vision–ECCV 2014*, pages 453–467. Springer, 2014.

[31] L. Wang, T. Liu, G. Wang, K. L. Chan, and Q. Yang. Video tracking using learned hierarchical features. *Image Processing, IEEE Transactions on*, 24(4):1424–1435, 2015.

[32] L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014.

[33] J. Yang, B. Price, S. Cohen, and M.-H. Yang. Context driven scene parsing with attention to rare classes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014.

[34] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pages 818–833. Springer, 2014.

[35] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. Panda: Pose aligned networks for deep attribute modeling. *arXiv preprint arXiv:1311.5591*, 2013.

[36] Y. Zhang and T. Chen. Efficient inference for fully-connected crfs with stationarity. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 582–589. IEEE, 2012.

[37] Z. Zuo, G. Wang, B. Shuai, L. Zhao, and Q. Yang. Exemplar based deep discriminative and shareable feature learning for scene image classification. *Pattern Recognition*, 2015.

[38] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, and X. Jiang. Learning discriminative and shareable features for scene classification. In *Computer Vision–ECCV 2014*, pages 552–568. Springer, 2014.