Multi-Scale Context Intertwining for Semantic Segmentation

Di Lin[1], Yuanfeng Ji[1], Dani Lischinski[2], Daniel Cohen-Or[1,3], and Hui Huang[1]*

[1]Shenzhen University [2]The Hebrew University of Jerusalem [3]Tel Aviv University
{ande.lin1988,jyuanfeng8,danix3d,cohenor,hhzhiyan}@gmail.com

Abstract. Accurate semantic image segmentation requires the joint consideration of local appearance, semantic information, and global scene context. In today's age of pre-trained deep networks and their powerful convolutional features, state-of-the-art semantic segmentation approaches differ mostly in how they choose to combine together these different kinds of information. In this work, we propose a novel scheme for aggregating features from different scales, which we refer to as *Multi-Scale Context Intertwining* (MSCI). In contrast to previous approaches, which typically propagate information between scales in a one-directional manner, we merge pairs of feature maps in a bidirectional and recurrent fashion, via connections between two LSTM chains. By training the parameters of the LSTM units on the segmentation task, the above approach learns how to extract powerful and effective features for pixel-level semantic segmentation, which are then combined hierarchically. Furthermore, rather than using fixed information propagation routes, we subdivide images into super-pixels, and use the spatial relationship between them in order to perform image-adapted context aggregation. Our extensive evaluation on public benchmarks indicates that all of the aforementioned components of our approach increase the effectiveness of information propagation throughout the network, and significantly improve its eventual segmentation accuracy.

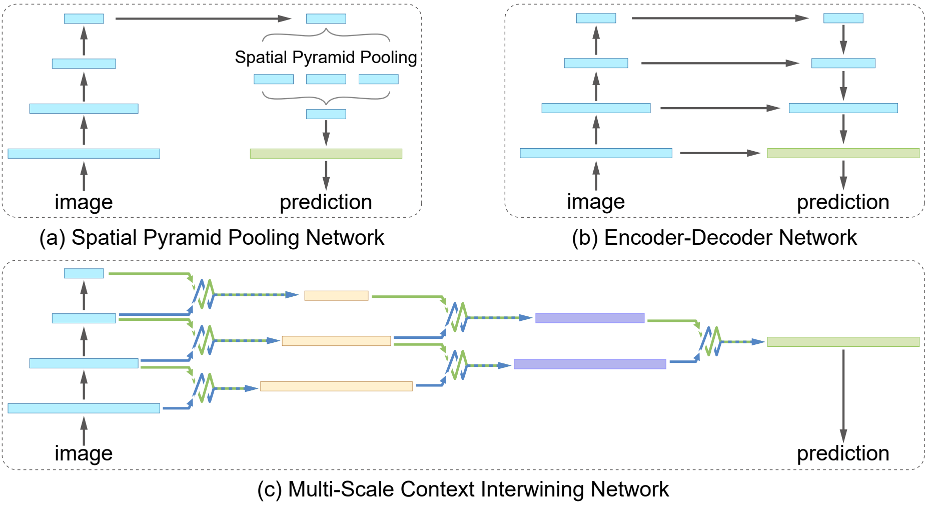
Keywords: Semantic Segmentation, Deep Learning, Convolutional Neural Network, Long Short-Term Memory

1 Introduction

Semantic segmentation is a fundamental task in computer vision, whose goal is to associate a semantic object category with each pixel in an image [1–4]. Many real-world applications, e.g., autonomous driving [4], medical analysis [5], and computational photography [6], can benefit from accurate semantic segmentation that provides detailed information about the content of an image.

In recent years, we have witnessed a tremendous progress in semantic segmentation accuracy. These advances are largely driven by the power of fully convolutional networks (FCNs) [7] and their derivatives [8, 9], which are pre-trained

[*] Hui Huang is the corresponding author of this paper.



(a) Spatial Pyramid Pooling Network

(b) Encoder-Decoder Network

(c) Multi-Scale Context Interwining Network

Fig. 1: Alternative approaches for encoding multi-scale context information into segmentation features for per-pixel prediction. The spatial pyramid pooling (SPP) network (a) and the encoder-decoder (ED) network (b) propagate information across the hierarchy in a one-directional fashion. In contrast, our multi-scale context intertwining architecture (c) exchanges information between adjacent scales in a bidirectional fashion, and hierarchically combines the resulting feature maps. Figure 2 provides a more detailed illustration of the multi-stage recurrent context intertwining process.

on large-scale datasets [10, 2]. It has also become apparent that accounting for the semantic context leads to more accurate segmentation of individual objects [11–14, 9, 15–19].

The feature maps extracted by the deeper layers of a convolutional network encode higher-level semantic information and context contained in the large receptive field of each neuron. In contrast, the shallower layers encode appearance and location. State-of-the-art semantic segmentation approaches propagate coarse semantic context information back to the shallow layers, yielding richer features, and more accurate segmentations [7, 9, 17–21]. However, in these methods, context is typically propagated along the feature hierarchy in a one-directional manner, as illustrated in Figure 1(a) and (b).

In this paper, we advocate the idea that more powerful features can be learned by enabling context to be exchanged between scales in a bidirectional manner. We refer to such information exchange as *context intertwining*. The intuition here is that semantics and context of adjacent scales are strongly correlated, and hence the descriptive power of the features may be significantly enhanced by such intertwining, leading to more precise semantic labeling.

Our approach is illustrated by the diagram in Figure 1(c). Starting from a collection of multi-scale convolutional feature maps, each pair of successive feature maps is intertwined together to yield a new enriched feature map. The intertwining is modeled using two chains of long short-term memory (LSTM) units [22], which repeatedly exchange information between them, in a bidirectional fashion, as shown in Figure 2. Each intertwining phase reduces the number of feature maps by one, resulting in a *hierarchical feature combination* scheme (the horizontal hierarchy in Figure 1(c)). Eventually, a single enriched high-resolution feature map remains, which is then used for per-pixel semantic label inference.

Furthermore, rather than using fixed information propagation routes for context aggregation, we subdivide images into super-pixels, and use the spatial relationship between the super-pixel in order to define *image-adapted feature connections.*

We demonstrate the effectiveness of our approach by evaluating it and comparing it to an array of state-of-the-art semantic segmentation methods on four public datasets (PASCAL VOC 2012 [1], PASCAL-Context [3], NYUDv2 [23] and SUN-RGBD [24] datasets). On the PASCAL VOC 2012 validation set, we outperform the state-of-the-art (with 85.1% mean IoU). On the PASCAL VOC 2012 test set, our performance (87.0% mean IoU) is second only to the recent result of Chen et al. [25], who uses a backbone network trained on an internal JFT dataset [26–28], while our backbone network is trained on the ImageNet dataset [10].

2 Related Work

Fully convolutional networks (FCNs) [7] have proved effective for semantic image segmentation by leveraging the powerful convolutional features of classification networks [29, 30, 27] pre-trained on large-scale data [10, 24]. The feature maps extracted by the different convolutional layers have progressively coarser spatial resolutions, and their neurons correspond to progressively larger receptive fields in the image space. Thus, the collection of feature maps of different resolutions encodes multi-scale context information. Semantic segmentation methods have been trying to exploit this multi-scale context information for accurate segmentation. In this paper, we focus on two aspects, i.e., *Feature Combination* and *Feature Connection*, which have also been explored by most of the recent works [7, 9, 18–20, 31–33] to make better use of the image context.

Feature Combination To capture the multi-scale context information in the segmentation features, many works combine feature maps whose neurons have different receptive fields. Various schemes for the combination of feature maps have been proposed. Spatial pyramid pooling (SPP) [34] has been successfully applied for combining different convolutional feature maps [9, 18, 20]. Generally, the last convolutional feature map, which is fed to the pixel-wise classifier, is equipped with an SPP (see Figure 1(a)). But the SPP-enriched feature maps have little detailed information that is missed by the down-sampling operations of an FCN. Though the atrous convolution can preserve the resolutions of feature maps for more details, it requires a large budget of GPU storage for computation [29, 30, 27]. To save the GPU memory and improve the segmentation performance, some networks [35, 17, 19, 21] utilize an Encoder-Decoder (ED) network to gradually combine adjacent feature maps along the top-down hierarchy of a common FCN architecture, propagating the semantic information from the low-resolution feature maps to the high-resolution feature maps and using the high-resolution feature maps to recover the details of objects (see Figure 1(b)). The latest work [25] further uses the ED network along with an atrous spatial pyramid pooling (ASPP) [20], and combines multi-resolution feature maps for information enrichment. In the ED network, each feature map of the decoder part only directly receives the information from the feature map at the same level of the encoder part. But the strongly-correlated semantic information, which is provided by the adjacent lower-resolution feature map of the encoder part, has to pass through additional intermediate layers to reach the same decoder layer, which may result in information decay.

In contrast, our approach directly combines pairs of adjacent feature maps in the deep network hierarchy. It creates new feature maps that directly receive the semantic information and context from a lower-resolution feature map and the improved spatial detail from a higher-resolution feature map. In addition, in our architecture the information exchange between feature maps is recurrent and bidirectional, enabling better feature learning. The pairwise bidirectional connections produce a second, *horizontal* hierarchy of the resulting feature maps, leading up to a full resolution context-enriched feature map (rightmost feature map in Figure 1(c)), which is used for pixel-wise label prediction.

Feature Connection Connections between feature maps enable the communication between neurons with different receptive field sizes, yielding new feature maps that encode multi-scale context information. Basically, FCN-based models [7–9, 17–20, 31] use separate neurons to represent the regular regions in an image. Normally, they use convolutional/pooling kernels with predefined shapes to aggregate the information of adjacent neurons, and propagate this information to the neurons of other feature maps. But traditional convolutional/pooling kernels only capture the context information in a local scale. To leverage richer context information, graphical models are integrated with FCNs [12, 13, 16]. Graphical models build dense connections between feature maps, allowing neurons to be more sensitive to the global image content that is critical for learning good segmentation features. Note that previous works use one-way connections that extract context information from the feature maps separately, which is eventually combined. Thus, the learned features at a given scale are not given the opportunity to optimally account for the multi-scale context information from all of the other scales.

In contrast to previous methods, our bidirectional connections exchange multi-scale context information to improve the learning of all features. We employ super-pixels computed based on the image structure, and use the relationship between them to define the exchange routes between neurons in different feature maps. This enables more adaptive context information propagation. Several previous works [31–33, 36] also use super-pixels to define the feature connections. And information exchange has been studied in [37, 38] for object detection. But

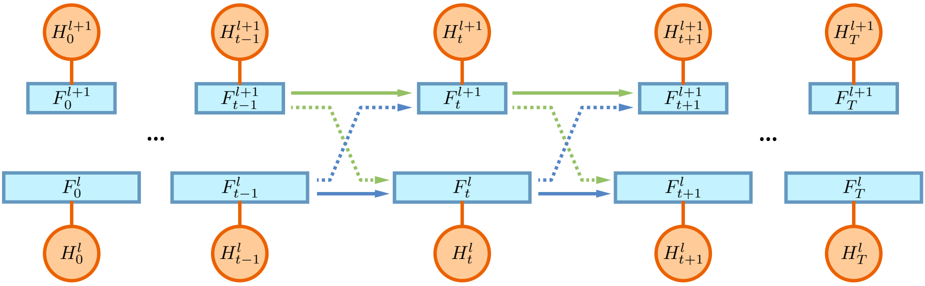


Fig. 2: Multi-scale context intertwining between two successive feature maps in the deep hierarchy. The green arrows propagate the context information from the lower-resolution feature map to the higher-resolution one. Conversely, the blue arrows forward information from the higher-resolution feature map to augment the lower-resolution one. The orange circle in each stage indicates the hidden features output by LSTMs, including the cell states and gates.

these works do not exchange the information between feature maps of different resolutions, which is critical for semantic segmentation.

3 Multi-Scale Context Intertwining

To utilize multi-scale context information, the common networks use one-way connections to combine feature maps of different resolution, following the top-down order of the network hierarchy (see Figure 1(a) and (b)). Here, we present a multi-scale context intertwining (MSCI) architecture, where the context information can be propagated along different dimensions. The first dimension is along the vertical deep hierarchy (see Figure 1(c)): our context intertwining scheme has connections to exchange the multi-scale context information between the adjacent feature maps. The connection is bidirectional with two different long short-term memory (LSTM) chains [22] that intertwines feature maps of different resolution in a sequence of stages. By training the LSTM units, the bidirectional connections learn to produce more powerful feature maps. The second dimension is along the horizontal hierarchy: the feature maps produced by our bidirectional connections are fed to the next phase of context intertwining, which can encode the context information memorized by our bidirectional connections into the new feature maps.

The overall MSCI architecture is illustrated in Figure 1(c). Initially, we use the backbone FCN to compute a set $\{F^l\}$ convolutional feature maps of different resolutions, where $l = 1, ..., L$ and F^1 has the highest resolution. Figure 2 provides a more detailed view of context intertwining between two successive feature maps F^l and F^{l+1}. To exchange the context information between F^l and F^{l+1}, we construct a bidirectional connection $\mathcal{L}$:

$$\{Q^l, C_T^{l\to l+1}, C_T^{l+1\to l}\} = \mathcal{L}(F^l, F^{l+1}, C^{l\to l+1}, C^{l+1\to l}, P^{l\to l+1}, P^{l+1\to l}, T). \quad (1)$$

Algorithm 1 Multi-Scale Context Intertwining

```
Input:
  1) the number of stages $T$ for each phase of the context intertwining;
  2) a set of convolutional feature maps $F = \{F^l\}$, where $l = 1, ..., L$;
  3) the trained parameter set $\{(P^{l \to l+1}, P^{l+1 \to l})\}$.
Initialization:
  1) a total $K$ phases for the context intertwining, where $K = L - 1$;
  2) a set $Q = \{Q_k\}$, where $Q_0 = \{Q_0^l | Q_0^l = F^l\}$; and $Q_k = \emptyset, k = 1, ..., K$;
  3) a set of cell states $\{(C^{l \to l+1}, C^{l+1 \to l})\}$, where $C^{l \to l+1}, C^{l+1 \to l} = 0$.
for $k = 1 \to K$ do
  for $l = 1 \to |Q_{k-1}| - 1$ do
    $\{Q_k^l, C_T^{l \to l+1}, C_T^{l+1 \to l}\} = \mathcal{L}(Q_{k-1}^l, Q_{k-1}^{l+1}, C^{l \to l+1}, C^{l+1 \to l}, P^{l \to l+1}, P^{l+1 \to l}, T)$
    $Q_k = Q_k \cup \{Q_k^l\}$, $(C^{l \to l+1}, C^{l+1 \to l}) = (C_T^{l \to l+1}, C_T^{l+1 \to l})$
  end for
end for
Output: the segmentation feature map $Q_K^1$.
```

The bidirectional connection $\mathcal{L}$ consists of two different LSTM chains. One chain has the parameter set $P^{l \to l+1}$. It extracts the context information from F^l and passes it to F^{l+1}. The other chain has the parameter set $P^{l+1 \to l}$ and passes context information from F^{l+1} to F^l. $C^{l \to l+1}$ and $C^{l+1 \to l}$ are the cell states of the two LSTMs, and they are initialized to zeros in the very beginning. As shown in Figure 2, the information exchange takes place over T stages. At each stage t, information is exchanged between the feature maps F_t^l and F_t^{l+1}, yielding the maps F_{t+1}^l and F_{t+1}^{l+1}. Note that the resulting feature map F_T^l has higher resolution than F_T^{l+1}. Thus, we deconvolve the feature map F_T^{l+1} with the kernel D_f^{l+1} and add it to F_T^l to obtain a combined high-resolution feature map Q^l:

$$Q^l = F_T^l + D_f^{l+1} * F_T^{l+1}. \tag{2}$$

Note that the feature map Q^l and the cell states $C_T^{l \to l+1}$ and $C_T^{l+1 \to l}$ can be further employed to drive the next phase of context intertwining (the next level of the horizontal hierarchy). Along the LSTM chains, the feature maps contain neurons with larger receptive fields, i.e., with richer global context. Besides, the cell states of LSTMs can memorize the context information exchanged at different stages. Due to the shortcut design of the cell states [22], the local context from the early stages can be easily propagated to the last stage, encoding the multi-scale context including the local and global information to the final feature map.

The entire MSCI process is summarized in Algorithm 1. We assume the MSCI process has K phases totally. Each phase of Algorithm 1 produces new feature maps. As each pair of feature maps is intertwined, the corresponding cell states $(C^{l \to l+1}, C^{l+1 \to l})$ are iteratively updated to provide the memorized context to assist the information exchange in the next phase. Finally, the output is the high-resolution feature map Q_K^1 that is fed to the pixel-wise classifier for segmentation. Algorithm 1 describes the feed-forward pass through the

LSTMs. We remark that the LSTM parameters are reusable, and the LSTMs are trained using the standard stochastic gradient descent (SGD) algorithm with back-propagation. Below, we focus on a single context intertwining phase, and thus omit the subscript k to simplify notation.

4 Bidirectional Connection

In this section, we describe in more detail the bidirectional connections that enable mutual exchange of context information between low- and high-resolution feature maps. Our bidirectional connections are guided by the super-pixel structure of the original image, as illustrated in Figure 3. Given an input image I, we divide it into non-overlapping super-pixels, which correspond to a set of regions $\{S_n\}$. Let F_t^l and F_t^{l+1} denote two adjacent resolution feature maps in our network, where l is the resolution level and t is the LSTM stage. The context information exchange between F_t^l and F_t^{l+1} is conducted using the regions defined by the super-pixels. Informally, at each of the two levels, for each region S_n we first aggregate the neurons whose receptive fields are centered inside S_n. Next, we sum together the aggregated features of S_n and all of its neighboring regions at one level and pass the resulting context information to the neurons of the other level that reside in region S_n. This is done in both directions, as shown in Figure 3(a) and 3(b). Thus, we enrich the locally aggregated context information of each neuron with that of its counterpart in the other level, as well as with the more global context aggregated from the surrounding regions. Our results show that this significantly improves segmentation accuracy.

Formally, given the feature map F_t^l and a region S_n, we first aggregate the neurons in S_n, yielding a regional context feature $R_{n,t}^l \in \mathbb{R}^C$:

$$R_{n,t}^l = \sum_{(h,w)\in\Phi(S_n)} F_t^l(h,w), \tag{3}$$

where $\Phi(S_n)$ denotes the set of centers of the receptive fields inside the region S_n. Next, we define a more global context feature $M_{n,t}^l$, by aggregating the regional features of S_n and of its adjacent regions $\mathcal{N}(S_n)$:

$$M_{n,t}^l = \sum_{S_m\in\mathcal{N}(S_n)} R_{m,t}^l. \tag{4}$$

The above features are propagated bidirectionally between F_t^l and F_t^{l+1} using a pair of LSTM chains, as illustrated in Figure 2. In the t^{th} stage, an LSTM unit

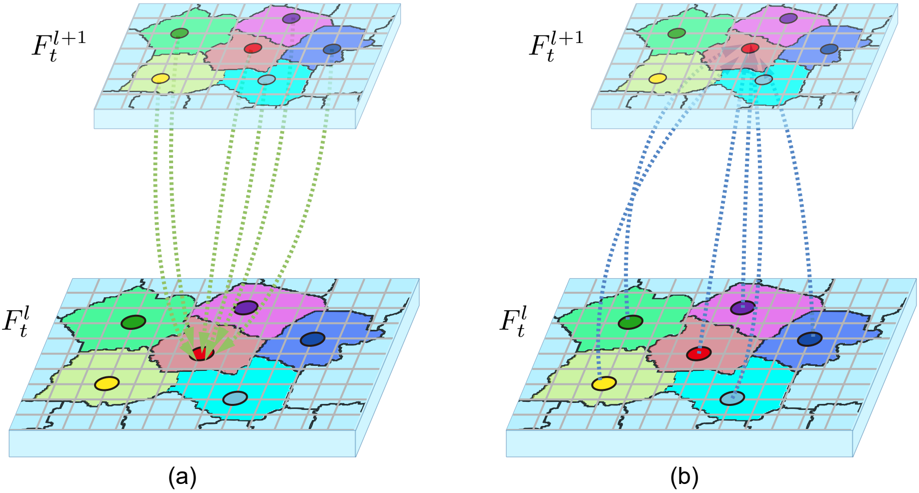


Fig. 3: Bidirectional context aggregation. The features are partitioned into different regions defined by super-pixels. We aggregate the neurons resided in the same region, and pass the information of the adjacent regions along the bidirectional connection (a) from a low-resolution feature to a high-resolution feature; and (b) from a high-resolution feature to a low-resolution feature.

generates a new feature F_{t+1}^{l+1} from F_t^{l+1}, $R_{n,t}^l$, and $M_{n,t}^l$, as follows:

$$\begin{aligned}
G_{i,t}^{l\to l+1}(h,w) &= \sigma(W_i^{l+1} * F_t^{l+1}(h,w) + W_{s,i}^l * R_{n,t}^l + W_{a,i}^l * M_{n,t}^l + b_i^{l+1}),\\
G_{f,t}^{l\to l+1}(h,w) &= \sigma(W_f^{l+1} * F_t^{l+1}(h,w) + W_{s,f}^l * R_{n,t}^l + W_{a,f}^l * M_{n,t}^l + b_f^{l+1}),\\
G_{o,t}^{l\to l+1}(h,w) &= \sigma(W_o^{l+1} * F_t^{l+1}(h,w) + W_{s,o}^l * R_{n,t}^l + W_{a,o}^l * M_{n,t}^l + b_o^{l+1}),\\
G_{c,t}^{l\to l+1}(h,w) &= \tanh(W_c^{l+1} * F_t^{l+1}(h,w) + W_{s,c}^l * R_{n,t}^l + W_{a,c}^l * M_{n,t}^l + b_c^{l+1}),\\
C_{t+1}^{l\to l+1}(h,w) &= G_{f,t}^{l\to l+1}(h,w) \odot C_t^{l\to l+1}(h,w) + G_{i,t}^{l\to l+1}(h,w) \odot G_{c,t}^{l\to l+1}(h,w),\\
A_{t+1}^{l\to l+1}(h,w) &= \tanh(G_{o,t}^{l\to l+1}(h,w) \odot C_{t+1}^{l\to l+1}(h,w)),\\
F_{t+1}^{l+1}(h,w) &= F_t^{l+1}(h,w) + A_{t+1}^{l\to l+1}(h,w), \qquad (5)
\end{aligned}$$

where $(h,w) \in \Phi(S_n)$. W and b are convolutional kernels and biases. In Eq. (5), convolutions are denoted by $*$, while $\odot$ denotes the Hadamard product. Respectively, G and C represents the gate and cell state of an LSTM unit. $A_{t+1}^{l\to l+1}$ is the augmentation feature for F_t^{l+1}, and they have equal resolution. We add the augmentation feature $A_{t+1}^{l\to l+1}$ with F_t^{l+1}, producing the new feature F_{t+1}^{l+1} for the next stage. The sequence of features F_t^l is defined in the same way as above (with the l superscripts replaced by $l+1$, and vice versa).

5 Implementation Details

We use the Caffe platform [39] to implement our approach. Our approach can be based on different deep architectures [29, 40, 30], and we use the ResNet-152 architecture [30] pre-trained on ImageNet dataset [10] as our backbone network. We randomly initialize the parameters of our LSTM-based bidirectional connections. Before training our network for the evaluations on different benchmarks, we follow [18, 17, 20, 25] and use the COCO dataset [2] to fine-tune the whole network.

Given an input image, we apply the structured edge detection toolbox [41] to compute super-pixels. Empirically, we set the scale of super-pixels to be 1,000 per image. The image is fed to the backbone network to compute the convolutional features. Following [31], we select the last convolutional feature map from each residual block as the initial feature maps fed into our context intertwining network. More specifically, we use the ResNet-152 network layers *res2*, *res3*, *res4* and *res5* as $\left\{F_0^1, F_0^2, F_0^3, F_0^4\right\}$, respectively. Successive pairs of these feature maps are fed into our LSTM-based context intertwining modules, each of which has 3 bidirectional exchange stages. We optimize the segmentation network using the standard SGD solver. We fine-tune the parameters of the backbone network and the bidirectional connections.

During training, we use the common flipping, cropping, scaling and rotating of the image to augment the training data. The network is fine-tuned with a learning rate of 1e−3 for 60K mini-batches. After that, we decay the learning rate to 1e−4 for the next 60K mini-batches. The size of each mini-batch is set to 12. With the trained model, we perform multi-scale testing on each image to obtain the segmentation result. That is, we rescale each testing image using five factors (i.e., $\{0.4, 0.6, 0.8, 1.0, 1.2\}$) and feed the differently scaled versions into the network to obtain predictions. The predictions are averaged to yield the final result.

6 Experiments

We evaluate our approach on four public benchmarks for semantic segmentation, which are PASCAL VOC 2012 [1], PASCAL-Context [3], NYUDv2 [23] and SUN-RGBD [24] datasets. The PASCAL VOC 2012 dataset [1] has been widely used for evaluating segmentation performance. It contains 10,582 training images along with the pixel-wise annotations for 20 object classes and the background. The PASCAL VOC 2012 dataset also provides a validation set of 1,449 images and a test set of 1,456 images. We use this dataset for the major evaluation of our network. We further use the PASCAL-Context, NYUDv2 and SUN-RGBD datasets for extensive comparisons with state-of-the-art methods. We report all the segmentation scores in terms of mean Intersection-over-Union (IoU).

Ablation Study of MSCI Our MSCI architecture is designed to enable exchange of multi-scale context information between feature maps. It consists of recurrent bidirectional connections defined using super-pixels. Below, we report an ablation study of our approach, which examines the effect that removing various key components has on segmentation performance. The results are summarized in Table 1.

Our approach is based on LSTMs, each of which can be regarded as a special recurrent neural network (RNN) unit with a cell state for memorization. By removing the RNNs and the cell states, we effectively disable the bidirectional connection between feature maps. In this case, our model degrades to a basic FCN, and obtains the segmentation score of 77.8 that lags far behind our full MSCI model.

RNN	cell states	super-pixels	mean IoU
			77.8
✓		✓	84.4
✓	✓		84.3
✓	✓	✓	**85.1**

Table 1: Ablation experiments on the PASCAL VOC 2012 validation set. Segmentation accuracy is reported in terms of mean IoU (%).

strategy	method	VOC 2012	CONTEXT
w/o combination	basic FCN [9]	77.8	41.2
w/o hierarchy	SPP [18]	81.1	43.6
	Encoder-Decoder [17]	81.4	44.3
	ASPP [20]	82.2	46.0
	Encoder-Decoder + ASPP [25]	82.5	47.4
w/ hierarchy	MSCI	**85.1**	**50.3**

Table 2: Comparison of different feature combination strategies. Performance is evaluated on the PASCAL VOC 2012 and PASCAL-Context validation sets. Segmentation accuracy is reported in terms of mean IoU (%).

Next, we investigate the importance of the cell states. The cell states are employed by our approach to memorize the local and global context information, which enriches the final segmentation feature map. With all the cell states removed from our bidirectional connections, our approach achieves an accuracy of 84.4%, which is significantly lower than the 85.1% accuracy of our full approach.

In our approach, the super-pixels adaptively partition the features into different regions according to the image structure, which are then used for context aggregation and exchange (Figure 3). We remove the super-pixels and interpolate the low-resolution feature maps [17, 18] to match with the high-resolution maps. Thus, each neuron aggregates context from a local regular window. Compared to our full model, the performance drops to 84.3%, demonstrating the effectiveness of using super-pixels to guide context aggregation.

Feature Combination Strategies Our approach combines in a hierarchical manner the features produced by the bidirectional connections. In Table 2, we compare our feature combination strategy to those of other networks [9, 18, 17, 20, 25]. For a fair comparison, we reproduce the compared networks by pretraining them with the ResNet-152 backbone model on the ImageNet dataset, and fine-tuning them on the COCO dataset and the PASCAL VOC 2012 training set. Without any combination of features, the backbone network FCN model achieves the score of 77.8%. Next, we compare our network to the SPP network [9, 18, 20] and Encoder-Decoder [35, 17, 19, 21, 25] network. For the SPP network, we chose a state-of-the-art model proposed in [18] for comparison. The ASPP network [20] is a variant of the SPP network, and it can achieve better results than the SPP network. For the Encoder-Decoder network, we select the model proposed in [17] for comparison here. We also compare our network with the latest Encoder-Decoder network with ASPP components [25]. These models combine the adjacent features that are learned with our bidirectional connections, which generally leads to $0.4 \sim 1.2$ improvement in the segmentation scores, compared to the counterparts without bidirectional connections. We find that our approach performs better than other methods. In Figure 4, we can also observe that MSCI provides better visual results than other methods.

val set		test set	
method	mean IoU	method	mean IoU
Chen et al. [9]	77.6	Wang et al. [42]	83.1
Sun et al. [43]	80.6	Peng et al. [19]	83.6
Wu et al. [44]	80.8	Lin et al. [17]	84.2
Shen et al. [45]	80.9	Wu et al. [44]	84.9
Peng et al. [19]	81.0	Zhao et al. [18]	85.4
Zhao et al. [18]	81.4	Wang et al. [46]	86.3
Lin et al. [17]	82.7	Fu et al. [47]	86.6
Chen et al. [20]	82.7	Luo et al. [48]	86.8
Chen et al. [25]	84.6	Chen et al. [20]	86.9
Fu et al. [47]	84.8	Chen et al. [25]	**89.0**
MSCI	**85.1**	MSCI	88.0

Table 3: Comparisons with other state-of-the-art methods. The performances are evaluated on the PASCAL VOC 2012 validation set (left) and test set (right). Segmentation accuracy is reported in terms of mean IoU (%).

Comparisons with State-of-the-Art Methods In Table 3, we report the results of our approach on the PASCAL VOC 2012 validation set and test set, and compare with state-of-the-art methods. On the validation set (see Table 3 (left)), MSCI achieves a better result than all of other methods. Specifically, given the same set of training images, it outperforms the models proposed in [18, 20, 17], which are based on SPP, ASPP and Encoder-Decoder networks, respectively. In addition, we also report our result on the test set. Our per-category results on

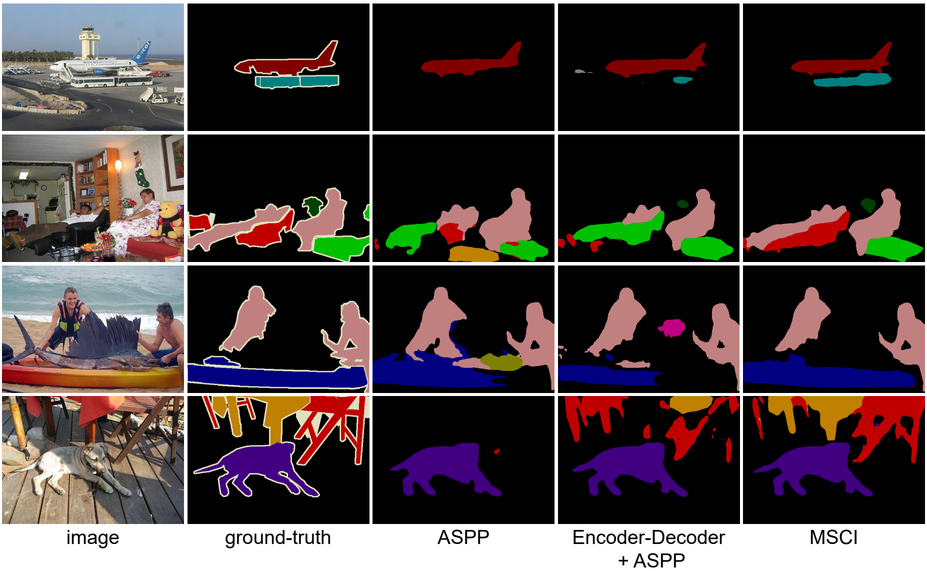


Fig. 4: The segmentation results of the ASPP model [20], Encoder-Decoder with ASPP model [25] and our MSCI. The images are taken from the PASCAL VOC 2012 validation set.

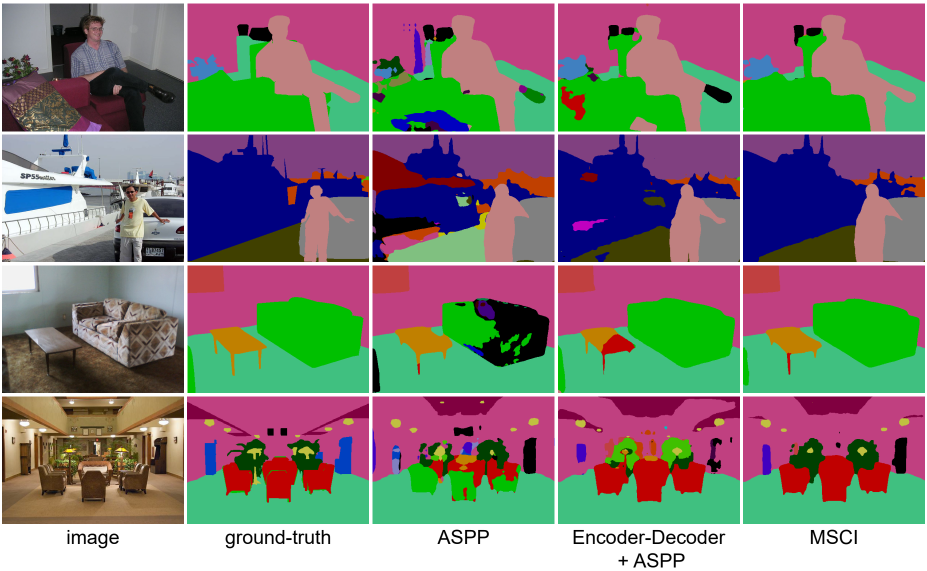


Fig. 5: The segmentation results of the ASPP model [20], Encoder-Decoder with ASPP model [25] and our MSCI. The images are scenes taken from the PASCAL-Context validation set.

the test set can be found on PASCAL VOC leaderboard[1]. Our result of 88.0% is second only to the score reported in [25], which leverages a stronger backbone network, trained on an internal JFT-300M dataset [26–28].

Experiments on Scene Labeling Datasets We perform additional experiments on three scene labeling datasets, including the PASCAL-Context [3], NYUDv2 [23], and SUN-RGBD [24]. In contrast to the object-centric PASCAL VOC 2012 dataset, these scene labeling datasets provide more complex pixel-wise annotations for objects and stuff, which require segmentation networks to have a full reasoning about the scene in an image. We use these datasets to verify if our network can label the scene images well.

The PASCAL-Context dataset [3] contains 59 categories and background, providing 4,998 images for training and 5,105 images for validation. In Table 2, we already used this dataset to compare MSCI to other feature combination strategies, and found that it works well on the scene labeling task. We provide several segmentation results in Figure 5. Table 4 shows that MSCI outperforms other state-of-the-art methods on this dataset.

CONTEXT		NYUDv2		SUN-RGBD	
method	mIoU	method	mIoU	method	mIoU
Dai et al. [49]	40.5	Long et al. [7]	34.0	Chen et al. [9]	27.4
Lin et al. [15]	42.0	Eigen et al. [50]	34.1	Kendall et al. [51]	30.7
Lin et al. [16]	43.3	He et al. [52]	40.1	Long et al. [7]	35.1
Wu et al. [53]	44.5	Lin et al. [16]	40.6	Hazirbas et al. [54]	37.8
Chen et al. [9]	45.7	Zhao et al. [18]	45.2	Lin et al. [16]	42.3
Lin et al. [17]	47.3	Lin et al. [17]	47.0	Lin et al. [17]	47.3
Wu et al. [44]	48.1	Lin et al. [55]	47.7	Lin et al. [55]	48.1
MSCI	**50.3**	MSCI	**49.0**	MSCI	**50.4**

Table 4: Comparison with other state-of-the-art methods. Performance is evaluated on the PASCAL-Context validation set (left), NYUDv2 validation set (middle) and the SUN-RGBD validation set (right). Segmentation accuracy is reported in terms of mean IoU (%).

We further evaluate our method on the NYUDv2 [23] and SUN-RGBD [24] datasets, originally intended for RGB-D scene labeling. The NYUDv2 dataset [23] has 1,449 images (795 training images and 654 testing images) and pixel-wise annotations of 40 categories. The SUN-RGBD dataset [24] has 10,335 images (5,285 training images and 5,050 testing images) and pixel-wise annotations of 37 categories. Unlike the PASCAL-Context dataset, the NYUDv2 and SUN-RGBD datasets consist of images of indoor scenes. We report the segmentation scores of MSCI and other state-of-the-art methods in Table 4. We note that the best

[1] http://host.robots.ox.ac.uk:8080/anonymous/F58739.html

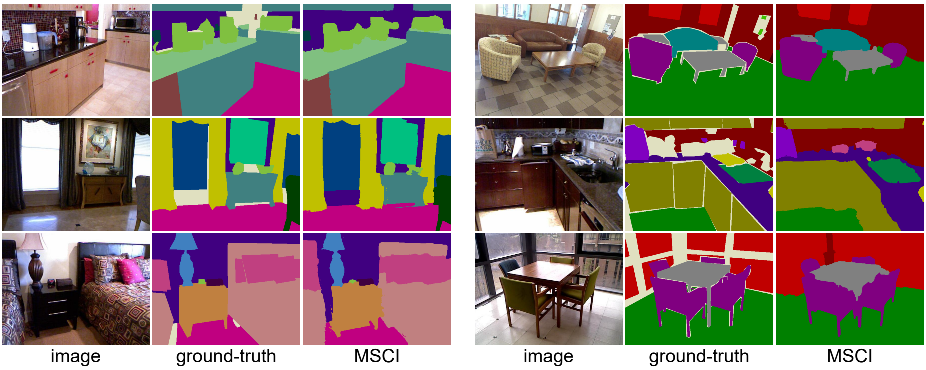


Fig. 6: MSCI segmentation results. The images are taken from the NYUDv2 validation set (left) and the SUN-RGBD validation set (right).

previous method proposed in [55] uses the RGB and depth information jointly for segmentation, and achieves the scores of 47.7 and 48.1 on the NYUDv2 and SUN-RGBD validation sets, respectively. Even without the depth information, MSCI outperforms the previous best results. We show some of our segmentation results on the NYUDv2 and SUN-RGBD validation sets in Figure 6.

7 Conclusions

Recent progress in semantic segmentation may be attributed to powerful deep convolutional features and the joint consideration of local and global context information. In this work, we have proposed a novel approach for connecting and combining feature maps and context from multiple scales. Our approach uses interconnected LSTM chains in order to effectively exchange information among feature maps corresponding to adjacent scales. The enriched maps are hierarchically combined to produce a high-resolution feature map for pixel-level semantic inference. We have demonstrated that our approach is effective and outperforms the state-of-the-art on several public benchmarks.

In the future, we plan to apply our MSCI approach to stronger backbone networks and more large-scale datasets for training. In addition, we aim to extend MSCI to other recognition tasks, such as object detection and 3D scene understanding.

Acknowledgments

We thank the anonymous reviewers for their constructive comments. This work was supported in part by NSFC (61702338, 61522213, 61761146002, 61861130365), 973 Program (2015CB352501), Guangdong Science and Technology Program (2015A030312015), Shenzhen Innovation Program (KQJSCX20170727101233642, JCYJ20151015151249564), and ISF-NSFC Joint Research Program (2472/17).

References

1. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. IJCV (2010)
2. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. (2014)
3. Mottaghi, R., Chen, X., Liu, X., Cho, N.G., Lee, S.W., Fidler, S., Urtasun, R., Yuille, A.: The role of context for object detection and semantic segmentation in the wild. In: CVPR. (2014)
4. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The Cityscapes dataset for semantic urban scene understanding. In: CVPR. (2016)
5. Chen, H., Qi, X., Yu, L., Dou, Q., Qin, J., Heng, P.A.: DCAN: Deep contour-aware networks for object instance segmentation from histology images. Medical Image Analysis (2017)
6. Yoon, Y., Jeon, H.G., Yoo, D., Lee, J.Y., Kweon, I.S.: Light-field image super-resolution using convolutional neural network. IEEE Signal Processing Letters (2017)
7. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015)
8. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: ICCV. (2015)
9. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. arXiv (2016)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. (2009)
11. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Hypercolumns for object segmentation and fine-grained localization. In: CVPR. (2015)
12. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: ICCV. (2015)
13. Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X.: Semantic image segmentation via deep parsing network. In: ICCV. (2015)
14. Papandreou, G., Chen, L.C., Murphy, K., Yuille, A.L.: Weakly-and semi-supervised learning of a DCNN for semantic image segmentation. arXiv preprint arXiv:1502.02734 (2015)
15. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. In: CVPR. (2016)
16. Lin, G., Shen, C., van den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: CVPR. (2016)
17. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. arXiv (2016)
18. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. arXiv (2016)
19. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. arXiv (2017)
20. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv (2017)

21. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. CVPR (2017)
22. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation (1997)
23. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV. (2012)
24. Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: A RGB-D scene understanding benchmark suite. In: CVPR. (2015)
25. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv preprint arXiv:1802.02611 (2018)
26. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. NIPS (2014)
27. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. CVPR (2017)
28. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: ICCV. (2017)
29. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: NIPS. (2012)
30. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CVPR (2016)
31. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. CVPR (2017)
32. Liang, X., Shen, X., Feng, J., Lin, L., Yan, S.: Semantic object parsing with graph LSTM. In: ECCV. (2016)
33. Liang, X., Shen, X., Xiang, D., Feng, J., Lin, L., Yan, S.: Semantic object parsing with local-global long short-term memory. In: CVPR. (2016) 3185–3193
34. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
35. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: MICCAI. (2015)
36. Gadde, R., Jampani, V., Kiefel, M., Kappler, D., Gehler, P.V.: Superpixel convolutional networks using bilateral inceptions. In: ECCV. (2016)
37. Bell, S., Lawrence Zitnick, C., Bala, K., Girshick, R.: Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In: CVPR. (2016)
38. Zeng, X., Ouyang, W., Yan, J., Li, H., Xiao, T., Wang, K., Liu, Y., Zhou, Y., Yang, B., Wang, Z., et al.: Crafting GBD-Net for object detection. PAMI (2017)
39. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM International Conference on Multimedia. (2014)
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv (2014)
41. Dollár, P., Zitnick, C.L.: Structured forests for fast edge detection. In: ICCV. (2013)
42. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G.: Understanding convolution for semantic segmentation. arXiv preprint arXiv:1702.08502 (2017)
43. Sun, H., Xie, D., Pu, S.: Mixed context networks for semantic segmentation. arXiv preprint arXiv:1610.05854 (2016)

44. Wu, Z., Shen, C., Hengel, A.v.d.: Wider or deeper: Revisiting the ResNet model for visual recognition. arXiv preprint arXiv:1611.10080 (2016)
45. Shen, F., Gan, R., Yan, S., Zeng, G.: Semantic segmentation via structured patch prediction, context CRF and guidance CRF. In: CVPR. (2017)
46. Wang, G., Luo, P., Lin, L., Wang, X.: Learning object interactions and descriptions for semantic image segmentation. In: CVPR. (2017)
47. Fu, J., Liu, J., Wang, Y., Lu, H.: Stacked deconvolutional network for semantic segmentation. arXiv preprint arXiv:1708.04943 (2017)
48. Luo, P., Wang, G., Lin, L., Wang, X.: Deep dual learning for semantic image segmentation. In: CVPR. (2017)
49. Dai, J., He, K., Sun, J.: BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: ICCV. (2015)
50. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV. (2015)
51. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv (2015)
52. He, Y., Chiu, W.C., Keuper, M., Fritz, M.: RGBD semantic segmentation using spatio-temporal data-driven pooling. arXiv (2016)
53. Wu, Z., Shen, C., Hengel, A.v.d.: High-performance semantic segmentation using very deep fully convolutional networks. arXiv preprint arXiv:1604.04339 (2016)
54. Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In: ACCV. (2016)
55. Lin, D., Chen, G., Cohen-Or, D., Heng, P.A., Huang, H.: Cascaded feature network for semantic segmentation of RGB-D images. In: ICCV. (2017)