

Deep Interactive Region Segmentation and Captioning

26 Jul 2017 Ali Sharifi Boroujerdi, Maryam Khanian & Brandenburg University of Technology,
Michael Breuß Cottbus-Senftenberg, Germany

arXiv:1707.08364v1 [cs.CV]

Abstract

With recent innovations in dense image captioning, it is now possible to describe every object of the scene with a caption while objects are determined by bounding boxes. However, interpretation of such an output is not trivial due to the existence of many overlapping bounding boxes. Furthermore, in current captioning frameworks, the user is not able to involve personal preferences to exclude out of interest areas. In this paper, we propose a novel hybrid deep learning architecture for interactive region segmentation and captioning where the user is able to specify an arbitrary region of the image that should be processed. To this end, a dedicated Fully Convolutional Network (FCN) named Lyncean FCN (LFCN) is trained using our special training data to isolate the User Intention Region (UIR) as the output of an efficient segmentation. In parallel, a dense image captioning model is utilized to provide a wide variety of captions for that region. Then, the UIR will be explained with the caption of the best match bounding box. To the best of our knowledge, this is the first work that provides such a comprehensive output. Our experiments show the superiority of the proposed approach over state-of-the-art interactive segmentation methods on several well-known datasets. In addition, replacement of the bounding boxes with the result of the interactive segmentation leads to a better understanding of the dense image captioning output as well as accuracy enhancement for the object detection in terms of Intersection over Union (IoU).

1 Introduction

As one of the main sources of the human knowledge, our visual system including eyes, optic nerves and brain is able to easily detect, separate and describe each object of a scene. Inspired by this natural ability, interactive region segmentation and captioning is the task of parallel detection, separation and description of the visual user interests. This procedure can be exploited in several complex applications such as automatic image annotation and retrieval [45, 53]. To approach the task, one needs to have a full understanding of the scene which is equivalent to recognize and also locate all the visible objects. To this end, several object recognition techniques [10, 49, 54] have been proposed to detect image objects in different scales. In most of the literature, detected objects are determined by drawing bounding boxes around them. Although this

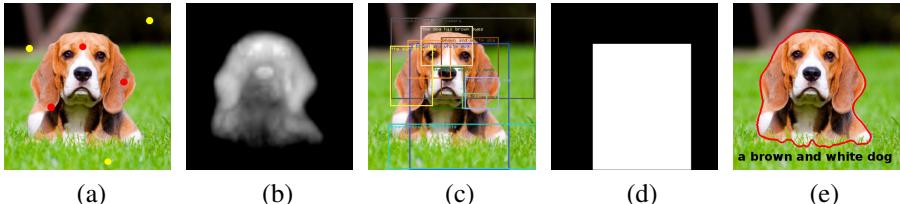


Figure 1: Input image including positive and negative user clicks (a), probability map of our LFCN considering user intention (b), output of the dense image captioning [28] (c), best match bounding box w.r.t. user intention (d), and the final output of the proposed model including determined user intention region (UIR) and its description (e).

notation is able to facilitate the detection process by decreasing its computational complexity, such an output is less informative when dealing with geometrical properties of the objects. As a more illustrative visual recognition technique, semantic segmentation [8, 23, 37, 38] aims to assign a label to each pixel of the image where the labels can be class-aware or instance-aware. While the multi-level nature of semantic segmentation increases the problem dimensionality, interactive image segmentation [8, 17, 48] tries to adjust the segmentation task with the user priorities in a simpler problem space. In reality, it sounds reasonable that human users may have a more restricted area of interest than the entire scope of the scene. Thus, the multi-dimensional semantic segmentation task can be shrunk to a binary segmentation problem aiming to separate the User Intention Region (UIR) as the foreground from other parts of the scene which requires less time and computations.

Equipped by the rich semantic memory of the visual data, the human observer is easily able to provide detailed explanation about different parts of an image which is a hard task in artificial intelligence. Thanks to recent developments of language models [29, 42], image captioning [8, 10, 13, 29, 39, 50] makes it possible to produce linguistic descriptions of an image through a multimodal embedding of the visual stimuli and the word representation [23] in a joint vector space [28].

In this paper, we propose a novel hybrid deep architecture for integrated detection, segmentation and captioning of the user preferences where the amount of the user interactions is limited to one or a few clicks. To this end, we designed a heuristic technique for the efficient generation of the synthetic user interactions. In addition, the new architecture of the proposed Lyncean Fully Convolutional Network (LFCN) leads to a better sight of the deep component that is responsible for interactive segmentation. Last but not least, as depicted in Fig. 1, our combination of interactive segmentation and dense captioning tasks introduces a new class of outputs where the user intention recognition meets linguistic interpretations and vice versa. Let us stress at this point that our main contributions are (i) to provide the first deep framework for combined interactive segmentation and captioning, and (ii) to achieve significant improvements in the interactive segmentation over other methods.

2 More Details on Related Works

With the increasing popularity of deep learning architectures [31, 35, 32], both detection and captioning procedures have attracted a new wave of considerations. Convolutional Neural Networks (CNNs) [31, 32] have presented the ability to construct numer-

ous visual features in different levels of abstraction through supervised learning. This property leads to feature generators that are able to reach near-human performance in various computer vision tasks [20]. In addition, the structure of Fully Convolutional Networks (FCNs) [21] made it feasible to apply inputs of any size to the network and generate associated output in the same spatial domain. In contrast to CNNs, FCNs are able to maintain spatial information which is crucial to perform a pixel-level prediction such as semantic segmentation, object localization [22], depth estimation [23] and interactive segmentation. Furthermore, Recurrent Neural Networks (RNNs) [24, 25] reveal potential for learning long term dependencies which is essential for simulating the continuous space of natural languages.

Recently, CNN-RNN models are proposed to wrap detection and captioning tasks in an end-to-end learnable platform [26]. However, up to now the results appear to be mostly an unorganized and overcrowded set of captions and bounding boxes. These results are not easily understandable especially in the presence of several overlapping region proposals cf. Fig. 1 (c). In addition, they do not involve user intentions.

Before the success of CNNs in object detection, some classical techniques such as Histogram of Gradients (HoG) [27], Deformable Part Models (DPM) [28] and selective search [29] (as an explicit region proposal method) were proposed. Later, in the Region-based CNN (R-CNN) model [15], each proposed region has been forwarded through a separate CNN for the feature extraction. This model had some drawbacks such as a complex multi-stage training pipeline and expensive training process. To overcome those obstacles, the Fast R-CNN model [20] is proposed where a combination of a CNN and the Region of Interest (RoI) pooling mechanism is used to produce better information for region proposal. In the Faster R-CNN [20], the CNN architecture is used not only for the feature extraction but also for region proposal itself. This leads to the invention of the Region Proposal Networks (RPNs) that are able to share full-image convolutional features with detection networks. The main achievement of this innovation is the parallel detection and localization of the objects in one forward pass of the network. In spite of these improvements, such models are not able to backpropagate through the bounding boxes information. Recently, Johnson et. al [26] proposed a localization layer based on Faster R-CNN architecture where the RoI pooling mechanism is replaced by bilinear interpolation [30, 31] that makes it possible to propagate backward through all the information.

The primary purpose of the image captioning was image annotation [24, 25] as the automatic assignment of some keywords to a digital image. By replacing keywords with some sentences that are able to describe not only the image objects but also the semantic relations in between, image captioning received more attention. The main problem in the automatic image description was the scarcity of the training data. Recent development of large datasets including images and their descriptions [27, 30, 36] makes it feasible to expand learning-based captioning techniques. Classical image captioning approaches produced image descriptions by generative grammars [24, 25] or pre-defined templates working on some specific visual features [2, 3]. In contrast, recently developed deep learning solutions apply an RNN-based language model that is conditioned on the output vectors of a CNN to generate descriptive sentences [8, 10, 28, 40, 58, 63].

With the growing popularity of interactive devices such as smart phones and tablets, interactive image processing attracts more attention. Interactive segmentation offers a pixel-wise classification based on user priorities. Among all the traditional approaches

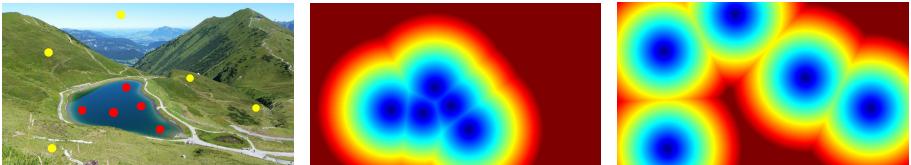


Figure 2: Input image including positive and negative clicks (left), and obtained positive (middle) and negative (right) Voronoi diagrams.

of the interactive segmentation, stroke-based techniques [54, 50, 52] are often based on graph cut techniques. In these methods, an energy function based on region/boundary division is optimized to find the segmentation result. Alternative approaches include random walks [10] and geodesics [8, 7], mostly relying on low-level features such as color, texture and shape information. These types of attributes can be difficult to apply when the image appearance is complicated due to complex lighting conditions or existence of intricate textures. Recently, deep learning models have been used for interactive segmentation where the information of the image will be considered in higher semantic levels. To this end, FCNs as the standard frameworks for pixel-wise end-to-end learning tasks, have been applied [53, 55, 51, 56].

3 Proposed Method

Our model receives an input image as well as user interactions in the form of positive/negative clicks and provides a seamless framework to generate accurate segmentation as well as expressive description of the UIR. In the preprocessing step, an efficient morphological technique will be used to provide a huge amount of training samples in the form of synthetic user interactions. Then, each set of positive/negative seeds will be transformed into separate Voronoi diagrams as shown in Fig. 2. Next, a sequence of dedicated LFCNs with different granularities are applied as the interactive segmentation modules. Afterwards, a dense captioning architecture inspired by [76] will be utilized to obtain a number of region proposals along with their captions. In the fusion step, a heuristic method will be provided to combine results of localization, segmentation and captioning procedures to acquire highlighted borders of the UIR along with its expositor caption. In the following, we will investigate all steps of our model in detail.

3.1 User Action Imitation

During interactive segmentation, the user will be asked to provide some general information about the position of the intended region. The requested information consists of some positive and negative seeds as depicted in Fig. 2 which are equivalent to internal and external points of the UIR, respectively. Next, each set of seeds will be used to shape a Voronoi diagram. We denote each seed by s_k , $k = \{1, \dots, n\}$. The value of pixel $v_{i,j}$ of the Voronoi diagram will be calculated by

$$v_{i,j} := \min\{D_1, D_2, \dots, D_n\} \quad (1)$$

where D_k is the Euclidean distance of $v_{i,j}$ to the seed s_k . To summarize, the value of each pixel in the Voronoi diagram is the Euclidean distance of that pixel to the nearest

seed. For the sake of clarity, there should be a minimum inter-cluster distance in each set of the seeds. In addition, a minimum intra-cluster distance is also required to retain boundary regions of the clusters as distinctive as possible. So:

- Every pair of seeds in each set should preserve a pre-defined distance from each other:

$$\exists d_1 \in \mathbb{R}^+ : \forall (s_i, s_j) \in S, \| (s_i, s_j) \|_2 > d_1 \quad (2)$$

- All the seeds of each set should preserve a minimum distance from boundary pixels of the UIR ($\partial(\text{UIR})$):

$$\exists d_2 \in \mathbb{R}^+ : \forall s_i \in S, u \in \partial(\text{UIR}), \| (s_i, u) \|_2 > d_2 \quad (3)$$

As expected, natural collection of such a data is unreasonably time consuming and expensive. Recently, Xu et al. [61] proposed some strategies for synthetic generation of user interactions. They ordained random generation for positive clicks inside the UIR while three distinct set of negative clicks are chosen as: 1) random background pixels with a certain distance to the UIR, 2) a point cloud inside the negative objects and 3) a uniform set of surrounding points of the UIR. Since their implementation is not publicly available, it seems their first and the second negative strategies do not obey natural interactions and the third one may be computationally expensive (see equation (2) in [61]).

Morphological Cortex Detection (MCD). While the inside of the UIR can be quite small, the background region is usually large enough to provide useful geometric information about the UIR. Consequently, it is beneficial to generate negative seeds that surround the UIR uniformly. To provide an efficient implementation for such an interaction, we replace third negative strategy proposed in Xu et al. [61] with a Morphological Cortex Detection (MCD) technique that noticeably improves computational efficiency. Moreover, this method is able to simulate UIR cortex in different scales that enables convolutional filter of the LFCN to track UIR geometry in different layers. To implement this idea, a 1-pixel-wide boundary shape of the UIR will be extracted by performing a dilation on the binary mask of UIR in training dataset. Then, the original mask is subtracted from the dilation result. In the next step, this boundary path will be completely traversed using a 3×3 window to transfer all the boundary points' coordinates into a 1-D array in which the requested negative seeds can be selected uniformly. As the result of the MCD process, a uniform set of negative seeds will be obtained that represents the cortex of the UIR perfectly. The visual illustration of this technique is shown in Fig. 3. During our experiments, positive clicks are simulated randomly inside the UIR while negative seeds are generated by MCD mechanism in three different levels.

3.2 Intention Recognition

For the task of intention recognition, we make use of a dedicated version of the standard FCN [38] where the last two fully connected layers are replaced with three convolutional layers containing decreasing kernel sizes of 7, 5 and 3. The impact of such an alternation is the gradual growth of the receptive field. This property improves network recognition of objects' geometry. Hence, we named this architecture as Lyncean

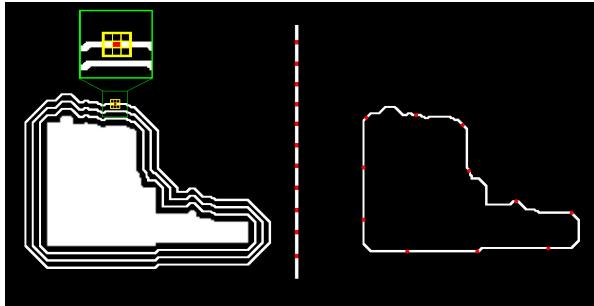


Figure 3: Our proposed Morphological Cortex Detection (MCD) technique.

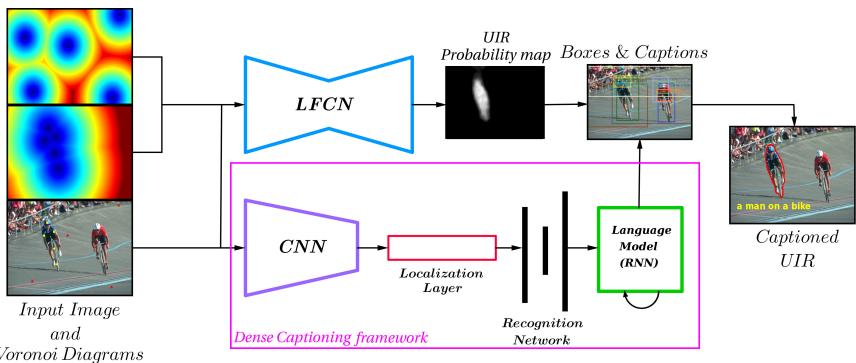


Figure 4: Architecture of the proposed deep interactive region segmentation and captioning model.

Fully Convolutional Network (LFCN). By a proper use of zero padding, all the extended convolutional layers have the same output size. At the end of the extended part, the aggregated output of the additional layers will be upsampled to the size of the input as elaborated in [38].

3.3 Fusion Approach

In order to supplement the result of the interactive segmentation with a proper linguistic commentary, we employ the dense image captioning framework [76]. The internal RPN of this architecture provides confidence scores for the existence of the object in proposed regions. After descending sort of the objectness scores, top-ranked region proposals include the most reasonable captions for the objects of the scene. With the comparison of the interactive segmentation result and the bounding boxes, the best match bounding box and the corresponding caption will be obtained (Fig. 4).

4 Experiments

Datasets. For fine tuning of the LFCN, we used the PASCAL VOC 2012 segmentation dataset [10]. The dataset includes 1464 images for training and 1449 images

for validation that are distributed in 20 different classes. We used the whole bunch of these samples to generate our special training pairs in the preprocessing step. For the final validation of the model as well as its comparison with state-of-the-art interactive segmentation, we utilized different well-known segmentation benchmarks including Alpha Matting [62], Berkeley segmentation dataset (BSDS500) [63], Weizmann segmentation evaluation database [64], image object segmentation visual quality evaluation database [65] and VOC validation subset.

Preprocessing. To generate all the necessary training pairs of the interactive segmentation process, we produced positive and negative Voronoi diagrams with respect to each object that is visible in VOC dataset. The positive seeds are selected randomly inside each object while the MCD approach is used to generate three distinct sets of negative seeds with different distances from the intended object. In the last step, each combination of positive/negative Voronoi diagrams, forms a unique training pair. This leads to production of 97,055 interaction patterns. We preserved 7,055 instances for the test and used the rest as the training data.

4.1 Fine Tuning of the Proposed LFCN Architecture

To reach the best quality for the interactive segmentation, our LFCN is trained in three different levels of granularity as proposed in [66]: LFCN32s, LFCN16s and LFCN8s. RGB channels of the input image should be concatenated with the corresponding Voronoi diagrams to form a training instance. Consequently, the first convolutional layer of our LFCN contains five channels. During the network initialization, the RGB-related channels will be initialized by the parameters of the original FCN [67]. For two extra channels that are associated with Voronoi diagrams, the zero initialization is the best choice as also mentioned in [61]. Learning parameters of the finer networks should be initialized from the coarser one. The global learning rates of the networks are 1e-8, 1e-10 and 1e-12, respectively while the extended convolutional layers exploit one hundred times bigger learning rates. The learning policy is fixed and we used the weight decay of 5e-3.

4.2 Metrics

In order to evaluate UIR localization accuracy of the proposed model, we calculated the well-known measure of Intersection over Union (IoU). To this aim, we computed the IoU of the detected UIR and the corresponding binary label of the validation samples. For the sake of complete comparison between our model and other interactive segmentation techniques, three performance metrics of pixel accuracy, mean accuracy and mean IoU are computed.

The segmentation task of the proposed approach can be considered as a binary segmentation where the classes are limited to foreground (UIR) and background. So, we used binary interpretation of the semantic segmentation metrics that are proposed by Long et al. [68]:

- **Pixel Accuracy (Pixel Acc.):**

This measure represents the proportion of the correctly classified foreground (C_f) and background (C_b) pixels (true positive rates) to the total number of

ground truth pixels in foreground (F) and background (B).

$$\frac{C_f + C_b}{F + B} \quad (4)$$

Unfortunately, this metric can be easily influenced by the class imbalance. Hence, high pixel accuracy does not necessarily mean that the accuracy is acceptable when one of the classes is too small or too large.

- **Mean Pixel Accuracy (Mean Acc.):**

This measure is computed as the mean of the separate foreground and background pixel accuracies:

$$\frac{\frac{C_f}{F} + \frac{C_b}{B}}{2} \quad (5)$$

This metric alleviates the imbalance problem but can be still misleading. For example when the great majority of pixels are background, a method that predicts all the pixels as background can still have seemingly good performance.

- **Mean Intersection over Union (Mean IoU):**

Intersection over union is the matching ratio between the result of object localization process and the corresponding ground truth label. This metric is the average of the computed intersection over union for the foreground and background regions:

$$\frac{\left(\frac{C_f}{C_f + FP + FN}\right)_f + \left(\frac{C_b}{C_b + FP + FN}\right)_b}{2} \quad (6)$$

Here FP and FN denote the number of the false positive and false negative rates of each class, respectively. This metric solves the previously described issues.

4.3 Results

Test of localization accuracy. In the first step of evaluation, we test our model with a random subset of unseen samples in validation datasets. The response of the model to some instances is shown in Fig. 5. It can be noticed that the output of our approach achieves a considerable rate of accuracy regarding the similarity of the model output with the corresponding ground truth. Furthermore, the confusing output of the dense image captioning is replaced with an explicit situation where the segmented UIR and its description are easily distinguishable. Fig. 6 (left diagram) presents a comparison between the localization accuracy of the proposed method and the internal RPN of the DenseCap [26] in terms of the obtained IoU for the samples presented in Fig. 5. As illustrated, our model provides a significant improvement regarding the localization accuracy. These results also demonstrate the proficiency of our model in combining interactive segmentation, region proposal and image captioning techniques.

Sensitivity analysis. In this part, we analysed variations of the model output quality against the number of user interactions. As it is shown in Fig 7, although IoU can be improved by applying more user interactions that facilitate boundary detection, our model still provides very good results even by minimum number of clicks. We also provided mean IoU accuracy of the proposed model for five different datasets in Fig. 6 (right diagram) that confirms satisfying performance of our model in the case of low

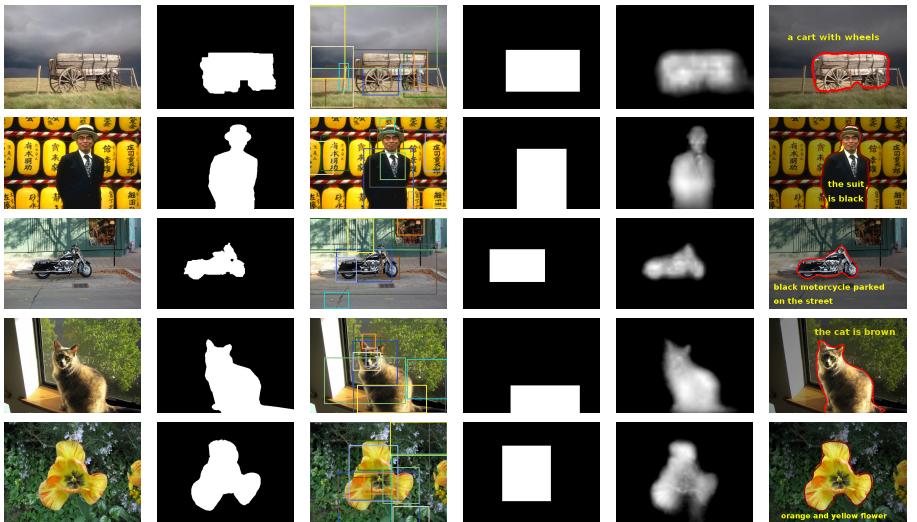


Figure 5: From left to right: input images, UIR ground truths, dense captioning bounding boxes [76], best match bounding boxes, our LFCN probability maps and the final outputs of our model including highlighted UIR and its description.

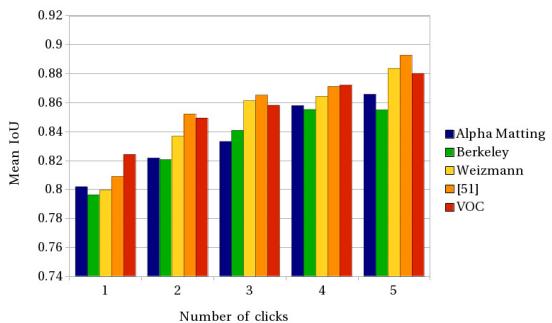
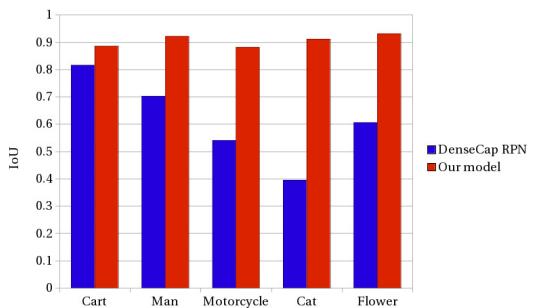


Figure 6: Localization accuracy comparison between our model and the internal RPN of the DenseCap [76] (up), and mean IoU accuracy of the proposed model on several segmentation benchmarks against different number of clicks (down).

interactive information. This noticeable property of our approach makes it convenient to be applied in real-world applications. During our experiments, the proposed method clearly achieves a satisfying segmentation outcome with just one click.

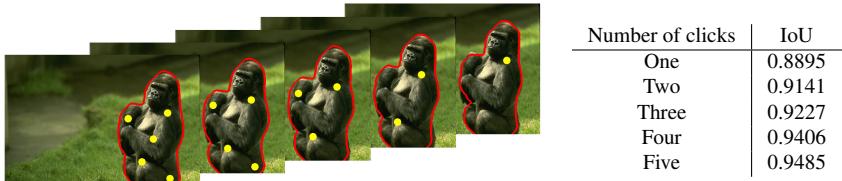


Figure 7: Model sensitivity analysis against number of clicks on a sample input image (left), and its associated IoU rates (right).

Comparison of the segmentation quality. In this part we performed an extensive evaluation on segmentation capabilities of the proposed method versus some prevalent segmentation techniques such as Geodesic Matting (GM) [1], GrowCut [2], Grabcut [3], Boykov Jolly (BJ) interactive graph cuts [4], Geodesic Star Convexity (GSC) [5], Geodesic Star Convexity with sequential constraints (GSCSEQ), Random Walker (RW) segmentation [6], Shortest Path-based interactive segmentation (SP) [7] and Matching Attributed Relational Graphs (MARG) [8]. In all the experiments we generated five positive and five negative clicks randomly. For some of the approaches where the user interactions were defined as points or scribbles, we determined click positions with five-pixel-wide circles. To observe the impact of the extended part of the LFCN on the output quality, we also report all the accuracy measures for the normal version of the FCN as well. Table 1 and 2 present quantitative results that confirm our approach superiority over several other segmentation techniques on five different benchmarks. As a qualitative comparison, Fig. 8 represents final segmentation output of the methods in Table 1 for two different samples. As it can be seen, our approach provides the most accurate segmentation result with respect to semantic interpretation of the scene using same number of interactions.

	Berkeley dataset [1]			Weizmann dataset [2]		
	Pixel Acc.	Mean Acc.	Mean IoU	Pixel Acc.	Mean Acc.	Mean IoU
GM [1]	0.8237	0.7645	0.6098	0.8328	0.7906	0.6615
GrowCut [2]	0.6639	0.7603	0.4520	0.6847	0.7025	0.4795
GrabCut [3]	0.6614	0.7582	0.4511	0.6896	0.7103	0.4818
BJ [4]	0.8138	0.8292	0.6488	0.8196	0.8416	0.6993
GSC [5]	0.8240	0.8465	0.6624	0.8253	0.8463	0.6977
GSCSEQ	0.8275	0.8425	0.6654	0.8290	0.8482	0.7014
RW [6]	0.8691	0.8046	0.6917	0.7953	0.7440	0.6207
SP [7]	0.7838	0.8405	0.6121	0.7712	0.8092	0.6272
MARG [8]	0.8067	0.6516	0.6180	0.7907	0.8110	0.6354
FCN8s	0.9379	0.8766	0.8352	0.9227	0.8978	0.8573
LFCN8s	0.9597	0.8989	0.8549	0.9549	0.9316	0.8837

Table 1: Segmentation accuracy comparison between the proposed method (LFCN8s), different types of interactive segmentation techniques and the original version of the FCN8s [8] on Berkeley [1] and Weizmann [2] segmentation benchmarks.

Alpha matting dataset [27]

	Pixel Acc.	Mean Acc.	Mean IoU
GM [8]	0.7958	0.7993	0.6602
GrowCut [25]	0.7924	0.7982	0.6465
GrabCut [48]	0.7908	0.7953	0.6513
BJ [9]	0.9117	0.9119	0.8569
GSC [29]	0.9076	0.9056	0.8476
GSCSEQ	0.9092	0.9066	0.8492
RW [20]	0.8568	0.8599	0.7561
SP [22]	0.8390	0.8466	0.7258
MARG [45]	0.9131	0.9152	0.8434
FCN8s	0.9193	0.9068	0.8438
LFCN8s	0.9320	0.9227	0.8656

Image object segmentation visual quality evaluation dataset [50]

	Pixel Acc.	Mean Acc.	Mean IoU
GM [8]	0.8313	0.7731	0.6226
GrowCut [25]	0.6600	0.6968	0.4425
GrabCut [48]	0.6571	0.6922	0.4391
BJ [9]	0.8108	0.8069	0.6406
GSC [29]	0.8170	0.8087	0.6458
GSCSEQ	0.8193	0.8092	0.6486
RW [20]	0.7787	0.7665	0.5810
SP [22]	0.7914	0.8030	0.6053
MARG [45]	0.7651	0.7871	0.5685
FCN8s	0.9526	0.9155	0.8710
LFCN8s	0.9689	0.9332	0.8927

VOC validation dataset [20]

	Pixel Acc.	Mean Acc.	Mean IoU
GM [8]	0.8165	0.7283	0.5787
GrowCut [25]	0.6268	0.6366	0.3999
GrabCut [48]	0.6282	0.6412	0.4028
BJ [9]	0.7559	0.7824	0.5794
GSC [29]	0.7707	0.7879	0.5903
GSCSEQ	0.7724	0.7883	0.5912
RW [20]	0.7014	0.7102	0.5095
SP [22]	0.7602	0.7809	0.5618
MARG [45]	0.7118	0.7218	0.5050
FCN8s	0.9527	0.9201	0.8723
LFCN8s	0.9630	0.9260	0.8801

Table 2: Segmentation accuracy comparison between the proposed method (LFCN8s), different types of interactive segmentation techniques and the original version of the FCN8s [48] on three more segmentation benchmarks.

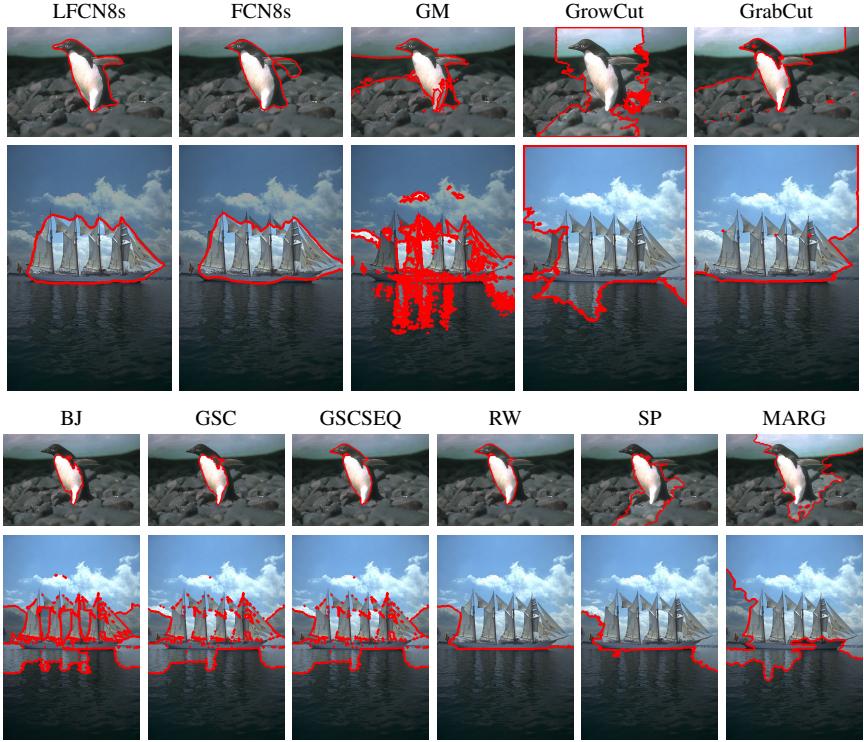


Figure 8: Segmentation quality comparison between our model (LFCN8s) and other interactive segmentation techniques. Our method clearly provides higher level of region understanding.

Dense interactive region captioning. In the final part of our experiments, we verified the ability of the proposed model to caption several regions of the image. Fig. 9 shows the result of such an experiment where multiple objects in different scales are detected via user interactions and described properly.

5 Conclusion

In this paper, we presented a novel hybrid deep learning framework which is capable of targeted segmentation and captioning as a response to user interactive actions. A wide variety of experiments confirmed our model superiority over various state-of-the-art interactive segmentation approaches. In addition, further experiments demonstrated our model capability to caption an arbitrary region of the image with one or few clicks, which is especially convenient for real-world interactive applications.

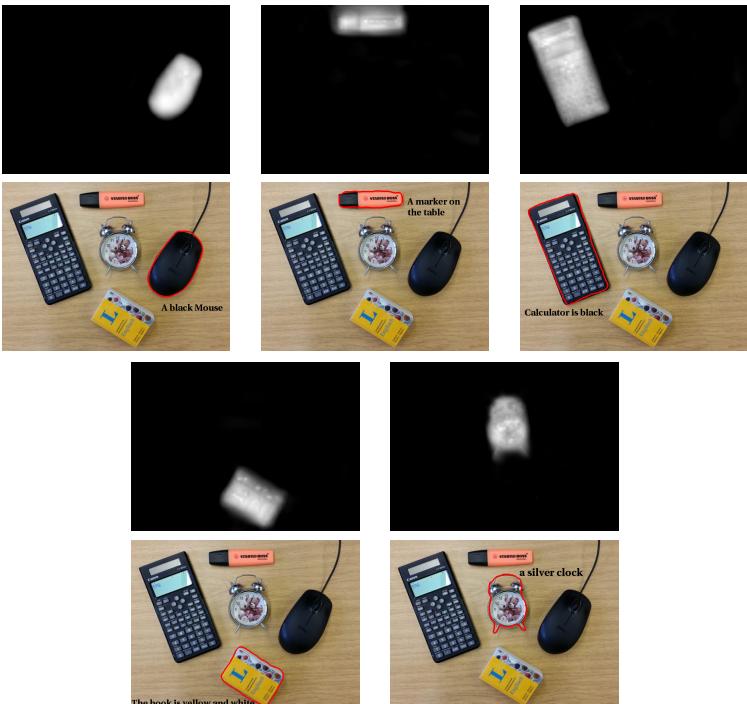


Figure 9: Probability maps and final outputs of the proposed model in response of user interactions over different parts of an image.

References

- [1] S. Alpert, M. Galun, A. Brandt, and R. Basri. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *Proc. CVPR*, 2007.
- [2] G. Ankush and M. Prashanth. From image annotation to image description. In *Proc. ICONIP*, 2012.
- [3] X. Bai and G. Sapiro. Geodesic matting: A framework for fast interactive image and video segmentation and matting. *International Journal of Computer Vision*, 82(2):113–132, 2009.
- [4] Y. Y. Boykov and M. P. Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images. In *Proc. ICCV*, 2001.
- [5] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint*, (arXiv: 1412.7062), 2014.
- [6] X. Chen and C. L. Zitnick. Mind’s eye: A recurrent visual representation for image caption generation. In *Proc. CVPR*, 2015.

-
- [7] A. Criminisi, T. Sharp, and A. Blake. Geos: Geodesic image segmentation. In *Proc. ECCV*, 2008.
 - [8] N. Dallal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.
 - [9] E. Desmond and K. Frank. Image description using visual dependency representations. In *Proc. EMNLP*, 2013.
 - [10] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. CVPR*, 2015.
 - [11] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proc. ICCV*, 2015.
 - [12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010.
 - [13] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *Proc. CVPR*, 2015.
 - [14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE T-PAMI*, 32(9):1627–1645, 2010.
 - [15] R. Girshick. Fast r-cnn. In *Proc. ICCV*, 2015.
 - [16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014.
 - [17] L. Grady. Random walks for image segmentation. *IEEE T-PAMI*, 28(11):1768–1783, 2006.
 - [18] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. Draw: A recurrent neural network for image generation. In *Proc. ICML*, 2015.
 - [19] V. Gulshan, C. Rother, A. Criminisi, A. Blake, and A. Zisserman. Geodesic star convexity for interactive image segmentation. In *Proc. CVPR*, 2010.
 - [20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. CVPR*, 2016.
 - [21] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
 - [22] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.
 - [23] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu. Spatial transformer networks. In *Proc. NIPS*, 2015.

-
- [24] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. ACM SIGIR*, 2003.
 - [25] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. ACM SIGIR*, 2003.
 - [26] J. Johnson, A. Karpathy, and F. F. Li. Densecap: Fully convolutional localization networks for dense captioning. In *Proc. CVPR*, 2016.
 - [27] J. L. Jones, X. Xie, and E. Essa. Image segmentation using combined user interactions. In *Proc. CVMC*, 2013.
 - [28] A. Karpathy and F. F. Li. Deep visual-semantic alignments for generating image description. In *Proc. CVPR*, 2015.
 - [29] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. In *Proc. TACL*, 2015.
 - [30] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. J. Li, D. A. Shamma, M. S. Bernstein, and F. F. Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, pages 1–42, 2017.
 - [31] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.
 - [32] 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.
 - [33] G. Li, Y. Xie, L. Lin, and Y. Yu. Instance-level salient object segmentation. *arXiv preprint*, (arXiv: 1704.03604), 2017.
 - [34] Y. Li, J. Sun, C. K. Tang, and H. Y. Shum. Lazy snapping. *ACM Transactions on Graphics (TOG)*, 23:303–308, 2004.
 - [35] D. Lin, J. Dai, J. Jia, K. He, and J. Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proc. CVPR*, 2016.
 - [36] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context. *arXiv preprint*, (arXiv: 1405.0312), 2014.
 - [37] Z. Liu, X. Li, P. Luo, C. C. Loy, and X. Tang. Semantic image segmentation via deep parsing network. In *Proc. ICCV*, 2015.
 - [38] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *arXiv preprint*, (arXiv: 1411.4038), 2014.
 - [39] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint*, (arXiv: 1410.1090), 2014.
 - [40] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proc. ICLR*, 2015.

-
- [41] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. ICCV*, 2001.
 - [42] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, 2010.
 - [43] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint*, (arXiv: 1301.3781), 2013.
 - [44] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. III Daume. Midge: Generating image descriptions from computer vision detections. In *Proc. EACL*, 2012.
 - [45] A. Noma, A. B. V. Graciano, R. M. Cesar Jr, L. A. Consularo, and I. Bloch. Interactive image segmentation by matching attributed relational graphs. *Pattern Recognition*, 45(3):1159 – 1179, 2012.
 - [46] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proc. NIPS*, 2015.
 - [47] C. Rhemann, C. Rother, J. Wang, M. Gelautz, P. Kohli, and P. Rott. A perceptually motivated online benchmark for image matting. In *Proc. CVPR*, 2009.
 - [48] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, 23(3): 309–314, 2004.
 - [49] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proc. ICLR*, 2014.
 - [50] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE T-PAMI*, 22(8):888–905, 2000.
 - [51] R. Shi, K. N. Ngan, S. Li, R. Paramesran, and H. Li. Visual quality evaluation of image object segmentation: subjective assessment and objective measure. *IEEE T-IP*, 24(12):5033–5045, 2015.
 - [52] R. Socher and F. F. Li. Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora. In *Proc. CVPR*, 2010.
 - [53] J. Soderberg and E. Kakogianni. Automatic tag generation for photos using contextual information and description logics. In *Proc. CBMI*, 2010.
 - [54] C. Szegedy, S. Reed, D. Erhan, and D. Anguelov. Scalable, high-quality object detection. *arXiv preprint*, (arXiv: 1412.1441), 2014.
 - [55] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, 2015.

-
- [56] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. In *Proc. IJCV*, 2013.
 - [57] V. Vezhnevets and V. Konouchine. Growcut: Interactive multi-label and image segmentation by cellular automata. In *Proc. Graphicon, Citeseer*, 2005.
 - [58] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proc. CVPR*, 2015.
 - [59] P. J. Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural Networks*, 1(4):339–356, 1988.
 - [60] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. ICML*, 2015.
 - [61] N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang. Deep interactive object selection. In *Proc. CVPR*, 2016.
 - [62] M. Yatskar, M. Galley, L. Vanderwende, and Zettlemoyer L. See no evil, say no evil: Description generation from densely labeled images. In *Proc. JCLCS*, 2014.
 - [63] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proc. CVPR*, 2016.
 - [64] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Proc. ECCV*, 2014.
 - [65] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. S. Torr. Conditional random fields as recurrent neural networks. *arXiv preprint*, (1502.03240), 2015.