Fixations Based Personal Target Objects Segmentation

Ran Shi
rshi@njust.edu.cn
School of Computer Science and Engineering,Nanjing University of Science and Technology
Nanjing, China

Gongyang Li
Weijie Wei
Zhi Liu
ligongyang@shu.edu.cn
codename1995@shu.edu.cn
liuzhisjtu@163.com
School of Communication and Information Engineering, Shanghai University
Shanghai, China

ABSTRACT

With the development of the eye-tracking technique, the fixation becomes an emergent interactive mode in many human-computer interaction study field. For a personal target objects segmentation task, although the fixation can be taken as a novel and more convenient interactive input, it induces a heavy ambiguity problem of the input's indication so that the segmentation quality is severely degraded. In this paper, to address this challenge, we develop an "extraction-to-fusion" strategy based iterative lightweight neural network, whose input is composed by an original image, a fixation map and a position map. Our neural network consists of two main parts: The first extraction part is a concise interlaced structure of standard convolution layers and progressively higher dilated convolution layers to better extract and integrate local and global features of target objects. The second fusion part is a convolutional long short-term memory component to refine the extracted features and store them. Depending on the iteration framework, current extracted features are refined by fusing them with stored features extracted in the previous iterations, which is a feature transmission mechanism in our neural network. Then, current improved segmentation result is generated to further adjust the fixation map and the position map in the next iteration. Thus, the ambiguity problem induced by the fixations can be alleviated. Experiments demonstrate better segmentation performance of our method and effectiveness of each part in our model.

CCS CONCEPTS

• **Computing methodologies** → **Image segmentation**; • **Human-centered computing** → **Human computer interaction (HCI)**.

MMAsia '20, March 7–9, 2021, Virtual Event, Singapore

ACM ISBN 978-1-4503-8308-0/21/03…$15.00
https://doi.org/10.1145/3444685.3446310

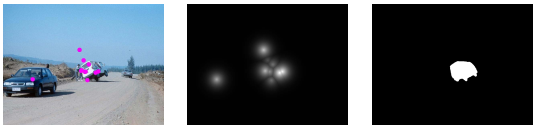
KEYWORDS

interactive segmentation, fixations, iteration

ACM Reference Format:
Ran Shi, Gongyang Li, Weijie Wei, and Zhi Liu. 2021. Fixations Based Personal Target Objects Segmentation. In *ACM Multimedia Asia (MMAsia '20), March 7–9, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3444685.3446310

1 INTRODUCTION

Object segmentation is one of challenging research topics in the image processing field. It aims at assigning a unique label to each pixel ("object" or "background") and plays an important role in object-aware image applications for their content understanding and manipulation, such as object-aware retrieval and cropping. Actually, different users read one image by different ways and have their own target objects. One user's target objects mean that they draw this user's main attention. Compared with the automatic segmentation [6] which segments the common objects only, interactive object segmentation [2, 8, 11, 15, 17, 25, 26] with manual inputs can fulfill the extraction of one user's personal target objects. So, the interactive object segmentation can make the object-aware image applications more individuation. Traditional interactive modes used in the interactive object segmentation are drawing some points [8, 11, 25], scribbles [2, 17] or bounding boxes [15, 26] in an image. All of them can be treated as explicit interactive modes. However, when users observe one image, their fixations indeed locate on certain regions of the image. Thanks to the development of the eye tracking technique, these fixations can already be recorded by an eye tracker device [12] in real time. Therefore, the fixation has large potential to be explored as an emergent and natural interactive mode. The advantages of this implicit mode is intuitional and to free our hands, so that it can be directly embedded in the subsequent applications without extra manual inputs. However, its main disadvantage is that the fixations provide ambiguous indications about objects and background. Since the fixations record the whole observation procedure, some fixations may locate in the background. Even more, there may be many objects obtaining the fixations in an image, but only some of them are one user's target



(a) (b) (c)

Figure 1: One example of the fixations, the fixation map and the corresponding target object. (a) The original image with the fixations indicated by the magenta dots,(b) The fixation map, (c) Ground truth of the target object.

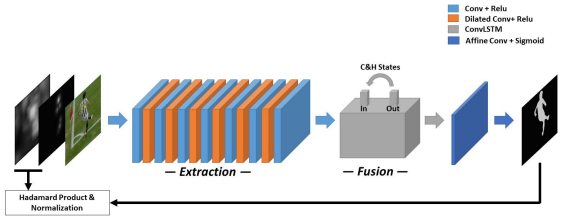


Figure 2: The architecture of our proposed neural network.

objects. So, it is hard to analyze the meanings of these fixations' indications. One example is shown in Figure.1, where the fixations are indicated by the magenta dots. Although there are fixations in the sky and the black car, the target object is the white car only. We can see that the fixations based personal target objects segmentation is quite different from the traditional explicit interaction modes providing definite indications about objects or background to direct the segmentation. Concretely, the points and scribbles with different labels are located in the part of the object and the background regions; the bounding box surrounding the object excludes the impossible object region. So, compared with the explicit interaction modes, the objects are more difficult to be segmented by the unlabelled fixations. Moreover, we do not only separate background and objects but also distinguish the target objects from non-target objects depending on the fixation information.

In this paper, our goal is to develop a personal target objects segmentation method using the fixation interactive mode. One critical problem is how to reduce the negative influence of the indication ambiguity induced by the fixations. We solve this problem by an extraction-to-fusion strategy based iterative neural network. The contributions of our work are as follows:

(1) Overall, we employ the iteration framework to gradually adjust inputs of the fixation map and the position map, and refine the extracted features. It can iteratively improve the segmentation result and make our neural network more lightweight.
(2) In the extraction part, we design a concise interlaced structure of the standard convolution layers and the progressively higher dilated convolution layers [27] to better extract and integrate the features of the local and the global contrast of the target objects.
(3) In the fusion part, we utilize the convolutional long short-term memory (ConvLSTM) [23] component to store and sufficiently fuse all features extracted in the whole iterations. Thus, the final segmentation result is generated depending on the features not only extracted in the final iteration but also the previous ones. It is different from the traditional iterative segmentation [15] that the features extracted in each iteration are less correlated.

This paper is organized as follows. Section II introduces some related works and Section III describes our fixations based iterative target object segmentation neural network in detail. Experimental results are presented in Section IV. We conclude our paper in Section V.

2 RELATED WORKS

Compared with the explicit interactive modes based object segmentation, fixation-based interactive methods have been studied less. In [11], it simply treats one drawing point equivalent to one fixation and use it as one center of the polar transformation, then combines monocular cues with motion and stereo to segment an object. Some methods attempt to infer the background information from the fixations [13, 16, 21] in some simple scenes. In [16], one image is segmented into several superpixels and these superpixels can be further divided into "object seed", "background seed" and "unknown region" according to the distribution of the fixations on them. In [13], the centroid of fixations and the mean distance of all fixations from the centroid are used to select the segmentation seeds. In [21], two aided saliency maps are used to estimate the background region. Following the above ideas, once the background cues can be obtained, the fixations based object segmentation is converted into the traditional interactive object segmentation by constructing object and background models [2]. Alternatively, [7] and [19] aim to establish a relationship between fixations of some users and their collective objects in one image. They extract their own designed hand-crafted features of the fixations' distribution on each object proposal [1], and then estimate a score to indicate the possibility of one object proposal belonging to the collective objects. Thus, the segmentation result can be generated according to the score ranking. So, these two methods convert the "segmentation" problem into the "selection" problem. Generally, there are two types of features used in these fixations based interactive object segmentation. One is used to distinguish fixations for different labels; The other is to describe object and background appearances. They can mutually improve each other rather than remaining independent. However, the main drawback of the methods mentioned above is that they consider these two types of features separately and the designed features are not good enough. Currently, convolutional neural networks

are widely used in traditional interactive object segmentations. [9, 25, 26] adopt Fully Convolutional Network [10] or Unet [14] as their basic frameworks, however they still rely on the drawing points with two different kinds of labels or the bounding box as their inputs. Using a different approach, [8] proposed a selection mechanism which makes it have potential to handle unlabelled drawing points. There are two coupled convolutional neural networks in [8], where the first one jointly extracts the features of the drawing points' information and the object appearance to synthesize a set of possible segmentation results and the second one selects the target objects from them.

3 THE PROPOSED METHOD

3.1 Input representation

Our proposed method is based on the neural network due to its powerful ability on features extraction and representing. The input of our neural network consists of an original image I, a fixation map FM and a position map PM. Each map plays its own role. Specifically, the fixation map as the interactive information and the position map as high level object information are combined to quickly locate the rough position of the target objects, and the original image can provide low level information to refine the target objects. Although one fixation is recorded by the eye tracker as a point in the image, it does not mean that the user merely looks at this location. So, a certain region around one fixation should be considered as an attention extent. Therefore, we convert the fixations into the fixation map to reflect the attention degrees of pixels. For one pixel j in the FM, its attention degree $FM(j)$ is estimated as:

$$FM(j) = \max_{k \in \Omega_{fix}} \left(\frac{T(k)/T_{max}}{1 + exp(D(j,k)/\tau)}\right) \tag{1}$$

where k is one fixation in the fixation set Ω_{fix}. There are two factors determining $FM(j)$: distance D and observation duration T. If j is closer to one fixation and the observation duration of this fixation itself is longer, it means that j also draws more attention. T_{max} indicates the longest observation duration of all fixations and τ controls the attention extent stimulated by the exponential function. τ is set to 20 in our experiments. $FM(j)$ is the maximum value among all pairs of j and k. After normalization and scaling $FM(j)$ to $[0, 255]$, the FM can be generated. The fixation map of Figure.1(a) is illustrated in Figure.1(b).

For the position map, we utilize feature maps "conv5_3" of the pretrained VGG-16 network [20] to compose it. These feature maps involve high level object semantic information which can be used to indicate rough positions of all potential objects in an image. In details, since there are 512 channels of "conv5_3", we select the maximum value of each channel and normalize them to compose our PM. The pixel with higher value in PM indicates its high possibility belonging to one object. Finally, we concatenate I, FM and PM as a five-channel input tensor.

3.2 Network Architecture

3.2.1 Iteration Framework. As mentioned above, our proposed method is under the iteration framework. In each iteration, our network generates a pixel-wise segmentation result S of the target objects. Then, this result is used to adjust FM and PM in the next iteration:

$$FM_t = FM \circ S_{t-1} \tag{2}$$

$$PM_t = PM \circ S_{t-1} \tag{3}$$

where t indicates the tth iteration and "$\circ$" is the Hadamard product. After the normalization, I, FM_t and PM_t are updated inputs for the tth iteration. By this adjustment, the fixation map and the position map are re-distributed. In the possible target objects regions indicated by S_{t-1}, the fixations' attention degree and the position possibilities are enhanced. Conversely, the fixations' attention degrees are reduced and the position possibilities are decreased in the impossible target objects regions. Thus, more accurate FM_t and PM_t can further assist the network to generate improved S_t in the next iteration.

3.2.2 Extraction Part. Our network consists of two main parts. The former extraction part is a convolutional neural network to extract the features of the target objects according to the current input. In the convolutional neural network, the standard convolution and the dilated convolution are two general convolution operations. The standard convolution is to extract features according to adjacent pixels. On the contrary, the dilated convolution extracts features using nonadjacent pixels with a dilation rate. The merit of the dilated convolution is to effectively enlarge its receptive field, while avoid the spatial information lost induced by reducing the image resolution [3].

In perspective of the personal target objects segmentation, the core is to explore the local contrast and the global contrast among the target objects and other regions [18]. Once the features of the local contrast and the global contrast can be properly extracted and well integrated, the ambiguity problem can also be alleviated accordingly. Corresponding to the convolutional neural network, the standard convolution and the dilated convolution can be treated to extract features of the local and the global contrast respectively due to their different receptive fields [27]. However, since the "global" is a relative concept, it is hard to estimate a very proper spatial extent for measuring the global contrast. In some convolutional neural networks [3, 27], they stack multiple dilated convolution layers in a row after standard convolution layers. In our opinion, although they can alleviate the problem of the spatial extent estimation, the local contrast and the global contrast cannot achieve a good balance because the global contrast with higher level features over dominates the contrast measure. In order to solve this problem, our neural network adopts a concise interlaced structure of the standard convolution layers and the progressively higher dilated convolution layers. We control the rates of the dilated convolutions to gradually increase their receptive fields. Thus, the local contrast and the global contrast in

multiple spatial extents can be mutually embedded closely from their low-level features to high-level features. It means that this concise interlaced structure can better extract and integrate the features of these two kinds of contrasts by roughly maintaining their equal importance in different spatial extents and in different feature levels throughout the whole features extraction.

3.2.3 Fusion Part. For the latter fusion part in our network, we introduce the ConvLSTM component [23]. Similar to traditional gated LSTM [4], the ConvLSTM uses the memory cells including the Cell state (C state) and the Hidden state (H state), and four gates i, f, c, o to control information flow. It extends traditional fully connected LSTM by substituting dot products with convolutional operations in the LSTM equations, which can preserve the spatial information of features [22]. Different from the traditional iteration based segmentation method [15] which only uses the previous segmentation result as the current input, our iteration framework also transmits the previous output states to the current iteration as the input states. It means that our method can use the ConvLSTM's strong ability to store the previous features and fuse them with current ones by the transmissions of the Cell state and the Hidden state to further refine the extracted features.

In terms of our model's architecture, the features transmission among iterations means that all features are correlated in the iterations and the effectiveness of their utilization are enhanced. Corresponding to the target objects segmentation, due to the interference of the ambiguity problem, it is difficult to generate a good segmentation result of the target objects by only once segmentation. However, even if the iterative segmentation, a simple iterative adjustment of the input may degrade the segmentation result rather than improving it. In our iterations, although the adjustments of FM and PM tend to make the segmentation result more precise, it is also possible that the iterative segmentation result becomes incomplete simultaneously. Then, the adjusted FM and PM may aggravate the incompletion of the segmentation results in the subsequent iterations. Actually, it is hard to achieve a perfect balance between the precision and the completeness. Considering this possibility, the ConvLSTM component is adopted to fuse the current features with transmitted previous ones by utilizing its memory mechanism during the iterations. It means that the current segmentation result is determined by the features extracted in two successive iterations to prevent over-segmentation. Consequently, the final segmentation result actually depends on all features extracted in our whole iterations, so it can achieve proper balance between the precision and the completeness.

Finally, the current segmentation result is generated by an affine convolution whose kernel size is 1×1 and a sigmoid activation function. Since the features extracted by the former part directly supply to the latter component, our network is an end-to-end architecture as shown in Figure. 2.

4 EXPERIMENTS

4.1 Dataset

To the best of our knowledge, there is no public dataset specifically designed for the fixations based personal target object segmentation. An only related one is OSIE dataset [24], which provides 15 users' fixations information (location and duration) of 700 images recorded by the eye tracker device (Eyelink 1000) and manual labeled masks of all objects in these images. Based on OSIE dataset, we re-forged it as an OSIE-Fixation based Personal Target Object Segmentation dataset (OSIE-FPTOS). For one user u and one image v , if one object in the image v can obtain more than Th_{uv} fixations of user u's, this object is chosen as this user' target object. The threshold Th_{uv} is calculate as $[N_{uv} \backslash N_R]$, where N_{uv} is the number of one user u's all fixations in the image v. If N_{uv} is less than a reference number N_R, the threshold is set to 1. By analyzing the numbers and the distributions of the fixations of all users in the whole dataset, we deliberately set N_R to 10, which is a proper trade-off between the numbers of one user's fixations and target objects. After the thresholding, we can generate one user u's target objects mask map for image v. We still call this mask map as ground truth. So, we can generate 15 different ground truths for one image according to different users' fixations. A tuple of one image, one user's fixations and corresponding ground truth is treated as one sample. Thus, in the OSIE-FPTOS, we randomly select 550 images with 8250 samples as a training set, 50 images with 750 samples as a cross validation set and 100 images with 1500 samples as a test set.

4.2 Implementation Details

For our proposed model, the original image, the fixation map and the position map are resized to 150×200. Each sample is iteratively trained five times. It means that the network is updated five times during one sample training. The initial values of Cell state and Hidden state in the ConvLSTM are set to 0 in the first iteration. In the tth iteration, given the ground truth G and the segmentation result S_t, we use the dice coefficient as the loss function:

$$L = 1 - \frac{2|G \cdot S_t|}{|G| + |S_t|} \tag{4}$$

The final binary segmentation result after five iterations is generated by a threshold as 0.5. The parameters of each type of the layer in our model are listed in Table.1. Especially, the Cell state and the Hidden state in the ConvLSTM are also set to four-channel tensors whose spatial dimensions are as same as those of the original image. We can see that our network is lightweight, whose number of the trainable parameters is about 0.8M only. Our model is trained using Adam [5], with single sample and learning rate 0.0001. Training proceeds for two stages. There are 20 epochs in each stage. In the second stage, we add the sum of all trainable variables as a regularization term in the loss function to further enhance the generalization ability of our model, where its

Table 1: The parameters of each type of the layer. M is the NO. of the layer.

Layer	Kernel	Rate	Outputs Num
Conv	3×3	1	64
Dilated Conv	3×3	$2^{M/2}$	64
ConvLSTM	3×3	1	4
Affine Conv	1×1	1	1

Table 2: Performances comparisons on OSIE-FPTOS test set.

Method	*Jaccard Index* ↑
GBOS	0.391
AVS	0.399
SOS	0.404
Unet : Iteration 1	0.578
Unet : Iteration 5	0.576
ISLD	0.592
ISLD_r : Iteration 1	0.613
ISLD_r : Iteration 5	0.609
Our(Stage 1): Iteration 1	0.612
Our(Stage 1): Iteration 5	0.632
Our(Stage 2): Iteration 1	0.625
Our(Stage 2): Iteration 5	**0.640**

weight is 0.00005. We validate the model on the cross validation set every epoch by *Jaccard Index* (i.e. *IoU*). The model which achieves the highest average *Jaccard Index* score on the cross validation set is chosen as the final model.

4.3 Overall Performance

In order to evaluate our proposed method's performance, we tested it on the OSIE-FPTOS test set. The average *Jaccard Index* (i.e. IoU) is used to indicate the segmentation result's accuracy. We evaluate our model's performances after different training stages, as well as different iterative times. Other related methods, i.e. collective objects segmentation using fixations (SOS [7], GBOS [19]), interactive object segmentation with unlabelled points (AVS [11], ISLD [8]) and retrained ISLD (ISLD_r) by OSIE-FPTOS training set, are compared against our methods. In addition, we also trained a wildly used segmentation network Unet [14] with our input as another baseline. Both of Unet and retrained ISLD_r are further involved in an iteration framework, which iteratively adjusts their inputs as our method does.

As shown in Table. 2, three hand-crafted features based methods GBOS, AVS and SOS cannot work on our topic, although their input modes are similar to ours. From the performance of Unet, we can see that since this general segmentation network does not design special strategy to handle the ambiguity problem, it cannot well segment the target objects. The ISLD especially after retrain, benefits from its selection mechanism to alleviate the ambiguity problem,

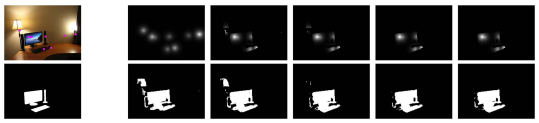


Figure 3: One example of the whole iteration procedure. The first column shows the original image with the fixations indicated by the magenta dots and the ground truth of the target objects. For other columns, the upper images are the fixation maps and the lower image are the corresponding segmentation results from the first iteration to the fifth iteration.

but its network with the stacking dilated convolution layers cannot properly measure and integrate the local and global contrasts of the target objects. Meanwhile, the iteration based Unet and ISLD_r show that the iterative adjustment of the inputs alone cannot guarantee the improvement of the segmentation result as mentioned above. Our method can outperform all other methods, especially adding the regularization term in the loss function. It demonstrates that our proposed neural network is reasonable and the extraction-to-fusion strategy can better match the iteration framework. The basic architecture of the interlaced convolution layers and the ConvLSTM in our model can guarantee the quality of the features of the target objects under the interference of the ambiguity problem. Moreover, according to the iterative segmentation result, the iteration framework can screen some fixations which do not locate in the target objects and also adjust the position map. In turn, the improved fixation map and position map, as well as the fused features in the previous iterations, can assist our model to generate better segmentation result in the further iteration. Figure. 3 shows all iterative fixation maps and corresponding segmentation results in the whole iteration procedure. We can see that the initial fixations sparsely locate and part of the lamp is mistakenly segmented out in the initial segmentation result due to the initial fixation map. Then, by the adjustments of the iterative results, the negative influence of some fixations which are not in the target objects' regions can be gradually removed accordingly and the segmentation result can also be further improved during the iterations. More ours and comparative segmentation results are illustrated in Figure. 4. Although there are diverse original images with various fixations' distributions, our method can generate better segmentation results of the corresponding target objects also in terms of visual quality. Especially, for two tough cases in the last two columns, our method can still roughly segment different target objects according to fixations' indications under the complicated illumination situation.

4.4 Ablation Study

In order to analyze the contribution of each part in our network, we perform an ablation study with two configurations by re-training two networks respectively following the same

Figure 4: Some examples of the segmentation results. The images from the first row to the fifth one are the original images with fixations indicated by color dots, our corresponding fixation maps, the ground truth of the target objects, the segmentation results generated by ISLD_r and the segmentation results generated by our best model reported in Table. 2.

Table 3: The ablation study of two different configurations of our model.

Configuration	*Jaccard Index* ↑
Our / Interlaced Convolutions: Iteration 5	0.616
Our / States Transmission: Iteration 5	0.618

procedure in the first training stage. The performances of these two configurations are shown in Table. 3.

4.4.1 The Interlaced Convolutions. The interlaced standard and dilated convolutions are re-ordered in this study. It means that the six dilated convolution layers follow the six standard convolution layers. As shown in the first row of Figure 5, the cloud is segmented out because our model without the interlaced structure cannot well balance the global contrast and the local contrast. Moreover, from Table. 3, we can see that our model without the interlaced structure makes its average *Jaccard Index* drop to 0.616. It demonstrates that the interlaced convolutions structure as the extraction part is useful to improve the segmentation result by extracting effective features to measure the local contrast and the global contrast of the personal target objects.

4.4.2 The States Transmission. In this study, all states are set to 0 in each iteration, so there is no features transmission in the ConvLSTM component. One pair of compared results are shown in the second row of Figure 5. We can see that the states transmission tends to guarantee the completeness of the target object. In terms of Table. 3, without the states transmission in the iterations, the performance drops to 0.618. The comparison presents that the ConvLSTM component as the fusion part is important to refine features by fusing them in the whole iterations. It makes our iteration framework more powerful and effective by not only the input adjustment but also the features fusion.

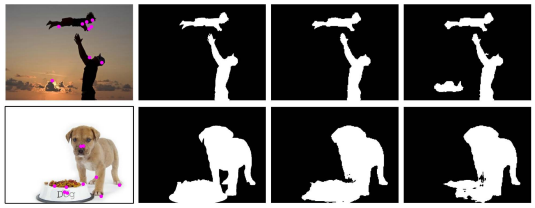


Figure 5: Two comparisons in the ablation study. The first row and the second row correspond to the two configurations of the interlaced convolutions and the states transmission respectively. The images from left to right are original images with fixations, ground truths, segmentation results of our proposed method and the ablation study.

4.5 Discussion

Although our method can achieve better performance, the fixations based personal target objects segmentation is still one tough topic. Even if the fixations can be screened by our proposed method to some extent, there are still two main problems for the segmentation accuracy: one is complicated semantic problem; the other is the size of the target object. On one hand, for the complicated semantic problem, it is induced by one object including prominent semantic regions. For example, a face is the prominent semantic information which can draw most fixations, but it is confused to predict that the target object is the face only or the whole person. On the other hand, the numbers and distributions of the fixations are quite different in the target objects with various sizes. It further aggravates the ambiguity problem and degrade the segmentation quality. So, the solutions of these two problems may be possible directions for improving our work in future.

5 CONCLUSION

In this paper, we explore the fixation as the emergent interactive mode to develop our iterative neural network for the personal target objects segmentation. The interlaced structure of the convolution layers and the ConvLSTM component compose the basic architecture of our neural network. Moreover, we take advantages of the iterative framework and the extraction-to-fusion strategy to handle the ambiguity problem induced by the fixations and improve the segmentation result mutually. The iteration framework also reduces the burden of our model to make it more effective and lightweight. The proposed method can be applied in the interactive object-aware image processing applications to fulfill personalized services.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grants 61801219 and 61771301.

REFERENCES

[1] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. 2014. Multiscale combinatorial grouping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 328–335.

[2] Yuri Y Boykov and Marie-Pierre Jolly. 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 1. IEEE, 105–112.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* (2017).

[4] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[5] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[6] Hongliang Li and King N Ngan. 2011. Learning to extract focused objects from low dof images. *IEEE transactions on circuits and systems for video technology* 21, 11 (2011), 1571–1580.

[7] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. 2014. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 280–287.

[8] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. 2018. Interactive image segmentation with latent diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 577–585.

[9] J. Liew, Y. Wei, W. Xiong, S. Ong, and J. Feng. 2017. Regional Interactive Image Segmentation Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2746–2754.

[10] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[11] Ajay Mishra, Yiannis Aloimonos, and Cheong Loong Fah. 2009. Active segmentation with fixation. In *2009 IEEE 12th international conference on computer vision*. IEEE, 468–475.

[12] Manoranjan Paul and Md Musfequs Salehin. 2018. Spatial and Motion Saliency Prediction Method using Eye Tracker Data for Video Summarization. *IEEE Transactions on Circuits and Systems for Video Technology* (2018).

[13] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. 2010. An Eye Fixation Database for Saliency Detection in Images. In *Computer Vision – ECCV 2010*. Springer Berlin Heidelberg, Berlin, Heidelberg, 30–43.

[14] O. Ronneberger, P.Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI) (LNCS)*, Vol. 9351. Springer, 234–241.

[15] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. Grabcut: Interactive foreground extraction using iterated graph cuts. In *ACM transactions on graphics (TOG)*, Vol. 23. ACM, 309–314.

[16] Anthony Santella, Maneesh Agrawala, Doug DeCarlo, David Salesin, and Michael Cohen. 2006. Gaze-based interaction for semi-automatic photo cropping. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 771–780.

[17] Jianbing Shen, Yunfan Du, and Xuelong Li. 2014. Interactive segmentation using constrained Laplacian optimization. *IEEE Transactions on Circuits and Systems for Video Technology* 24, 7 (2014), 1088–1100.

[18] Ran Shi, King Ngi Ngan, Songnan Li, and Hongliang Li. 2018. Interactive object segmentation in two phases. *Signal Processing: Image Communication* 65 (2018), 107–114.

[19] Ran Shi, Ngi King Ngan, and Hongliang Li. 2017. Gaze-based object segmentation. *IEEE Signal Processing Letters* 24, 10 (2017), 1493–1497.

[20] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[21] Xiaoliang Tian and Cheolkon Jung. 2015. Point-cut: Fixation point-based image segmentation using random walk model. In *Proceedings of the IEEE International Conference on Image Processing*. IEEE, 2125–2129.

[22] Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. 2018. Salient object detection driven by fixation prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1711–1720.

[23] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*. 802–810.

[24] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. 2014. Predicting human gaze beyond pixels. *Journal of vision* 14, 1 (2014), 28–28.

[25] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. 2016. Deep interactive object selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, 373–381.

[26] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. 2017. Deep grabcut for object selection. *arXiv preprint arXiv:1707.00243* (2017).

[27] Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* (2015).