Object Aware Contrastive Prior for Interactive Image Segmentation

Praful Mathur, Shashi Kumar Parwani, Mrinmoy Sen,
Roopa Sheshadri, Aman Sharma
Samsung R&D Institute India - Bangalore
{p.mathur, p.shashi, mrinmoy.sen, roopa, aman.sharma}@samsung.com

Abstract

Interactive Image Segmentation is a process of separating a user selected object from the background. This task requires building an effective class-agnostic segmentation model that performs well even on unseen categories. To achieve good accuracy with limited training dataset, it is important that the model has robust prior understanding of features of similar class objects. The model should also have good distinguishing capabilities of foreground objects with the background. In this paper, we propose Object Aware Click Embeddings (OACE) that represents user click aware foreground object features. OACE is obtained based on a prior network trained using the Contrastive Learning paradigm. The single-click object selection accuracy of our base interactive segmentation network is vastly improved with the OACE input. Additionally, we propose a Multi-Stage fusion approach to better utilize user click information. With the proposed method, we outperform existing state-of-the-art approaches by 21% on publicly available test-sets for click-based Interactive Image Segmentation.

1. Introduction

Interactive image segmentation methods allow users to segment any object in an image based on user inputs. User inputs can be provided as clicks, drawing contour around objects or through lines and scribbles. Click based interactions are more intuitive and provide a simple way to mark the object. In this paper, we focus on click-based interactive segmentation.

Conventional interactive segmentation methods rely on the concept of iteratively providing positive or negative clicks to select the unselected region from the first click or deselect the over-selected region. This problem of under-segmentation and over-segmentation is natural in complex class-agnostic segmentation paradigms like interactive segmentation. It is primarily because the segmentation model focuses on determining decision boundaries to segment object without explicit understanding of intra-object feature similarities and object-background feature dissimilarities.

The need for more than one click to segment an object completely and correctly, degrades user-experience. Moreover, in conventional interactive segmentation approaches, the segmentation mask output changes vastly based on user click locations. The reason for varied output is the method used to represent user clicks in the interactive segmentation network. Conventional methods [20, 21] use Euclidean distance transforms or disk of fixed radius to represent user click location. However, distance transform based or disk based representations vary drastically with user click thereby resulting in different segmentation outputs.

Inspired by the advances in contrastive learning, to improve single-click accuracy of interactive segmentation models, we propose Object Aware Click Embeddings (**OACE**) as an additional input to the interactive segmentation model. A contrastive prior network is proposed, that uses the image and the user click to generate OACE (Figure 1 (a)). The prior network is trained in contrastive fashion to maximize the similarity between user clicked intra-object features and minimize the similarity between features of object and background regions. OACE provides two-fold advantage: 1.) It represents a novel way to learn object-aware user click information with increased robustness for different touch-points on an object; and 2.) It comprises of rich foreground features that are distinguished from the background features thus facilitating learning of class-agnostic segmentation masks. In addition, we propose **MSFNet**, a novel interactive segmentation network that incorporates OACE as an input to the network and uses the Multi-Stage Fusion module to inject the user click information at multiple stages to improve segmentation accuracy (Figure 1 (b)). The major contributions of the paper are as follows:

1.) We propose OACE, obtained using a contrastive learning framework to represent user click aware foreground object features that are robust towards different user click locations on an object.

2.) We propose MSFNet that utilizes OACE and Multi-Stage Fusion module to outperform existing state-of-the-art methods by 21% on seen and unseen object categories on

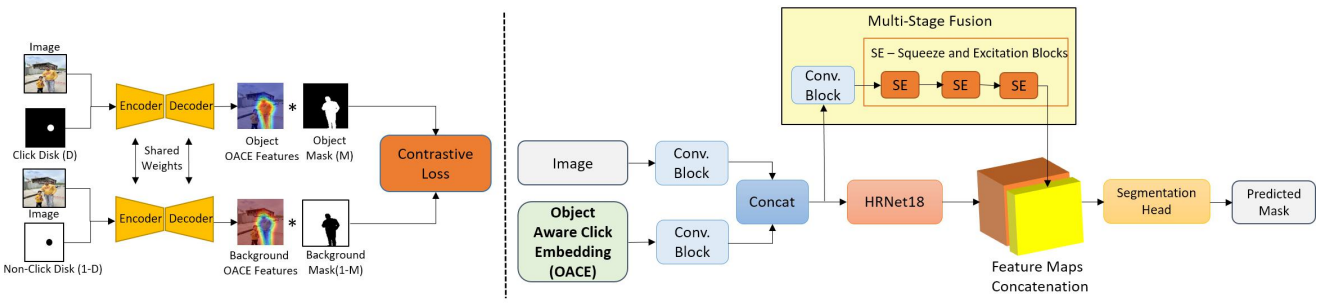

Figure 1. (a) Object Aware Click Embeddings (OACE), (b) Proposed MSFNet Architecture for Interactive Segmentation.

publicly available test datasets.

3.) The proposed framework is efficient and lightweight and is suitable for deployment on resource constrained embedded devices.

2. Related Works

Early works [10, 18] in interactive segmentation used optimization based techniques to solve the problem. With the advent of deep learning, data-driven methods like DIOS [25] that leverages large-scale datasets were proposed that improved the accuracy significantly. Jang *et al*. introduced Backpropagation Refinement Scheme (BRS) [8] that adopts an iterative optimization approach to refine the outputs based on the user-interaction maps. Sofiiuk *et al*. [20] improved the computational cost of this approach by adopting the technique in the intermediate layers of the network in the feature space. Recently, many coarse-to-fine approaches like RITM [21] have been proposed, that progressively improve the results based on the previous output and new user clicks. Chen *et al*. [3] improved the accuracy and efficiency of such coarse-to-fine methods by refining local patches and using morphology analysis to change the predicted mask only in the vicinity of the subsequent user clicks. Most of these works have two major limitations: 1.) Low single-click object selection accuracy thereby relying on iterative method of providing positive or negative clicks to correct the segmentation; and 2.) High variations in segmentation output based on user click location. Positive click refers to the user click made inside object region to select the object or select any undersegmented part of the object. Negative click refers to the user click made outside object region to deselect the oversegmented region.

In this work, we design object aware click embeddings (OACE) to represent user click. OACE eliminates the concept of negative clicks and utilize only single-positive click to generate object aware click representation. Moreover, OACE representation is robust towards different user click locations thus providing similar segmentation outputs irrespective of different user click locations.

Recently, a foundation model SAM [11] was proposed for a variety of image segmentation related tasks including interactive segmentation that has achieved state-of-the-art performance. While the capabilities of foundation models like SAM are remarkable, they are not ideal for all applications due to their high computation cost and size ($>$ 600M parameters). In this work, we aim to design an efficient and practical network for interactive segmentation that can be deployed in resource constrained environments like embedded devices.

Contrastive learning is a self-supervised machine learning paradigm where data samples are contrasted against each other so that samples from the same distribution are near to each other in the latent space and samples from different distribution are separated in the latent space. Some prior works [24, 26] have proposed the use of contrastive learning for dense prediction tasks like image segmentation. In [23] the authors demonstrate that adopting contrastive learning, where a contrastive loss is formulated to increase intra-class feature similarity and decrease inter-class feature similarity, improve the performance of object segmentation on novel categories. Contrastive Learning has also gained popularity in generative AI task like text-to-image generation models based on CLIP (Contrastive Language Image Pre-training) [16] embeddings wherein the text-encoder and image-encoder are trained in a contrastive fashion.

3. Proposed Method

The proposed framework has two main components: (a) Contrastive Prior network to generate object aware click embeddings (**OACE**) and (b) Multi-Stage Fusion Interactive Segmentation Network (**MSFNet**).

3.1. Contrastive Prior Network for OACE

The network comprises of a fully convolutional encoder network based on HRNetV2-W18-Small-v2 [22] backbone. The network is trained on labeled (image-object mask) dataset and simulated user clicks inside the object mask region. User clicks are represented in the form of bi-

nary disks with radius of 5 pixels. The image stacked with click disk forms the input to the network.

To train the network in contrastive fashion, we perform two forward passes and a single backward pass for weight updation. As depicted in Figure 1 (a), the first forward pass takes the image and the user click, represented in the form of a click disk (D) as input and the second forward pass takes the image and the non-click region represented in the form of a non-click disk (1-D). The latent space outputs from two forward passes are then elementwise multiplied with object mask (M) and background mask (1-M) respectively to extract the object and background features. This ensures that the first forward pass representing user click region, focuses on foreground object features and the second forward pass focuses on background features.

During training, random patches are selected from object features and background features obtained from two forward passes respectively. Let $P_o = \{p1_o, p2_o, \dots pn_o\}$ denote the set of object features and $P_b = \{p1_b, p2_b, \dots pm_b\}$ denote the set of background features. The loss function L tries to maximize the cosine similarity ϕ between intra-object and intra-background features and minimize the similarity between object-background features.

$$L = -\sum_{i=0,j=0}^{n} \phi(pi_o, pj_o) - \sum_{i=0,j=0}^{m} \phi(pi_b, pj_b) + \sum_{i=0,j=0}^{n,m} \phi(pi_o, pj_b) \quad (1)$$

The first term in the loss function ensures that foreground objects belonging to similar classes have similar representation in the latent space. This method of learning latent representations helps the network to better understand objects and its features resulting in highly accurate segmentation even on unseen object categories. The second term in the loss function ensures to maximize the similarity between background regions. The third term helps the network to distinguish foreground and background regions in latent representation.

The contrastive prior network is trained as explained above and the object aware click embeddings (OACE) obtained from prior network is used to train MSFNet. Visualizations of OACE are depicted in Figure 2.

3.2. Multi-Stage Fusion Interactive Segmentation Network (MSFNet)

As shown in Figure 1 (b), for MSFNet we use the standard HRNet18 backbone with a CNN based Segmentation Head. The input to this network is the image and OACE. Note that we used click disk based representation of user input only to train OACE prior network and our segmentation network MSFNet uses image and OACE as inputs.

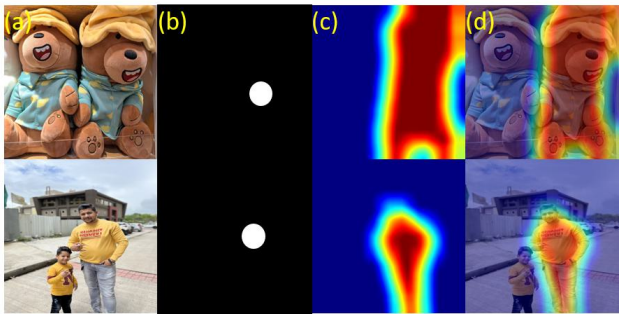

Figure 2. (a) Input Image, (b) User click disk, (c) Object aware click embedding (OACE), (d) Overlaid OACE on image.

There are two methods to fuse the two inputs (i.e. image and OACE) before passing to network backbone. The first method is known as Early Fusion where both inputs are first concatenated and processed by a convolution block, the output of which is fed to the backbone. Second method is known as Late Fusion where both inputs are processed separately by two convolution blocks and then concatenated and passed to backbone. Authors in [21] have shown that late fusion techniques works better in interactive segmentation tasks. In designing MSFNet architecture, we used late fusion method to fuse Image and OACE.

To prevent dilution of the user click information represented through OACE during CNN processing, we propose multi-stage fusion (MSF) module. The core idea is to fuse the OACE information at multiple locations in the network. MSF module comprises of a convolution block followed by three Squeeze-and-Excitation (SE) inception blocks. An SE-Inception block re-calibrates each channel feature adaptively by computing interdependencies explicitly between channels. Output of MSF module is fused with HRNet18 features before the segmentation head. The segmentation head comprises of upsampling and convolutional blocks resulting in output segmentation mask.

3.3. Training Datasets

The majority of Interactive Segmentation algorithms are trained on PASCAL and SBD. Together these two datasets contain thousands of images with annotated masks. However, as discussed in [4] these datasets are imprecise and may result in poor prediction quality. Also, [21] points out the limitation on the variety of predictable classes of these datasets. Extensive segmentation datasets like OpenImages [12], LVIS [5] and COCO [13] are available. These datasets have a wide range of labeled examples as well as a large variety of labeled classes. LVIS dataset contains approximately 100k images with 1.2M instance-level masks, OpenImages contains around 944k images with 2.6M instance-level masks and the COCO dataset contains 118k images with 1.2M instance-level masks. For our experiments, we

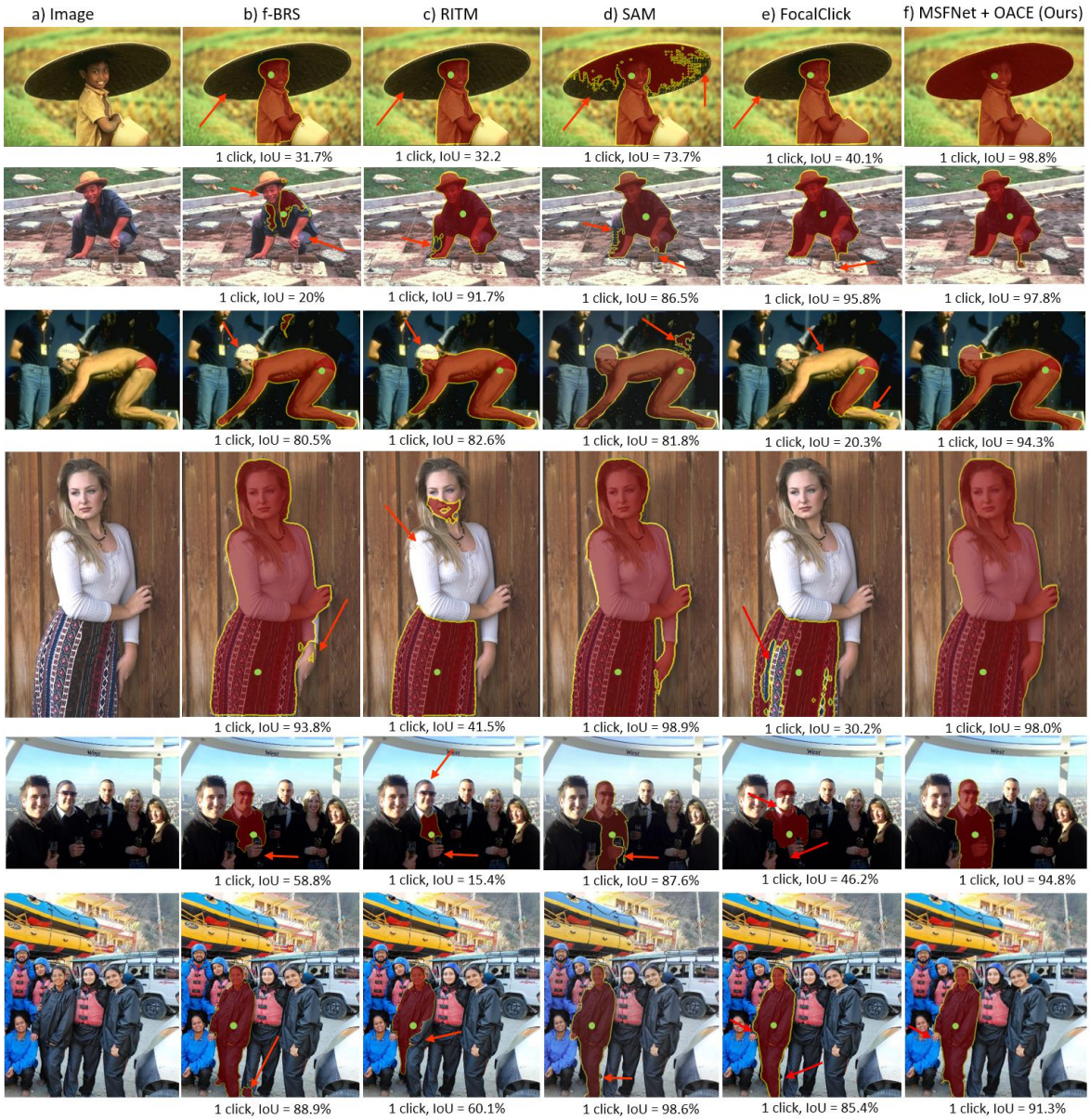

Figure 3. Qualitative comparison of single-click segmentation accuracy of different interactive segmentation methods. The green dot on each image denotes the user click location. Red arrows point to the erroneous regions.

use a combination of COCO+LVIS as proposed in [21].

3.4. Training Paradigm

For training prior network for OACE, we simulate user clicks on-the-fly during training. User clicks are simulated based on random sampling strategy that samples a point inside the object based on ground-truth mask region. As described in subsection 3.1, we used HRNetV2-W18-Small-v2 backbone followed by decoder block comprising of three upsampling and convolutional blocks. Input images are resized to the size of 512 $\times$ 512 and concatenated with user click disk. The dimensions of output embedding is same as the input image. The network is lightweight and comprises of 5.2M parameters. OACE prior network is trained first using COCO+LVIS dataset with a batch size of 64 for 210 epochs. Once the training is complete, the network's weights are frozen and the embeddings obtained using OACE prior network is used as an input to train MSFNet.

We train the MSFNet to minimize Normalized Focal loss [19]. NFL handles class imbalance problems by assigning more weight to erroneous regions, concurrently the

total gradient of NFL doesn't fade over time. It can be formalized as:

$$NFL(k,l,\hat{M}) = -\frac{1}{P(\hat{M})}(1-p_{k,l})^{\gamma}\log p_{k,l} \quad (2)$$

$$P(\hat{M}) = \sum_{k,l}(1-p_{k,l})^{\gamma} \quad (3)$$

Here $\hat{M}$ denote the output of the network and $p_{k,l}$ denotes the confidence of prediction at the point (k,l).
Network is trained using Adam optimizer and a batch size of 64. The model is trained with a learning rate of 5×10^{-4} for 210 epochs . For the HRNet18 backbone in MSFNet, the network weights are initialized from a model pre-trained on ImageNet. MSFNet comprises of 10 million parameters. Horizontal flipping, brightness and contrast shifts are used as augmentations.

4. Results

4.1. Evaluation Dataset

We evaluate our methods on four publicly available datasets that are commonly used for benchmarking Interactive Image segmentation algorithms.

1. The GrabCut [17] dataset contains 50 images with a single object mask for each image.

2. The Berkeley [14] is a 96 image dataset with 100 ground truth masks.

3. The SBD [6] dataset is divided into training set of 8498 images and a validation set of 2820 images. We use the validation set which contains 6671 instance-level masks from 2820 images.

4. The DAVIS [15] dataset is used for benchmarking video object segmentation algorithms. For our evaluation we use 345 randomly sampled frames from the videos sequences that was proposed in [9].

4.2. Evaluation Metrics

We evaluate our method on two metrics. First, the standard mean Number of Clicks (NoC@X) that is required to achieve X% Intersection over Union (IoU) between predicted segmentation mask and ground truth segmentation mask. For example, NoC@90 represent the number of clicks required to achieve the IoU of 90% . Lower value of NoC@X is better. Secondly, the average Intersection over Union for a single user click (denoted as mIoU@1) is used for evaluation. mIoU@1 represents single click accuracy of interactive segmentation model and higher value of mIoU@1 signifies superior performance.

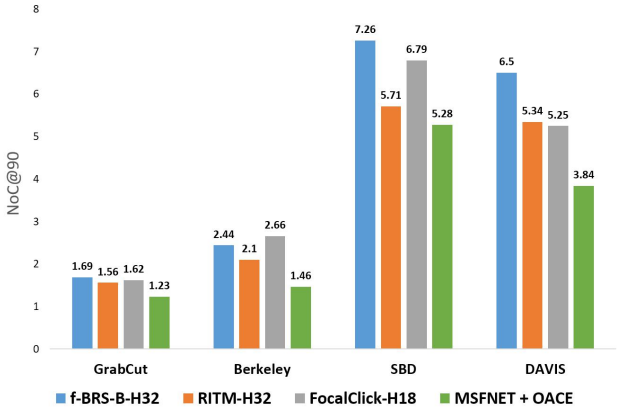

Figure 4. Quantitative comparison with existing interactive segmentation methods. NoC@90 refers to number of clicks required to achieve 90% of mean Intersection over Union (mIoU).

4.3. Qualitative and Quantitative Results

Qualitative comparison of our method (MSFNet+OACE) with existing interactive segmentation methods is presented in Figure 3. Figure 4 presents the quantitative comparison of NoC@90 with other methods. The NoC@90 numbers shows that proposed method reduce the number of clicks required to achieve 90% accuracy by **21%** as compared to the present state-of-the-art method RITM-H32, thereby enhancing the performance.

To evaluate single click mIoU, fixed set of user click points are used for all methods. As shown in Table 1, the mIoU@1 numbers highlights the effectiveness of our proposed light-weight model against the existing heavier networks. We evaluated Segment Anything Model (SAM) [11] trained on SA-1 dataset comprising of 11 million images and 1 billion instance-level masks. Our proposed method is comparable with SAM in terms of mIoU numbers on test sets despite having $40\times$ lesser parameters than SAM and trained on a smaller dataset. Our method also outperforms RITM-H32 model that is based on the heavier HRNet32 backbone. Evaluation on unseen categories in the test sets showcase the generalization ability of the proposed method.

5. Ablation Study

5.1. Network Architecture Ablation

In subsection 3.1 and subsection 3.2, we presented our proposed architecture for OACE prior network and MSFNet. Conventional approaches like [21] based on HRNet backbone directly takes as input an image and a user click represented as binary disk. Proposed framework has two major novelties. Firstly, addition of Multi-stage fusion blocks and secondly, utilizing OACE as input to rep-

Method	#Params	Training Dataset	mIoU@1		
			Berkeley	GrabCut	DAVIS
f-BRS-B (HRNet-32)	30.9M	COCO + LVIS	80.1	84.2	74.1
RITM (HRNet-18)	10.03M	COCO + LVIS	83.2	88.3	71.2
RITM (HRNet-32)	30.2M	COCO + LVIS	85.4	89.9	73.6
FocalClick (HRNet-18s)	4.22M	COCO + LVIS	81.1	85.4	76.32
SAM	632M	SA-1B	89.6	93.1	**84.6**
MSFNet+OACE (Ours)	15.2M	COCO + LVIS	**91.8**	**93.9**	80.2

Table 1. Comparison of click based interactive segmentation methods in terms of number of parameters and mIoU on different test sets.

	Berkeley	GrabCut	DAVIS
Baseline	82.7	87.9	70.2
MSFNet	85.6	89.7	73.6
MSFNet+OACE	**91.8**	**93.9**	**80.2**

Table 2. Comparison of mIoU@1 for different network architectures.

resent object aware user click locations. We ablated on these two changes in our network. We created a baseline architecture from proposed MSFNet by removing Multi-stage fusion block. We define this as Baseline. We compared the Baseline with MSFNet designed to take binary click disk based user input. Next we draw comparisons with MSFNet+OACE. Table 2 presents the details of mIoU with single click across Baseline, MSFNet and MSFNet+OACE.

Multi-stage fusion module designed using squeeze and excitation blocks [7] provides channel level attention that helps in improving the segmentation accuracy by 3% over baseline. Utilizing OACE embeddings as input in MSFNet improves the accuracy by 10% over baseline. In addition to object attention, OACE also provides good distinguishing capabilities between foreground and background regions. Additionally, OACE embeddings helps the network to identify similar class objects thus helping network to perform better on object categories with limited representation in training dataset.

5.2. Effectiveness of OACE towards touch point locations

One major limitation in existing interactive segmentation methods is they produce different segmentation outputs for different user click locations. The error is highest when the user click locations are near the boundary of the objects. The primary reason for this error is due to disk based representation or distance transform based representation of user clicks. Disk based representation or distance transform based representation changes vastly when user click location changes for an object. This vast change in input representation leads to vast change in segmentation network's output.

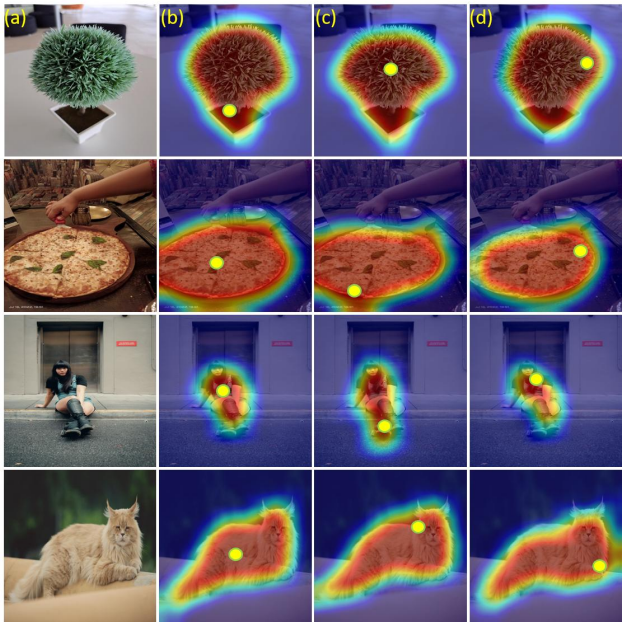

Figure 5. Visualization of OACE with different user click points.

Our proposed method solves this problem by utilizing OACE as inputs to interactive segmentation network. OACE represents object aware features that are distinguished from background features. The prior network trained to generate OACE is robust towards variations in user click locations and generate similar OACE representations for different click locations as depicted in Figure 5. This leads to similar OACE embeddings for different click locations on an object. Using OACE as input for interactive segmentation network leads to generation of consistent segmentation output despite variations in user click locations as shown in Figure 6.

5.3. Effectiveness of OACE for limited object classes

LVIS [5], COCO [13] authors have categorized the distribution of object categories into three classes namely rare, common and frequent based on the frequency of occurrence of object categories. LVIS dataset has over 40% of the cate-

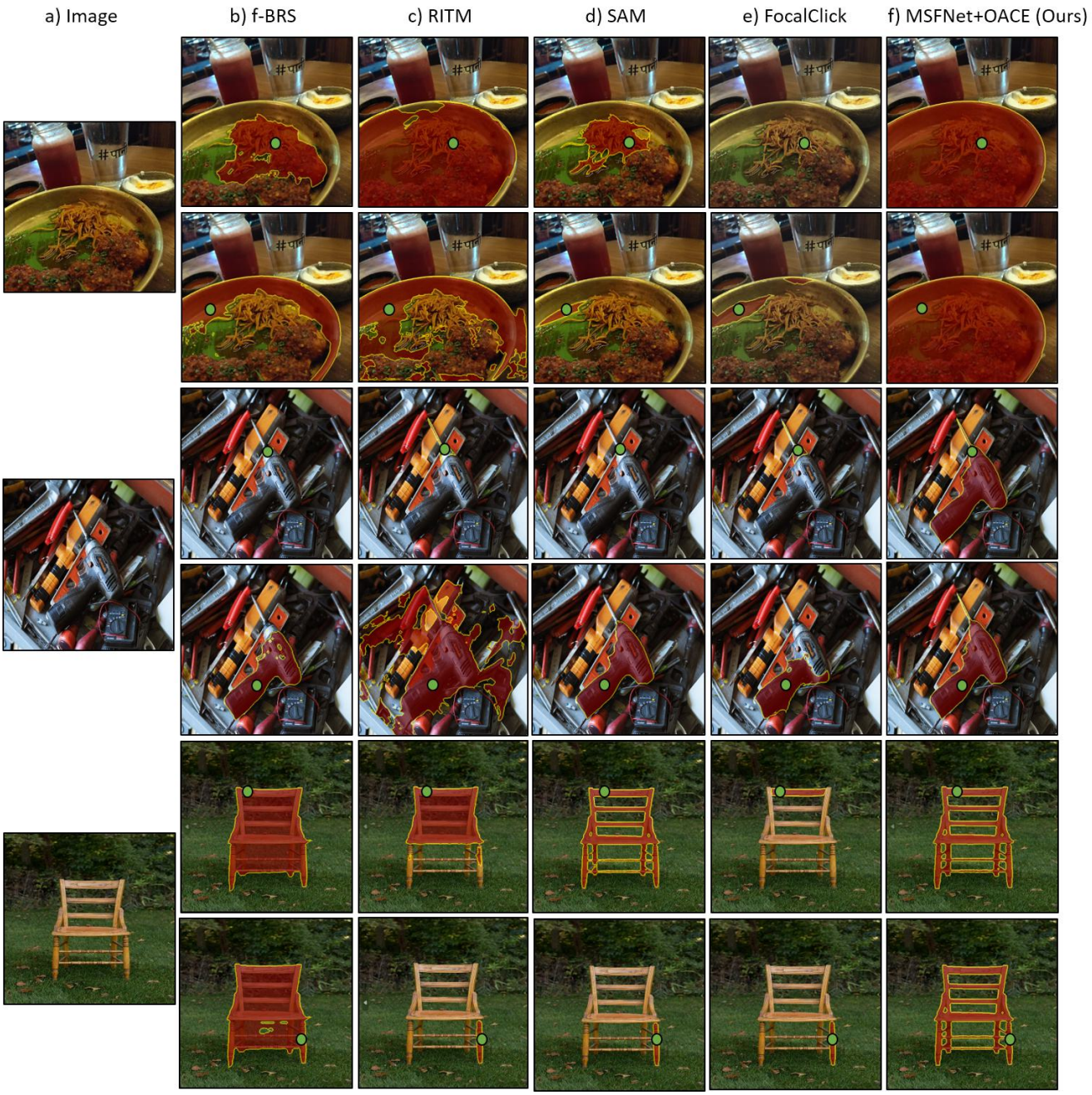

Figure 6. Qualitative comparison of different interactive segmentation methods in terms of segmentation output for different user click locations. The green dot on each image denotes the user click location.

gories marked as rare occurring out of a total of about 1000 object categories in training set. Categories like pottery, scissors, goldfish, drones, appliances like toaster, hair dryers are rare occurring object categories in LVIS dataset. Person, pet like cats and dogs, food items like pizza, vehicles like bicycle and cars come under frequent occurring object categories.

Most of the segmentation networks trained on LVIS, COCO datasets inherently gets biased towards frequent categories and tend to perform poorly on rare occurring categories in testing phase. One common method to tackle this is by resampling rare occurring object categories by repeating them in training set. However, authors in [2] have shown that such approaches do not fare well for object de-

	Rare Object Test-set	Frequent Object Test-set
MSFNet	68.8	88.4
MSFNet+OACE	**82.4**	**94.3**

Table 3. Comparison of mIoU@1 on rare category object and frequent category object custom test set.

tection or object segmentation tasks. Resampling of rare object categories leads to severe over-fitting thereby deteriorating the performance of segmentation networks.

Conventional class agnostic segmentation networks learns to separate foreground objects from background without explicit understanding of features of similar class objects. Also, there is no additional constraint that distinguishes the foreground object features from the background. These limitations of conventional approaches leads to poor segmentation accuracy on objects that have limited or no representation in the training dataset. On the contrary, OACE comprises of rich object aware features that helps the class agnostic interactive segmentation network to develop an explicit understanding of similar class object features. Additionally, OACE provides capabilities to explicitly distinguish foreground and background features, thus easing the task of interactive segmentation network.

In order to prove the effectiveness of OACE on limited and unseen object categories, we compared the performance of MSFNet designed with conventional input method using binary click disk; and MSFNet with OACE based input. Figure 7 showcases that MSFNet+OACE has better object selection accuracy on rare category objects. MSFNet trained with OACE input has better foreground object - background separation capabilities even in cases where foreground object and background have similar colors and textures.

To further evaluate the efficacy of OACE on limited object categories, we constructed two custom test sets with rare occurring objects and frequently occurring objects respectively. Each test set comprise of 100 images procured and labeled with manual efforts. The rare category object test images and frequent category object test images comprise of objects that are marked as rare and frequent in LVIS train set, respectively. Table 3 presents the mIoU numbers of MSFNet model trained without OACE and with OACE input on rare category object test set and frequent category object test set. OACE input results in about 20% improvement in rare object segmentation accuracy whereas the frequent object segmentation accuracy is improved by 7%.

6. Discussion

The proposed framework (MSFNet+OACE) is a lightweight model with 15.2M parameters and a model size of about 15 MB with TFLite [1] int8 quantized format. The proposed framework takes about 25ms to load and 250ms to infer (including preprocessing and model runtime) when evaluated on a modern day flagship level smartphone - Samsung's Galaxy S23, thus proving the efficiency of the proposed framework.

Figure 7. (a) Input Image with rare category object, (b) Segmentation output of MSFNet without OACE, (c) Segmentation output of MSFNet with OACE.

The proposed framework overcomes the shortcomings of existing works by improving single click object selection accuracy thereby eliminating the need of iteratively providing positive or negative clicks to correct the segmentation. Moreover, the proposed method provides robustness towards different user click locations and generate similar segmentation output for different user click location on an object. However, one open challenge is to segment very thin objects accurately. As the proposed network processes the image on low resolution (512×512), thin parts of the objects like flying hair strands gets missed from segmentation. High resolution post-processing or matting based solutions can be explored in future to improve thin object selection.

7. Conclusion

This paper presents a novel method to generate object aware click embeddings (OACE). Additionally, the proposed MSFNet uses OACE inputs to significantly improve the single-click accuracy of interactive image segmentation. We demonstrate that with limited training data for limited seen categories, our class-agnostic segmentation framework achieves good performance even on unseen category objects. The proposed framework is lightweight and can be deployed in resource constrained environments like embedded devices.

References

[1] TensorFlow Lite. https://www.tensorflow.org/lite/. 8

[2] Nadine Chang, Zhiding Yu, Yu-Xiong Wang, Animashree Anandkumar, Sanja Fidler, and Jose M Alvarez. Image-level or object-level? a tale of two resampling strategies for long-tailed detection. In *International conference on machine learning*, pages 1463–1472. PMLR, 2021. 7

[3] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299, 2022. 2

[4] Marco Forte, Brian Price, Scott Cohen, Ning Xu, and François Pitié. Getting to 99% accuracy in interactive segmentation. *arXiv preprint arXiv:2003.07932*, 2020. 3

[5] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5356–5364, 2019. 3, 6

[6] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998. IEEE, 2011. 5

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 6

[8] Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5292–5301, 2019. 2

[9] Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5297–5306, 2019. 5

[10] Tae Hoon Kim, Kyoung Mu Lee, and Sang Uk Lee. Nonparametric higher-order learning for interactive segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3201–3208, 2010. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 2, 5

[12] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, pages 1–26, 2020. 3

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 6

[14] Kevin McGuinness and Noel E O'connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, 2010. 5

[15] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 5

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2

[17] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 5

[18] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "grabcut": Interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3):309–314, aug 2004. 2

[19] Konstantin Sofiiuk, Olga Barinova, and Anton Konushin. Adaptis: Adaptive instance selection network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7355–7363, 2019. 4

[20] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. F-brs: Rethinking backpropagating refinement for interactive segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8620–8629, 2020. 1, 2

[21] Konstantin Sofiiuk, Ilya A. Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145, 2022. 1, 2, 3, 4, 5

[22] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 2

[23] Xuehui Wang, Kai Zhao, Ruixin Zhang, Shouhong Ding, Yan Wang, and Wei Shen. Contrastmask: Contrastive learning to segment every thing. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11594–11603, 2022. 2

[24] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8372–8381, 2021. 2

[25] N. Xu, B. Price, S. Cohen, J. Yang, and T. Huang. Deep interactive object selection. In *2016 IEEE Conference on Com-*

puter Vision and Pattern Recognition (CVPR), pages 373–381, Los Alamitos, CA, USA, jun 2016. IEEE Computer Society. 2

[26] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7253–7262, 2021. 2