Object Aware Contrastive Prior for Interactive Image Segmentation (Supplementary Material)

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

1. Contrastive Prior Network for OACE

The role of Contrastive Prior Network is to generate a representation of user click that will be fed into the segmentation network MSFNet for producing object segmentation mask as an output. The user click representation defined as Object Aware Click Embeddings (OACE) is obtained by contrasting the features of user clicked foreground object and rest of the background features.

The contrastive prior network uses HRNetV2-W18-Small-v2 backbone as encoder followed by three convolution blocks as decoder. The core idea of HRNet is to maintain high resolution representation throughout the network with multiple stages and each stage having parallel branches operating at different resolutions. In our experiments, HRNetV2-W18-Small-v2 comprise of three stages and the final stage produces output at four different resolutions (512×512, 256×256, 128×128 and 64×64). Except for original scale output, other scales outputs are upsampled to original resolution. All the outputs are then concatenated and passed through the decoder block to generate OACE as an output.

During training, random patches of dimension $3 \times 3 \times 3$ are selected from object features and background features obtained from two forward passes respectively. Note that the dimensions of OACE is same as input image ($512 \times 512 \times 3$ in our experiments). Let $P_o = \{p1_o, p2_o, \dots pn_o\}$ denote the set of object feature patches selected within the object region and $P_b = \{p1_b, p2_b, \dots pm_b\}$ denote the set of background features patches selected within the background region. For our experiments, we used the values n = 10 and m = 15. 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 object and background feature patches of dimension $3 \times 3 \times 3$ are flattened first to obtain 27 dimensional vectors and then passed to cosine similarity function to compute the similarity score.

$$CosineSimilarity(\phi)[a,b] = \frac{a.b}{|a||b|} = \frac{\sum_1^k a_i b_i}{\sqrt{\sum_1^k a_i^2}\sqrt{\sum_1^k b_i^2}} \quad (2)$$

where a, b are two vectors and k refers to the dimension of the vectors (k = 27 in our case). The range of cosine similarity function is between -1 and +1, where +1 value denotes highest similarity between two vectors and -1 value denotes highest dissimilarity. The loss function L is minimized when the cosine similarity between pi_o and pj_o is maximized (i.e. +1), cosine similarity between pi_b and pj_b is maximized, and cosine similarity between pi_o and pj_b is minimized (i.e. -1).

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

MSFNet uses standard HRNet18 backbone with a CNN based Segmentation Head. Each input (Image and OACE) is first passed through two separate convolution layers individually, and then the outputs are concatenated and passed to HRNet18 backbone. HRNet18 comprise of four stages and the final stage produces output at four different resolutions. The outputs are upsampled to original resolution and

concatenated with the output from multi stage fusion block. The concatenated feature maps are processed by the segmentation head to produce segmentation probability map as an output.

2.1. Loss Function ablation - MSFNet

We experimented by training MSFNet with three different loss functions independently: 1.) Normalized Focal Loss (NFL), 2.) Dice Loss, 3.) Binary Cross-Entropy (BCE) Loss. As the output of MSFNet is a binary mask, there exists only two classes, i.e. foreground and background. However, the variations in the ratio of foreground pixels to background pixels is large based on the size of the user interacted object with respect to the entire image. BCE Loss treats each pixels independently whereas both NFL and Dice loss takes spatial relationship of pixels into account resulting in more coherent segmentations. Moreover, NFL allows fine-tuning the balance between easy and hard examples during training, thus providing more flexibility in capturing subtle patterns and improving model convergence. NFL combines the benefits of both Dice loss and BCE loss by considering both overlapping region (Dice) and the class probabilities (BCE) resulting in a robust loss function. Our experiments have shown that NFL outperforms both Dice loss and BCE loss for interactive image segmentation task by achieving better single click mean IoU (mIoU) as shown in Table 1.

Loss Function	mIoU@1		
	Berkeley	GrabCut	DAVIS
Normalized Focal Loss	**91.8**	**93.9**	**80.2**
Dice Loss	88.5	89.2	76.3
BCE Loss	85.7	88.1	74.2

Table 1. Comparison of mIoU@1 for MSFNet + OACE trained with different loss functions.

3. Analysis of user click location on segmentation output

Figure 1 depicts the segmentation outputs of different existing click-based interactive segmentation methods with two different click locations. Most of the existing interactive segmentation methods do not produce consistent outputs with two different click locations to segment an object. There could be several possible reasons resulting in different outputs. One possible reason could be lack of robustness towards different click locations due to inappropriate simulation of click points during trainings of the network. However, the most important and logical reason is the way existing networks represents and process the user click information. Existing methods use disk based or distance transform based representation of user click location that changes drastically when the click locations changes. The drastic variations in representation of user clicks, that forms the input to the existing interactive segmentation networks, leads to drastic variations in the output segmentation mask.

In this paper, we have proposed a novel way of representing user click information through object aware click embeddings (OACE). User click representations using OACE do not vary drastically with change in click location on an object to be segmented, thus resulting in consistent outputs with different click locations.

4. Rare and Frequent Category Object Test Set

LVIS version 0.5 dataset comprises a total of 1230 object categories, that is further classified into 454 rare category object, 461 common category objects, and 315 frequent category objects. LVIS dataset defines object categories with > 100 images as frequent, object categories with > 10 but < 100 images as common, and object categories with ≤ 10 images as rare. Segmentation networks trained with LVIS dataset gets biased towards frequent category objects due to higher number of representation in training set and have inferior performance on rare category object due to lower number of images in training set. In order to evaluate the performance of our proposed framework on rare and frequent category objects, we created two custom test sets each comprising of 100 images.

Rare object test set comprises of categories like armor, gameboard, bow (weapon), boxing glove, horse buggy, casserole, cassette, chessboard, chocolate mousse, compass, dagger, drone, first-aid kit, gemstone, goldfish, harmonium, hot-air balloon, joystick, limousine, microscope, pendulum, piggy bank, roller skate, space shuttle, army tank etc.. Frequent object test set comprises of categories like airplane, baseball cap, bath towel, bear, bed, bicycle, cake, duck , egg, goggles, horse, jacket, laptop, magazine, mug, newspaper, pen, person, pizza, shoe, sofa, strawberry, teddy bear, television, tennis ball etc.. These object categories are selected from LVIS train set and images related to these object categories are procured through open-source and annotated with manual efforts. Figure 2 depicts the samples of rare category objects in the test set and Figure 3 depicts the samples of frequent category objects in the test set.

The visual results comparison of different interactive segmentation methods on rare category object test set and frequent category object test set are presented in Figure 4 and Figure 5.

5. Limitations of Proposed Method

The proposed framework operates on inputs with resolution of 512×512 and produces segmentation output at the same resolution. The inference resolution of 512×512 was

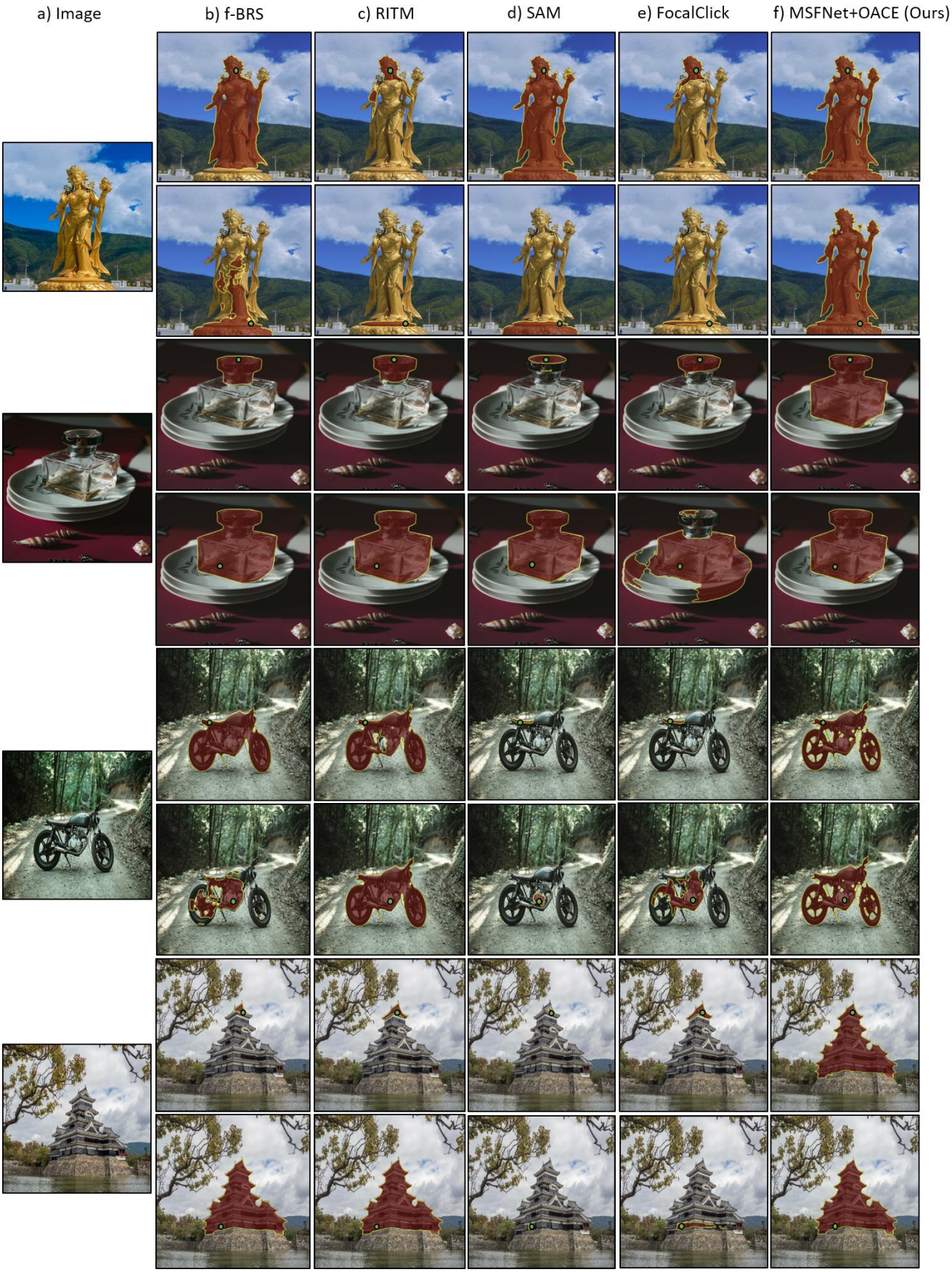

Figure 1. Qualitative comparison of different interactive segmentation methods in terms of segmentation output for different user click locations. The green dot on each image denotes the user click location.

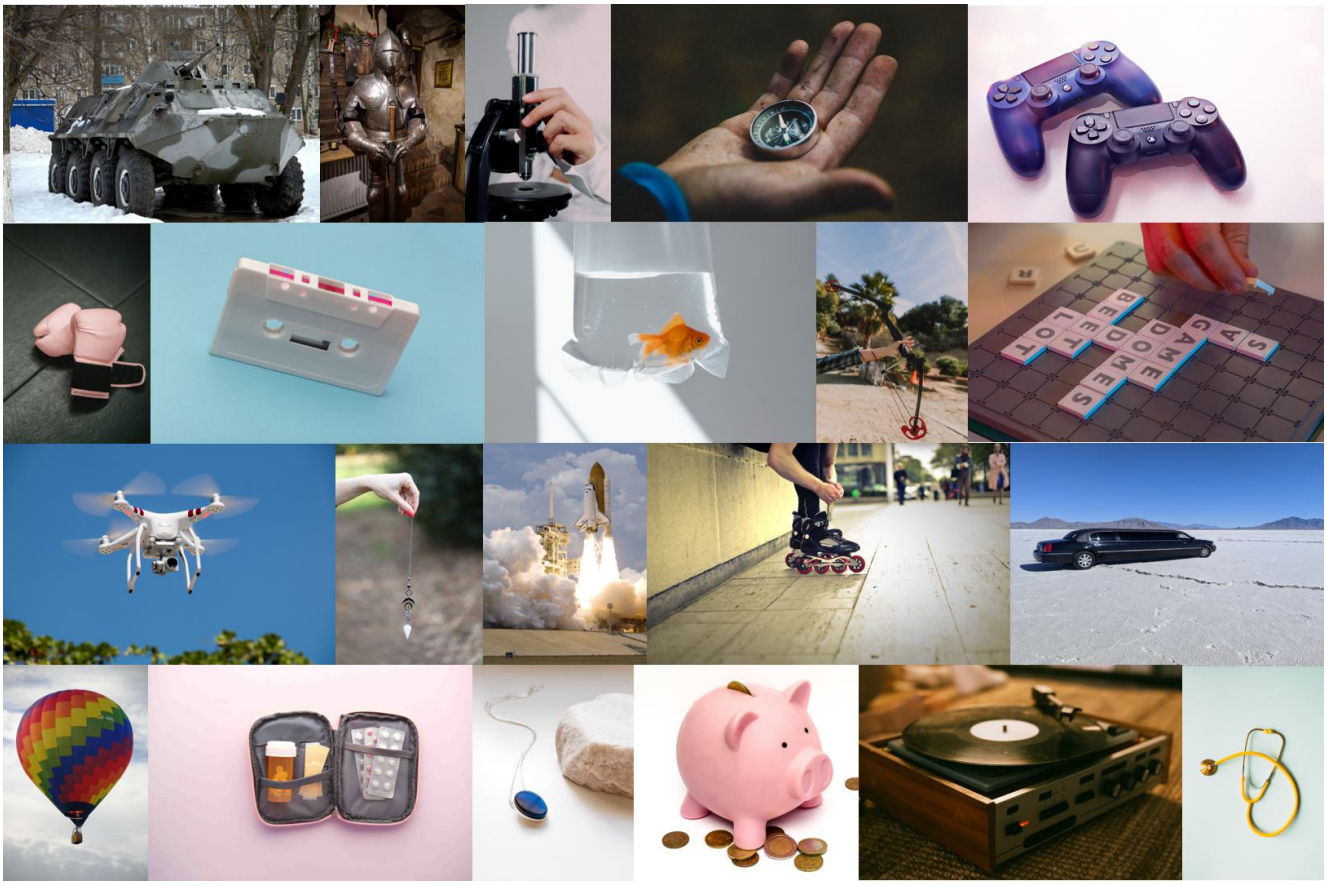

Figure 2. Rare category object sample images.

decided keeping in mind the time for inference on resource constrained mobile devices. However, there are certain limitations when operating at such lower resolutions. Firstly, finer details on an object such as thin hair strands in case of human segmentation or furry texture on animals, whiskers of cat or tiger are not segmented properly as shown in Figure 6. Moreover, small gaps between hairs are oversegmented. Secondly, in cases of very small objects, i.e. the objects whose dimensions in the image are very small compared to the dimension of the image, problems like imprecise segmentation around boundaries of the object are observed as shown in Figure 7.

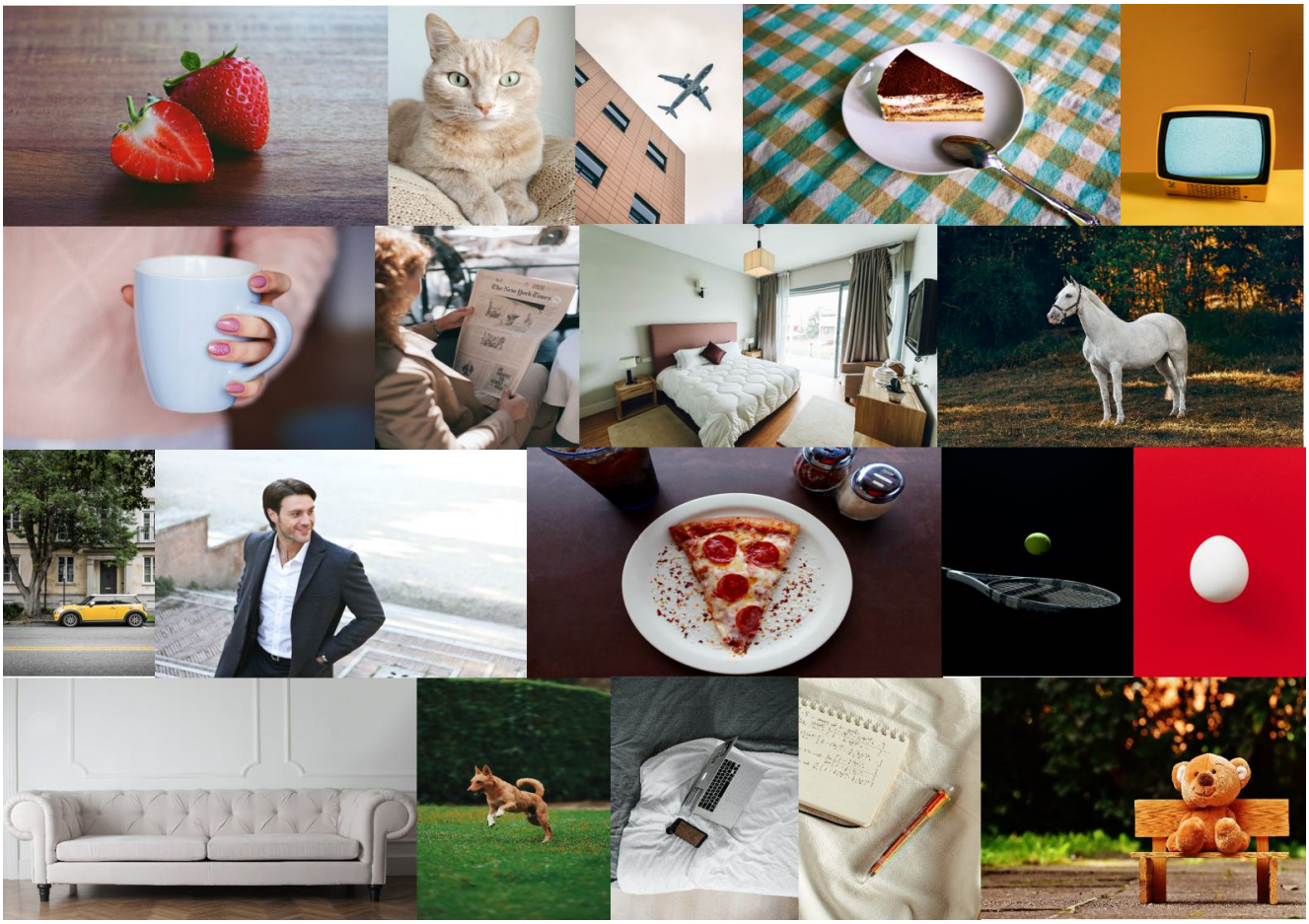

Figure 3. Frequent category object sample images.

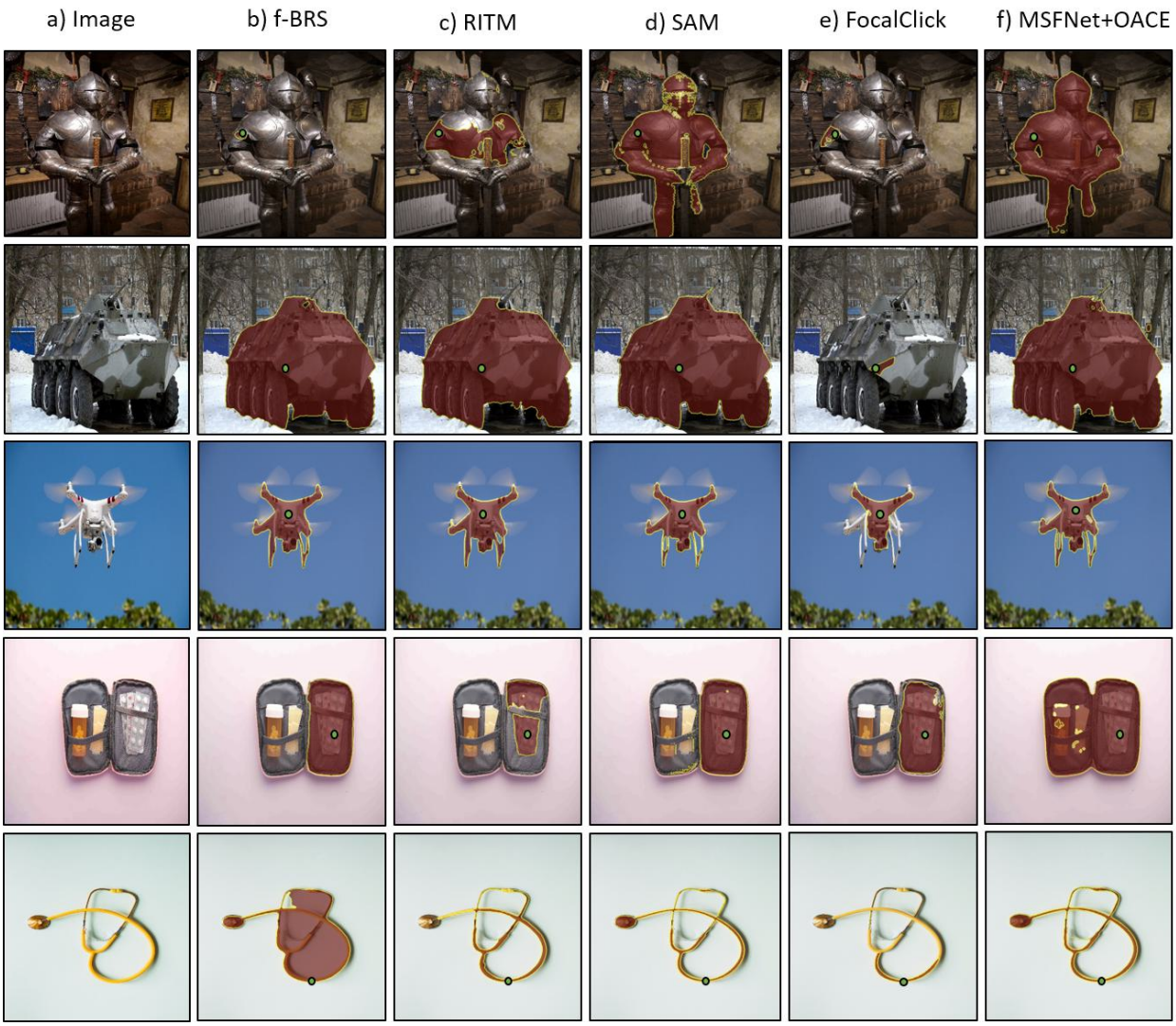

Figure 4. Visualization of results on rare category test set. The green dot represents the user-click location.

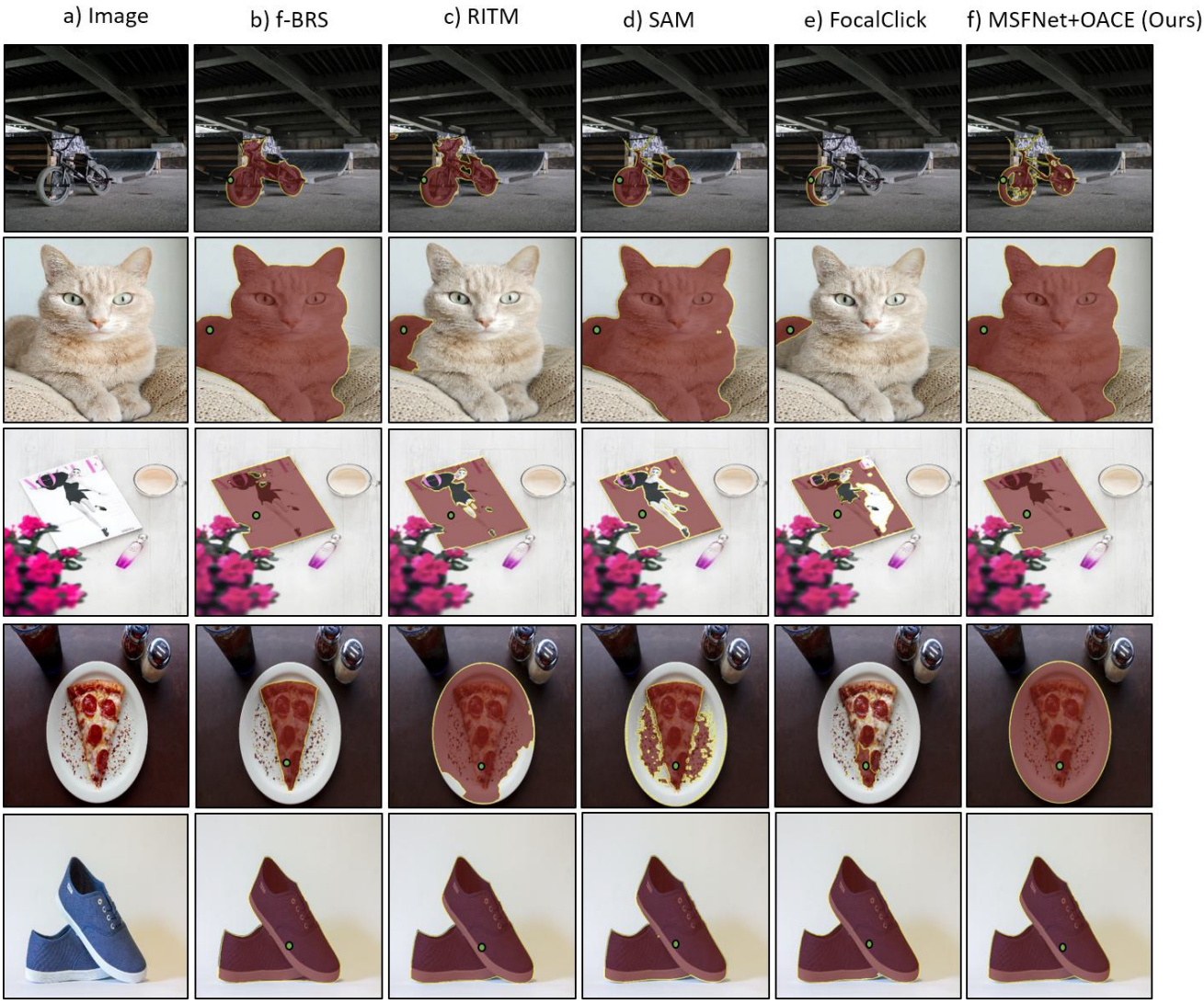

Figure 5. Visualization of results on frequent category test set. The green dot represents the user-click location.

Figure 6. Visualization of failure case scenarios. The green dot represents the user-click location. Red arrows points to erroneous region.

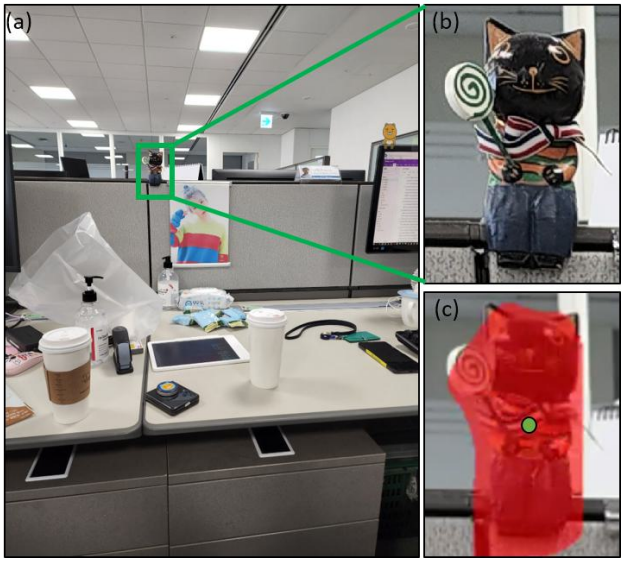

Figure 7. Visualization of failure case scenario. (a) Image, (b) Zoomed-in view of small object in image, (c) Zoomed-in view of segmentation output.