Supplementary material for Diverse Explanations for Object Detectors with Nesterov-Accelerated iGOS++

Mingqi Jiang
jiangmi@oregonstate.edu
Saeed Khorram
khorrams@oregonstate.edu
Li Fuxin
lif@oregonstate.edu

Collaborative Robotics and Intelligent Systems Institute
Oregon State University
Corvallis, USA

1 Results for Faster R-CNN

In Table 1, we present more quantitative comparisons using Faster R-CNN on the validation set of PASCAL VOC 2012. Our methods still have superior performance over previous work at the respective resolutions of each method. NAG is still twice as fast as LS.

Method	16×16			25×25			100×100		
	Del ↓	Ins ↑	Time(s)	Del ↓	Ins ↑	Time(s)	Del ↓	Ins ↑	Time(s)
D-RISE	0.4916	0.7428	220	– –	– –	– –	– –	– –	– –
Grad-CAM	– –	– –	– –	0.6094	0.3739	**4**	– –	– –	– –
LS-iGOS++	0.2963	0.8044	47	0.2045	0.8043	44	0.1049	0.7895	39
NAG-iGOS++	0.2801	**0.8055**	**22**	0.1880	0.8044	22	0.1034	0.8001	**22**
Best-NAG-iGOS++	**0.2521**	0.8050	88	**0.1721**	**0.8048**	88	**0.0950**	**0.8200**	88

Table 1: Quantitative comparison on Deletion (lower is better), insertion (higher is better) and runtime on the PASCAL VOC dataset using Faster R-CNN. The top row shows the different resolutions.

2 Accumulating Integrated Gradients with different heads

In Sec. 4 of the main paper, we used the score head of the proposal region to calculate the integrated gradient for Mask R-CNN. Here we compare the performance of accumulating integrated gradients using different heads.

For box head, we set the output to the final prediction box and the intersection-over-union (IoU) value of the object's fixed area. This enables us to identify which features contribute to fixing the box onto the target object area. For the mask head, we consider the edge of the predicted mask by multiplying the output mask by 1 inside the predicted mask and -1 outside the predicted mask, and then computing the mean of the resulting mask. Table 2 presents

the results obtained from calculating integrated gradients with different heads. It is evident that the classification head exhibits the lowest deletion score and the highest insertion score, while maintaining the same resolution. This shows that the classification head requires the least amount of information inside the bounding box. We chose not to show results of the box and mask heads in the main paper because of their unclear meanings – e.g. for mask head, if one already have the mask to start with, what would be making sense for a heatmap algorithm to output except to output the mask itself?

Head	16×16		25×25		100×100	
	Del ↓	Ins ↑	Del ↓	Ins ↑	Del ↓	Ins ↑
Score	**0.5577**	**0.6760**	**0.4641**	**0.6285**	**0.2380**	**0.5478**
Box	0.6649	0.5599	0.5844	0.4991	0.3567	0.3625
Mask	0.6192	0.6080	0.5292	0.5393	0.3043	0.4144

Table 2: Comparison for integrated gradient accumulation using different heads of Mask R-CNN. Here we use NAG-iGOS++ method.

3 More Visualizations using Mask R-CNN

In Sec 4.4, we provide additional visualizations of images in Figure 7 of the main paper. In Fig 1, we present some more visualizations obtained by NAG-iGOS++ with different initializations for the insertion task. It is noteworthy that knots on ties and wheels of the buses almost always contribute to the predicted scores in the insertion task, regardless of the initialization, showing their importance in the classification. Whereas, the algorithm could be relatively robust to other regions in the image and different region combinations can achieve similar results. From the two middle columns of Figure 1, we observe that two similar objects are simultaneously used to provide predicted scores, potentially leading to the erroneous bounding box prediction. In the last two columns where the network made an erroneous category prediction, the visualizations can help guide humans to understand which image features led to those erroneous predictions. For example, in the case of the umbrella missed as a bicycle, after the heatmap occluded certain regions, the rest of the regions look more similar to a bicycle wheel with multiple spokes. And in the case of the bus missed as the truck, after occlusion one can see that the railings on top, the frontal part, the jumpboard and the wheels are features that made the classifier think the bus is more similar to a truck. Such interpretations can help further improving data augmentation routines to avoid making such mistakes and improve the performance of the network.

We present additional results in Fig 2 and Fig 3, showcasing the contributions of specific image features to the predictions and their role in positioning. In these visualizations, we can observe how the ears of cats, the heads of people, the faucets of sinks, the knots of ties, and the tires of buses consistently influence the predicted outcomes. These features play a significant role in the model's decision-making process and provide valuable information for the understanding of their inner workings.

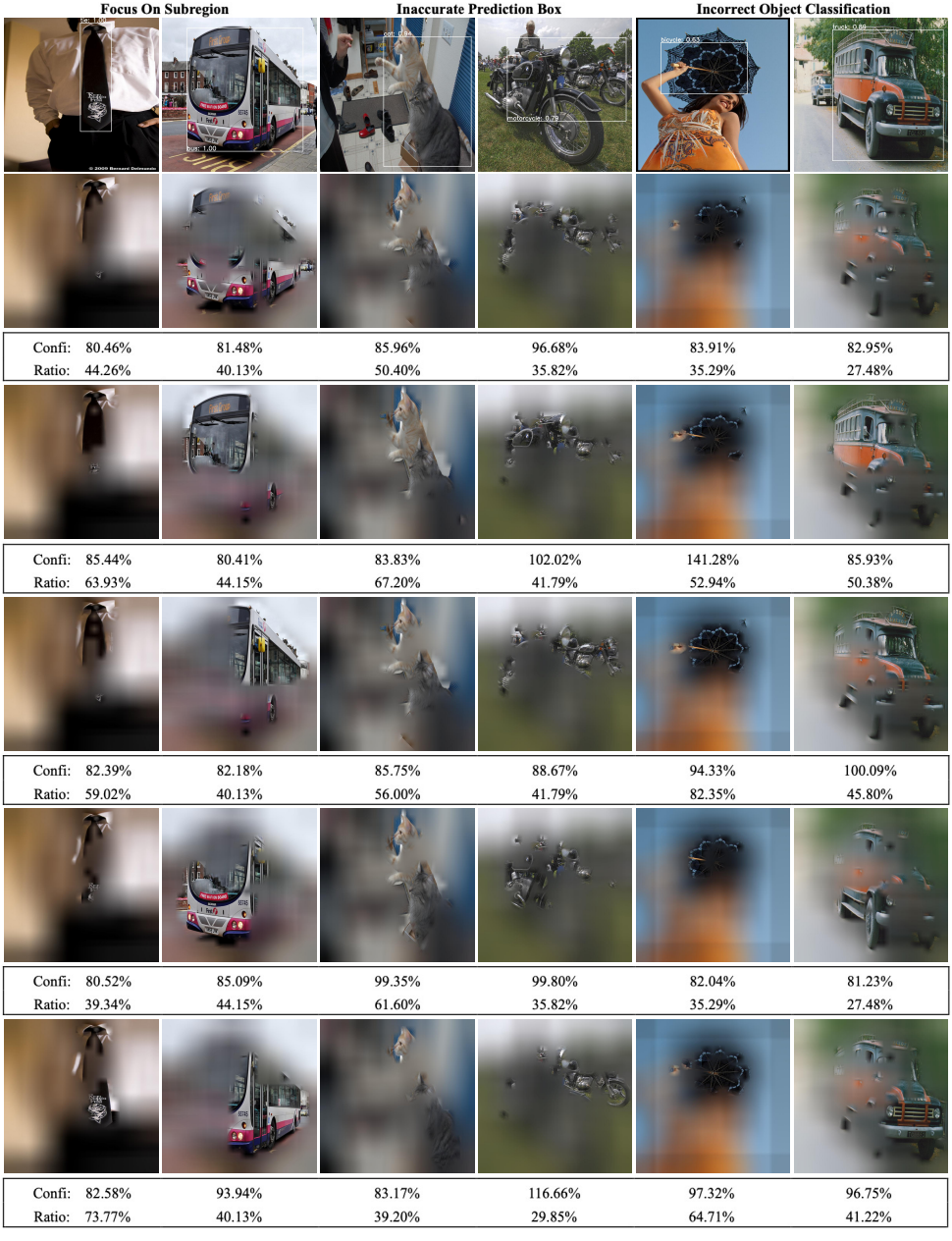

Figure 1: More visualizations with different initializations in the insertion task of Fig. 5 in main paper. From top to bottom is the Original, using the Top Left corner, using the Top Right corner, using the Bottom Left corner, using the Bottom Right corner to initialize.

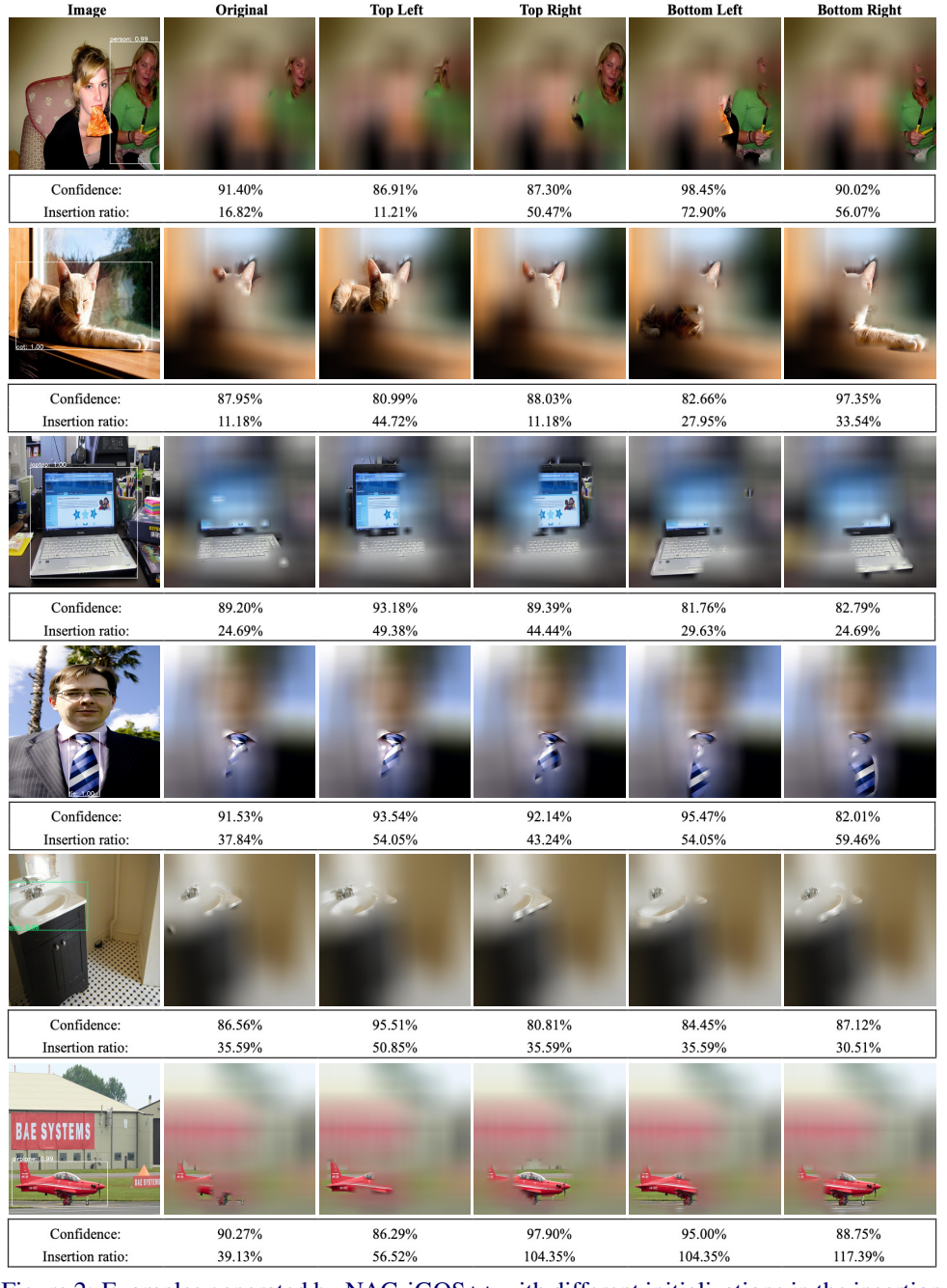

Figure 2: Examples generated by NAG-iGOS++ with different initializations in the insertion task using Mask R-CNN. The regions not highlighted on the heatmap are blured.

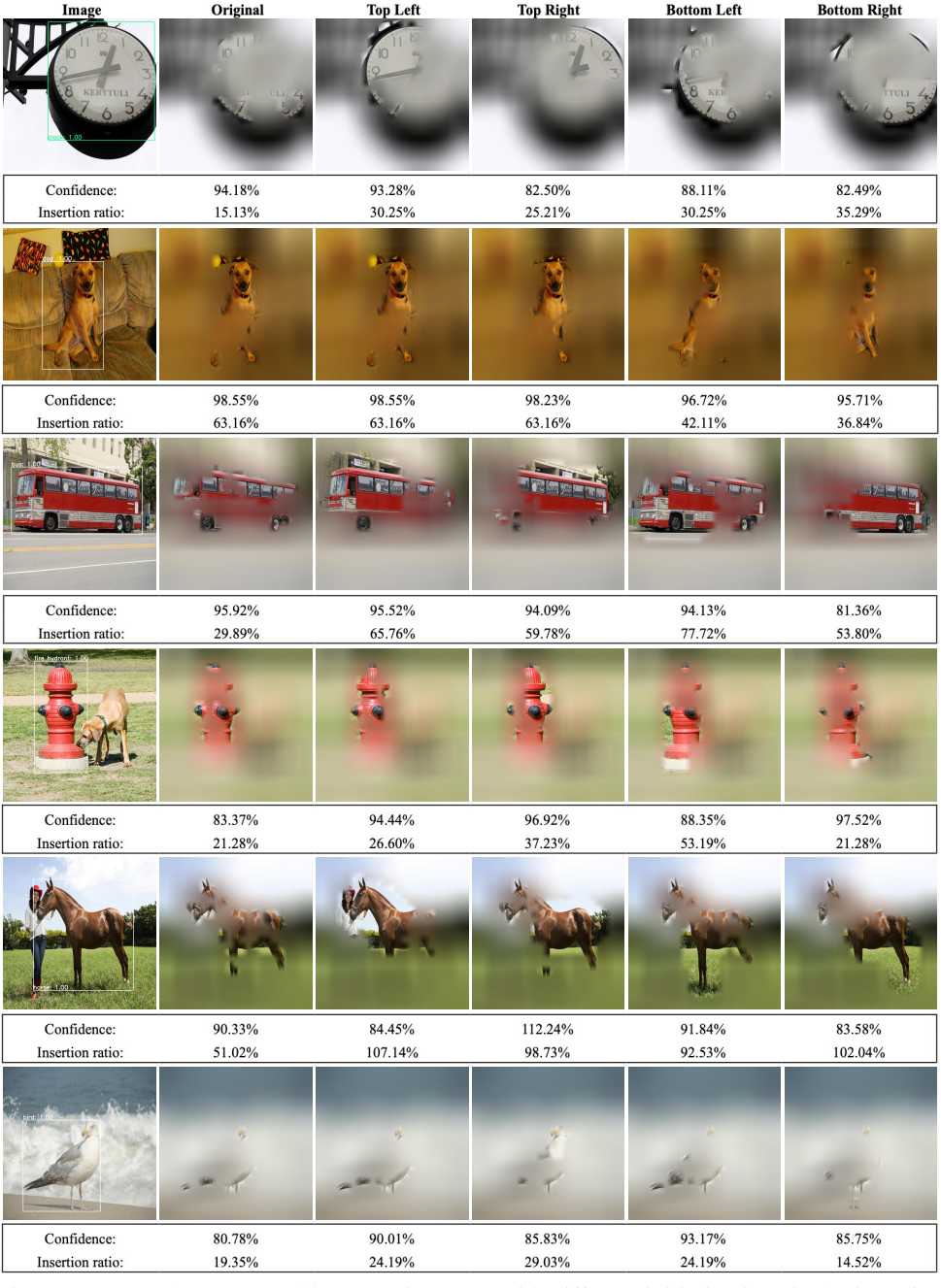

Figure 3: Examples generated by NAG-iGOS++ with different initializations in the insertion task using Mask R-CNN. The regions not highlighted on the heatmap are blured.