—— Supplementary Material ——
Learning to Recognize Occluded and Small Objects with Partial Inputs

Hasib Zunair
CIISE, Concordia University
Montreal, QC, Canada
hasibzunair@gmail.com

A. Ben Hamza
CIISE, Concordia University
Montreal, QC, Canada
hamza@ciise.concordia.ca

1. Implementation details

Data preprocessing. Images and masks are resized to 448×448 and normalized to have values in $[0, 1]$. For ViT [3] based models, the input resolution is set to 224×224 to leverage ImageNet-21k and ImageNet pretrained weights. We use simple data augmentation techniques such as random flip and random resize crop. Unlike previous works [1, 8], we do not employ complex data augmentation strategies such as CutMix, GPU Augmentations, or RandAugment.

Architecture. We apply MSL on two CSRA [8] based backbones, a convolutional backbone ResNet-cut which is a ResNet-101 pretrained on ImageNet with CutMix [7] augmentation strategy. It is worth mentioning that we do not use CutMix [7] augmentation strategy when applying MSL, to demonstrate its effectiveness. Note that here CutMix is for the pretrained model and not during fine-tuning on VOC2007 and MS-COCO datasets. To demonstrate the generality of MSL, we use a transformer backbone ViT-L16 [3] pretrained on ImageNet-21k and fine-tuned on ImageNet with the 224×224 resolution. We drop class tokens and use the final output embeddings as features, and we also interpolate positional embeddings when the models are fine-tuned on the higher resolution datasets. We refer to these MSL variants as MSL-C and MSL-V, where C and V denote convolutional and vision transformer, respectively.

Model Training. MSL models are trained in a single stage, requiring a training set comprised of images and labels. We use the SGD optimizer to minimize the loss function. Following previous work [8], we apply simple data augmentation such as random flip and random resize crop. For training both the baseline and MSL models, we set the learning rate, momentum and weight decay to 0.01, 0.9 and 0.0001, respectively. The models are trained for 60 epochs with a batch size of 6, and the best weights according to the mAP score on the test set are recorded. We follow CSRA [8] models and set $H = 1$, $\lambda = 0.1$ for VOC2007, and $H = 6$, $\lambda = 0.4$ for MS-COCO.

Model Testing. After training, given an image as input, the model simply makes a prediction by assigning multiple label(s) among the defined classes.

Hardware and software details. Our experiments were conducted on a Linux workstation running 4.8Hz and 64GB RAM, equipped with a single NVIDIA RTX 3080Ti GPU packed with 12GB of memory. All algorithms are implemented in Python using PyTorch.

2. Additional Results

In this section, we provide additional experimental results on VOC2007, MS-COCO and WIDER-Attribute datasets, showing the effectiveness of MSL in recognizing small and occluded objects.

Runtime Analysis. MSL incurs a minor computational overhead compared to traditional supervised learning. This is primarily due to the masking operation and the computation of predictions on the masked images. It is important to mention that this extra cost is only present during the training phase, and during inference, there is no masking involved. Instead, predictions are directly computed on the original input images. When compared to previous approaches, our method stands out for its simplicity and ease of training. Unlike other methods, MSL does not require multiple stages of training, the combination of multiple learnable networks, the utilization of large language models, high input resolution, complex data augmentation strategies, or the inclusion of additional data.

Discussion on MLIR for small objects. Upon analyzing recent MLIR methods, we noticed that MCAR [4] stands out as the only method that explicitly tackles the problem of small-sized and occluded objects. Comparatively, our MSL model achieves higher scores in terms of mean Average Precision (mAP), with values of 96.1% and 86.4% on the VOC2007 and MS-COCO datasets, respectively. On the other hand, MCAR's performance falls slightly behind, scoring 94.8% and 84.5% on the same datasets. Note that MCAR employs an input resolution of 576×576, while MSL operates at a resolution of 448×448. MSL explicitly addresses the problem of small and occluded objects through the Masked Branch since that task of the branch is to recognize masked objects, which are partial inputs. We further illustrate the effectiveness of MSL in handling small objects and heavily occluded objects through visual examples presented in Figures 1 and 2. These examples demonstrate MSL's ability to accurately predict such challenging instances.

MSL is model-agnostic. In Tables 1 and 2, we show recent state-of-the-art methods, as well as convolutional and transformer backbones, all of which were trained using MSL. As can be seen, MSL consistently improves performance of various methods, demonstrating that MSL is model-agnostic.

Table 1. **Comparison of recent architectures trained using MSL.** MSL is a versatile approach that enhances the performance of different methods. Note that MCAR models are trained using 576×576 input resolution, while the others utilized a resolution of 448×448. MSL improves performance of recent MLIR methods.

Method	VOC2007, mAP (%)
MCAR [4]	94.8
MCAR [4] w/ MSL	**95.6**
SST [2]	94.5
SST [2] w/ MSL	**95.8**

Table 2. **Comparison of different architectures trained using MSL on VOC2007.** MSL improves performance of both convolutional and transformer baselines in terms of mAP and other metrics.

Method	mAP	CR	CF1
ViT [3]	94.4	86.9	89.6
+ MSL	**95.0**	**84.8**	**89.5**
ResNet-cut [8]	93.7	87.5	88.3
+ MSL	**96.1**	**92.4**	**91.6**

WIDER-Attribute dataset results. Table 3 shows that MSL outperforms strong baselines on the WIDER-Attribute dataset [5].

Comparison with CSRA variants. In Table 4, a comparison is made between CSRA variants and MSL variants on VOC2007 and MS-COCO. Specifically, we train CSRA and MSL with two pretrained backbones, namely ViT-L16 and ResNet with CutMix. Note that in the main body of the paper, we use CSRA-based backbones in MSL with MSL-C

Table 3. **Performance comparison of MSL and baselines on WIDER-Attribute dataset.** MSL outperforms all baselines. † indicates our reproduced result. Other results are taken from [8].

Method	mAP	CF1	OF1
DHC	81.3	-	-
VA	82.9	-	-
SRN	86.2	75.9	81.3
VAC	87.5	77.6	82.4
VIT-B16	86.3	75.9	81.5
VIT-L16	87.7	78.1	82.8
VIT-L16 + CSRA†	89.6	80.4	84.9
VIT-L16 + MSL	**90.6**	**80.5**	**85.3**

and MSL-V notations. Here, we test CSRA and MSL independently to highlight the contributions of MSL. We find that MSL improves performance for both transformer and convolutional backbones on both datasets. For fair comparison, we run CSRA variants on our working environment and conduct all experiments with a batch size of 6, whereas the CSRA results reported in the paper [8] use a batch size of 64. Hence, the results we report here do not exactly match those in [8]. To analyze the effect of batch size on the performance of CSRA and MSL, we conduct a small experiment on VOC2007 by varying the batch size from 4 to 12, which maximizes our GPU usage, and we found that both CSRA and MSL improve in terms of performance. Therefore, we argue that the performance of MSL could be further improved using a higher batch size.

Table 4. **Performance comparison of MSL and CSRA variants in terms of mAP (%) on VOC2007 and MS-COCO.** MSL outperforms CSRA variants on both datasets.

Method	VOC2007, mAP (%)	MS-COCO, mAP (%)
VIT-L16	92.1	75.6
VIT-L16 w/ CSRA	94.4	76.8
VIT-L16 w/ MSL	**94.9**	**77.4**
ResNet-Cut	92.4	81.0
ResNet-Cut w/ CSRA	93.7	84.3
ResNet-Cut w/ MSL	**94.4**	**85.5**

Analysis of masking in MSL. In Table 5, we report the impact of low and high masking on the performance of MSL-C and MSL-V. As can be seen, better results are achieved with high masking on different backbones tested on both VOC2007 and MS-COCO. High masking enables the network to learn better context when training using MSL. Low masking, on the other hand, does not result in significant performance improvements, partly due to learning redundant features. In other words, low masking does not significantly change the original image. Hence, learning very similar features does not help to learn useful representations.

Table 5. **Ablation analysis in mAP (%) of high- and low-masked pixels during MSL training on VOC2007 and MS-COCO.** MSL with high-masked pixels yields better performance.

Method	Masking	VOC2007	MS-COCO
MSL-V	Low	94.6	77.8
MSL-V	High	**95.0**	**79.0**
MSL-C	Low	95.0	85.1
MSL-C	High	**96.1**	**86.4**

References

[1] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proc. IEEE International Conference on Computer Vision*, 2021. 1

[2] Zhao-Min Chen, Quan Cui, Borui Zhao, Renjie Song, Xiaoqin Zhang, and Osamu Yoshie. SST: Spatial and semantic transformers for multi-label image recognition. *IEEE Transactions on Image Processing*, 31:2570–2583, 2022. 2

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 2

[4] Bin-Bin Gao and Hong-Yu Zhou. Learning to discover multi-class attentional regions for multi-label image recognition. *IEEE Transactions on Image Processing*, 30:5920–5932, 2021. 1, 2

[5] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *Proc. European Conference on Computer Vision*, pages 684–700, 2016. 2

[6] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proc. European Conference on Computer Vision*, pages 85–100, 2018. 8

[7] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. CutMix: Regularization strategy to train strong classifiers with localizable features. In *Proc. IEEE International Conference on Computer Vision*, pages 6023–6032, 2019. 1

[8] Ke Zhu and Jianxin Wu. Residual attention: A simple but effective method for multi-label recognition. In *Proc. IEEE International Conference on Computer Vision*, pages 184–193, 2021. 1, 2, 4

Table 6. **Performance comparisons of MSL and CSRA ResNet-cut [8] as baseline in terms of mAP (%) and other metrics when provided randomly masked images at test time on VOC2007 and MS-COCO datasets**. Boldface numbers indicate the best performance. MSL is more robust to partial inputs.

Method	mAP	CP	CR	CF1	OP	OR	OF1
Baseline (VOC2007)	67.9	85.5	40.6	55.1	75.2	48.7	59.1
+ w/o MSL	**86.7**	**88.6**	**71.2**	**79.0**	**91.8**	**73.7**	**81.8**
Baseline (MS-COCO)	54.2	73.6	36.0	48.4	69.1	44.8	54.4
+ w/o MSL	**74.8**	**81.9**	**59.3**	**68.8**	**84.5**	**64.0**	**72.9**

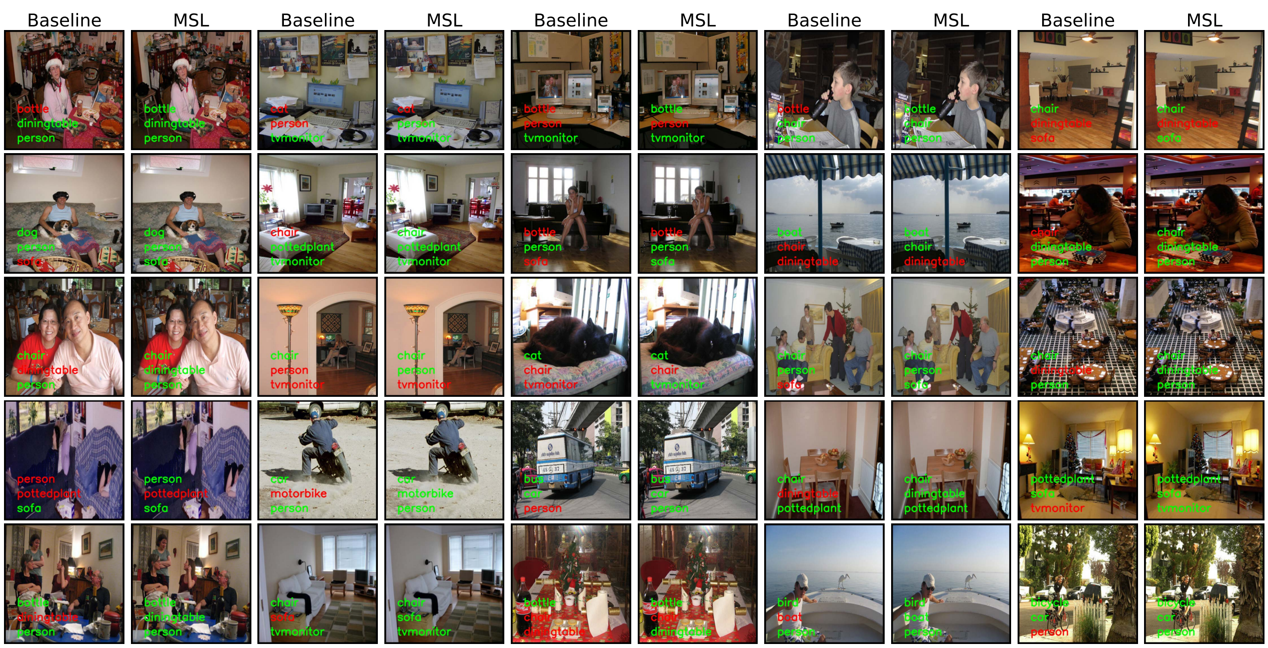

Figure 1. **Visual comparison of predictions of MSL and baseline on the VOC2007**. MSL is able to accurately predict small objects as well as objects under heavy occlusions.

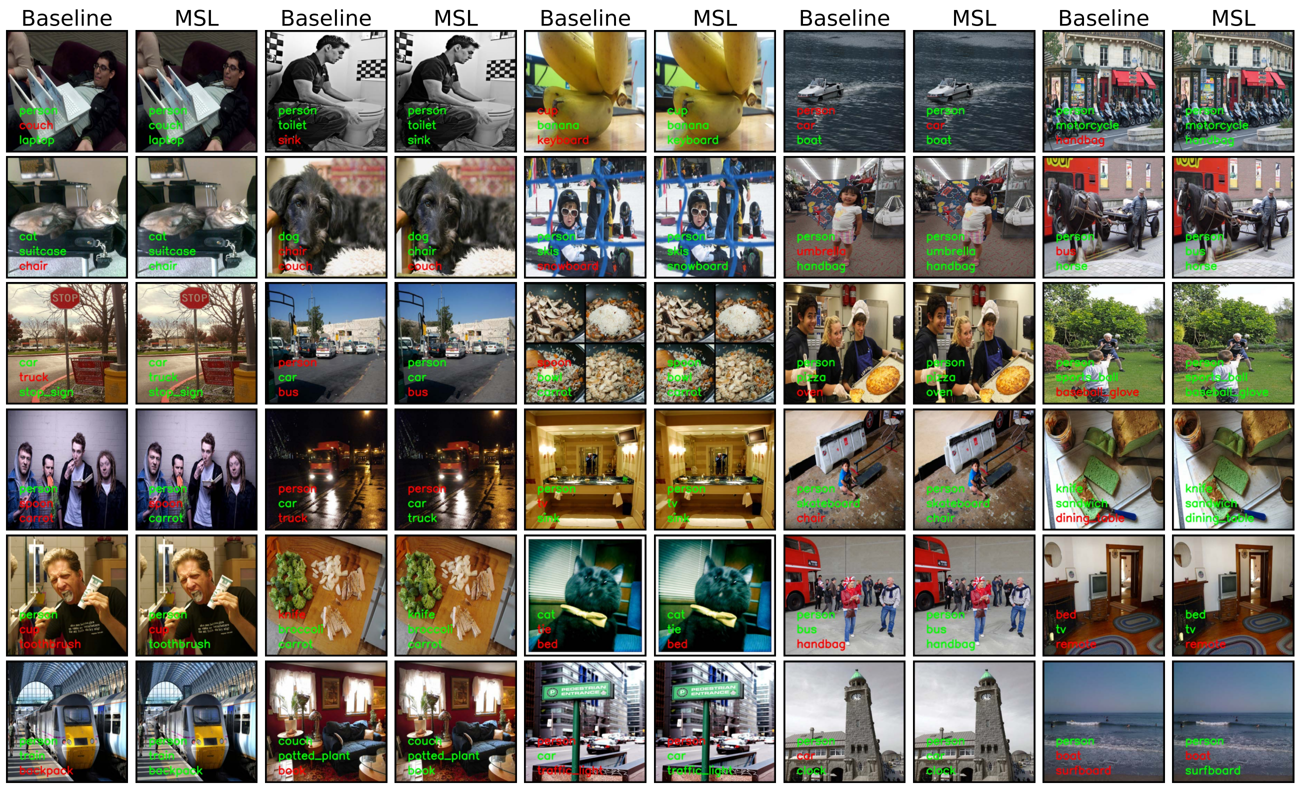

Figure 2. **Visual comparison of predictions of MSL and baseline on the MS-COCO test set**. MSL is able to accurately predict small objects as well as objects under heavy occlusions.

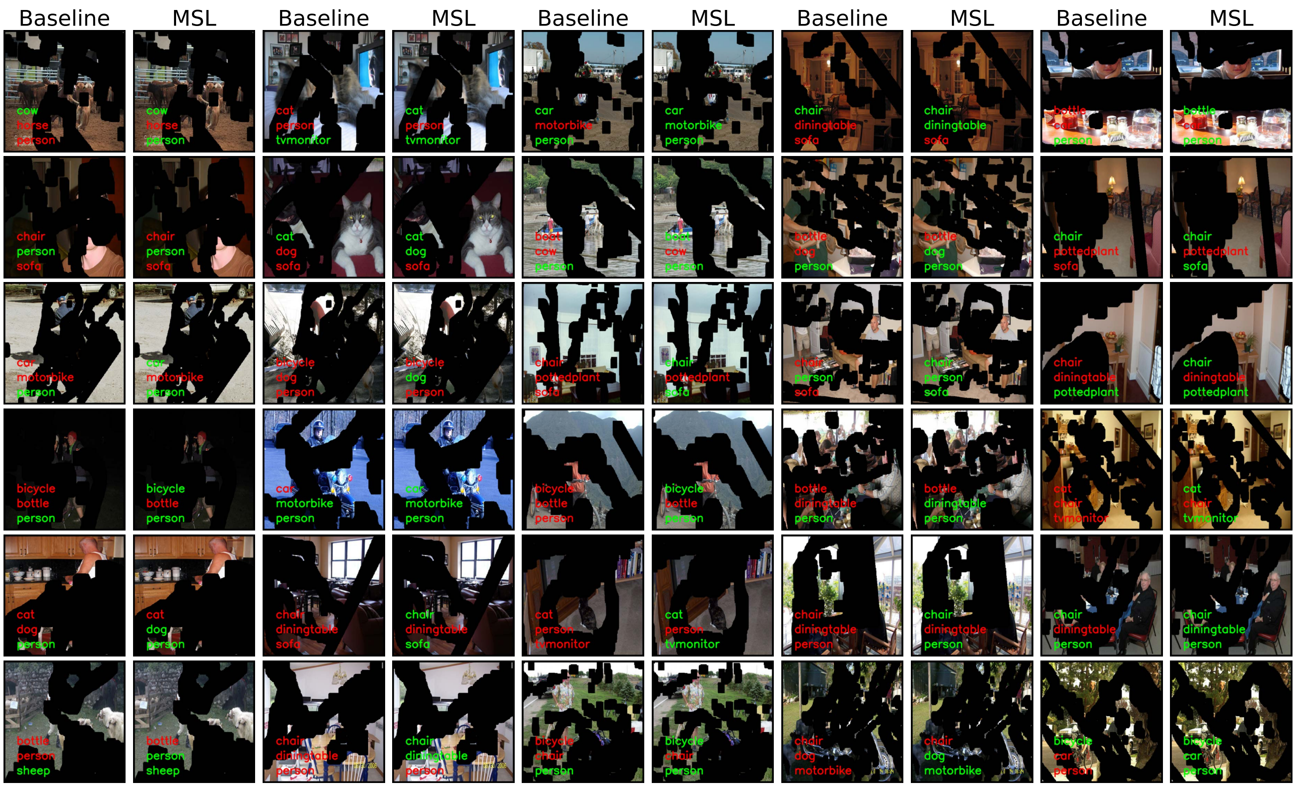

Figure 3. **Visual comparison of predictions of MaskSup and baseline on the VOC2007 test set when provided with masked regions as input**. MSL is able to recognize objects that are heavily masked and even recognize objects that are almost completely masked.

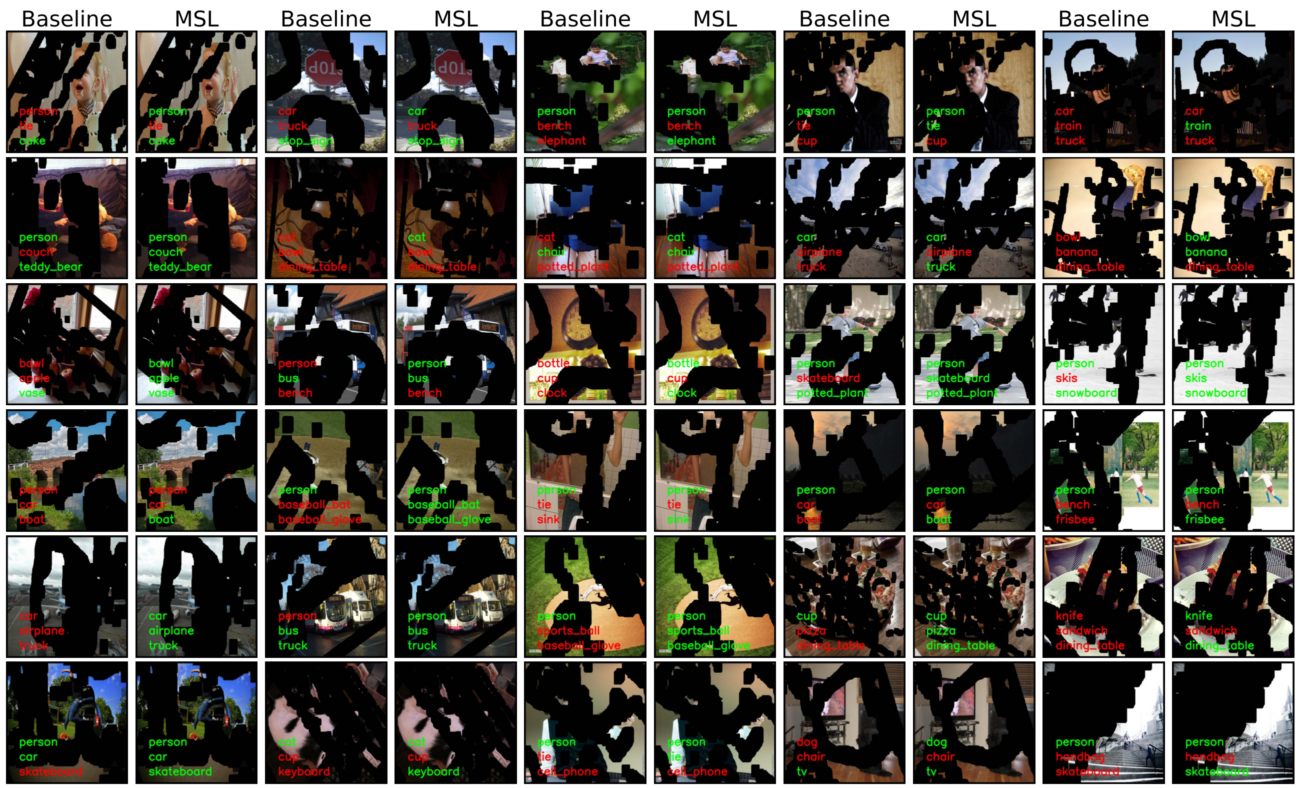

Figure 4. **Visual comparison of predictions of MaskSup and the strongest baseline on the MS-COCO test set when provided with masked regions as input**. MSL is able to recognize objects that are heavily masked and even recognize objects that are almost completely masked..

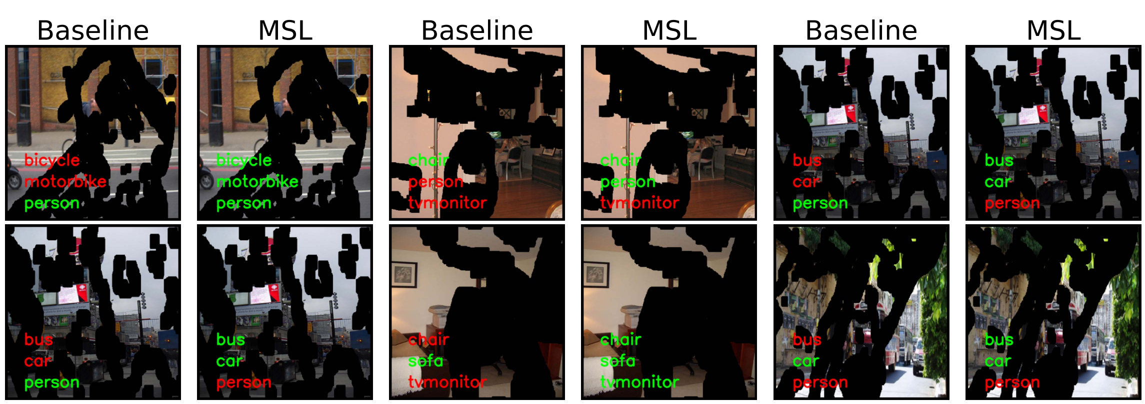

Figure 5. **Comparison of MSL and the strongest baseline on the VOC2007 and MS-COCO test sets in first and second rows**. It is worth noting that MSL is able to predict non-masked objects that the baseline model often fails to detect.

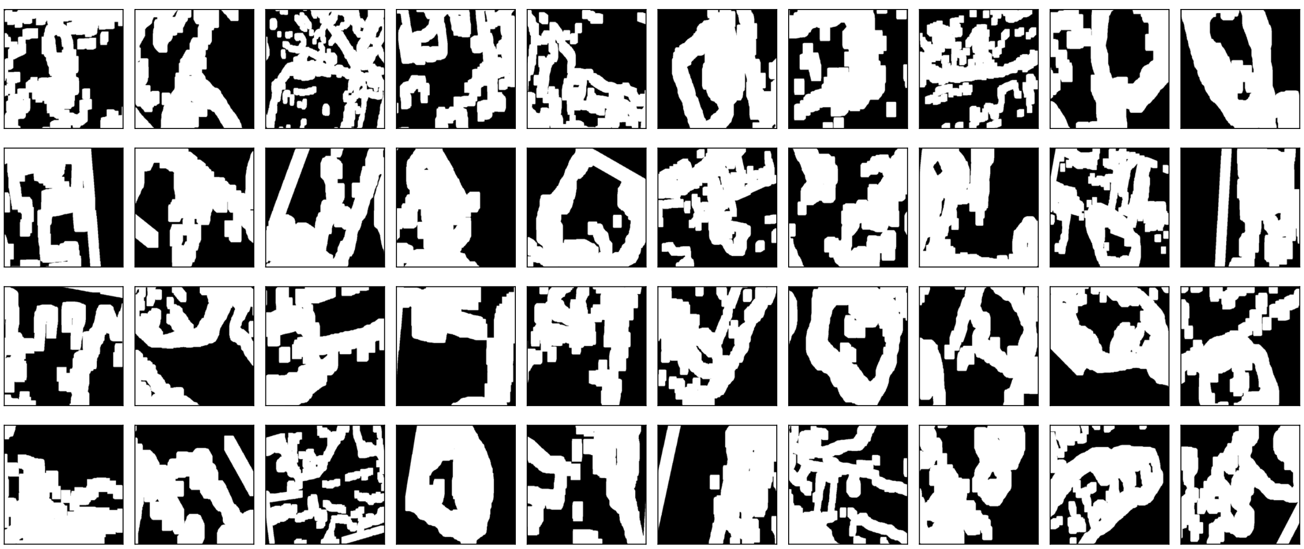

Figure 6. **Visual of masks during training in MSL**. Some masks cover more than 50% of the image. Images are from Irregular Masks Dataset [6] after applying binary thresholding.