Learning to Recognize Occluded and Small Objects with Partial Inputs

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

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

Abstract

Recognizing multiple objects in an image is challenging due to occlusions, and becomes even more so when the objects are small. While promising, existing multi-label image recognition models do not explicitly learn context-based representations, and hence struggle to correctly recognize small and occluded objects. Intuitively, recognizing occluded objects requires knowledge of partial input, and hence context. Motivated by this intuition, we propose Masked Supervised Learning (MSL), a single-stage, model-agnostic learning paradigm for multi-label image recognition. The key idea is to learn context-based representations using a masked branch and to model label co-occurrence using label consistency. Experimental results demonstrate the simplicity, applicability and more importantly the competitive performance of MSL against previous state-of-the-art methods on standard multi-label image recognition benchmarks. In addition, we show that MSL is robust to random masking and demonstrate its effectiveness in recognizing non-masked objects. Code and pretrained models are available on GitHub.

1. Introduction

Multi-label image recognition (MLIR) is a fundamental and challenging task in a variety of computer vision applications such as automatic tagging of images on social media platforms and object detection in autonomous vehicles [3]. The aim is to recognize multiple objects or attributes in an image. A major challenge in MLIR is how to effectively tackle the issue of large variations in the size and spatial locations of objects. This issue becomes more pronounced when the objects are occluded and small.

Recent MLIR approaches, including graph convolutional networks and their variants [7,9,27], focus primarily on capturing semantics and label co-occurrence among objects. While powerful, most of these methods require the combination of multiple networks, resulting in high computation cost. Also, methods that deal with both semantics of objects and label relations often consist of multiple stages of training [21,22], rely on large language models [17], and operate on high input resolution [6, 13, 16, 27]. Moreover, they require additional data for pretraining [4], even with models already pretrained on large datasets such as ImageNet-1k and ImageNet-21k, and also rely on complex data augmentation strategies [1, 31]. In addition, these methods do not explicitly address the occlusion problem and fail to accurately recognize small objects, leading to suboptimal performance on images containing small and occluded objects. In practical real-world applications of MLIR such as object detection in self-driving cars, images are usually comprised of multiple objects of different sizes (e.g., small) and shapes that co-exist and are densely cluttered (e.g., occluded), and hence it is of vital importance to develop MLIR approaches that can effectively recognize small objects even under heavy occlusions.

Intuitively, we can consider occluded objects as *partial inputs*, and hence accurate recognition requires knowledge of *partial inputs*, and hence context. Motivated by this intuition, we propose *Masked Supervised Learning (MSL)*, a single-stage, model-agnostic learning paradigm for MLIR tasks. Given a base recognition network, MSL uses a masked branch to predict the labels for a heavily masked version of the input image, which is a good cue for learning context-based representations. We also propose to use label consistency to model label co-occurrence by maximizing the similarity between the predictions from the recognition and masked branches. The main contributions of this work can be summarized as follows:

- We propose a simple yet effective single-stage, model-agnostic learning paradigm that aims to learn context-based representations and to better model label co-occurrence from *partial inputs* via masking.
- We demonstrate through experimental results and ablations that MSL yields competitive performance in comparison with single- and multi-stage approaches, especially for small and occluded objects.
- We show that MSL is not only robust to *partial in-*

puts, but also predicts objects that are almost entirely masked, while yielding improved recognition of non-masked objects.

2. Related Work

Hybrid Methods. These methods leverage a combination of convolutional, graph, transformer or recurrent neural networks [5, 8, 9, 16, 31]. Graph based networks, for instance, leverage semantic relations between object classes [7, 9], but tend to incur heavy computation costs. ADD-GCN [27] dynamically generates graphs for an image by first generating a category-aware representation, followed by modeling the relationship between the representations. ADD-GCN [27] operates on high resolution in the same vein as SSGRL [6], C-Tran [16] and MCAR [13]. KGGR [4] operates on knowledge graphs and requires additional data for pretraining. Our method does not require the combination multiple networks, high input resolution, or additional data.

Model-Agnostic Methods. This class of approaches are not architecture dependent, and include ASL [1] and CSRA [31], which can be applied to any architecture, but require an exhaustive hyperparameter tuning. Moreover, they achieve competitive results only when using complex data augmentation techniques such as CutMix, GPU Augmentations, or RandAugment [1, 31]. By comparison, our proposed model achieves state-of-the-art performance without relying on complex data augmentation strategies.

Multistage and Bimodal Frameworks. Query2Label (Q2L) [21] is a two-stage framework that focuses on class-specific attention. KSS-Net [22] is a knowledge distillation based method comprised of a two-stage training scheme with teacher and student models. BMML [17] is a bimodal learning approach that not only uses a convolutional neural network and a recurrent neural network, but also relies on large language models [10] and additional data. Our work differs from these frameworks in that it does not require multiple stages of training and also does not rely on large language models.

Transformer-Based Methods. TDRG [30] consists of convolutional neural network, a transformer, as well as a graph neural network that is used to capture long-term contextual information and to build position-wise relationships at different scales. C-Tran [16] is a transformer based method that relies on an additional image feature extractor and high input resolution. It exploits the dependencies among both visual features and labels using a single transformer encoder. By comparison, our work is significantly different from C-Tran. First, during training we mask images, whereas C-Tran masks labels. Second, the input to the transformer encoder in C-Tran consists of an image and a masked label (i.e., token), whereas our model requires only an image. Also, our method is model-agnostic and can be applied to any kind of network for MLIR tasks.

Overall, our work differs from previous MLIR approaches in that we propose a simple yet effective single-stage learning paradigm that is model-agnostic. Most notably, our model does not require multiple stages of training, the combination of multiple networks, large language models, high input resolution, complex data augmentation strategies, or additional data for pretraining.

3. Masked Supervised Learning

In this section, we begin by formulating the task at hand and subsequently introduce the fundamental components that make up the proposed MSL paradigm. The overall framework of MSL is depicted in Figure 1.

Problem Statement. Let $\mathcal{D} = \{(\mathbf{I}_i, \mathbf{y}_i)\}_{i=1}^N$ be a training set of N labeled images $\mathbf{I}_i \in \mathcal{X}$ and their ground-truth multi-label vectors $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,K})^\intercal \in \mathcal{Y} = \{0,1\}^K$, with $y_{i,k} = 1$ indicating the presence of the k-th label (i.e., object or attribute) in the image, and $y_{i,k} = 0$ indicating its absence. In other words, each image $\mathbf{I}_i$ is associated with multiple labels chosen from a set of K possible classes (i.e., object categories). The task of multi-label image recognition is to learn a multi-label recognition model $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathcal{Y}$, where $\boldsymbol{\theta}$ is a set of learnable parameters. Given a test image $\mathbf{I}$, the trained model predicts the corresponding multi-label vector $\mathbf{y}_{\text{p}} = \sigma(f_{\boldsymbol{\theta}}(\mathbf{I}))$, where $\sigma(\cdot)$ is the sigmoid activation function applied element-wise.

3.1. Masked Inputs

For masked image generation, we leverage the Irregular Mask dataset [19], which is commonly used in image inpainting [19, 23] and is comprised of roughly $20,000$ masks with random streaks and holes of arbitrary shapes. From this dataset, we generate *low* and *high* mask subsets, each of which is comprised of 1000 samples. The process for creating these two subsets is as follows: For a given mask sampled from the Irregular Mask dataset, we first compute the percentage p of zero pixels in the mask. If p is greater than 50%, then the mask is included in the high mask subset. Otherwise, the mask is placed in the low mask subset. In our experiments, we find that *high* masks generally improve performance. Intuitively, image masking can be viewed as "simulating" images with partial inputs. During training, we randomly sample a mask from the *high* mask subset and perform binary thresholding, where the pixel values are either 0 or 1, and we denote this mask by $\mathbf{M}_{\text{holes}}$. Then, we follow the masking procedure in [19, 32] to create a masked image $\mathbf{I}_{\text{masked}} = \mathbf{I} \odot \mathbf{M}_{\text{holes}}$, where $\mathbf{I}$ is the input image, and $\odot$ denotes element-wise multiplication. The masked image has a similar layout as the input image, but with roughly 50% of pixels randomly removed.

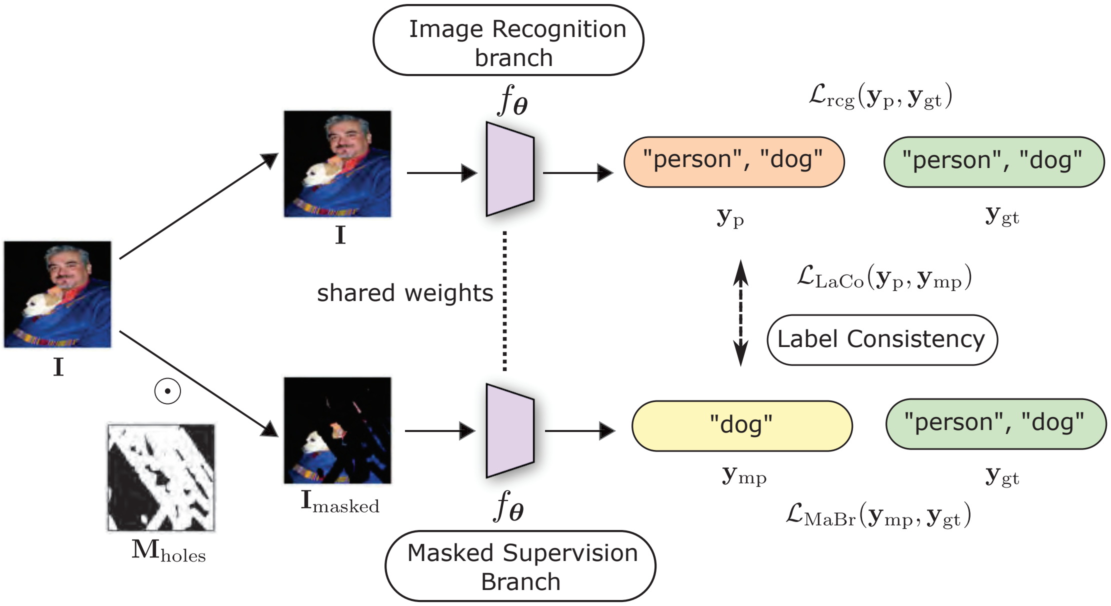


Image Recognition branch
f_θ
$\mathcal{L}_{\text{rcg}}(\mathbf{y}_{\text{p}}, \mathbf{y}_{\text{gt}})$
"person", "dog"
"person", "dog"
$\mathbf{y}_{\text{p}}$
$\mathbf{y}_{\text{gt}}$
I
I
shared weights
$\mathcal{L}_{\text{LaCo}}(\mathbf{y}_{\text{p}}, \mathbf{y}_{\text{mp}})$
Label Consistency
$\odot$
"dog"
"person", "dog"
$\mathbf{y}_{\text{mp}}$
$\mathbf{y}_{\text{gt}}$
$\mathbf{I}_{\text{masked}}$
f_θ
$\mathcal{L}_{\text{MaBr}}(\mathbf{y}_{\text{mp}}, \mathbf{y}_{\text{gt}})$
$\mathbf{M}_{\text{holes}}$
Masked Supervision Branch

Figure 1. **Overview of Masked Supervised Learning (MSL)**. A single-stage model-agnostic training scheme with a Masked Branch (MaBR) and Label Consistency (LaCo) for MLIR tasks where f_θ is a base network. The recognition and masked branches are identical and share weights. After training, f_θ is used to obtain the multi-label prediction.

3.2. Masked Branch

The goal of Masked Branch (MaBr) is to explicitly learn context-based representations, as this branch is tasked to predict labels of heavily masked inputs (i.e., *partial inputs*), translating into better multi-label predictions. **The masked branch has the ability to learn short-range context even when objects in the image are densely cluttered. Similarly, it can also learn long-range context when objects are more spaced apart**.

Given an input image $\mathbf{I}$ and its masked version $\mathbf{I}_{\text{masked}}$, we train a base recognition network f_θ to predict both the output $\mathbf{y}_{\text{p}}$ of the image recognition branch and the output $\mathbf{y}_{\text{mp}}$ of the masked branch. Here, f_θ is a Siamese network like architecture [2], where the branches are identical and share weights.

We train f_θ by minimizing the following combined loss function of the recognition branch and masked branch

$$\mathcal{L}_{\text{inter}} = \mathcal{L}_{\text{rcg}}(\mathbf{y}_{\text{p}}, \mathbf{y}_{\text{gt}}) + \mathcal{L}_{\text{MaBr}}(\mathbf{y}_{\text{mp}}, \mathbf{y}_{\text{gt}}), \tag{1}$$

where $\mathcal{L}_{\text{rcg}}$ and $\mathcal{L}_{\text{MaB}}$ are binary cross-entropy losses between the ground truth and the outputs of the recognition and masked branch, respectively. Application-specific loss functions can also be used in lieu of cross-entropy.

3.3. Label Consistency

As objects generally co-exist in an image (e.g., *chair* is more likely to co-occur with *table* than a *sportsball*), it is of vital importance to model this label co-occurrence to help improve the recognition performance. **To perceive this label-level feature, we propose to use Label Consistency (LaCo) that maximizes the similarity between the predictions from the recognition and masked branch**. Since we use a Siamese style architecture, where the network is the same with shared weights, maximizing the predictions helps the network learn to predict heavily occluded objects (e.g., partial inputs) from the presence of other target objects, thereby effectively utilizing masked branch. More specifically, we maximize the similarity between the predictions from the recognition branch $\mathbf{y}_{\text{p}}$ and the masked branch $\mathbf{y}_{\text{mp}}$ by minimizing the L_2-loss $\mathcal{L}_{\text{LaCo}} = \|\mathbf{y}_{\text{p}} - \mathbf{y}_{\text{mp}}\|^2$.

3.4. Overall Loss Function

Using the recognition branch, masked branch and label consistency, we define the overall loss function for the proposed MSL model as follows

$$\begin{aligned}\mathcal{L}_{\text{total}} = \alpha_1 \mathcal{L}_{\text{rcg}}(\mathbf{y}_{\text{p}}, \mathbf{y}_{\text{gt}}) + \alpha_2 \mathcal{L}_{\text{MaBr}}(\mathbf{y}_{\text{mp}}, \mathbf{y}_{\text{gt}}) \\ + \alpha_3 \mathcal{L}_{\text{LaCo}}(\mathbf{y}_{\text{p}}, \mathbf{y}_{\text{mp}}),\end{aligned} \tag{2}$$

where the scalars α_1, α_2 and α_3 are nonnegative trade-off hyperparameters, which control the contribution of each loss term.

During training, $\mathcal{L}_{\text{total}}$ is minimized between predictions and ground-truth labels for several epochs using stochastic gradient descent to learn the parameters of f_θ using a labeled training set. For inference, the trained network f_θ is used in multi-label image recognition to obtain multi-label predictions given a test image $\mathbf{I}$. Hence, MSL is simple in structure (i.e., model-agnostic) and easy to implement (i.e., single-stage training). When $\alpha_1 = 1$ and $\alpha_2 = \alpha_3 = 0$, we obtain the recognition loss, which is basically the loss function for the vanilla network.

4. Experiments

In this section, we demonstrate the performance of MSL in comparison with state-of-the-art methods. Details on the implementation, architecture and training, as well as additional results are included in the supplementary material.

4.1. Experimental Setup

Datasets. We conduct experiments on two MLIR benchmarks: VOC2007 [12] and MS-COCO [18].

- **VOC2007.** This is a widely-used dataset for MLIR tasks, and is comprised of 9,963 images with 20 classes, where the *train-val* set has 5,011 images and the *test* set has 4,952 images. Following previous work [8, 31], we use the *train-val* for training and *test* for testing. We also set the input resolution to 448×448, unless otherwise specified.

- **MS-COCO.** This is a standard benchmark for training and evaluating image recognition, segmentation, and detection algorithms. In our experiments, we use COCO-2014, which consists of 82,081 and 40,137 training and validation images, respectively, with 80 different classes. For fair comparison with previous work [16, 30, 31], we use the same training and evaluation procedures, and evaluation metrics.

Baselines. We compare MSL against several state-of-the-art graph-based methods that use different learnable networks such as ML-GCN [9], P-GCN [7], ADD-GCN [27] and TDRG [30]. We also compare against model-agnostic methods such as ASL [1] and CSRA [31], which rely on complex data augmentation. Moreover, we compare against methods that require large language models and additional data for pretraining such as BMML [17] and KGGR [4], as well as methods that operate on high input resolution such as SSGRL [6], C-Tran [16], MCAR [13], and IDA [20]. Finally, we compare against multi-stage frameworks such as KSS-Net [22] and Query2Label [21].

Evaluation Metrics. We use the mean average precision (mAP) as primary evaluation metric [8, 31]. We set positive threshold to 0.5 and report overall performance results of MSL and baselines using other evaluation metrics, including overall precision (OP), overall recall (OR), overall F1-measure (OF1), per-category precision (CP), per-category recall (CR), and per-category F1-measure (CF1).

4.2. Comparison with State-Of-The-Art

Comparisons on VOC2007. We compare the performance of MSL against several state-of-the-art methods, and the results are reported in Table 1. All scores are averaged over 3 runs. We employ MSL with two CSRA-based backbones: ResNet-cut, which is a ResNet-101 [15] pretrained on ImageNet-1k with CutMix [29], and ViT-L16 [11], which is a large vision Transformer pretrained on ImageNet-1k with 224×224 resolution. We refer to these MSL variants as MSL-C and MSL-V, respectively. The classification head of these backbones differs from the typical fully connected or global average pooling layer by utilizing a CSRA module [31]. This module generates class-specific features for each category, and then combines the intermediate results to produce the final logits. As shown in the table, MSL-C outperforms all previous state-of-the-art models, achieving relative improvements of 1.1%, 5.6% and 3.9% in terms of mAP, CR and CF1, respectively, over the strongest baseline. MSL-C performs better than graph-based methods such as ML-GCN and ADD-GCN. MSL-C also achieves a relative improvement of 2.8% in terms of mAP over SSGRL, which is trained on input resolution of 640×640 and uses both a convolutional feature extractor and a graph neural network. Notably, MSL-C is also efficient and more accurate than KGGR and BMML, which use additional data (MS-COCO) consisting of 82,081 images for pretraining on top of ImageNet-1k pretraining, and also rely on large language model BERT [10] and operate on label-level attentions (i.e, multiple images), making them compute intensive.

The first two rows of Figure 2 show visual examples of predictions made by MSL-C and CSRA ResNet-cut as baseline. In the first row, we can see that the baseline fails to recognize small objects such as *motorbike*, *person*, *chair* and *tvmonitor*. The second row shows instances where the baseline model fails to recognize target objects under heavy occlusions, such as *sports ball*, *person* and *vase*. In contrast, MSL-C is able to recognize small objects, as well as objects that are heavily occluded. The masked branch, which is responsible for recognizing target object(s) from partial inputs through masking, can acquire context-based representations. This ability is likely responsible for its success in recognizing objects under challenging conditions. Label consistency, on the other hand, helps model label co-occurrence by maximizing the similarity between the predictions made by the recognition and masked branches.

Comparisons on MS-COCO. In Table 2, we report results on MS-COCO, where all scores are averaged over 3 runs and MSL is applied on CSRA-based ResNet-cut backbone. As can be seen, MSL-C outperforms all baselines operating on input resolution 448×448 by 1.4% in terms of mAP. MSL-C also outperforms complicated and time-consuming methods such as KSSNet and MCAR, as well as methods that operate on higher input resolution 575×576 such as ADD-GCN, SSGRL, and C-Tran. In particular, MSL-C outperforms MCAR by a relative improvement of

Table 1. **Performance comparison of MSL and baselines on VOC2007 using mAP, CR and CF1 metrics**. Boldface numbers indicate the best performance, whereas the best baselines are underlined. † indicates the results reproduced by the corresponding released codes or their modified versions. "*pre*" means pretrained on the MS-COCO dataset.

Method	mAP	CR	CF1
ResNet [15]	92.9	-	-
FeV+LV [25]	92.0	-	-
Atten-Reinforce [5]	92.0	-	-
RCP [24]	92.5	-	-
SSGRL [6]	93.4	-	-
SSGRL (*pre*) [6]	95.0	-	-
ML-GCN [9]	94.0	-	-
ADD-GCN [27]	93.6	-	-
BMML† (*pre*) [17]	95.0	-	-
IDA-R101 [20]	94.3	-	-
ASL [1]	94.6	-	-
MCAR [13]	94.8	-	-
CSRA† [31]	93.7	87.5	88.3
KGGR [4]	93.6	-	-
KGGR (*pre*) [4]	95.0	-	-
SST [8]	94.5	-	-
MSL-V	95.0	84.8	89.5
MSL-C	**96.1**	**92.4**	**91.6**

2.2% in terms of mAP. MCAR has two network streams that are trained jointly, and at inference predictions are fused from the two streams to generate a final prediction, whereas MSL has two streams with same weights in a Siamese-style network, which is much easier to optimize, and at inference a single network is used to make predictions. Moreover, MSL-C outperforms ADD-GCN, which uses a CNN and a GCN, by a relative improvement of 1.4% in terms of mAP.

In the last two rows of Figure 2, we show visual examples of predictions made by MSL-C and CSRA ResNet-cut as baseline on MS-COCO. A similar pattern can be observed, where MSL can recognize small objects and also objects under heavy occlusions compared to the baseline. It is worth mentioning that the variation of objects and their shapes or sizes are more complex in MS-COCO than those in VOC2007.

Overall, MSL is able to **learn context-based representations and to better model label co-occurrence** by masked branch and label consistency, thereby translating to better predictions in comparison with the baselines. MSL can **better recognize small objects and also objects under heavy occlusions**. MSL is also **very simple and much easier to train**, as **it does not require multiple stages of training, the combination of multiple learnable networks, large language models, high input resolution, complex data augmentation strategies, or additional data**.

4.3. Ablation Study

We analyze how each of the key components of the proposed MSL framework affects the final performance. We also perform hyperparameter sensitivity analysis.

Effectiveness of Masked Branch. Table 3 illustrates the benefit of using masked branch tasked to make predictions, given *partial inputs* by random masking. We adopt CSRA with ResNet-cut backbone as our baseline, and evaluate performance on VOC2007. We find that the masked branch improves performance in terms of mAP and other evaluation metrics. It helps learn useful representations, especially for small and occluded objects due largely to the fact that the branch is tasked to recognize **masked objects** (i.e., *partial inputs*), thereby leveraging information from neighboring objects.

Effectiveness of Label Consistency. As shown in Table 3, label consistency helps improves performance in terms of mAP and other metrics. This constraint essentially guides the model to make accurate predictions on masked inputs by minimizing the distance between the predictions made by the recognition and masked branches. Basically, we push the predictions of the masked branch predictions and recognition branch predictions together and learn representations for *partial inputs*. As can be seen, the best performance is achieved when combining masked branch and label consistency. Also, Table 4 shows that MSL is model-agnostic, and can also improve performance of not only classical convolutional backbones, but also modern transformer backbones.

Effectiveness of Binarization. Table 5 shows the benefit of using binarization of masking during training. We find that applying binary thresholding to the masks significantly improves performance of the baseline in terms of all metrics. This is attributed to the fact binarization yields true masking, dropping certain pixels while retaining the rest, thereby resulting in better numerical stability. Without binarization, the image is slightly offset in the pixel space when multiplied by 0.884 instead of 1, yielding a different representation in the feature space that degrades performance.

Amount of Masking. In Table 6, we report the effect of the amount of masking on MSL performance. We adopt CSRA with ResNet-cut backbones, and evaluate performance on VOC2007 and MS-COCO, respectively. We find that applying extensive image masking during the training process leads to improved performance.

Hyperparameter Sensitivity Analysis. We adopt CSRA with ResNet-cut as a base model and apply MSL to evaluate its performance for various values of the trade-off hyperparameters α_1, α_2 and α_3 on VOC2007. Table 7 shows the ef-

Table 2. **Performance comparison of MSL and baselines on MS-COCO in terms of mAP and other evaluation metrics**. Boldface numbers indicate the best performance, whereas the best baselines are underlined. † indicates the results reproduced by the corresponding released codes or their modified versions.

Method	Input Resolution	mAP	CP	CR	CF1	OP	OR	OF1
ResNet [15]	448 × 448	79.4	83.4	66.6	74.0	86.8	71.1	78.2
PLA [26]	228 × 228	-	80.4	68.9	74.2	81.5	73.3	77.2
ResNet-cut† [15]	448 × 448	82.1	86.2	68.7	76.4	88.9	73.1	80.3
ML-GCN [9]	448 × 448	83.0	85.1	72.0	78.0	85.8	75.4	80.3
MS-CMA [28]	448 × 448	83.8	82.9	74.4	78.4	84.4	<u>77.9</u>	81.0
KSSNet [22]	448 × 448	83.7	84.6	73.2	77.2	87.8	76.2	81.5
MCAR [13]	448 × 448	83.8	85.0	72.1	78.0	88.0	73.9	80.3
TDRG† [30]	448 × 448	84.6	86.0	73.1	79.0	86.6	76.4	81.2
CSRA† [31]	448 × 448	84.3	83.5	74.3	78.6	85.1	77.2	81.0
Q2L-R101† [21]	448 × 448	84.0	82.0	75.8	78.8	83.3	78.8	81.0
IDA-R101 [20]	448 × 448	83.8	-	-	-	-	-	-
SST† [8]	448 × 448	84.2	86.1	72.1	78.5	87.2	75.4	80.8
P-GCN† [7]	448 × 448	83.2	84.9	72.7	78.3	85.0	76.4	80.5
KGGR† [4]	448 × 448	84.3	85.6	72.7	78.6	87.1	75.6	80.9
ADD-GCN [27]	576 × 576	<u>85.2</u>	84.7	75.9	80.1	84.9	79.4	<u>82.0</u>
SSGRL [6]	576 × 576	83.8	<u>89.9</u>	68.5	76.8	<u>91.3</u>	70.8	79.7
C-Tran [16]	576 × 576	85.1	86.3	<u>74.3</u>	<u>79.9</u>	87.7	76.5	81.7
MCAR [13]	576 × 576	84.5	84.3	73.9	78.7	86.9	76.1	81.1
MSL-C	448 × 448	**86.4**	**90.1**	**76.3**	**80.4**	**89.1**	**80.0**	**82.2**

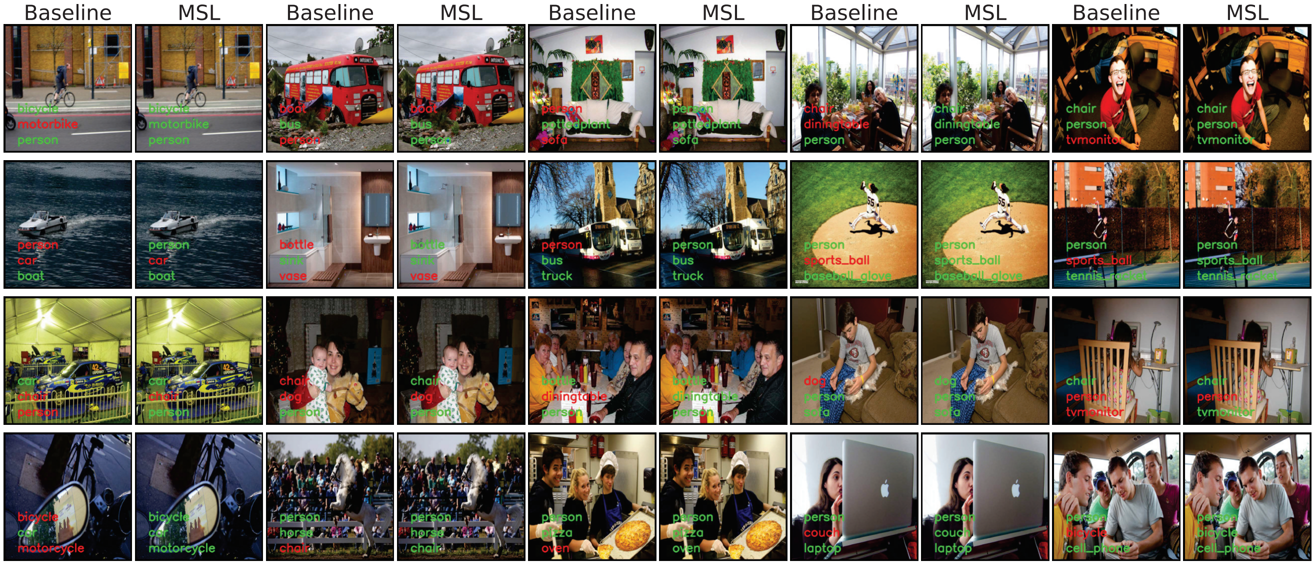


Figure 2. **Visual comparison of MSL and CSRA [31] on VOC2007 and MS-COCO**. First two rows show samples from VOC2007: the first row shows cases where MSL can better recognize small objects and the second row shows cases of heavy occlusions. The last two rows shows samples from MS-COCO. For both datasets, MSL is effective at recognizing small and occluded objects compared to the CSRA baseline. Zoom-in for better details.

fect of each hyperparameter on MSL performance in terms of mAP, CR and CF1. Interestingly, the best performance is achieved when the trade-off hypeparameters α_1 and α_2 are weighted almost equally. Moreover, using label consistency with $\alpha_3 = 0.5$ gives the best results. This suggests that the learned representations for partial inputs contribute to the

Table 3. **Effectiveness of masked branch and label consistency on MSL performance using VOC2007**. Using both masked branch and label consistency significantly improves the baseline performance.

Method	MaBr	LaCo	mAP	CR	CF1
Baseline			93.7	87.5	88.3
MSL	✓		94.0	89.1	88.4
MSL		✓	94.3	88.1	88.9
MSL	✓	✓	**96.1**	**92.4**	**91.6**

Table 4. **Comparison of different architectures trained using MSL on VOC2007 and MS-COCO**. MSL is model-agnostic and improves performance of different architectures.

Architecture	VOC2007, mAP (%)	MS-COCO, mAP (%)
ViT	94.4	76.8
+ MSL	**95.0**	**79.0**
ResNet	93.7	84.3
+ MSL	**96.1**	**86.4**

Table 5. **Ablation analysis of binarization in MSL using VOC2007**. Binarization helps achieve better numerical stability.

Binarization	mAP	CR	CF1
Baseline	93.7	87.5	88.3
+ w/o Binarization	93.8	88.2	88.2
+ w/ Binarization	**96.1**	**92.4**	**91.6**

Table 6. **Ablation analysis of *high*- and *low*-masked pixels during MSL training using VOC2007 and MS-COCO**. MSL with *high*-masked pixels yields better performance.

Masking	VOC2007, mAP (%)	MS-COCO, mAP (%)
Low	95.0	85.1
High	**96.1**	**86.4**

improvement of the overall performance.

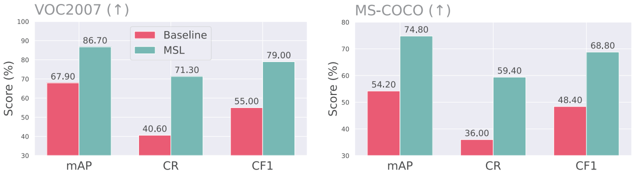
Table 7. Effect of hyperparameters on MSL performance using VOC2007.

α_1	α_2	α_3	mAP	CR	CF1
1	1	1	93.6	87.7	88.2
0.2	0.2	0.6	94.7	88.6	89.5
0.3	0.3	0.4	95.0	89.0	89.9
0.4	0.4	0.2	94.6	89.1	89.5
0.3	0.2	0.5	**96.1**	**92.4**	**91.6**

4.4. Robustness

We now examine the robustness of MSL against partial inputs and showcase its ability to predict non-masked objects.

Quantitative Results. We evaluate MSL against partial inputs by deliberately masking the input images before making a prediction. In Figure 3(a), we show comparison results of MSL against CSRA with ResNet-cut on VOC2007. Using MSL, mAP is improved by 19.8%, while CR and CF1 are improved by 30.7% and 24%, respectively. A similar trend is observed when comparing MSL against CSRA with ResNet-cut on MS-COCO, as depicted in Figure 3(b), achieving a mAP improvement of 20.6%. This shows that MSL is robust to heavily masked inputs, and hence occlusions.



(a) ResNet-cut on VOC2007. (b) ResNet-cut on MS-COCO.

Figure 3. **Performance comparisons when provided randomly masked images at test-time on VOC2007 and MS-COCO**. MSL is robust to heavily masked inputs, and hence occlusions.

Qualitative Results. In Figure 4, we show visual comparisons of the top three predictions by our approach compared to the baseline when making predictions on masked inputs. We can see that the baseline fails to make predictions when the input image is masked. Even in cases where the object is slightly masked, the baseline fails to make a prediction. By comparison, our model is able to **recognize objects that are heavily masked** thanks to the masked branch. Also, there are cases where the **object is almost completely masked, but still our method is able to make a prediction**. This is largely attributed to label consistency, where the target label can be inferred from the other predicted labels.

Non-Masked Objects. We also highlight an interesting property of MSL predictions in Figure 5, which shows that our model **predicts non-masked objects** better than the baseline. We hypothesize that this is due in part to the initial features or cues that the model needs to focus on.

Comparison with random masking strategy. While the Masked Autoencoder (MAE) [14] is a well-established masking strategy frequently employed in self-supervised learning, the key novelty of our MSL framework lies in the application of a masking strategy within the context of supervised learning. This novel utilization of masking during supervised learning sets our approach apart from existing methods. Moreover, MAE follows a two-step process: first,

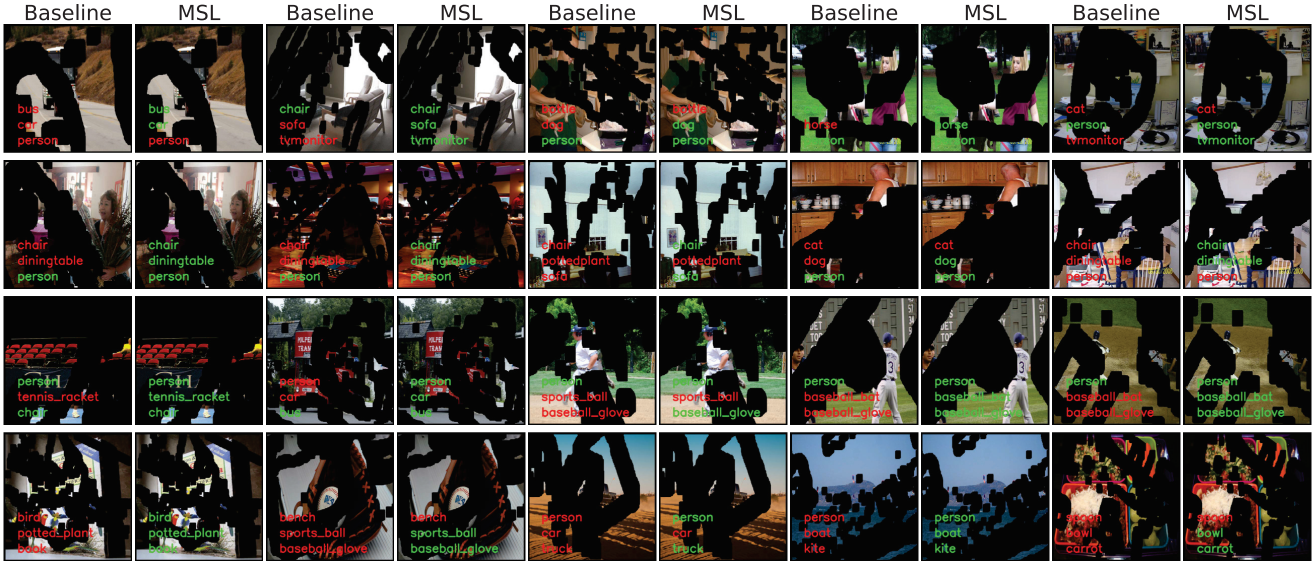


Figure 4. **Visual comparison of predictions made by MSL and baseline for masked input images on VOC2007 (first two rows) and MS-COCO (last two rows) datasets**. MSL is able recognize objects that are heavily masked and performs well in cases where the object is almost completely masked.

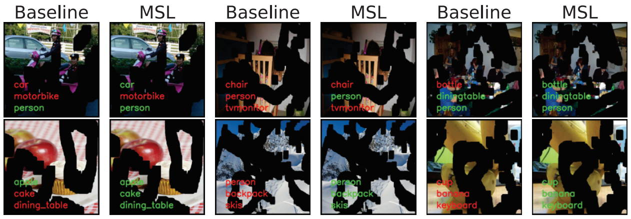


Figure 5. **Comparison of MSL and baseline on VOC2007 (first row) and MS-COCO (second row)**. Interestingly, MSL yields better prediction of non-masked objects.

it undergoes pre-training for 800 epochs exclusively on images, and then it proceeds to fine-tune for an additional 50 epochs using both images and labels. In contrast, MSL requires only a single stage of training, lasting 60 epochs, utilizing both images and labels. To compare the performance of MSL and MAE, we present the results in Table 8, which demonstrates the superiority of MSL over MAE in terms of mAP on both VOC2007 and MS-COCO datasets.

Table 8. **Performance comparison of MSL and MAE in mAP.**

Masking	VOC2007	MS-COCO
MAE [14]	95.3	85.5
MSL	**96.1**	**86.4**

Comparison with CSRA variants. In Table 9, we compare CSRA variants and MSL variants on VOC2007 and MS-COCO. As can be seen, MSL yields improved performance for both transformer and convolutional backbones.

Table 9. **Performance comparison of MSL and CSRA variants**.

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**

5. Conclusion

In this paper, we presented a single-stage, model-agnostic learning paradigm using masking. The proposed paradigm, which is motivated by the intuition that occluded objects are *partial inputs*, enables models to explicitly learn context-based representations and to model the label co-occurrence. We showed through extensive experiments that our method surpasses state-of-the-art models that heavily depend on multiple stages of training, high input resolution, the combination of multiple networks, large language models, complex data augmentation strategies, and additional data. We also demonstrated that MSL is robust to masked partial inputs for large and small objects, which is a strong indicator of its ability to handle challenging cases of small and occluded objects. Our method distinguishes itself from previous approaches due to its simple and straightforward training process, with the added benefit of incurring only a minor computational overhead compared to those methods. For future work, we aim to adapt the proposed framework to other computer vision tasks such as object detection.

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, 4, 5

[2] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, 1993. 3

[3] Long Chen, Wujing Zhan, Wei Tian, Yuhang He, and Qin Zou. Deep integration: A multi-label architecture for road scene recognition. *IEEE Transactions on Image Processing*, 28:4883–4898, 2019. 1

[4] Tianshui Chen, Liang Lin, Xiaolu Hui, Riquan Chen, and Hefeng Wu. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 4, 5, 6

[5] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. Recurrent attentional reinforcement learning for multi-label image recognition. In *Proc. AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2, 5

[6] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proc. IEEE International Conference on Computer Vision*, pages 522–531, 2019. 1, 2, 4, 5, 6

[7] Zhaomin Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Learning graph convolutional networks for multi-label recognition and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1, 2, 4, 6

[8] 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, 4, 5, 6

[9] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proc. IEEE International Conference on Computer Vision*, pages 5177–5186, 2019. 1, 2, 4, 5, 6

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186, 2019. 2, 4

[11] 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. 4

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 4

[13] 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, 4, 5, 6

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 7, 8

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4, 5, 6

[16] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021. 1, 2, 4, 6

[17] Peng Li, Peng Chen, Yonghong Xie, and Dezheng Zhang. Bi-modal learning with channel-wise attention for multi-label image classification. *IEEE Access*, 8:9965–9977, 2020. 1, 2, 4, 5

[18] 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 *Proc. European Conference on Computer Vision*, pages 740–755. Springer, 2014. 4

[19] 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. 2

[20] Ruyang Liu, Jingjia Huang, Thomas H Li, and Ge Li. Causality compensated attention for contextual biased visual recognition. In *International Conference on Learning Representations*, 2023. 4, 5, 6

[21] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021. 1, 2, 4, 6

[22] Yongcheng Liu, Lu Sheng, Jing Shao, Junjie Yan, Shiming Xiang, and Chunhong Pan. Multi-label image classification via knowledge distillation from weakly-supervised detection. In *Proc. ACM International Conference on Multimedia*, pages 700–708, 2018. 1, 2, 4, 6

[23] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proc. IEEE Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 2

[24] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proc. IEEE International Conference on Computer Vision*, pages 464–472, 2017. 5

[25] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Exploit bounding box annotations for multi-label object recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–288, 2016. 5

[26] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2020. 6

[27] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *Proc. European Conference on Computer Vision*, pages 649–665. Springer, 2020. 1, 2, 4, 5, 6

[28] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proc. AAAI Conference on Artificial Intelligence*, volume 34, pages 12709–12716, 2020. 6

[29] 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. 4

[30] Jiawei Zhao, Ke Yan, Yifan Zhao, Xiaowei Guo, Feiyue Huang, and Jia Li. Transformer-based dual relation graph for multi-label image recognition. In *Proc. IEEE International Conference on Computer Vision*, pages 163–172, 2021. 2, 4, 6

[31] 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, 5, 6

[32] Hasib Zunair and A. Ben Hamza. Masked supervised learning for semantic segmentation. In *Proc. British Machine Vision Conference*, 2022. 2