Learning to Recognize Occluded and Small Objects with Partial Inputs

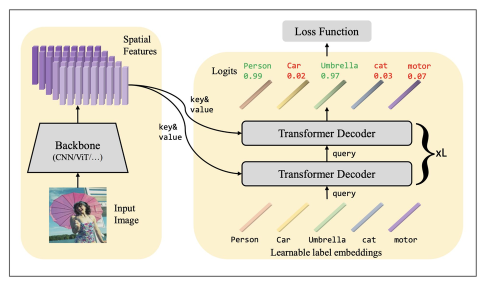
Authors: *Hasib Zunair* and A. Ben Hamza | Affiliation: Concordia University, Montreal, Canada | Paper ID: 678

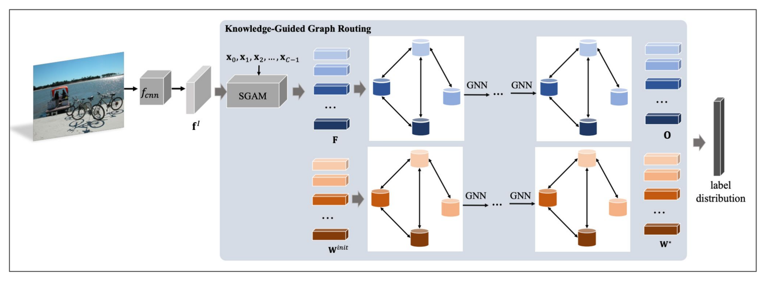
Abstract

Motivation: Multi-label image recognition is challenging due to large variations in the size and spatial location of objects, even harder when objects are occluded and small.

Specific Problem: Existing methods are complex in terms of both model and data; they do not explicitly address problem of small objects and occlusions

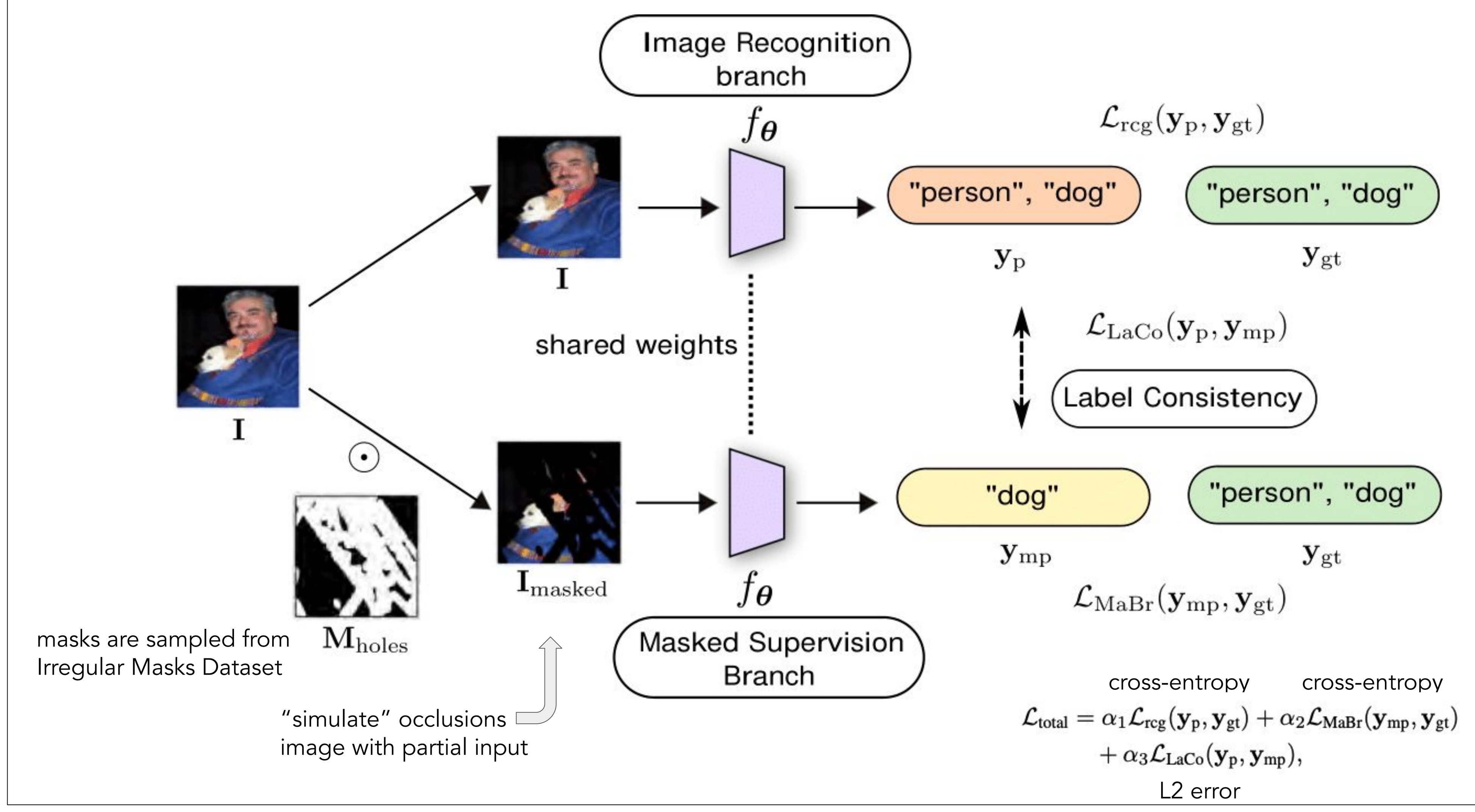
Model
- Multiple stages of training
- Combination of multiple learnable networks
- Relies on large language models



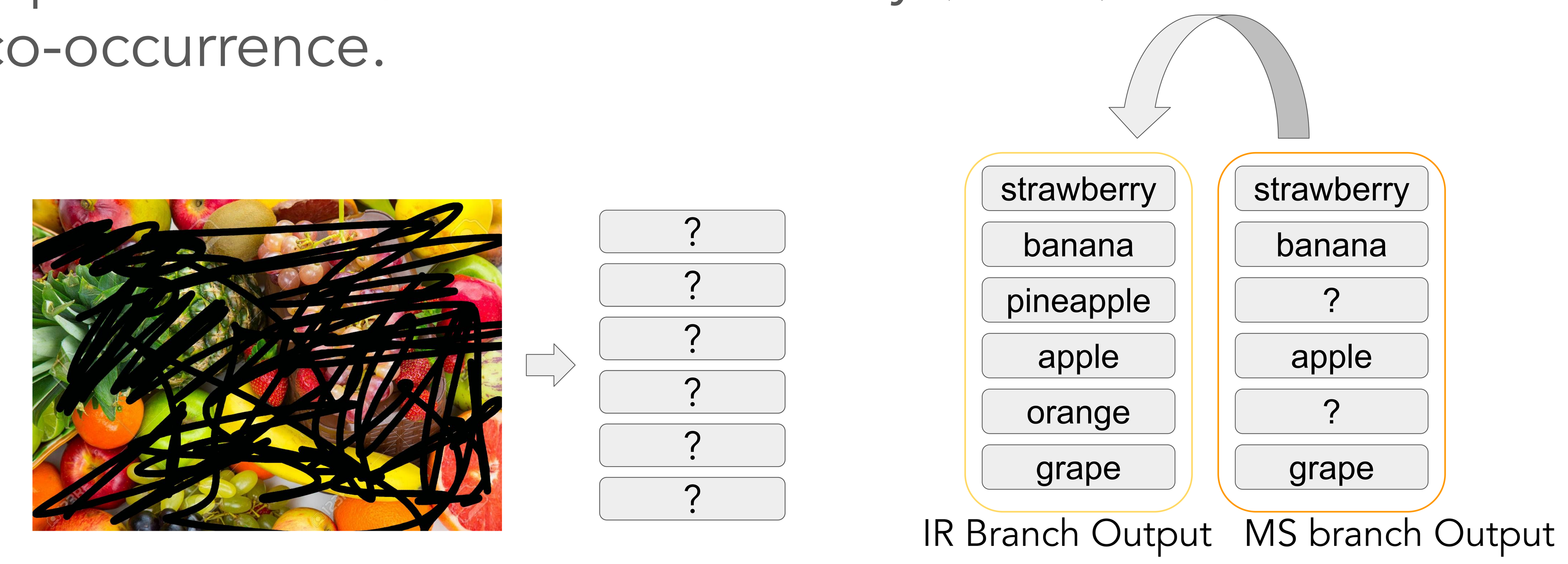


Data:
- High input resolution
- Complex data augmentation
- Additional data

Masked Supervised Learning (MSL)



Masked Branch (MaBr) aims to learn context based representation; Label Consistency (LaCo) models label co-occurrence.



MaBr: uses nearby non-masked regions around objects to make predictions for partly visible/masked objects

LaCo: learn a distribution of association across different classes "use context to infer presence of a class"

Results

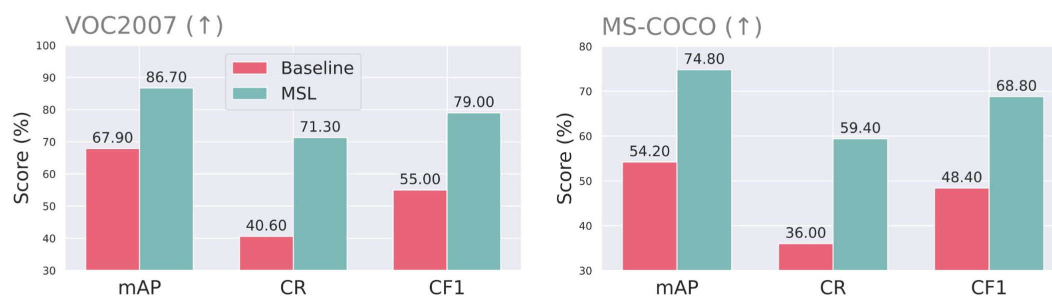
Outperforms SOTA methods on VOC2007, MS-COCO and WIDER-Attribute.

Method	mAP	CR	CF1
ResNet [15]	92.9	-	-
FeV+LV [26]	92.0	-	-
Atten-Reinforce [5]	92.0	-	-
RCP [25]	92.5	-	-
SSGRL [6]	93.4	-	-
SSGRL (*pre*) [6]	95.0	-	-
ML-GCN [9]	94.0	-	-
ADD-GCN [28]	93.6	-	-
BMML† (*pre*) [17]	95.0	-	-
IDA-R101 [21]	94.3	-	-
ASL [1]	94.6	-	-
MCAR [13]	94.8	-	-
CSRA† [32]	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**

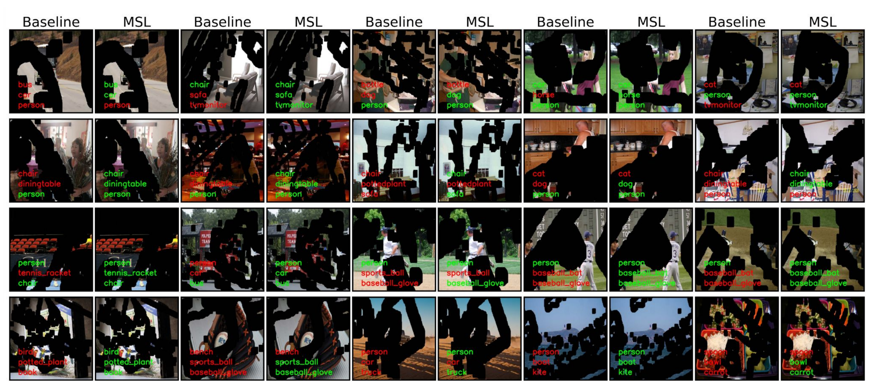
Method	Input Resolution	mAP	CP	CR	CF1	OP	OR	OF1
ResNet [15]	448 × 448	79.4	83.4	66.6	74.0	86.8	71.1	78.2
PLA [27]	228 × 228	-	80.4	68.9	74.2	81.5	73.3	77.2
ResNet-cut† [15]	448 × 448	82.1	86.2	68.7	76.4	88.9	73.1	80.3
ML-GCN [9]	448 × 448	83.0	85.1	72.0	78.0	85.8	75.4	80.3
MS-CMA [29]	448 × 448	83.8	82.9	74.4	78.4	84.4	77.9	81.0
KSSNet [23]	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† [31]	448 × 448	84.6	86.0	73.1	79.0	86.6	76.4	81.2
CSRA† [32]	448 × 448	84.3	83.5	74.3	78.6	85.1	77.2	81.0
Q2L-R101† [22]	448 × 448	84.0	82.0	75.8	78.8	83.3	78.8	81.0
IDA-R101 [21]	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 [28]	576 × 576	85.2	84.7	75.9	80.1	84.9	79.4	82.0
SSGRL [6]	576 × 576	83.8	89.9	68.5	76.8	91.3	70.8	79.7
C-Tran [16]	576 × 576	85.1	86.3	74.3	79.9	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**

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**

Robust to heavily masked inputs; would be good at occlusions naturally.



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



Better than self-supervised method; better at heavy masking; generic and applicable to any model architecture.

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

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**

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**

Method	VOC2007, mAP (%)
MCAR [13]	94.8
MCAR [13] w/ MSL	**95.6**
SST [8]	94.5
SST [8] w/ MSL	**95.8**

Conclusion

MSL is a simple yet effective single-stage, model-agnostic learning paradigm using masking. Use context on both image- and label-level to infer presence of multiple objects, even in cases where object are small and occluded.

Compare to recent methods, MSL does not require:
- multiple stages of training, combination of multiple networks, reliance on large language models (LLMs)
- high input resolution, complex data augmentation strategies, additional data for pretraining

TLDR: MSL is a simple and effective single-stage model-agnostic learning paradigm for visual learning tasks. It is similar to how us humans use context to perceive the visual world. Do try it!