Learning Contextual Vision Representations via Masking

Hasib Zunair
Concordia University, Montreal, Canada

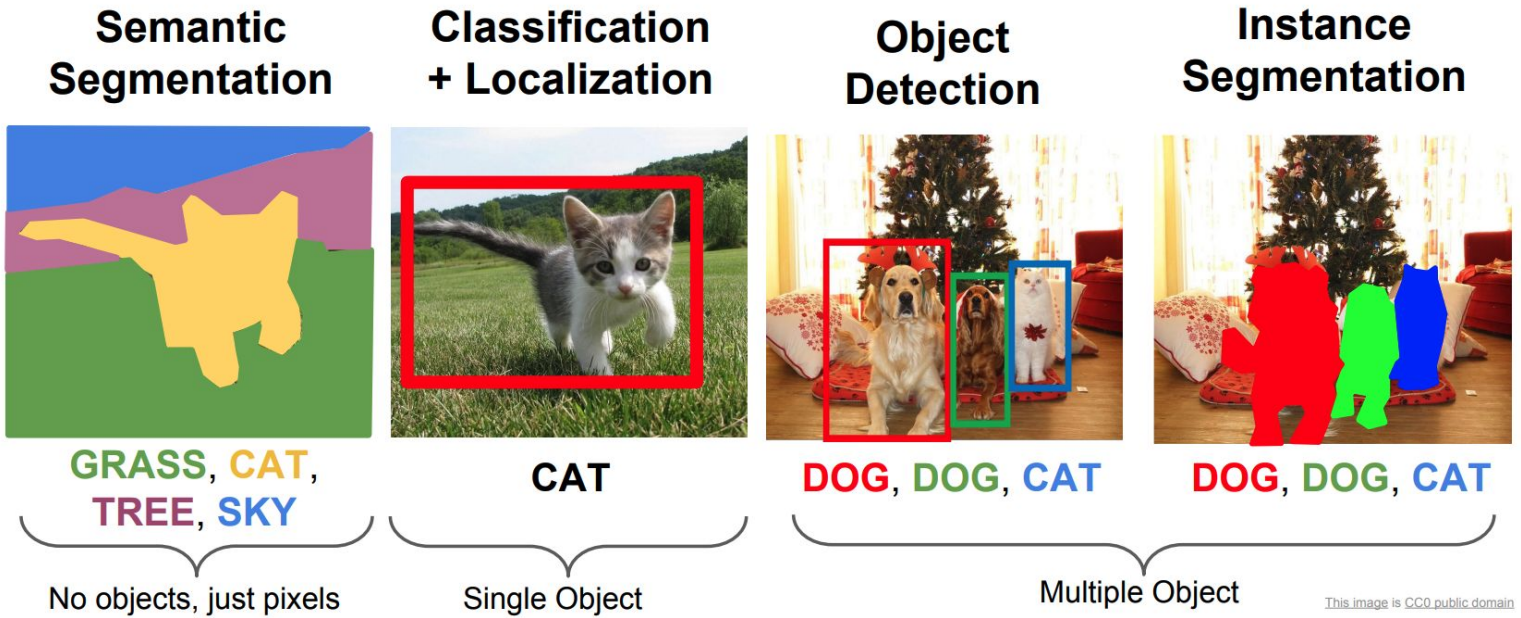
PhD Oral Exam, December 12, 2024

Outline

- Computer vision, approaches and problem
- Masked Supervised Learning for Semantic Segmentation
- Learning to Recognize Occluded and Small Objects with Partial Inputs
- Hiding Parts of an Image for Unsupervised Object Localization
- Next steps

Université Concordia University

Computer vision makes sense of data such as images and videos, just like humans do from examples



Deep learning uses neural networks to learn from data, allows for more accurate and versatile systems.

Image credit: Link

Applications are everywhere from healthcare, robotics, retail, agriculture, to entertainment etc.

AR/VR and Mixed Reality for Gaming, Immersive experience

Sort and filter photos of friends and family

Industrial and domestic robots

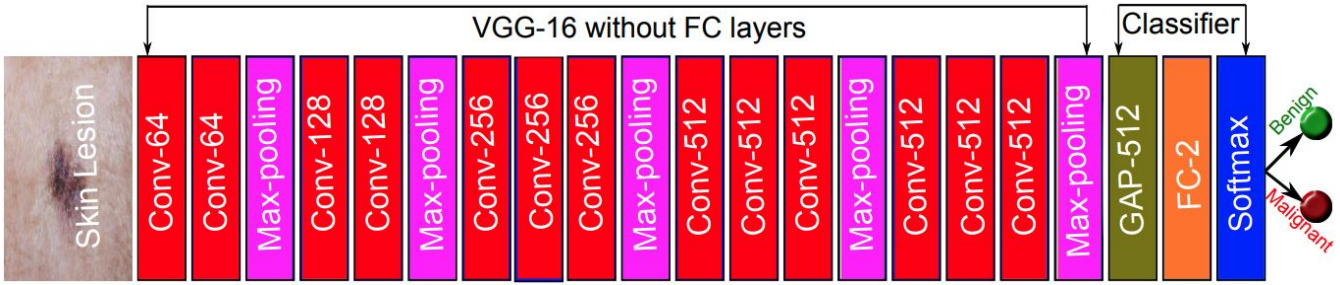
Image credit: Link

Computer vision using deep learning in real-world is hard, even for humans!

Mainly due to viewpoint difference, intraclass variability/shapes/size, deformation, illumination, background clutter and occlusion.

Image credit: Link

Supervised Learning is a common approach used to solve computer vision tasks



An overview of supervised learning for an image-level visual task.

- Convolutional Neural Networks (ResNet, ConvNeXt)
- Transformers (Vision & Swin Transformer)
- Encoder-Decoder Networks (U-Net, U-Net++)
- Generative Models (Diffusion, CGANs, StyleGAN)

Zunair, et al, https://arxiv.org/abs/2004.06824

Supervised Learning heavily relies on extensive input and human-annotated output pairs to work well

Example of novel (in ~2020) COVID-19 infected lung chest X-ray scans.

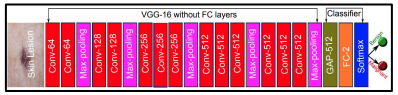
- Building large labeled datasets are time-consuming, costly and error-prone
- Sometimes infeasible (e.g. medical scans, need doctors to label, privacy issues)

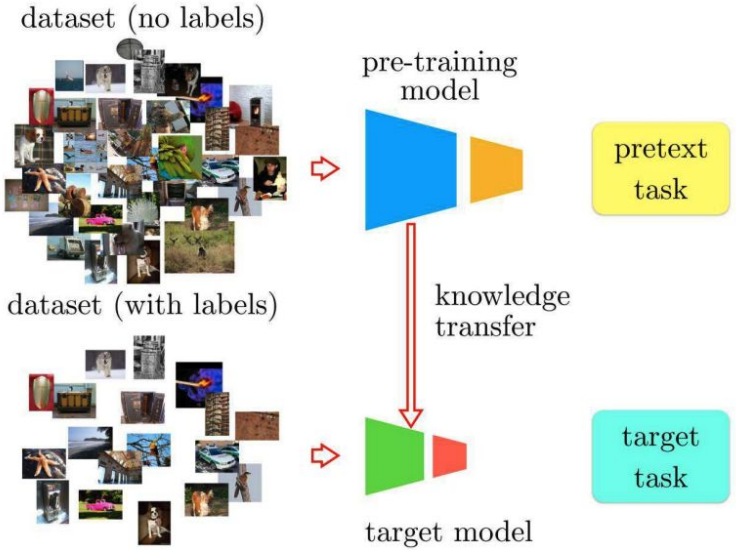
Zunair, et al, https://arxiv.org/abs/2106.09759

Self-supervised learning (SSL) leverages the semantics and structure within the data to learn useful representations without labels

Stage 1:
Pre-training
or *pretext* task

Stage 2: Fine-tuning
or *downstream* task
(e.g. classify skin lesions)





Made up or proxy task with "pseudo labels" for free to learn image representations from unlabeled data. (e.g., rotation prediction, colorization)

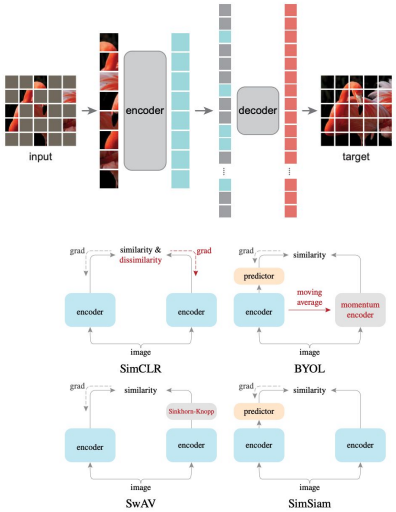
Two-stage method to address the challenges posed by over-dependence on labeled data for supervised learning.

Noroozi, et al, https://arxiv.org/abs/1805.00385

We can use SSL with unlabeled images and few labeled images to train models for our tasks, but..

Questions?

- What pretext tasks works well for a problem?
- How many unlabeled samples are needed?



Drawbacks

- Heavy compute demand due to two stages
- Discrepancy in pre-training and fine-tuning can lead to overfitting

He, et al, https://arxiv.org/abs/2111.06377
Chen, et al., https://arxiv.org/abs/2011.10566
Bardes, et al, https://arxiv.org/pdf/2105.04906

Research questions asked in this thesis?

- To achieve good predictive performance, recent works rely on time-consuming training strategies, computationally demanding model architectures and large volumes of data.
  - Can we develop a learning technique that is **simple**, **accurate** as well as **compute**- and **data-efficient**?

- Humans use context that significantly enhance their ability to perceive and recognize objects within a scene.
  - Can we learn **context** in supervised settings for artificial vision systems?

Wit, et al, https://doi.org/10.3389/fpsyg.2022.903960
Willems, et al, https://doi.org/10.1016/j.isci.2021.102392

Masked Supervised Learning for Semantic Segmentation

BMVC, 2022 (Oral Presentation)

We tackle the task of semantic segmentation; applications in medical imaging, agriculture etc.

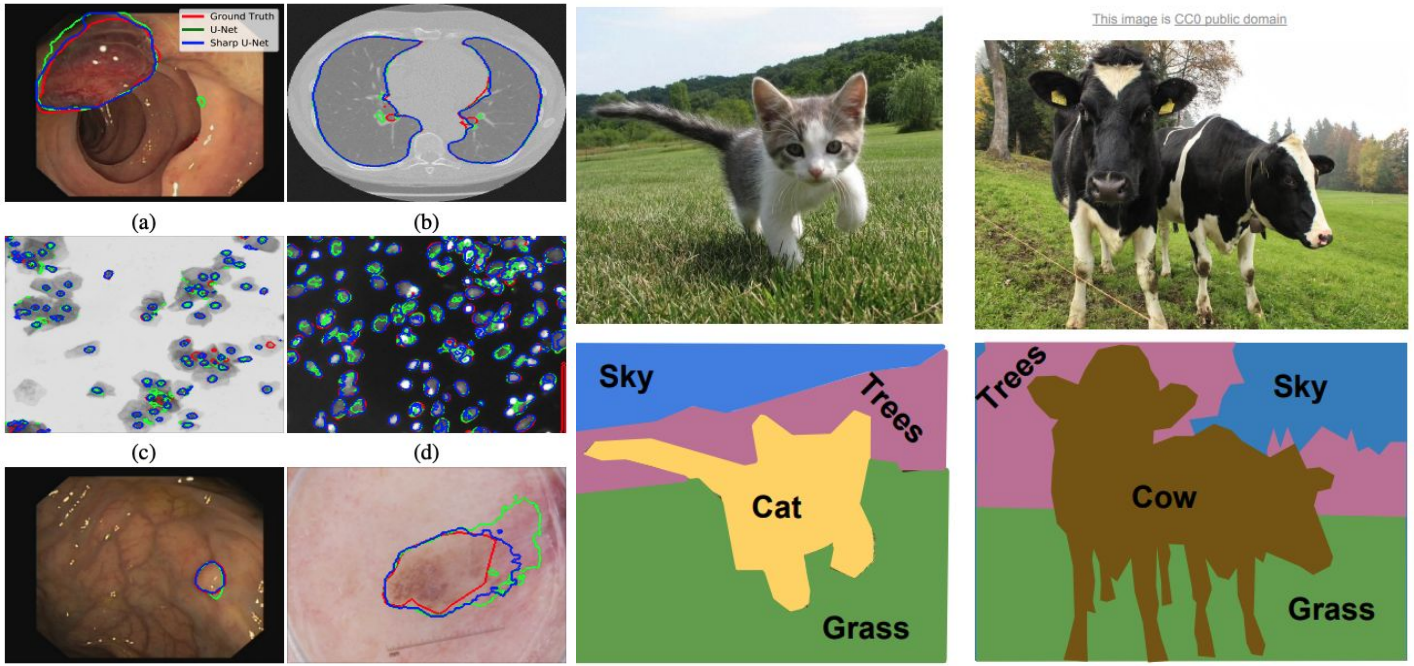
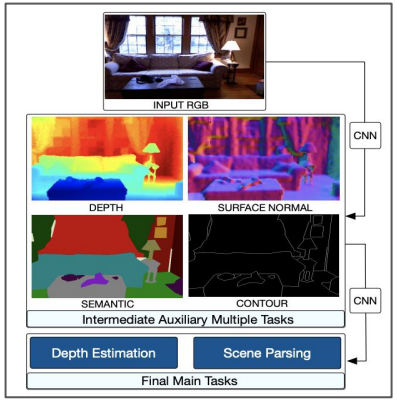


Image credit: Link 1 and Link 2.

Existing segmentation methods are complex and fail in several real-world edge cases

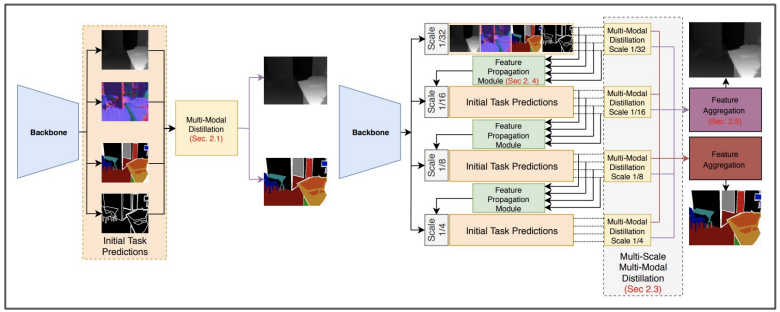
Method Complexity

- Multi-task learning (PAD-Net, HybridNet, MTI-Net)
- Self-supervised learning (Masked Autoencoders)
- Additional annotated data



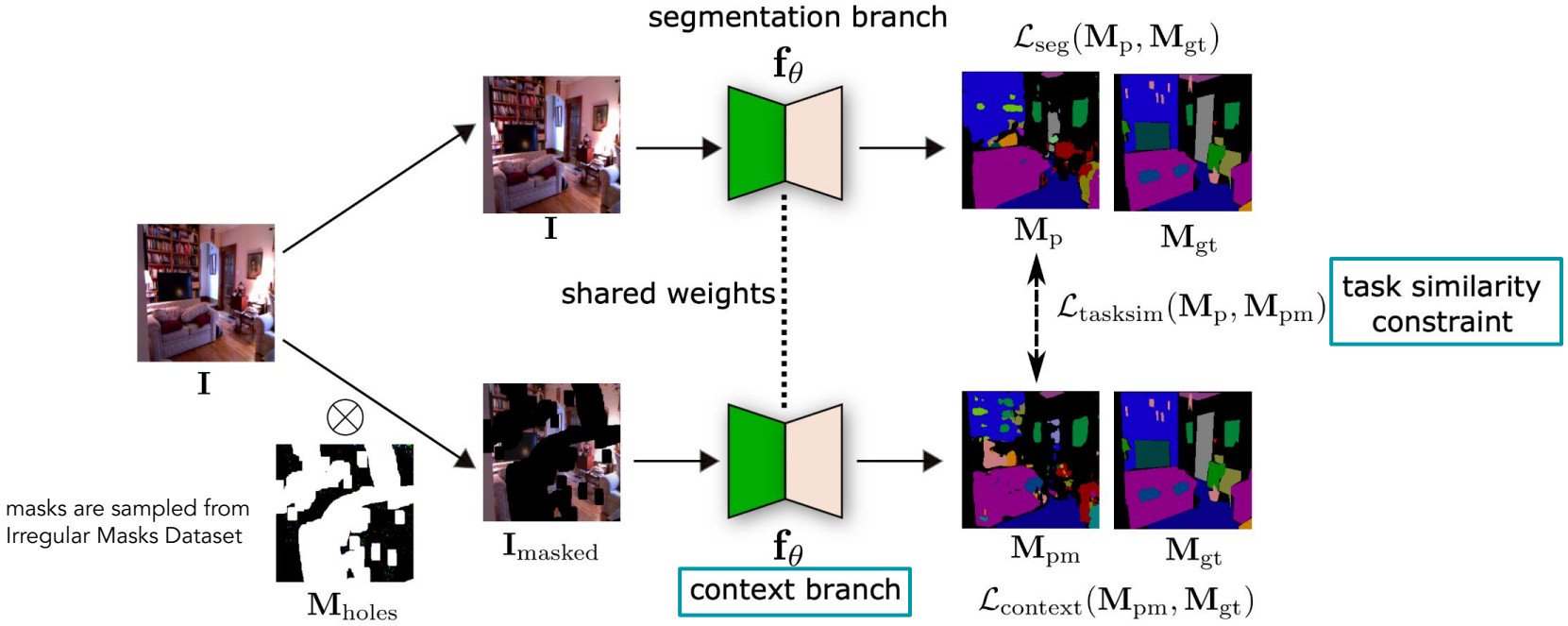
Fails to segment

- Small and ambiguous regions
- Minority class instances



Xu, et al, https://arxiv.org/abs/1805.04409
Vandenhende, et al, https://arxiv.org/abs/2001.06902

We propose Masked Supervised Learning as a new deep visual learning paradigm

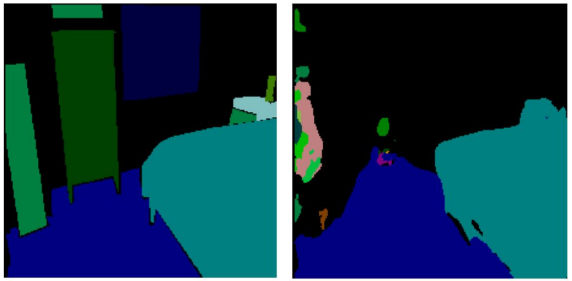


$$\mathcal{L}_{\text{total}} = \alpha_1 \mathcal{L}_{\text{seg}}(\mathbf{M}_\text{p}, \mathbf{M}_\text{gt}) + \alpha_2 \mathcal{L}_{\text{context}}(\mathbf{M}_\text{pm}, \mathbf{M}_\text{gt}) + \alpha_3 \mathcal{L}_{\text{tasksim}}(\mathbf{M}_\text{p}, \mathbf{M}_\text{pm})$$

cross-entropy | cross-entropy | L2 error

Context Branch predicts class of masked pixel
Task Similarity maximizes predictions from both unmasked and masked inputs

CB: uses nearby non-masked pixels to make predictions for masked pixels



TS: learns shape level representations by ensuring consistency between predictions of unmasked and masked object

- CB learns short-range context from using nearby pixels
- TS models long-range context by learning shape of class instance(s)

MaskSup outperforms state-of-the-art medical image segmentation and multi-task learning methods

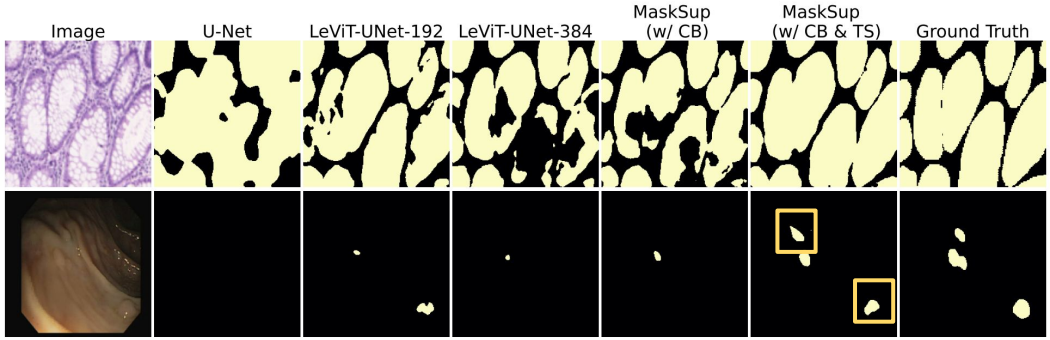
Method	GLaS, mIoU (↑)	CVC-Clinic-DB, mIoU (↑)	NYUDv2 (↑)
U-Net [18]	67.41	69.74	33.60
FCN [15]	50.84	-	29.20
U-Net++[30]	69.10	72.90	34.74
HRNet-18 [26]	-	-	33.18
ResU-Net [27]	65.95	-	-
ResU-Net++ [9]	-	79.60	-
SFA [7]	-	60.70	-
Attention U-Net [17]	-	82.70	-
Axial Attention U-Net [25]	63.03	-	-
MedT [21]	69.61	-	-
KiU-Net [22]	72.78	-	-
LeViT-UNet-128 [29]	70.45	-	-
LeViT-UNet-192 [29]	71.83	79.16	-
LeViT-UNet-384 [29]	73.88	81.38	-
PAD-Net [28] △	-	-	33.10
HybridNet A2 [13] △	-	-	34.30
MTI-Net [23] △	-	-	37.49
MaskSup (**Ours**)	**76.06**	**84.02**	**39.31**

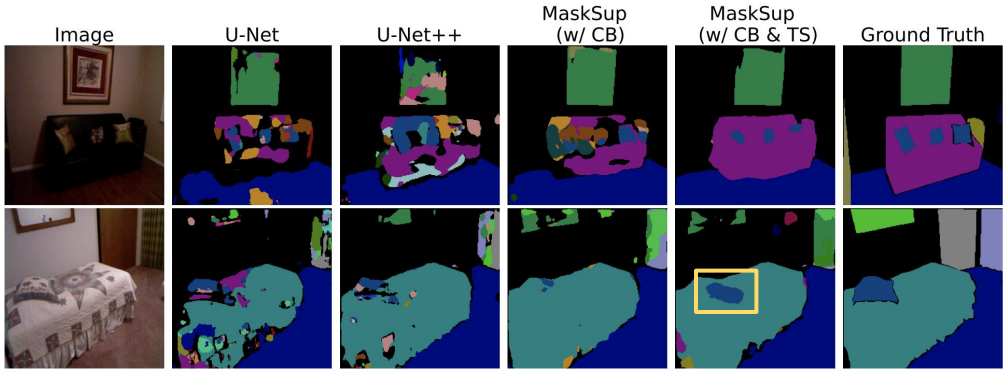
Method	GLaS, mIoU (↑)	CVC-Clinic-DB, mIoU (↑)	NYUDv2 (↑)
MAE [8]	75.04	82.50	37.42
MaskSup (**Ours**)	**76.06**	**84.02**	**39.31**

- MAE : 800 epochs of pre-training on unlabeled images, fine-tune for 50 epochs (Two stages)
- MaskSup: 200 epochs on labeled images and masks (Single stage)

Zunair, et al, https://arxiv.org/abs/2210.00923

MaskSup enables better segmentation of small and ambiguous regions as well as minority class instances

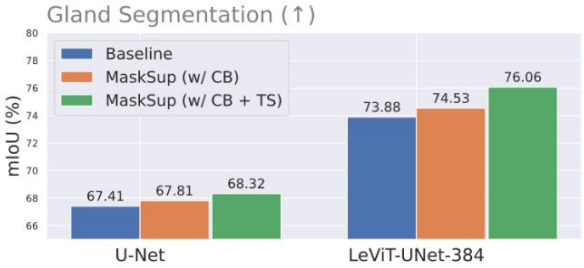


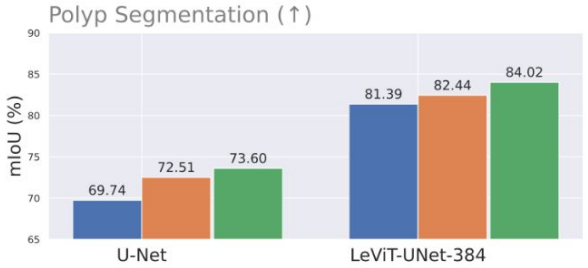


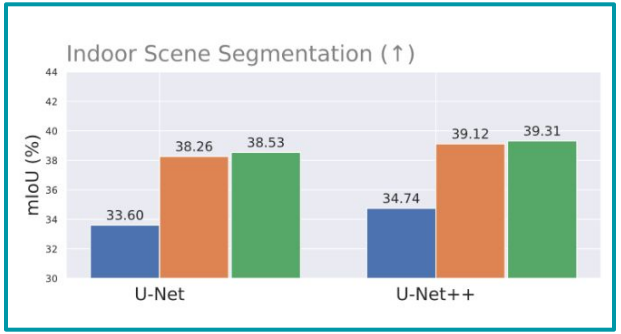
See paper for more!

Zunair, et al, https://arxiv.org/abs/2210.00923

MaskSup is a generic learning paradigm and is applicable to any model architecture



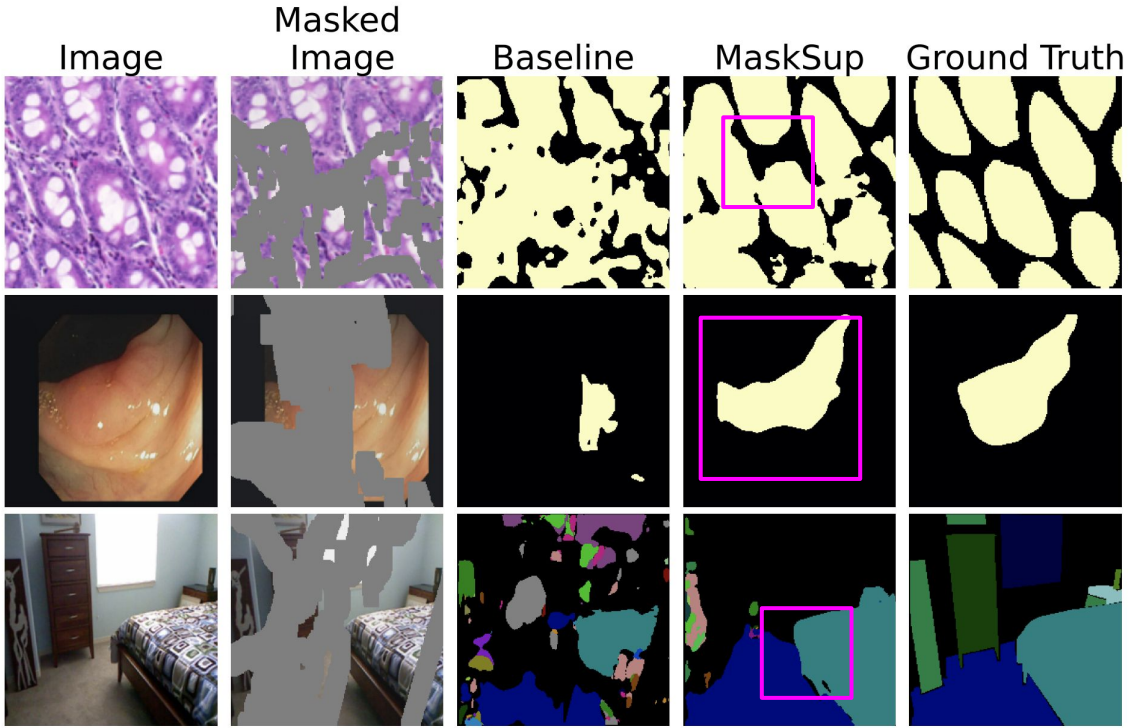




Context Branch and Task Similarity both improves performance of different segmentation models across multiple datasets and tasks (e.g. binary, multi-class segmentation).

Zunair, et al, https://arxiv.org/abs/2210.00918

MaskSup is shape aware, as it predicts even with heavily masked inputs



Zunair, et al, https://arxiv.org/abs/2210.00918

MaskSup is computationally efficient; achieves 10% mIoU improvement with 3x less parameters

Method	Params (M) (↓)	GLaS, mIoU (↑)	CVC-Clinic-DB, mIoU (↑)	NYUDv2 (↑)
LeViT-384 [29]	51	73.88	81.38	-
MaskSup (LeViT-192)	**19**(2.6x)	**74.44**(+0.75)	**82.17**(+0.97)	-
U-Net++ [30]	9	-	-	34.74
MaskSup (U-Net)	**3**(3x)	-	-	**38.54**(+10.91)

Zunair, et al, https://arxiv.org/abs/2210.00918

Learning to Recognize Occluded and Small Objects with Partial Inputs

WACV, 2024

We tackle the task of multi-label image recognition (MLIR); applications in media & entertainment, e-commerce, visual search engines etc.

	Multi-Class	Multi-Label
Image		
Labels	cat ✓ dog hamster rabbit fish	✓ dog long-haired ✓ glasses ears up ✓ collar

Image credit: Link

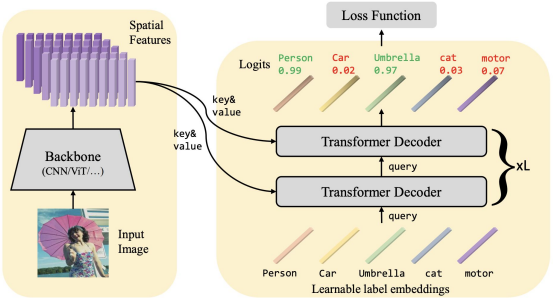
Recent methods are complex in terms of both model and data; they do not explicitly address problem of small objects and occlusions

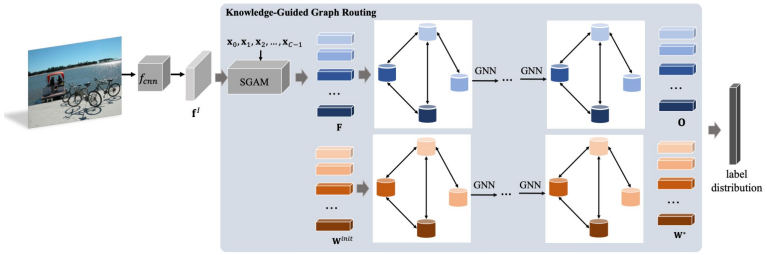
Model

- Multiple stages of training
- Combination of multiple learnable networks
- Relies on large language models (LLMs)

Data:

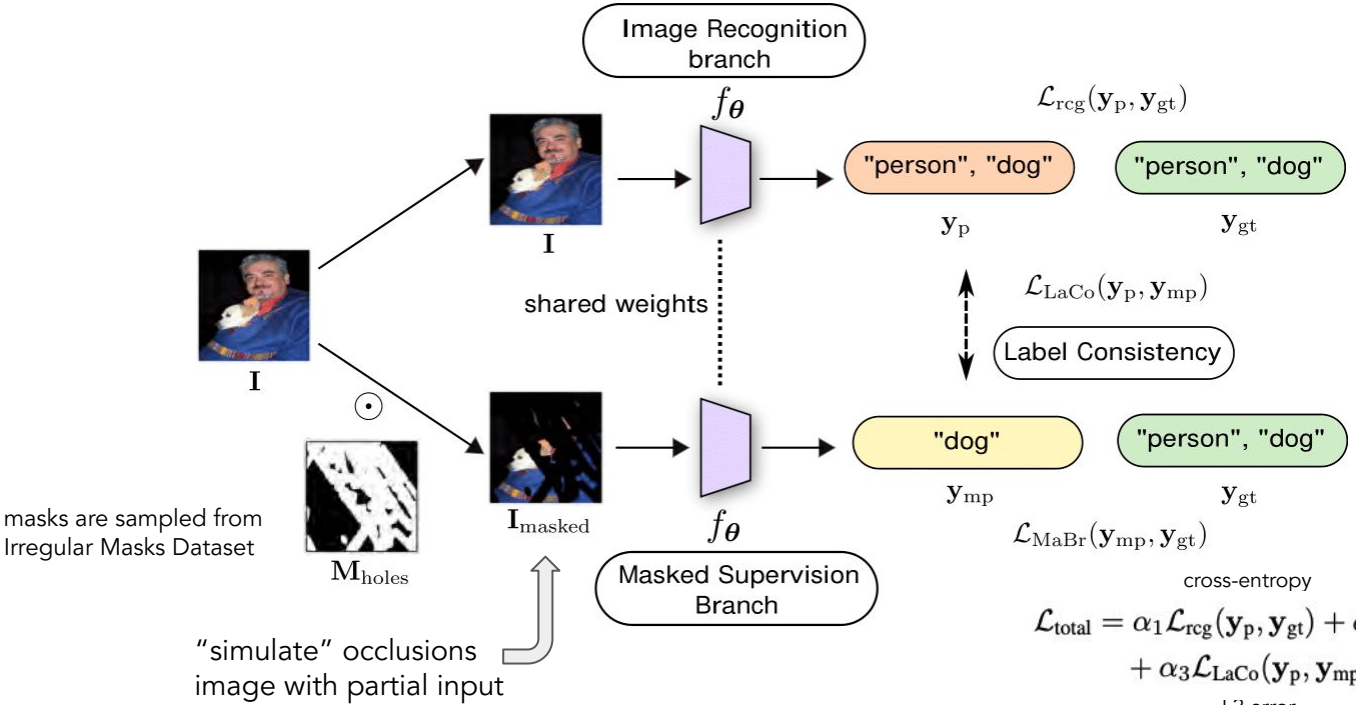
- High input resolution
- Complex data augmentation
- Additional data to train





Liu, et al, https://arxiv.org/abs/2107.10834
Chen, et al, https://arxiv.org/abs/2009.09450

We use Masked Supervised Learning as a training strategy for MLIR models

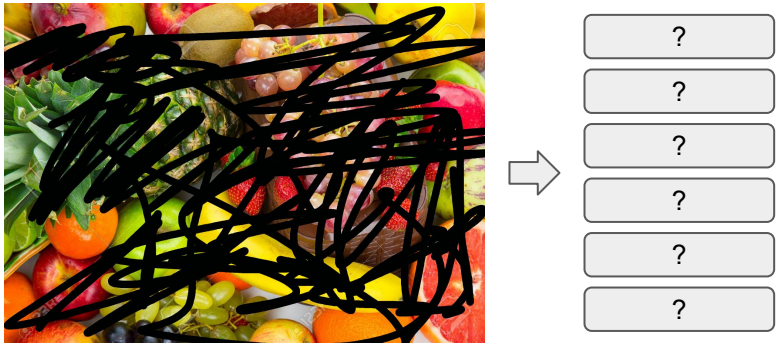


cross-entropy cross-entropy

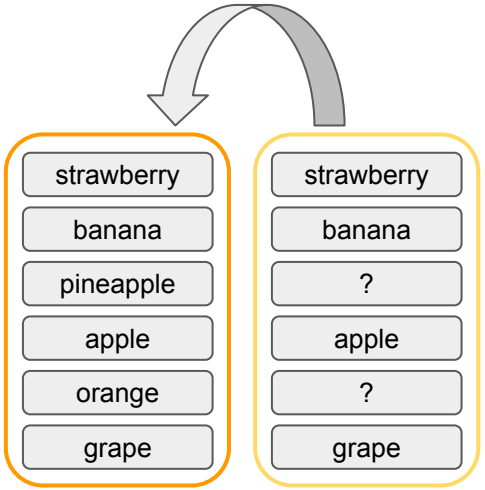
$$\mathcal{L}_{\text{total}} = \alpha_1 \mathcal{L}_{\text{rcg}}(\mathbf{y}_\text{p}, \mathbf{y}_\text{gt}) + \alpha_2 \mathcal{L}_{\text{MaBr}}(\mathbf{y}_\text{mp}, \mathbf{y}_\text{gt}) + \alpha_3 \mathcal{L}_{\text{LaCo}}(\mathbf{y}_\text{p}, \mathbf{y}_\text{mp}),$$

L2 error

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



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



IR Branch Output MS branch Output

LaCo: learn a distribution of association across different classes

"use context to infer presence of a class"

MSL outperforms SOTA MLIR and SSL methods on VOC2007, MS-COCO and WIDER-Attribute datasets

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**

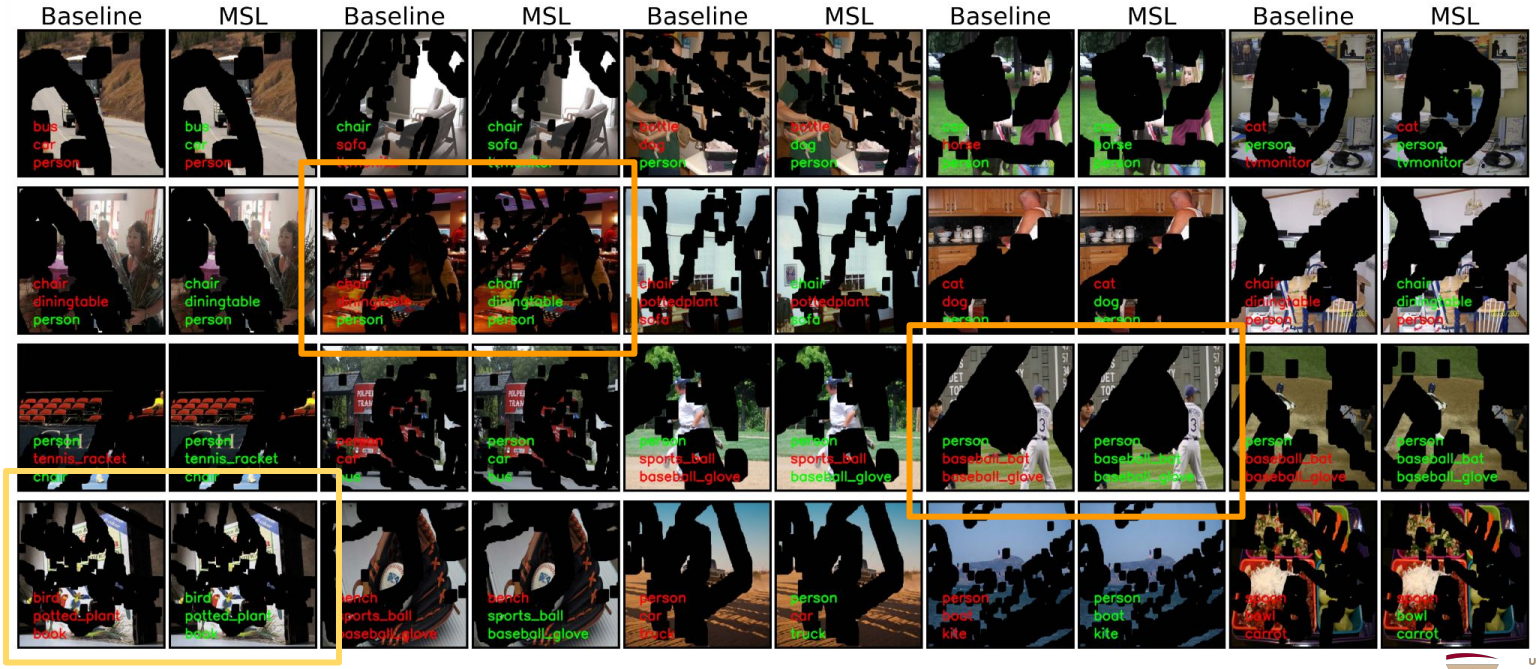
Masking	VOC2007	MS-COCO
MAE [14]	95.3	85.5
MSL	**96.1**	**86.4**

Other metrics: overall precision (OP), overall recall (OR), overall F1- measure (OF1), per-category precision (CP), per-category recall (CR), and per-category F1-measure (CF1)

MSL is accurate and computationally efficient compared to recent methods!

MSL is effective at recognizing small and occluded objects compared to baseline

MSL is able to recognize heavily masked objects and also works in cases where the object is almost completely masked



MaBr and LaCo improve performance, enjoys heavy masking; generic to any model architecture

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**

Both MaBr and LaCo improves performance.

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	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**

Increasing masking improves results.

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

Applicable to any model architecture.

PEEKABOO: Hiding Parts of an Image for Unsupervised Object Localization.

BMVC, 2024

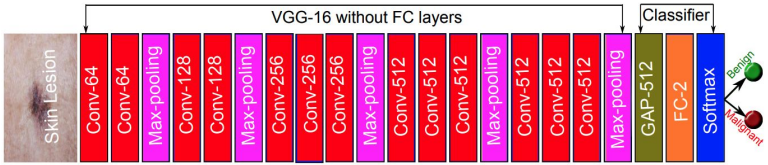
We tackle the task of object localization; applications in robotics, AR/VR headsets, etc.

Saliency Detection: foreground/background separation, binary mask

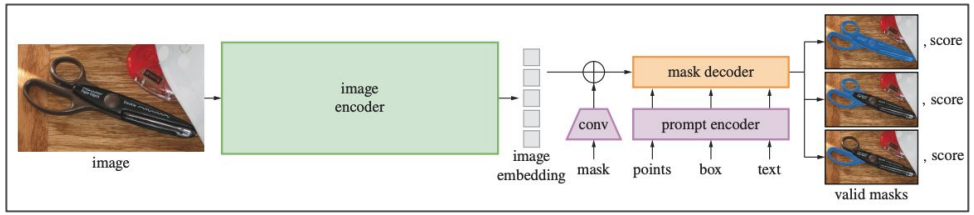
Single-object Discovery: enclose the *main* or *one of the main objects* of interest in a box

Image credit: Link

We aim to find unfamiliar and salient objects without any class specific training and prompting



An overview of supervised learning for an image-level visual task.



Segment Anything Model (SAM), Meta AI

- Already know it is time-consuming, costly to acquire annotated data
- Fails in cases of novel objects due to finite and pre-defined nature of object classes.
- Specialists not generalists

- Trained on 11M images with human annotations with 1B masks
- Need prompts like points, boxes or masks to indicate objects to segment
- Can't use for saliency detection or single-object discovery, requires training

Kirillov, et al, https://arxiv.org/abs/2304.02643

Recent unsupervised methods are complex and do not explicitly model visual context; fails when objects are small, or against complex or dim backgrounds

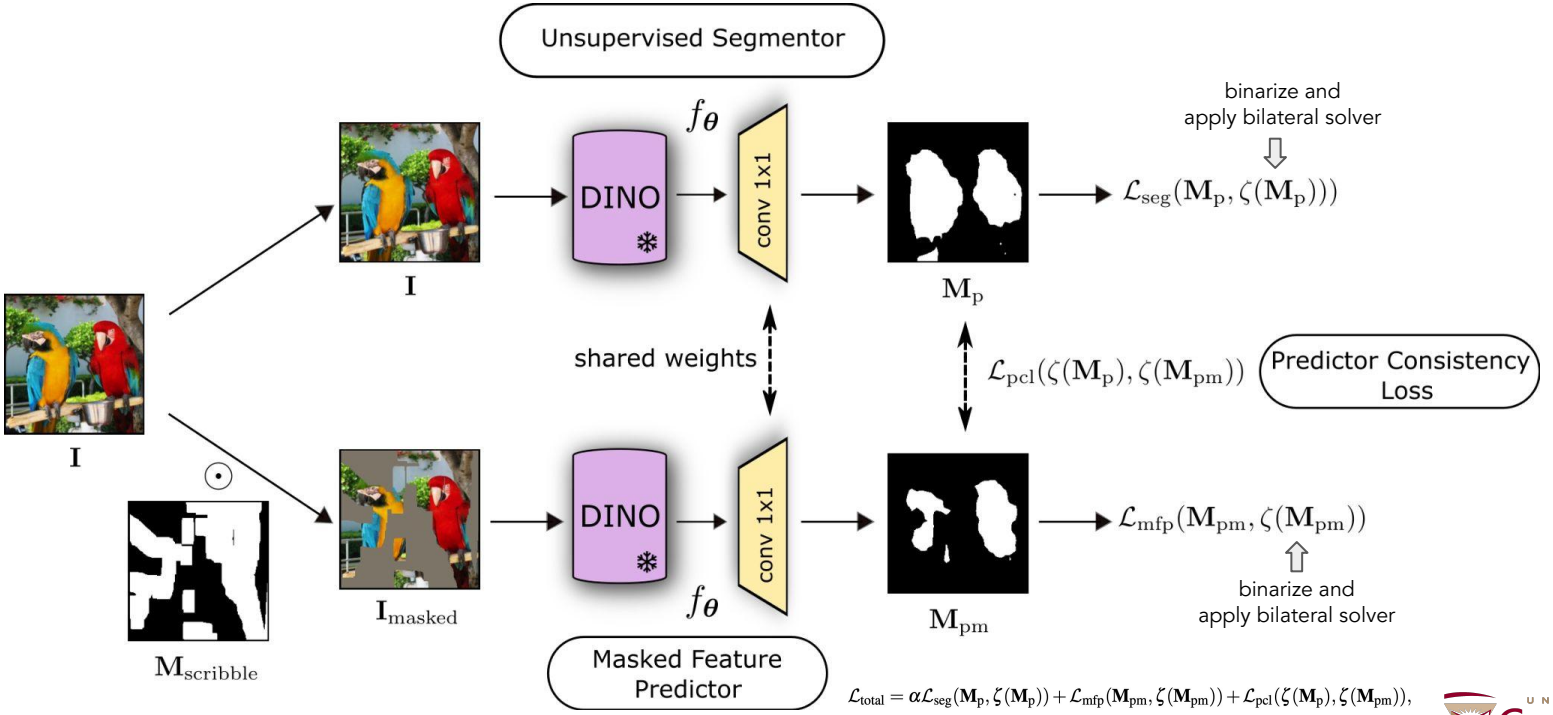
Model

- Multiple stages of training (e.g., COMUS, three days of training)
- Millions of learnable parameters (e.g., FreeSOLO has 66M params)
- Ensembles (e.g., Self-Mask with three vision encoders)

Data:

- Uses both large-scale real world and synthetic data (e.g., DINOSAUR uses 300K images)

Peekaboo is a self-supervised single-stage method to localize novel & salient objects, entirely unsupervised



Peekaboo outperforms SOTA methods on Saliency Detection and Single Object Discovery tasks

Method	Learning	DUT-OMRON Acc	DUT-OMRON IoU	DUT-OMRON max F_β	DUTS-TE Acc	DUTS-TE IoU	DUTS-TE max F_β	ECSSD Acc	ECSSD IoU	ECSSD max F_β
HS [42]		84.3	43.3	56.1	82.6	36.9	50.4	84.7	50.8	67.3
wCtr [67]		83.8	41.6	54.1	83.5	39.2	52.2	86.2	51.7	68.4
WSC [18]		86.5	38.7	52.3	86.2	38.4	52.8	85.2	49.8	68.3
DeepUSPS [25]		77.9	30.5	41.4	77.3	30.5	42.5	79.5	44.0	58.4
BigBiGAN [37]		85.6	45.3	54.9	87.8	49.8	60.8	89.9	67.2	78.2
E-BigBiGAN[37]		86.0	46.4	56.3	88.2	51.1	62.4	90.6	68.4	79.7
Melas-Kyriazi et al. [23]		88.3	50.9	-	89.3	52.8	-	91.5	71.3	-
LOST [30]		79.7	41.0	47.3	87.1	51.8	61.1	89.5	65.4	75.8
DSM [24]		80.8	42.8	55.3	84.1	47.1	62.1	86.4	64.5	78.5
TokenCut [40]		88.0	53.3	60.0	90.3	57.6	67.2	91.8	71.2	80.3
SelfMask [29]	✓	90.1	**58.2**	-	92.3	62.6	-	94.4	78.1	-
FOUND† [30]	✓	90.7	57.1	79.9	93.5	63.7	85.2	**94.9**	**80.6**	95.1
DeepCut [0]	✓	-	-	-	-	59.5	-	-	74.6	-
WSCUOD [22]	✓	89.7	53.6	64.4	91.7	59.9	73.1	92.2	72.7	85.4
PEEKABOO (Ours)	✓	**91.5**	57.5	**80.4**	**93.9**	**64.3**	**86.0**	94.6	79.8	**95.3**
LOST + BS [30]	✓	81.8	48.9	57.8	88.7	57.2	69.7	91.6	72.3	83.7
DSM + CRF [24]	✓	87.1	56.7	64.4	83.8	51.4	56.7	89.1	73.3	80.5
WSCUOD + BS [22]	✓	90.9	58.5	68.3	92.5	63.0	76.4	92.8	74.2	89.6
TokenCut + BS [40]	✓	89.7	61.8	69.7	91.4	62.4	75.5	93.4	77.2	87.4
SelfMask + BS [29]	✓	91.9	**65.5**	-	93.3	66.0	-	**95.5**	**81.8**	-
FOUND + BS† [30]	✓	91.7	60.9	69.1	94.0	66.1	75.0	95.2	81.7	93.0
PEEKABOO + BS (Ours)	✓	**92.4**	61.2	**71.4**	**94.4**	**66.3**	**77.4**	94.9	80.6	**93.7**

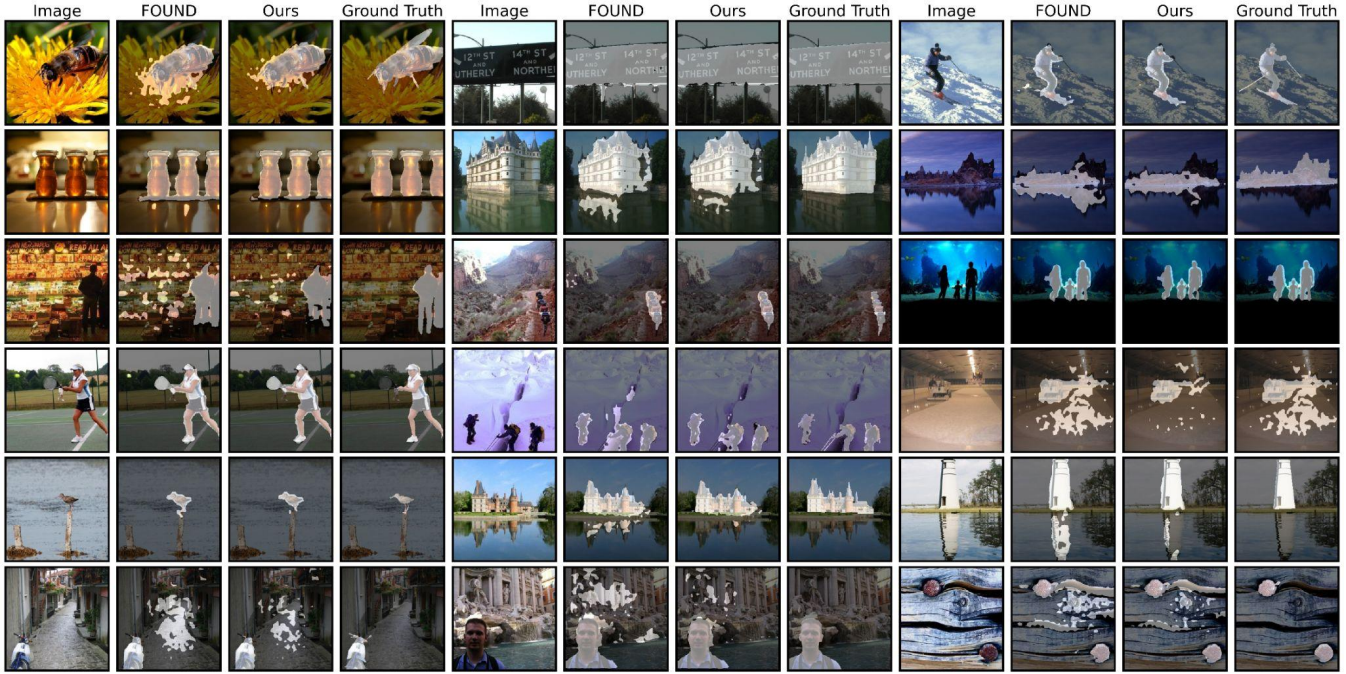
CorLoc metric

Method	Learning	VOC07	VOC12	COCO20K
Zhang et al. [64]		46.2	50.5	34.8
DDT+ [61]		50.2	53.1	38.2
rOSD [55]		54.5	55.3	48.5
LOD [56]		53.6	55.1	48.5
DINO [0]		45.8	46.2	42.1
LOST [30] (ViT-S/16)		61.9	64.0	50.7
LOST + CAD [30]		65.7	70.4	57.5
DSM [24] (ViT-S/16)		62.7	66.4	52.2
TokenCut [40] (ViT-S/16)		68.8	72.1	58.8
TokenCut + CAD [40]		71.4	75.3	62.6
SelfMask [29]	✓	72.3	75.3	62.7
FOUND† [30]	✓	71.7	75.6	61.1
FreeSOLO [59]	✓	56.1	56.7	52.8
DeepCut [0]	✓	69.8	72.2	61.6
WSCUOD [22]	✓	70.6	72.1	63.5
DINOSAUR [27]	✓	-	70.4	**67.2**
PEEKABOO (ViT-S/8) (Ours)	✓	**72.7**	**75.9**	64.0

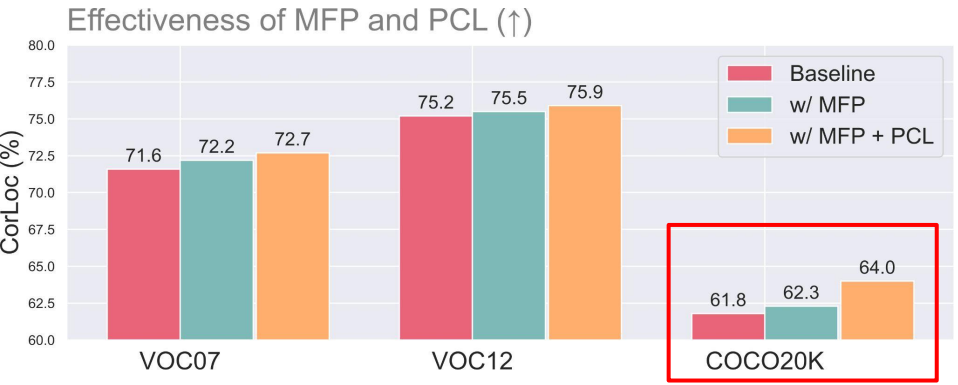
Uses 24x more data & 85,714x more params than Peekaboo!

Peekaboo has less than 1K trainable params, frozen DINO with ~21M params.

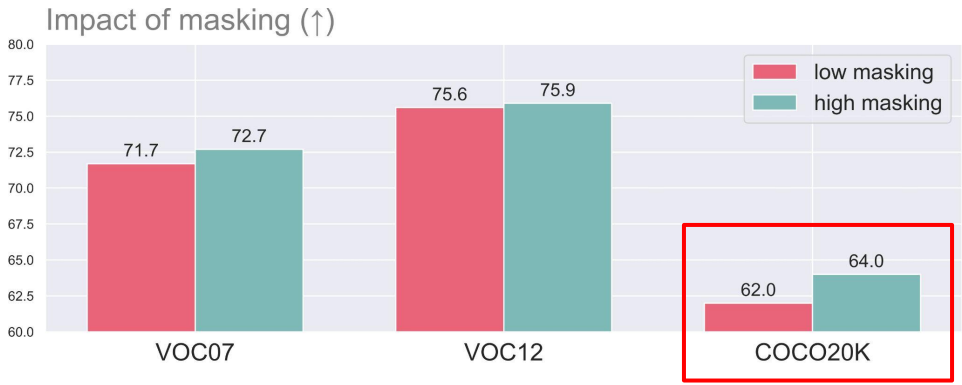
Peekaboo can localize objects that are small, reflective, or against complex or dim backgrounds



Peekaboo components complement each other; is better with high masking of images



Significant improvement observed on COCO20K, which contains a diverse range of images with multiple objects in an image.



Heavily masked inputs improve performance; low masking preserves most info leading to learning redundant features.

Peekaboo can detect multiple unfamiliar objects of different shapes and scales; basically which are not background

These objects that are out-of-domain of ImageNet and DUT-TR. Specifically, octopus, dinosaurs and spaceships are not present in ImageNet/DUT-TR, Peekaboo can still detect them.

Contributions

- Masked Supervised Learning for Semantic Segmentation
  - A single-stage learning paradigm that models local and global context via masking
  - Good at segmenting *small*, *ambiguous regions* and *minority classes* in cluttered scenes
  - Is accurate and computationally efficient compared to existing methods (e.g., multi-task)

- Learning to Recognize Occluded and Small Objects with Partial Inputs
  - An approach that learns contextualized representations and models label co-occurrence
  - Works well in cases of *small objects* and *occlusions*
  - Outperforms state-of-the-art methods that are complex in terms of model and data

- PEEKABOO: Hiding Parts of an Image for Unsupervised Object Localization
  - A self-supervised single-stage zero-shot model that learns context at pixel- and shape-level
  - Can localize novel objects that are *small*, *reflective*, or against *complex* or *dim backgrounds*
  - Better than methods that require computationally intensive training processes, resulting in high resource demands in terms of compute, learnable parameters, and data

Publications

Hasib Zunair, A. Ben Hamza, *PEEKABOO: Hiding Parts of an Image for Unsupervised Object Lo* BMVC 2024.

Hasib Zunair, A. Ben Hamza, *Learning to Recognize Occluded and Small Objects with Partial Inputs*, WACV 2024.

Hasib Zunair, A. Ben Hamza, *Masked Supervised Learning for Semantic Segmentation*, BMVC 2022.

Hasib Zunair, Shakib Khan, A. Ben Hamza, *RSUD20K: A Dataset for Road Scene Understanding In Autonomous Driving*, ICIP 2024.

Hasib Zunair, Yan Gobeil, Samuel Mercier, A. Ben Hamza, *Fill in Fabrics: Body-Aware Self-Supervised Inpainting for Image-Based Virtual Try-On*, BMVC 2022.

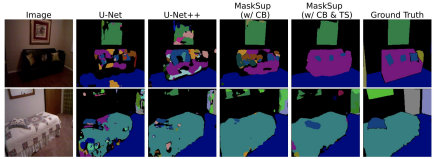
What's next?

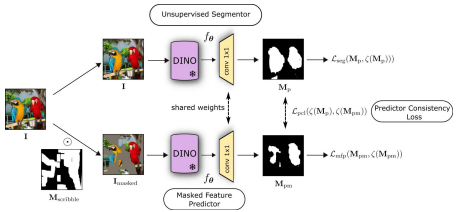
- Real-world applications of Masked Supervised Learning
  - Improve predictive accuracy of products or services using computer vision, driving higher user engagement and retention
  - Easy to implement, efficient and competitive against more advanced learning paradigms (SSL), save ML teams development time & reduce costs for large-scale model training

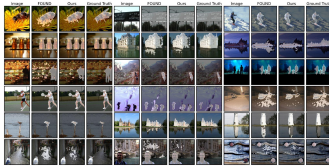
- Dynamic Masked Supervision
  - Generate masks programmatically during training using generative models, perlin noise, or semantically-meaningful masks, adding more randomness and diversity
  - Change size or level of masking while training, similar to on-the-fly data augmentation

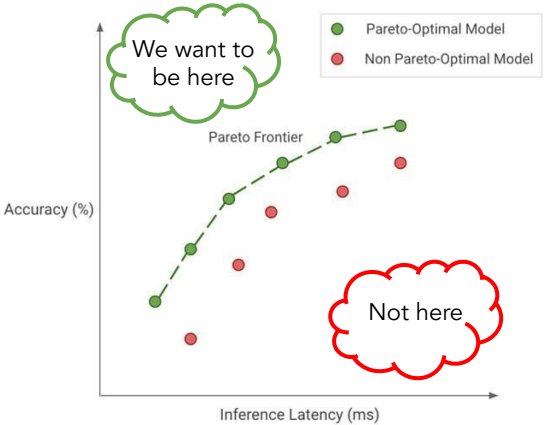
- Adapting Masked Supervised Learning to other domains
  - 3D data (e.g., videos, point cloud, CT scans, MRI etc.) & Image generation
  - Training Vision-Language Models (VLMs), context to capture multiple levels of granularity, comprehend spatial relationships between objects

Make things simpler and get state-of-the-art results, what's there not to like?









Goal of Efficient Deep Learning
To train and deploy pareto-optimal models that cost lesser resources to train and/or deploy while achieving similar results.

Masked Supervised Learning is…

- a simple and efficient learning technique that improves on the pareto frontier
- a small step towards bridging the gap between biological and artificial systems "feel the AGI"

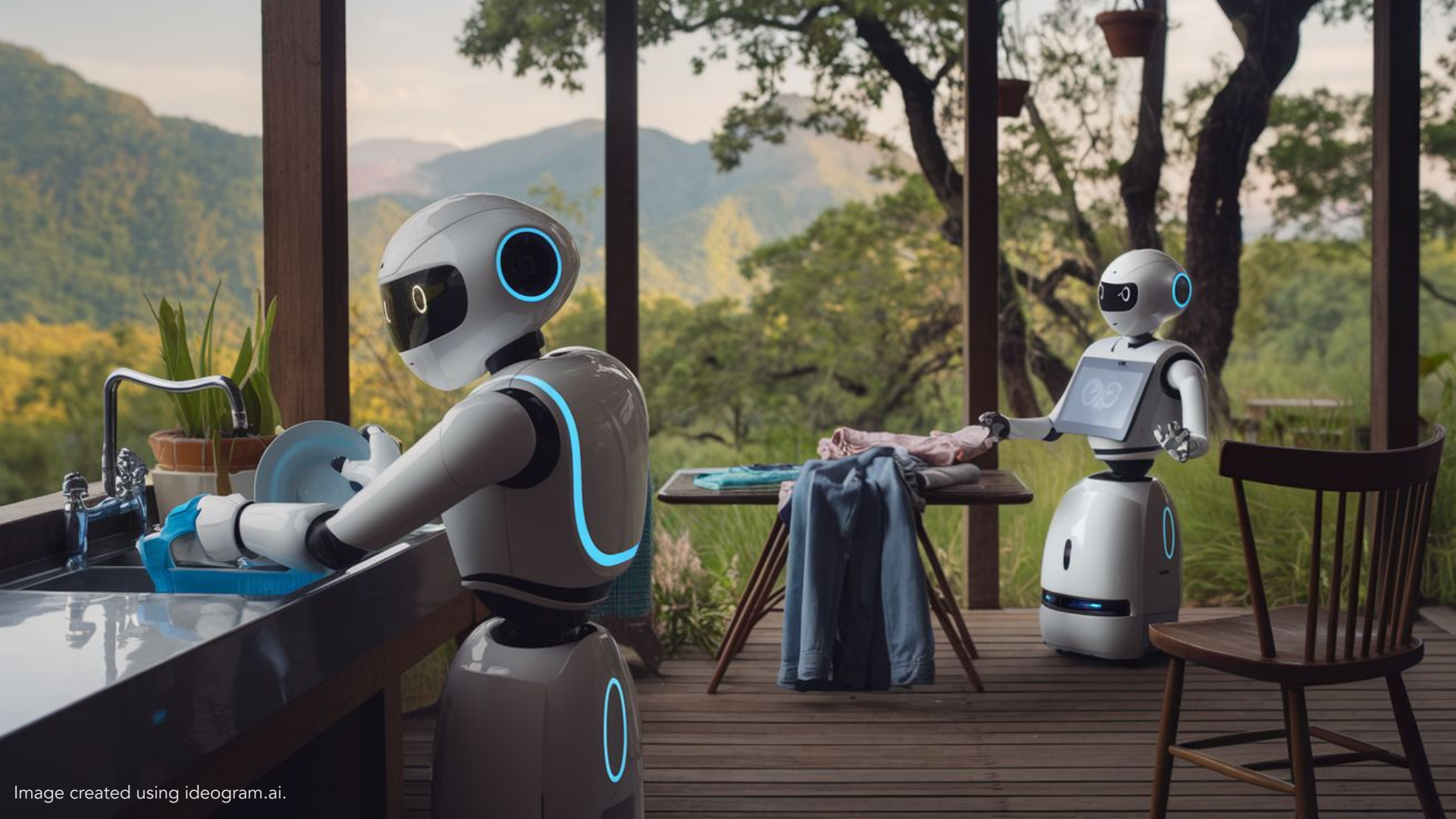


Image created using ideogram.ai.