Think before you Learn:
Image Segmentation with Weak Supervision

Andrei-Bogdan Florea

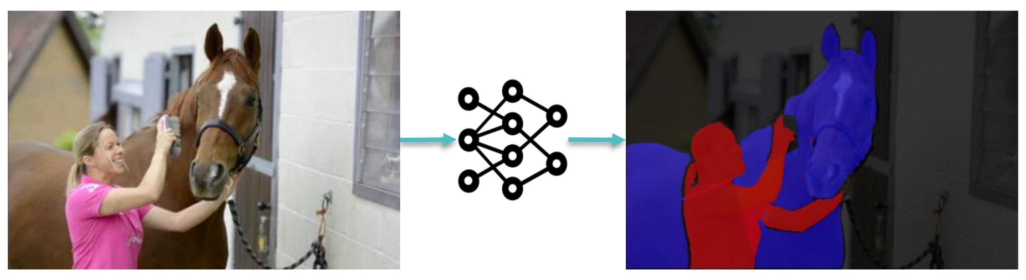
KU LEUVEN

DTAI Declarative Languages & Artificial Intelligence

Supervisor: Jaron Maene
Promoter: Prof. Dr. Luc de Raedt

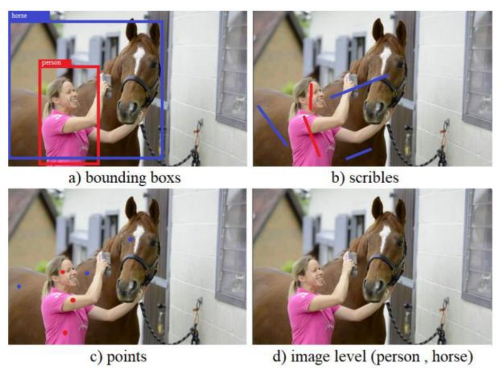
Context

- Image semantic segmentation aims to assign a class label to each pixel of an image

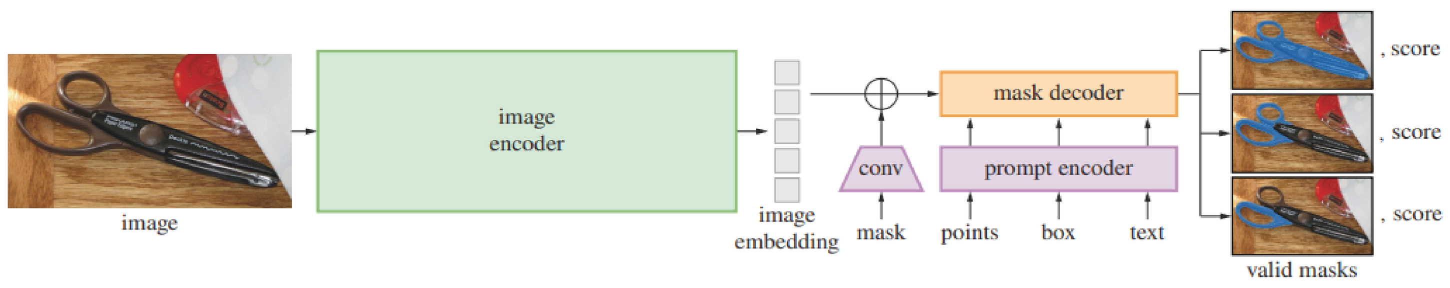


Full supervision
- Requires a manually annotated ground-truth map

Weak supervision
- Less informative labels
- Easier to annotate



- Recent trend: from training end-to-end models to prompting a large, foundational model
  - Segment Anything (SAM) [1]



Motivation

- In image segmentation, pixel-wise annotations are extremely labor-intensive to acquire
- Previous methods exploit only a single type of weak label
- Although it offers great segmentation results, SAM cannot be used standalone because it requires at least one prompt

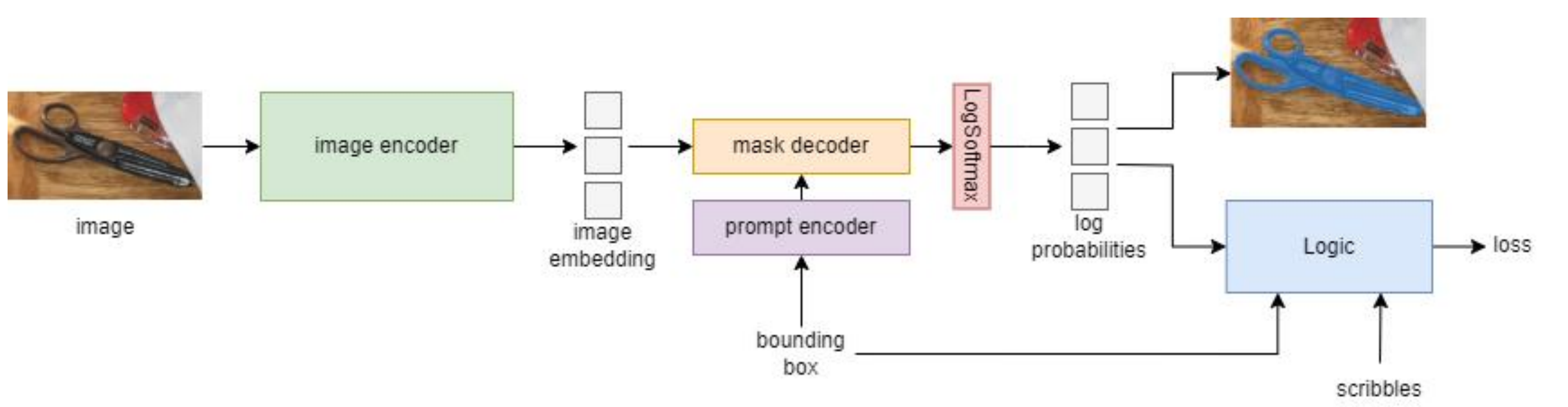
Objectives

- Developing a weakly supervised framework for training a segmentation network
- Using logic to express and learn from any classic weak label or other constraints
- Leveraging SAM indirectly to align the method to the state-of-the-art performance

Method

- Standard two-stage procedure:
  1. Obtaining pseudo-labels with a weakly supervised method
  2. Training fully supervised using the pseudo-labels as ground truth data

Stage 1: Weakly-supervised fine-tuning of SAM



- The **inputs** - images along with weak labels in the form of bounding boxes and scribbles for the objects of interest

- The **logic** - a tensorized implementation of the product t-norm fuzzy logic in log-space
  - $\neg x \longrightarrow \log(1-\exp(x))$
  - $x \wedge y \longrightarrow x + y$, with an additional operation $\wedge(x) \longrightarrow \text{sum}(x)$
  - $x \vee y \longrightarrow \neg(\neg x \wedge \neg y)$, with an additional operation $\vee(x)$ that performs $\neg(\wedge(\neg x))$

- The **loss** - a large expression as an "and" operation between constraints:
  - Background constraint: all pixels not within a bounding box belong to the background class
  - Bounding boxes tightness prior: any row/column of a bounding box has at least one pixel of the target class
  - Scribble-based terms: all pixels within a scribble belong to the scribble's target class
  - Neighborhood constraints
    - if a pixel has a class C, then at least one of its neighbors should have class C
    - if all surrounding pixels have the same class C, then the pixel in the middle should have class C
  - Border precision constraint: If two adjacent pixels belong to different classes, they must belong to different superpixels

- The **training**
  - Freezing the image encoder and the prompt encoder
  - Fine-tuning the mask decoder

- The **aim** - producing high-quality pseudo-labels for the training images

Stage 2: Fully supervised training of a segmentation network

- SAM requires prompts to produce segmentations, but a prompt-less network is desired
- Therefore, a segmentation network that only takes the image as input is trained in this stage, supervised by the segmentation masks produced in the first stage
- Intuition: higher quality pseudo labels obtained in the first stage translate to better performance in the second stage training
- Networks trained: Mask2Former [2] and DeepLabV2 [3]

Results

🏆 **This method sets a new benchmark on Pascal VOC 2012**
- highest mIoU among all weakly-supervised methods
- sometimes beats the fully supervised method

Method	Labels	mIoU (%)
Sun et al. [4]	Image-level	88.3
FMA-WSS [5]	Image-level	80.4
SemPLeS [6]	Image-level	78.4
SAM prompted with scribbles [7]	Scribbles	89.7
SAM prompted with boxes [7]	Boxes	91.5
This method	Boxes, Scribbles	**94.3**

Pseudo-labels quality on the *train* set

Pseudo-masks produced in the first stage, overlaid on the images

Method	Labels	$mIoU_v$	$mIoU_t$
Fully supervised	Pixel-level ground-truth	87.1	86.7
FMA-WSS [5]	Image-level	82.6	81.6
SemPLeS [6]	Image-level	83.4	82.9
This method	Boxes, Scribbles	**88.4**	**87.9**

Mask2Former performance on the *val* and *test* sets

Method	Labels	$mIoU_v$	$mIoU_t$
Fully supervised [3]	Pixel-level ground-truth	76.3 (77.7*)	**79.7***
SemPLeS [6]	Image-level	73.9	73.8
Sun et al. [4]	Image-level	77.2*	77.1*
SAM prompted with boxes [7]	Boxes	76.3	75.8
Box2TagBack	Boxes	77.1	76.1
AGMM	Scribbles	74.2	75.7
TEL	Scribbles	75.2	75.6
Scribble hides class	Scribbles	75.9	76.0
Chan et al.	Scribbles	76.2	-
SAM prompted with scribbles [7]	Scribbles	75.9	76.6
This method	Boxes, Scribbles	**77.6 (79.1*)**	78.2 (79.6*)

DeepLabV2 performance on the *val* and *test* sets

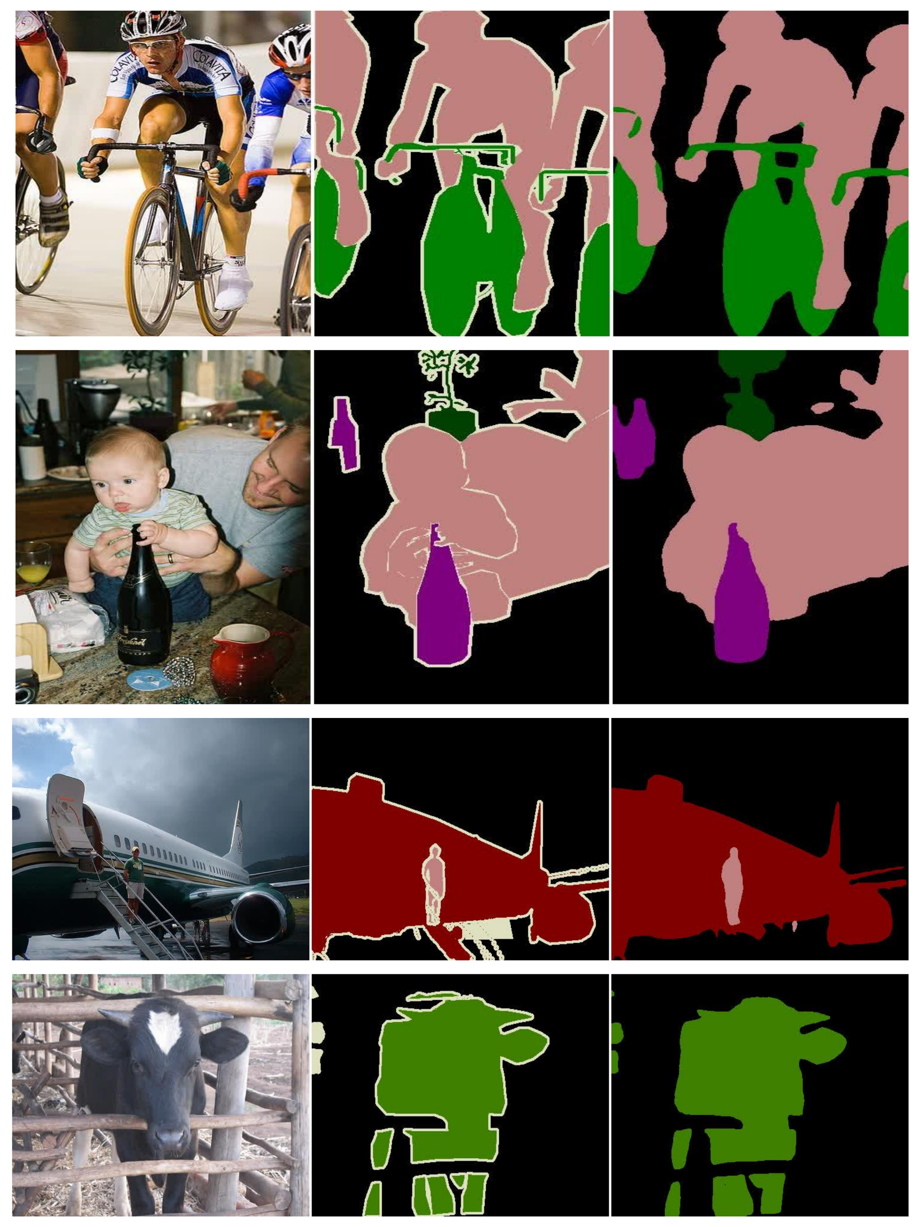


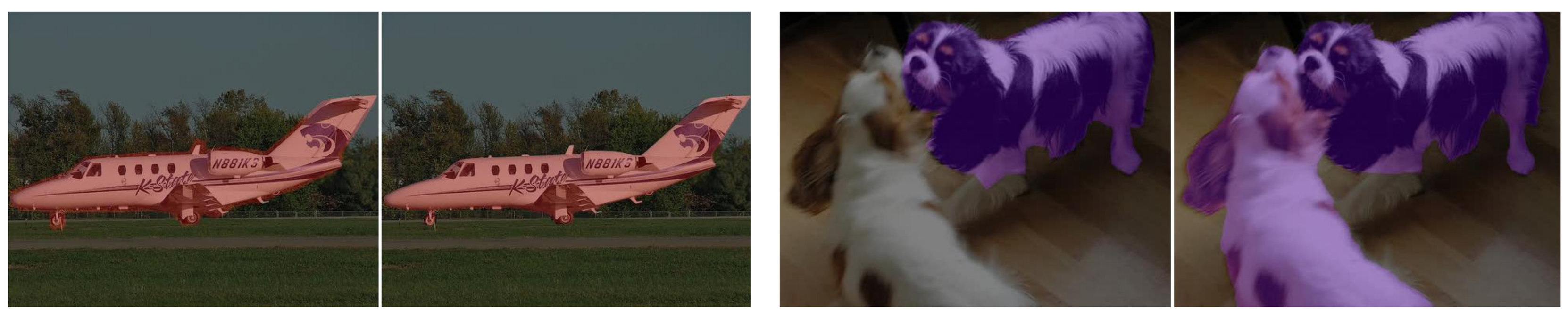
Image | Ground truth | Mask2Former prediction

* after applying DeepLabV2's CRF post-processing

Conclusion

- Trade-off between segmentation accuracy and information required for training
  - \+ this method sets a state-of-the-art performance
  - \- benefits from the most amount of information during training (both bounding box and scribble annotations)

- In some cases, this weakly-supervised method outperforms full supervision
  - can be attributed to a poor quality of some segmentation maps from the dataset
  - Segment Anything is a strong baseline in datasets with common classes



Ground-truth | SAM produced pseudo-label | Ground-truth | SAM produced pseudo-label

- More work can be done to apply the method in a practical scenario such as in medical segmentation

References

[1] Kirillov, Alexander, et al. "Segment anything." Proceedings of the IEEE/CVF international conference on computer vision. 2023.
[2] Cheng, Bowen, et al. "Masked-attention mask transformer for universal image segmentation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[3] Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2017): 834-848.
[4] Sun, Weixuan, et al. "An alternative to wss? an empirical study of the segment anything model (sam) on weakly-supervised semantic segmentation problems." arXiv preprint arXiv:2305.01586 (2023).
[5] Yang, Xiaobo, and Xiaojin Gong. "Foundation model assisted weakly supervised semantic segmentation." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024.
[6] Lin, Ci-Siang, et al. "SemPLeS: Semantic prompt learning for weakly-supervised semantic segmentation." arXiv preprint arXiv:2401.11791 (2024).
[7] Jiang, Peng-Tao and Yuqi Yang. "Segment Anything is A Good Pseudo-label Generator for Weakly Supervised Semantic Segmentation." ArXiv abs/2305.01275 (2023)