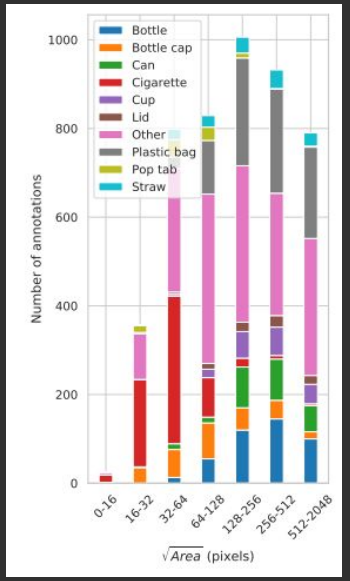
TACO: Trash Annotations in Context for Litter Detection

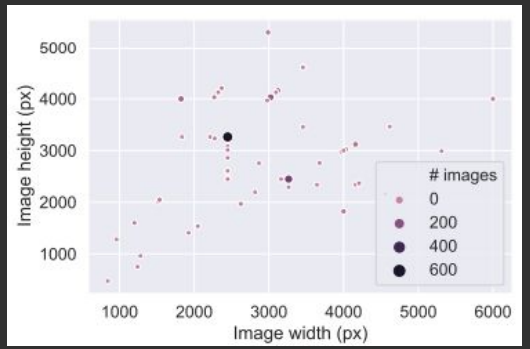
Benchmarking and Explainability of Instance Segmentation Models

Authors of this study:
Dell'Olio Domenico · Delvecchio Giovanni Pio · Disabato Raffaele

Dataset Description

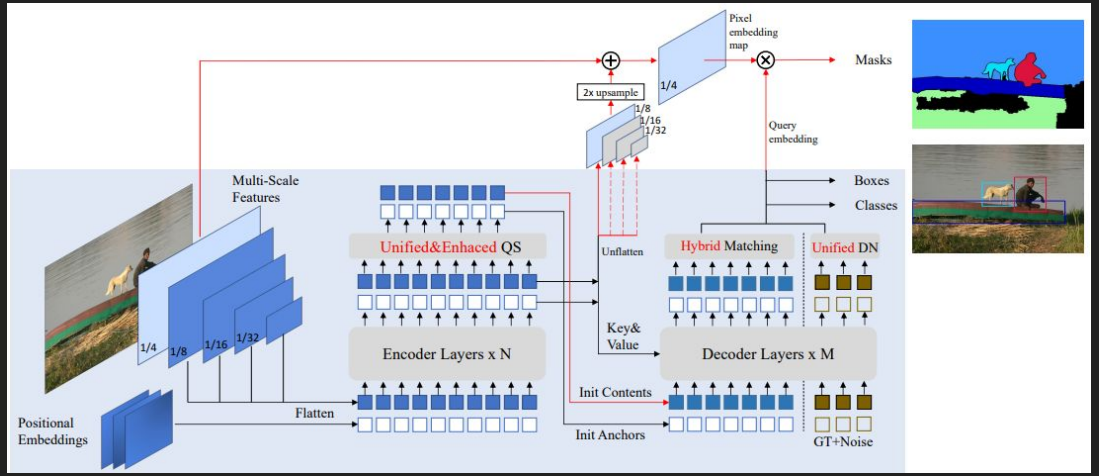
- 1500 hi-res images
- 4784 annotations
- 10 classes
- Baseline:
  Mask R-CNN 17.6 ± 1.6 AP
- Train/Val/Test: 80/10/10%





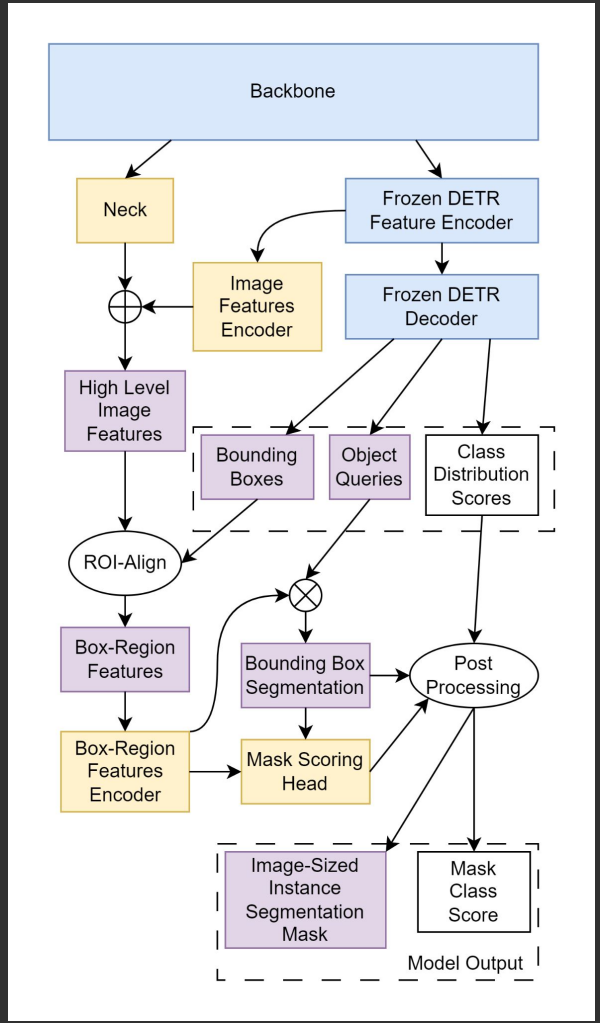
Benchmarks: MaskDINO

- Backbone: ResNet50
- Encoder: FPN + Positional Embeddings + Unified and Enhanced Query Selection
- Decoder: Transformer Decoder (with Deformable Attention) + Hybrid (BB and Mask) Matching + Unified Denoising
- Segmentation branch: $m = q_c \otimes M(T(C_b)+F(C_e))$
- 33,5 AP test (segm.)



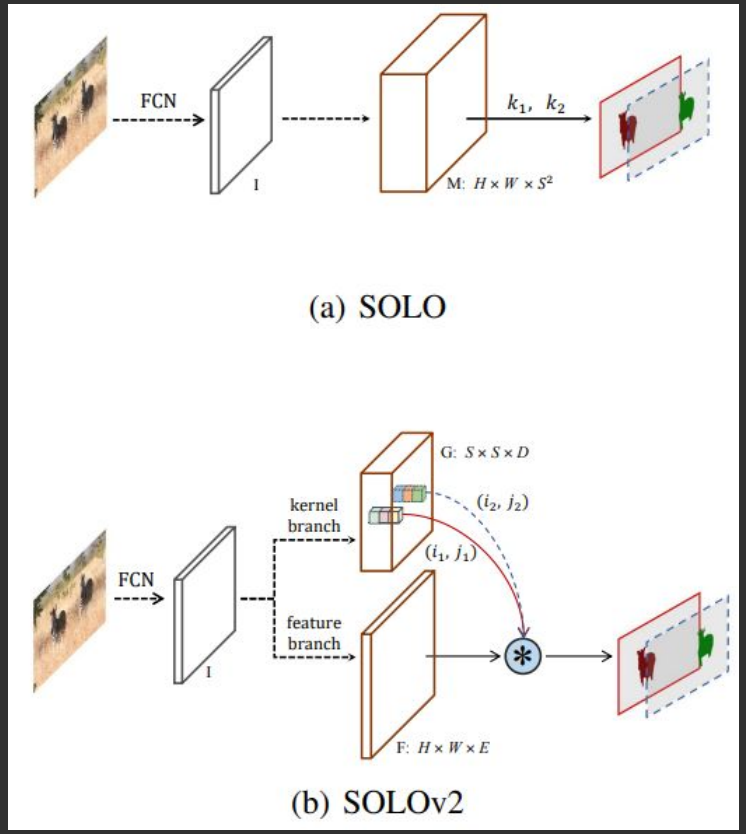
Benchmarks: Mask Frozen-DETR

- Backbone (Swin-Tiny) and DETR are kept frozen
- Box Region Features are a downscaled representation of the content of each BB
- Image Feature and Box Feature Encoder two-layer Deformable Encoders
- Masks are obtained by multiplying object queries and Box Region Features, later resized to BB dimension and pasted on a blank mask.
- 30.7 AP test (segm.)

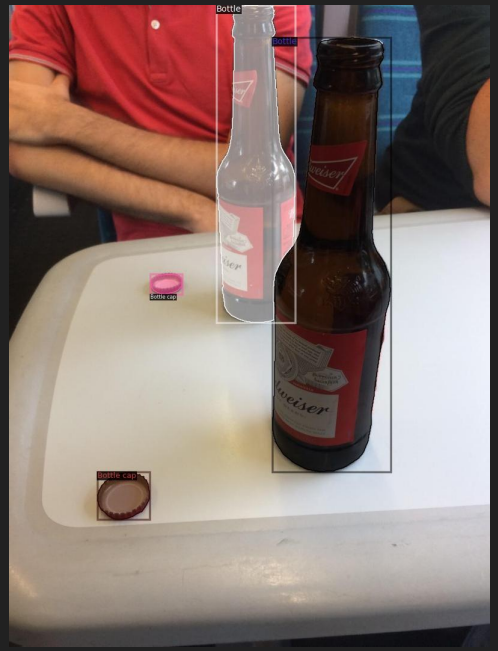


Benchmarks: SOLOv2

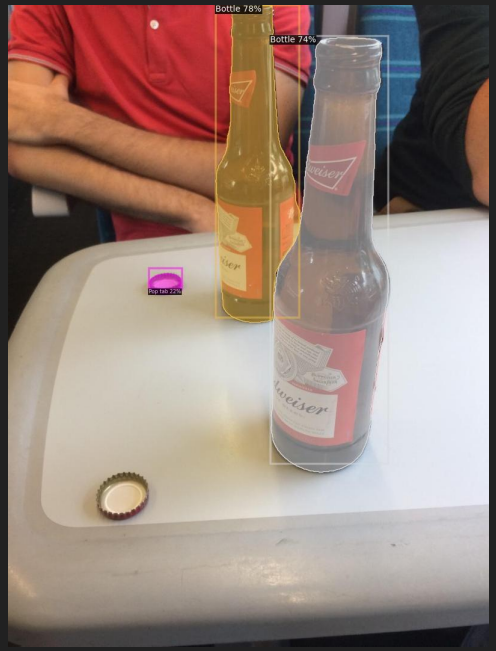
- Backbone: ResNet-50
- Has FPN and divides the image in a SxS patches regular grid
- Convolutional Layers to extract feature tensor F
- Kernel G is obtained from Image + Pixel Coordinates (normalized)
- G*F outputs the tensor M containing all instance masks (shape: $HxWxS^2$)
- 12.4 AP validation (segm.)



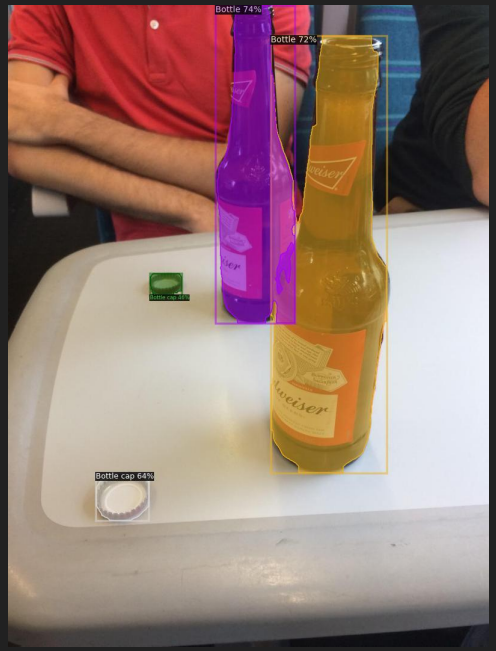
Some results



Ground Truth



MaskDINO (th. 0.20)

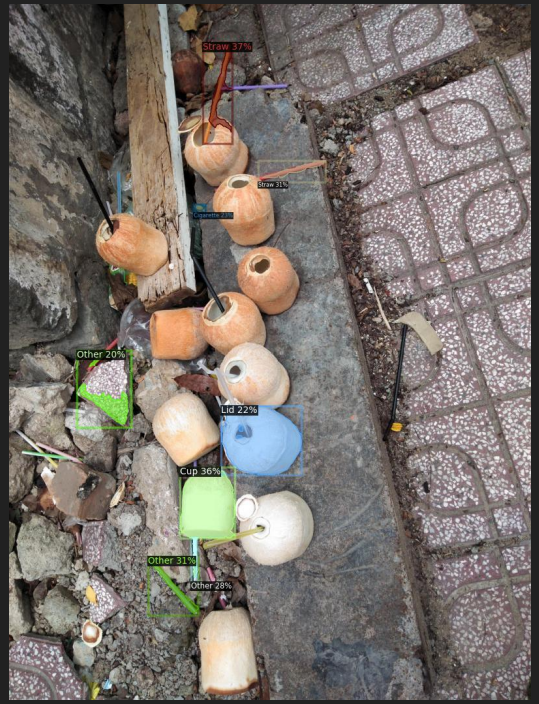


Mask-Frozen DETR
(th. 0.20)

Some results (cont.)

Ground Truth

MaskDINO (th. 0.20)



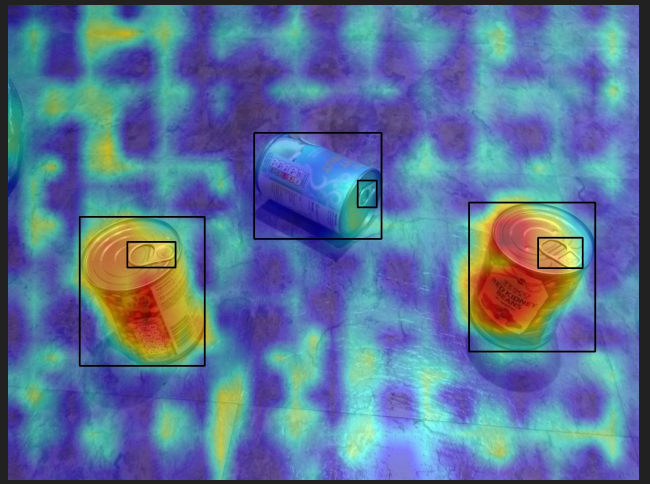
Mask-Frozen DETR (th. 0.20)

Explainability: CAM methods

- To the best of our knowledge, **no specific XAI methods for IS** are present in literature.
- We decided to **split** the Instance Segmentation Task into its two main sub-tasks (**Object Detection and Semantic Segmentation**) and to apply methods for each of them.
- We chose **Class Activation Map (CAM) methods**, that try to explain what a model learns from the data or why it behaves poorly in a given task by manipulating its **intermediate activations**.
- Most of the methods are tested only on **Mask-Frozen DETR** due to computation resource limitations.

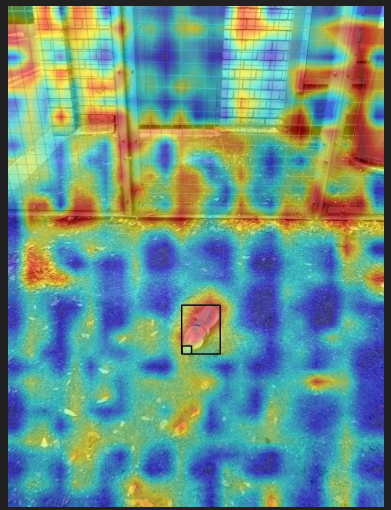
Grad-CAM

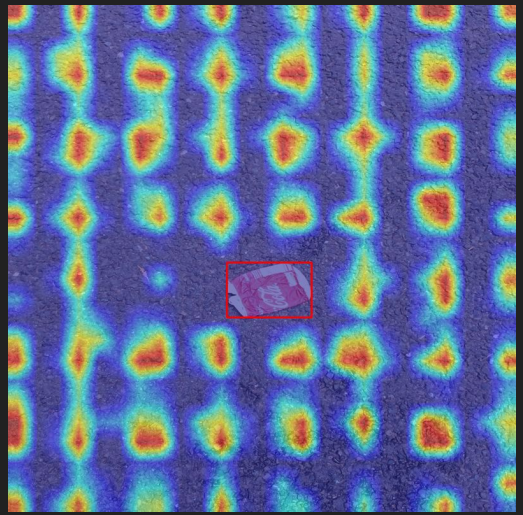
- **Grad-CAM** uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

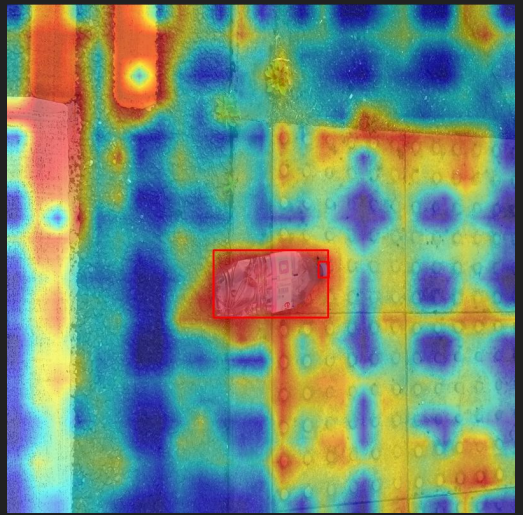


Ablation-CAM

- **Ablation-CAM** works by *zeroing out* individual features or groups of features (such as channels) from the activation maps. By observing the impact of this “ablation” on the output, it identifies which features are most important for a particular prediction. No gradients are required.

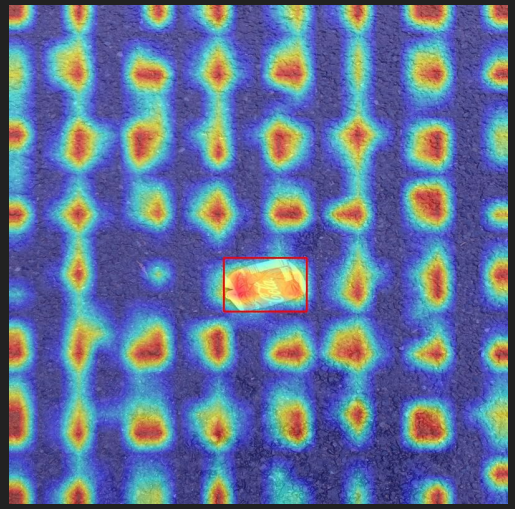


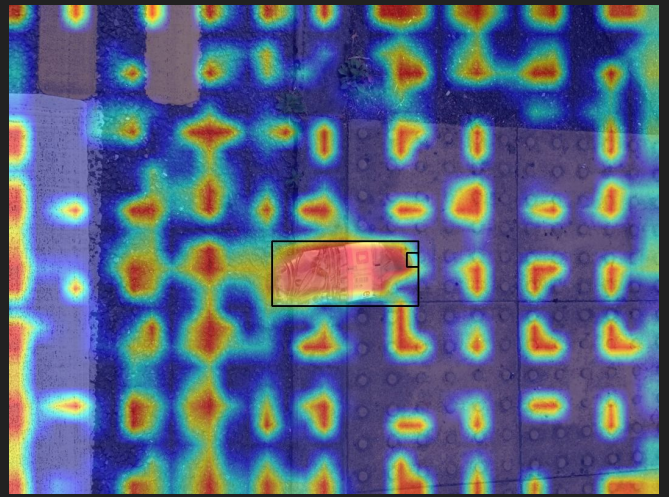




Eigen-CAM for Detection

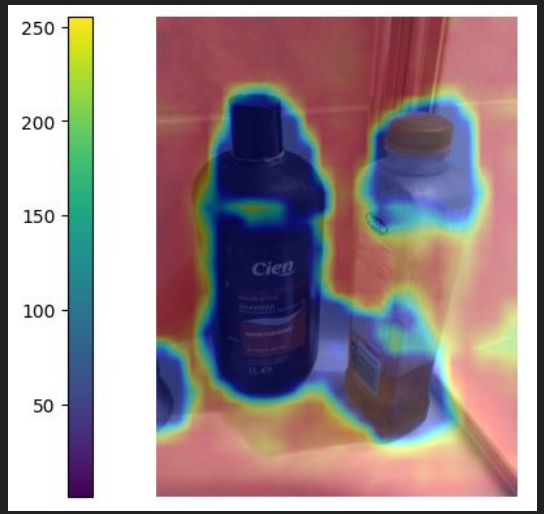
- **Eigen-CAM** visualizes the *principle components* of the learned features/representations on the input image, no gradients are required. Explanations are Class-indepenent.

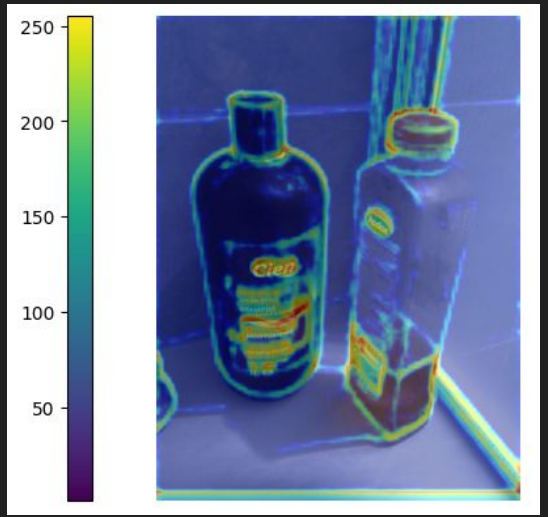


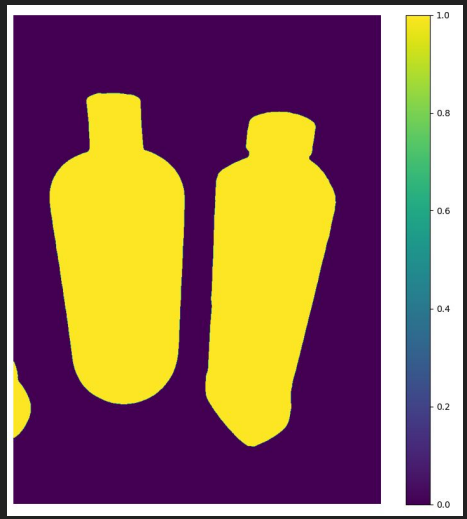


Eigen-CAM for Segmentation

- Gizzini et Al. **extended** CAM methods to semantic segmentation tasks in order to focus only on segmented region of interest.
- We managed to test only this CAM with this extension on different layers of **MaskDINO**.

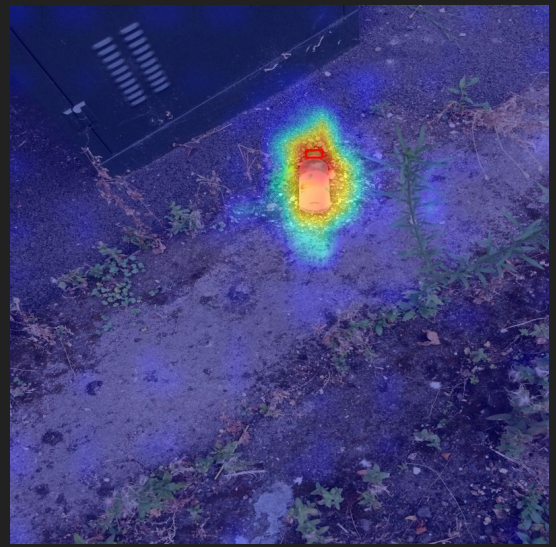


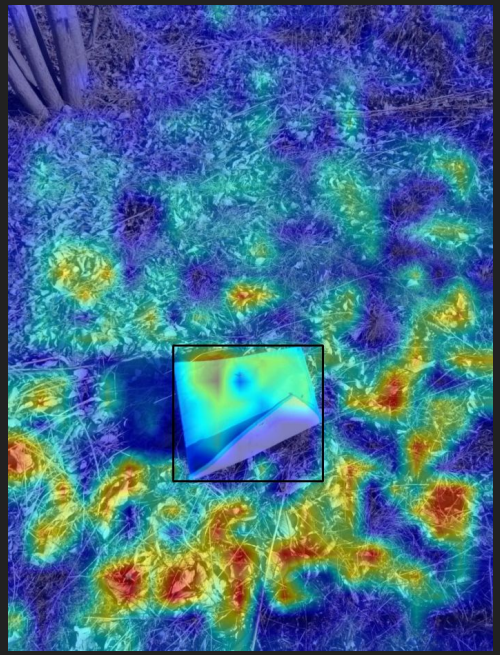




Score-CAM

- **Score-CAM** exploits the activation maps by using them as *masks on the input image*. Each masked image is fed to the network to compute the class score. These scores are used to compute weighted sum of activations that will form the final map.





Conclusion and Future works

- In this work we benchmarked some architectures among the **best scoring ones on COCO** to approach a **litter instance segmentation task**, as well as testing some **explainability methods** to get some insights on them.

- Both **MaskDINO** and **M-FDETR outperformed** the results of Mask R-CNN in the TACO paper, with the former returning **better segmentation maps** and the latter providing **more precise detections**.

- Among the Explainability methods, the **Score-CAM** is the best considering resource requirements and explanation quality.

- As **future works**, the models may be run at **higher image resolution** in order to solve the deficiencies in **small object detection**, as well as develop a more **specific XAI method** for the task.

TACO

Thank you for the Attention

Questions?