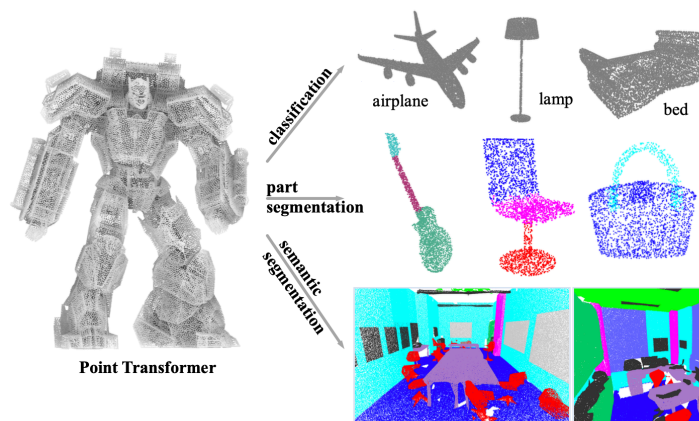


Back to Segmentation: *Transformers*

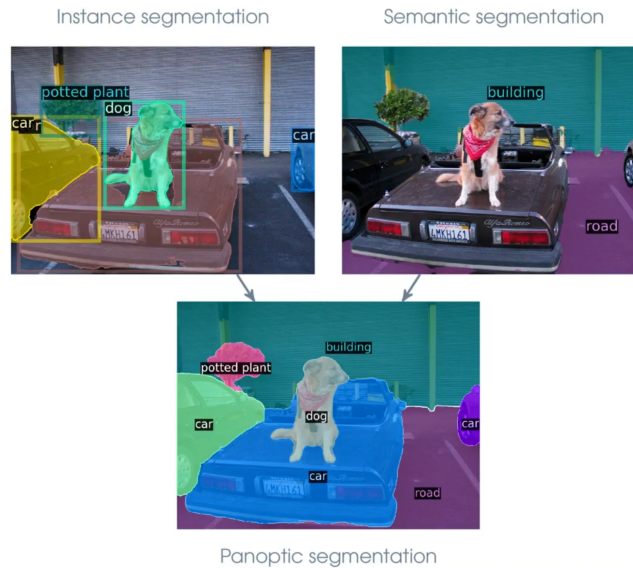
Luiz Velho
IMPA

Point Transformers

- Point Transformers to the Rescue
 - Many Tasks



Segmentation Duality



2D Object Segmentation with Transformers

Integrated End-2-End

- DETR
 - Object Detection
- Mask2Former
 - Panoptic Segmentation

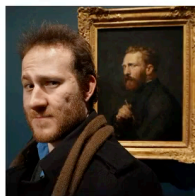
DETR:

End-to-End Object Detection with Transformers

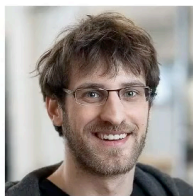
End-to-End Object Detection with Transformers



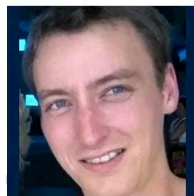
Nicolas Carion
Facebook AI



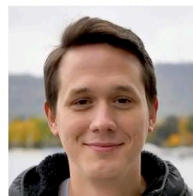
Francisco Massa
Facebook AI



Gabriel Synnaeve
Facebook AI



Nicolas Usunier
Facebook AI



Alexander Kirillov
Facebook AI



Sergey Zagoruyko
Facebook AI

Video Presentation

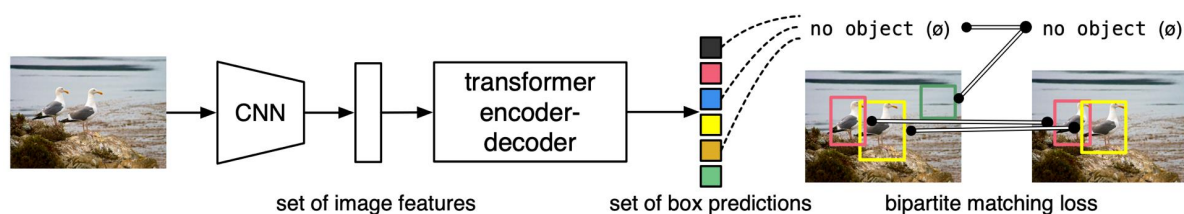
FACEBOOK AI

Object **D**etection with **T**ransformers: DETR

Simple object detection pipeline: directly output a set of boxes from a Transformer

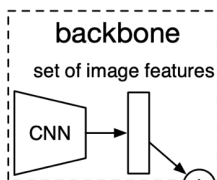
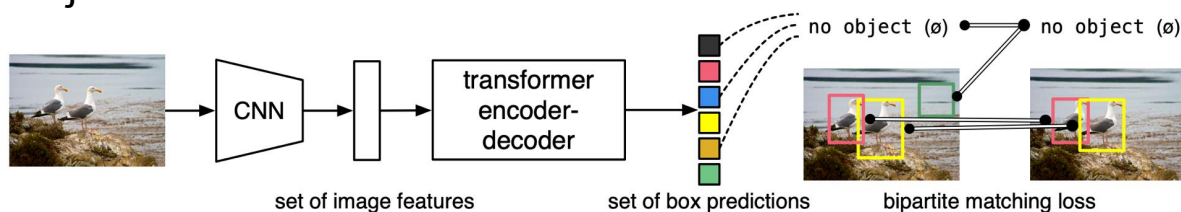
No anchors, no regression of box transforms

Match predicted boxes to GT boxes with bipartite matching; train to regress box coordinates



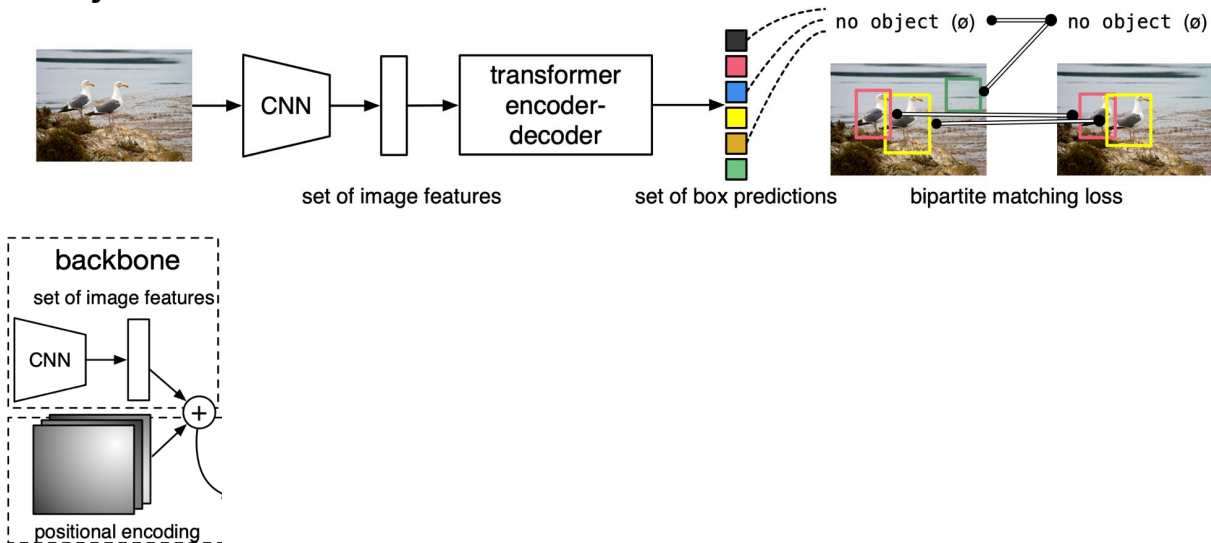
Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

Object Detection with Transformers: DETR



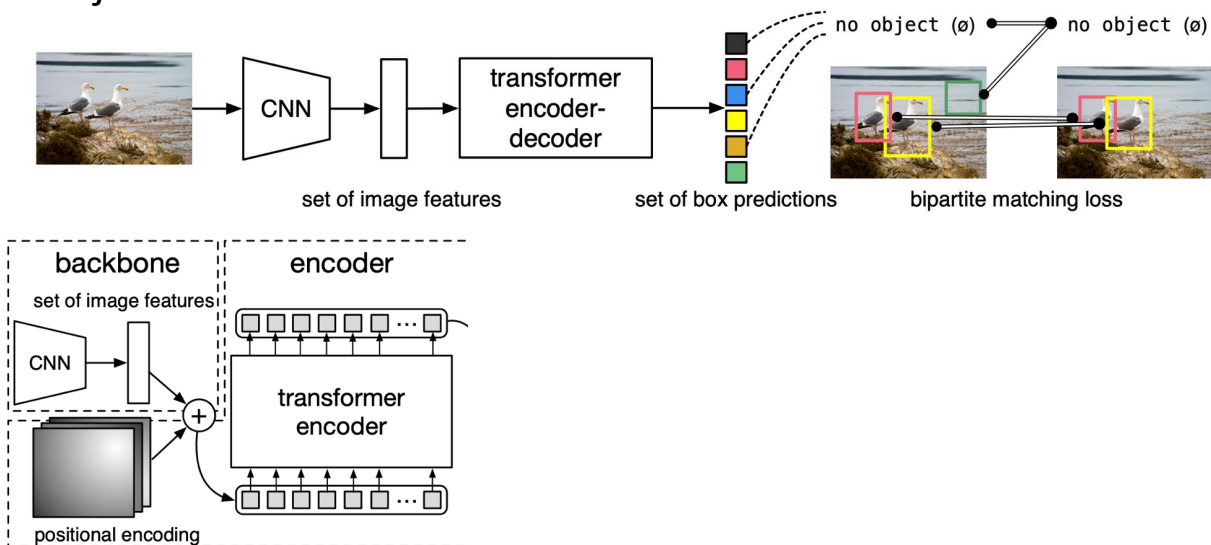
Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

Object Detection with Transformers: DETR



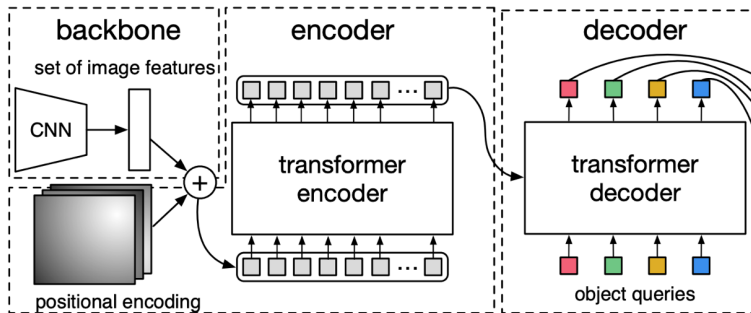
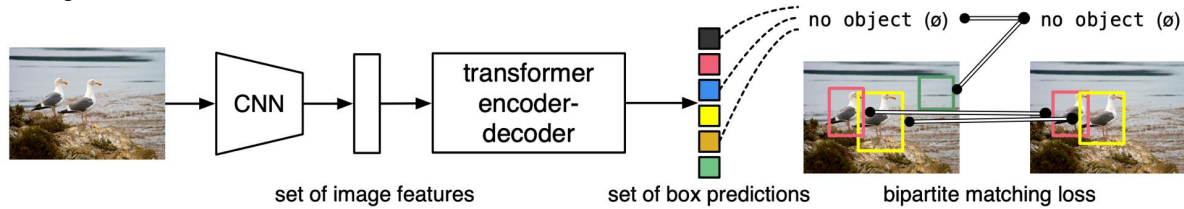
Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

Object Detection with Transformers: DETR



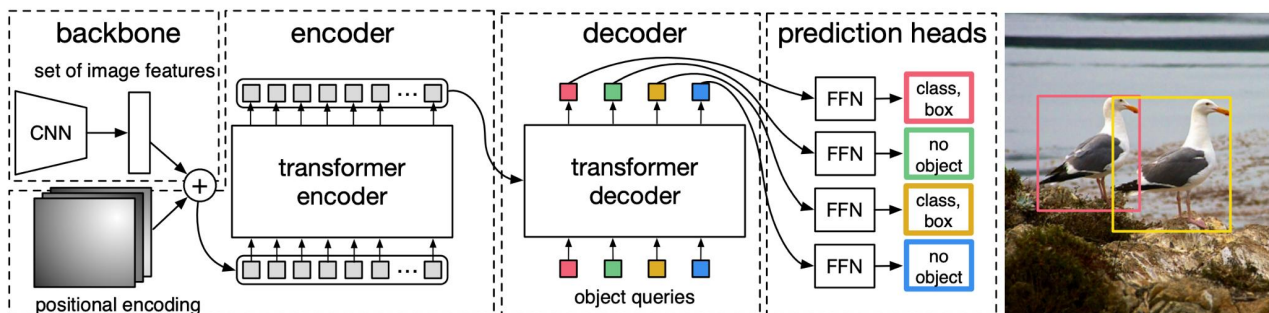
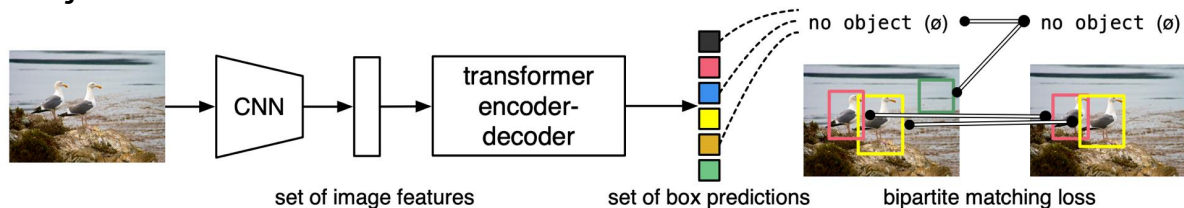
Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

Object Detection with Transformers: DETR



Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

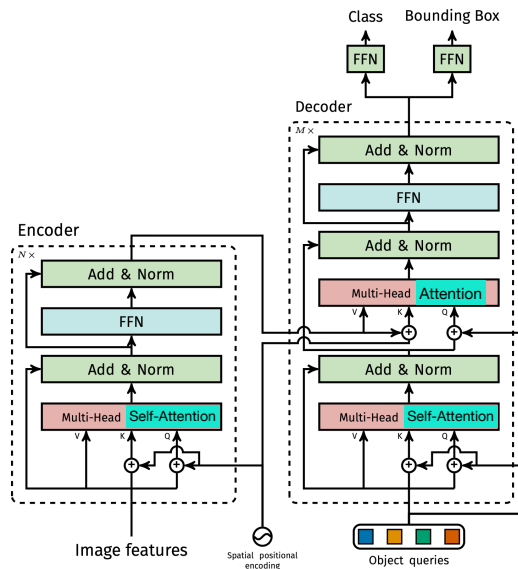
Object Detection with Transformers: DETR



Carion et al, "End-to-End Object Detection with Transformers", ECCV 2020

DETR Transformer

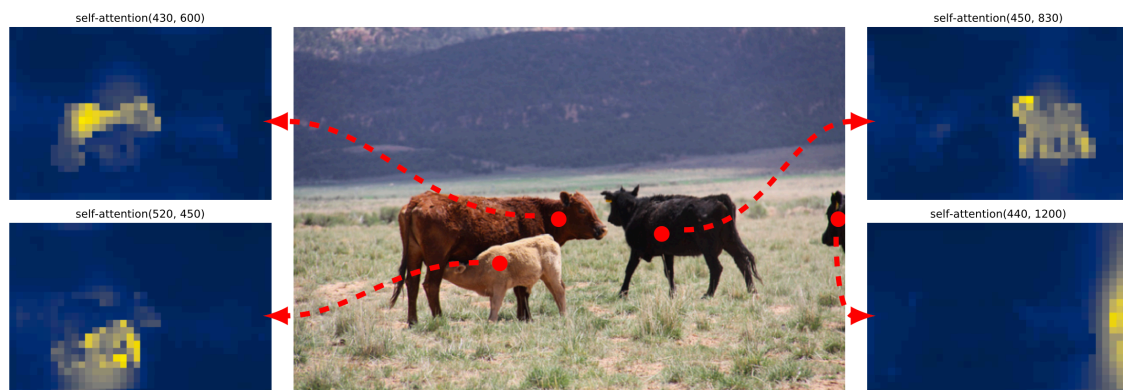
- Encoder / Decoder



- **Queries:** learnable in the decoder
 - act as slots seeking specific objects
- **Keys:** from encoder features (CNN)
 - encode spatial locations and visual content
- **Values:** paired with keys
 - Feats aggregated from Query-Key interactions
- * **Attention:**
 - match Keys against Queries via dot-product

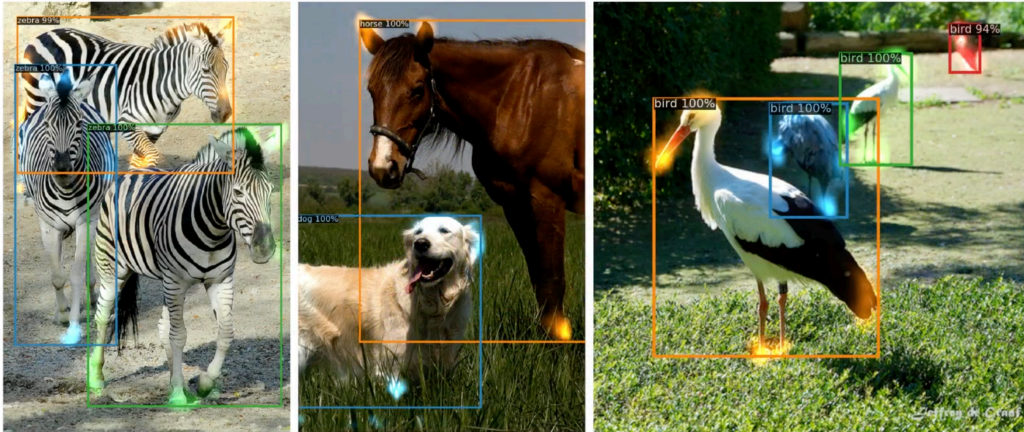
Role of Encoder

- Self-Attention to Object



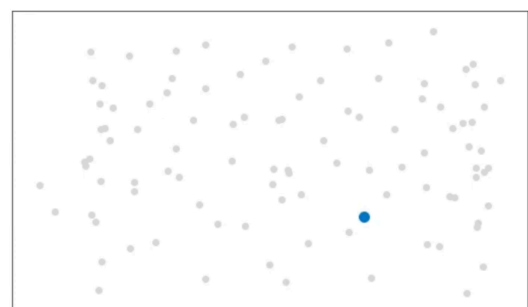
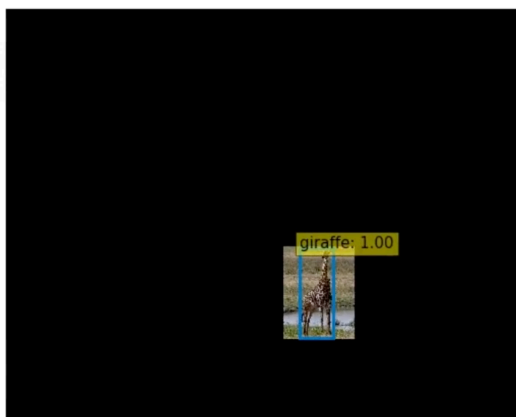
Role of Decoder

- Mask Extremities



Object Queries

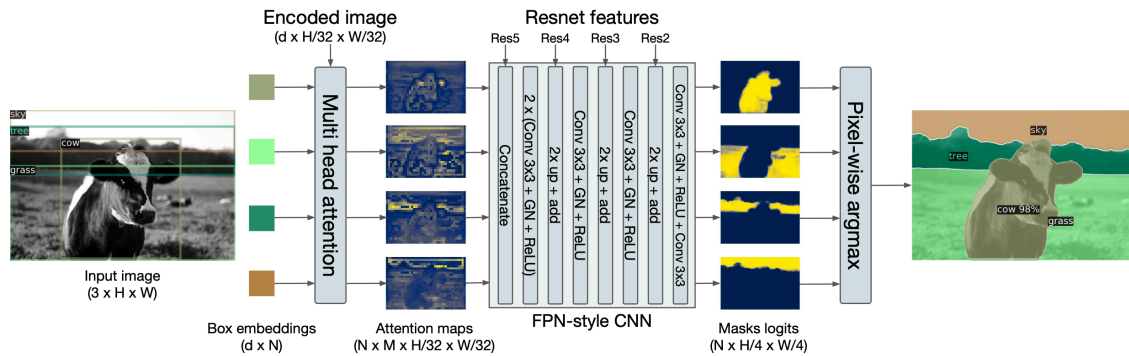
- Uniformly Distributed



Active object query

Panoptic Segmentation

- Panoptic Head

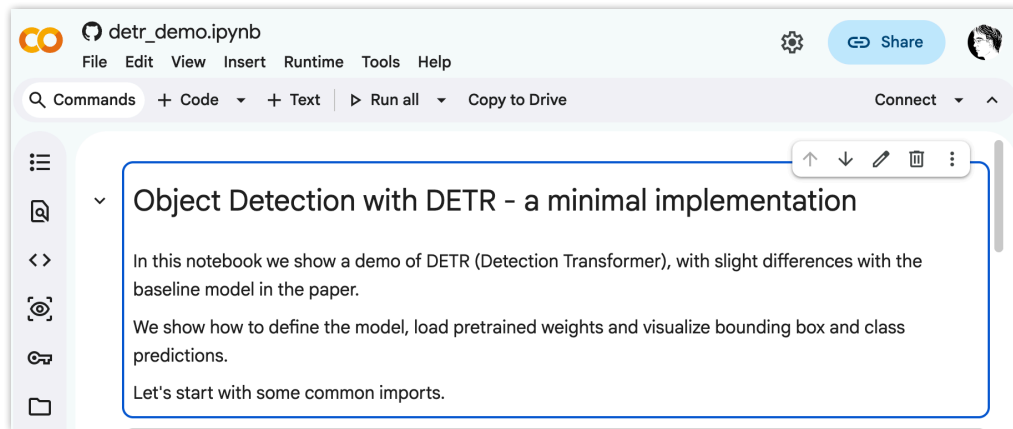


Panoptic Results

- COCO dataset: 53 stuff categories and 80 things categories



Demo



Mask2Former

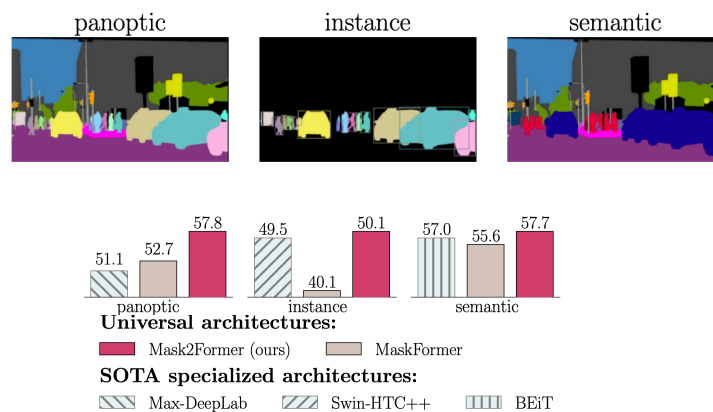
Universal Segmentation

Masked-attention Mask Transformer for Universal Image Segmentation

Bowen Cheng^{1,2*} Ishan Misra¹ Alexander G. Schwing² Alexander Kirillov¹ Rohit Girdhar¹

¹Facebook AI Research (FAIR) ²University of Illinois at Urbana-Champaign (UIUC)

<https://bowenc0221.github.io/mask2former>

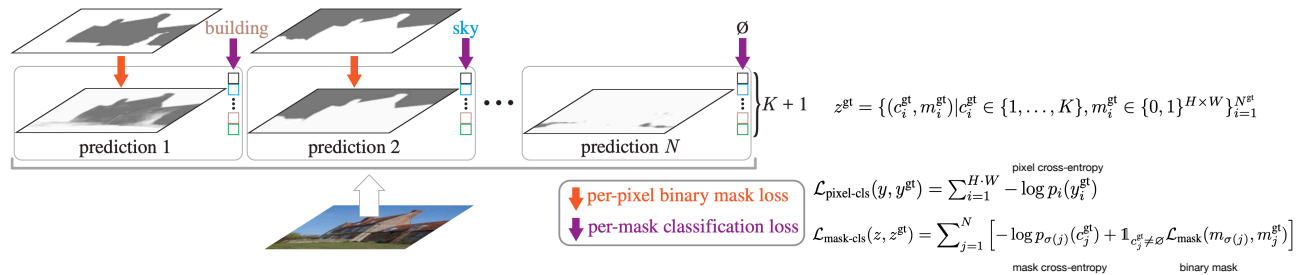


Mask2Former

- Unified Transformer Architecture
- Three tasks with the same model
 - *Instance segmentation*: detect and segment each object instance.
 - *Semantic segmentation*: classify every pixel.
 - *Panoptic segmentation*: combine instance + semantic segmentation.
- Predicts a set of segmentation masks with class labels

Joint Semantic and Instance Segmentation

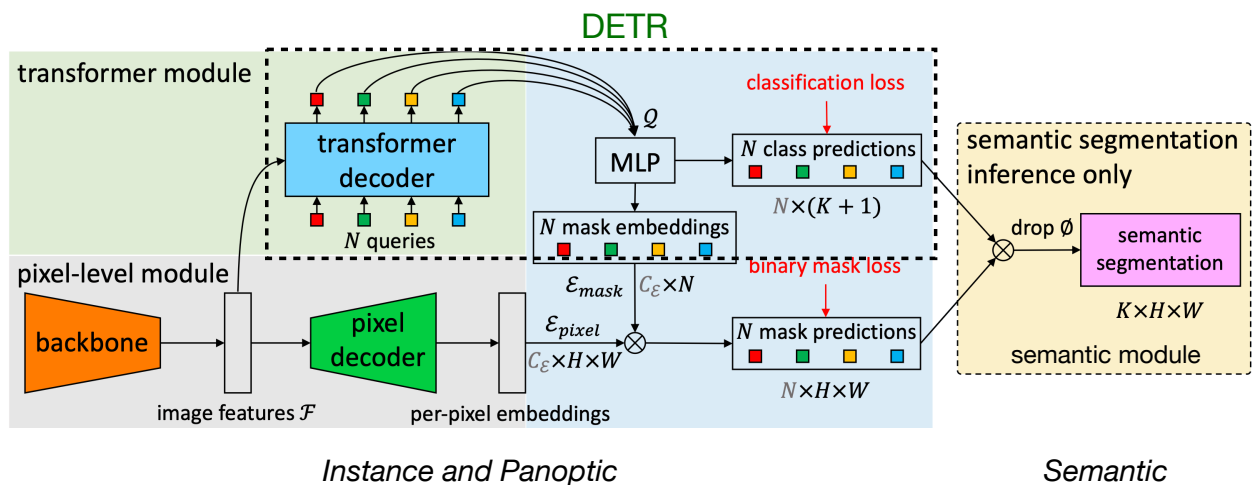
- Segmentation Mask with Class Labels



Matching between the set of predictions and ground truth segments is done via bipartite matching similarly to DETR

MaskFormer Architecture

- Baseline

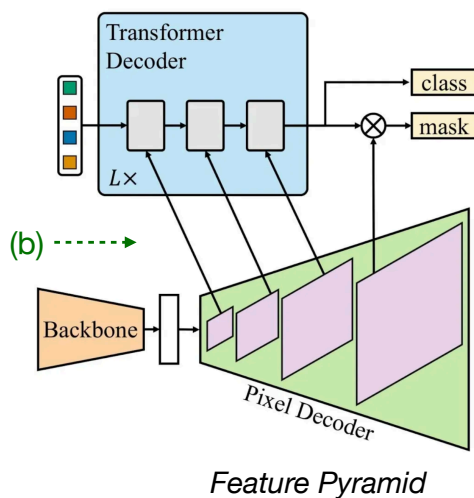


MaskFormer -> Mask2Former

- MaskFormer Drawbacks
 - Expensive attention over full feature maps
 - Inefficient query interaction
- Mask2Former Improvements
 - (a) Masked Cross Attention to improve efficiency and accuracy
(*query* → *attends ONLY inside its predicted mask*)
 - (b) Multiscale Features
(*handling of large + small objects*)

Mask2Former Architecture

- Network



- Decoder Module

