TRENDS IN COMPUTER VISION CONFERENCES

Prasad Magdum (MT2022078)

Rittik Panda (MT2022090)

- Panoptic Segmentation
- 3D Shape Reconstruction from 2D Images
- Transformer-based GAN for High-resolution Image Generation
- Text to Image Generation

Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models

Task: -

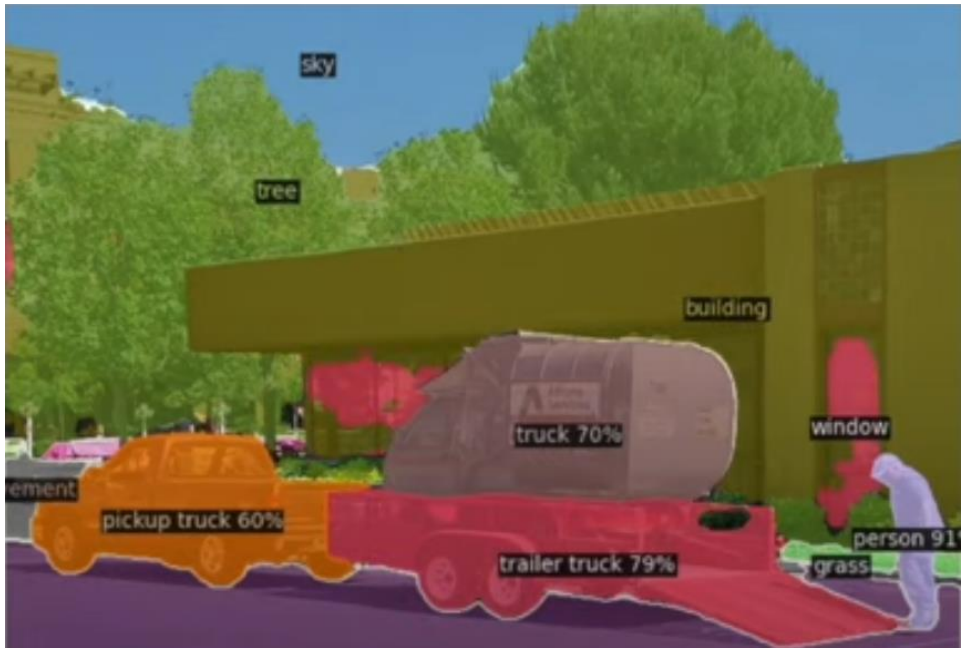
- The goal is to perform panoptic segmentation on images while allowing the model to recognize an unlimited number of object categories, even those not seen during training (open-vocabulary).
- Panoptic segmentation is an advanced computer vision task that combines two traditional tasks: semantic segmentation and instance segmentation.
- In Panoptic segmentation every pixel in the image is assigned both a semantic label and an instance ID. This means that each pixel not only receives a class label, but it is also associated with a specific instance of that class.
- This paper proposes a novel approach to achieve open-vocabulary panoptic segmentation by integrating "text-to-image diffusion models" and "discriminative models."

Problem Setting: -

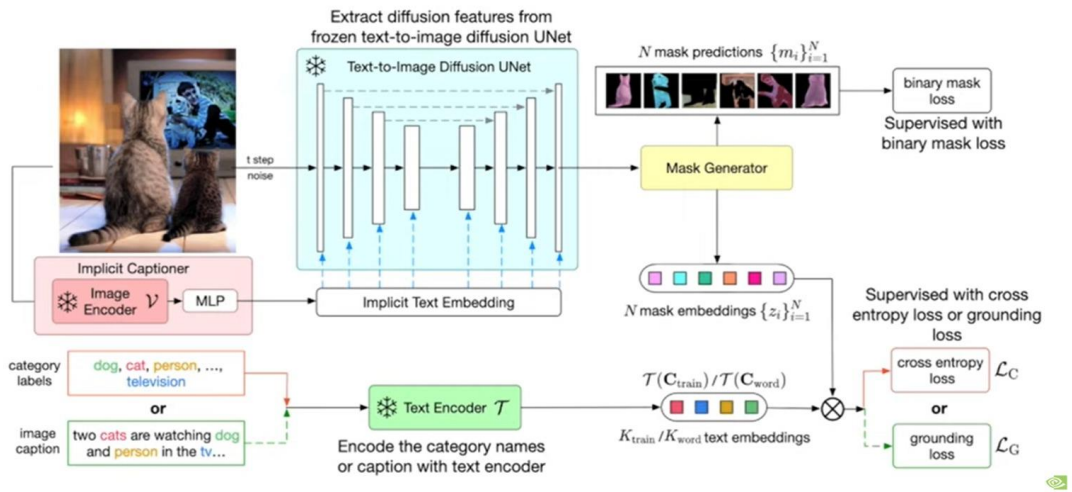
- **Input:**
  - Image: The input to the model is an image that needs to be segmented. The model takes this image as a visual representation for analysis.
  - Caption: A language description (caption/category levels) associated with the input image. This caption is used to guide the text-to-image diffusion model in generating relevant image features.

- **Models Used:**
  - Pre-trained Text-to-Image Diffusion Model: These models can generate high-quality images based on diverse open-vocabulary language descriptions.
  - Frozen Image & Text Encoder
  - Mask Generator
  - Mask Classification Module

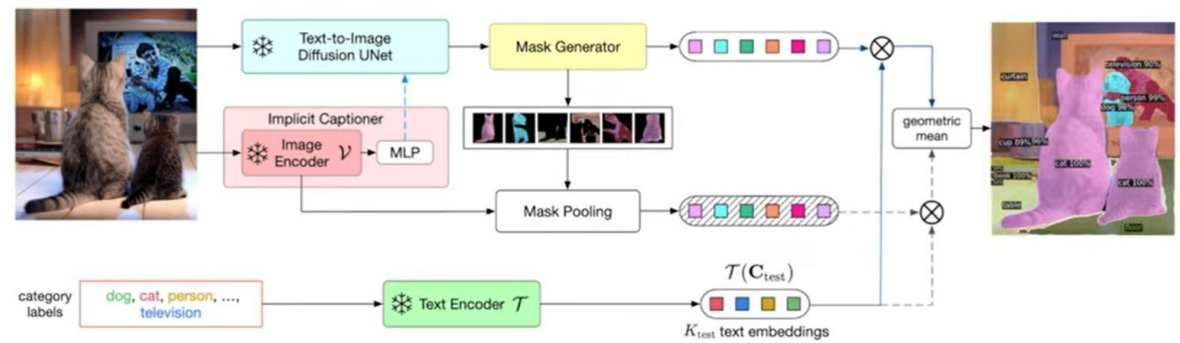
- **Output:**
  - the output of the task provides a comprehensive and detailed segmentation of the input image, with each pixel being labeled according to its semantic class and instance


sky
tree
building
truck 70%
window
pickup truck 60%
trailer truck 79%
grass

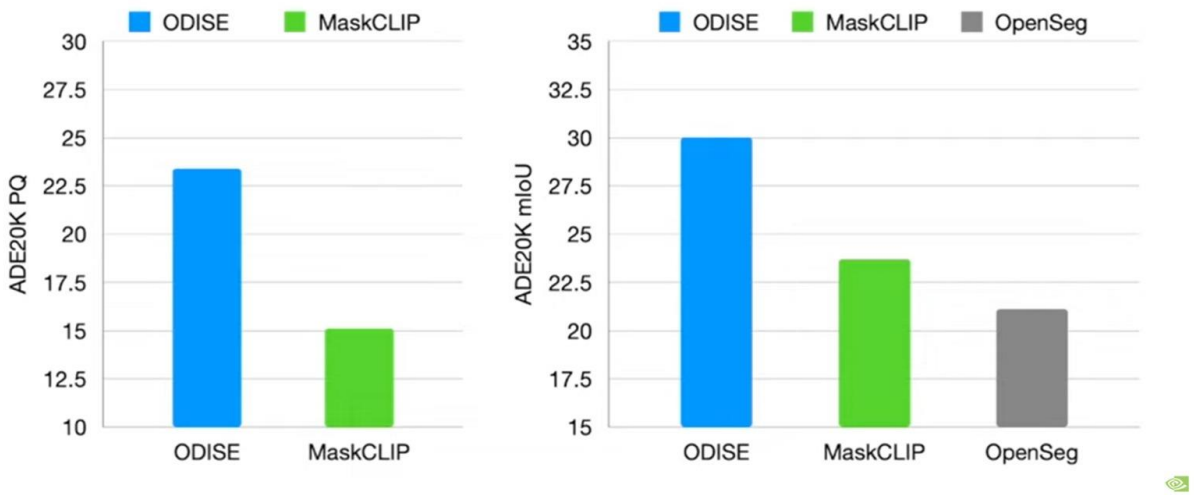
ODISE Training Pipeline



ODISE Inference Pipeline

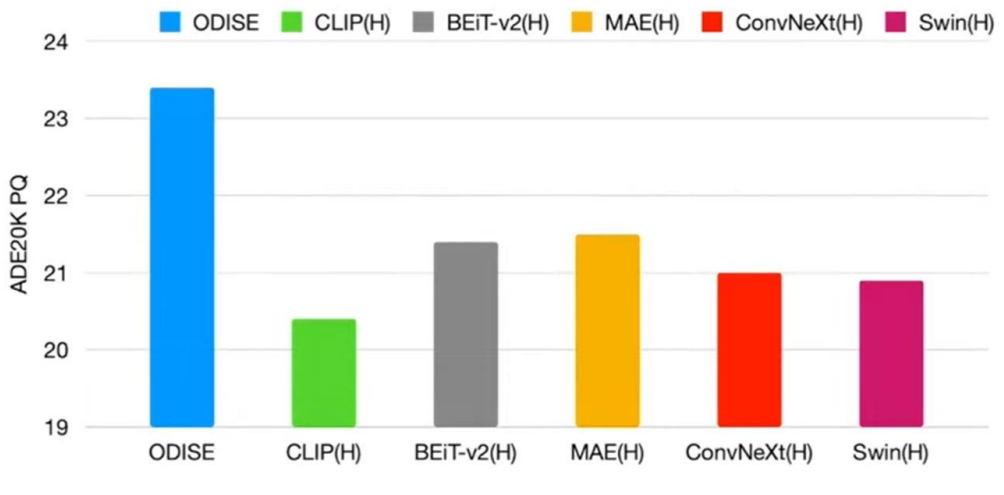


ODISE state-of-the-art performance



Ablation Study

Compare with SOTA representations



Summary: -

- Leverage the frozen representation of larger-scale text-to-image diffusion models for downstream recognition tasks.
- Demonstrate great potential of text-to-image generative models in open-vocabulary segmentation..