JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

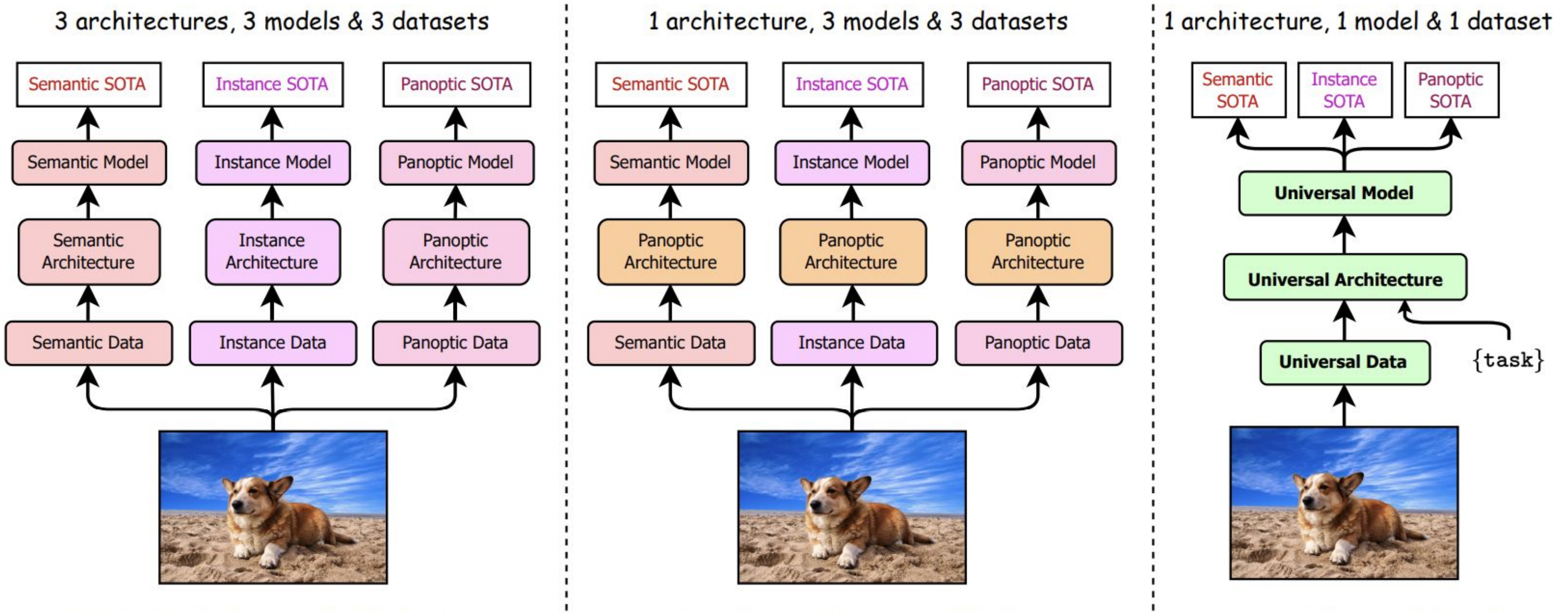
TUE-AM-284

OneFormer: One Transformer to Rule Universal Image Segmentation

Jitesh Jain[1,2], Jiachen Li[1]*, MangTik Chiu[1]*, Ali Hassani[1], Nikita Orlov[3], Humphrey Shi[1,3]

[1]SHI Labs @ U of Oregon & UIUC, [2]IIT Roorkee, [3]Picsart AI Research

OneFormer: *Preview*



(a) Specialized Architectures, Models & Datasets

(b) Panoptic Architecture BUT Specialized Models & Datasets

(c) Universal Architecture, Model and Dataset

- ✅ Universal Architecture
- ✅ Multiple Tasks
- ✅Single Set of Annotations
- ✅Single Model
- ✅Single Training Process
- ✅SOTA Performance

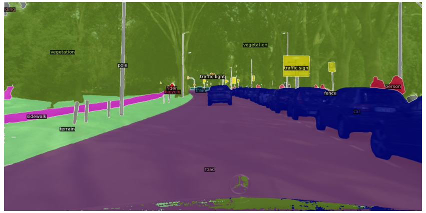
Introduction

Image Segmentation

- Image Segmentation is the task of **grouping pixels into multiple segments**.
- The grouping can be:
  - **Semantic:** One binary segment for **each category irrespective of shape** (*e.g. road, sky, etc.)*.
  - **Instance:** Distinct segments for **each object with well-defined shape** (*e.g. car, person, etc.)*.
  - **Panoptic:** An **amorphous segment for amorphous background regions** (labeled “stuff”) and **distinct segments** for objects with well-defined shape (labeled “thing”).

Input Image

Semantic Segmentation



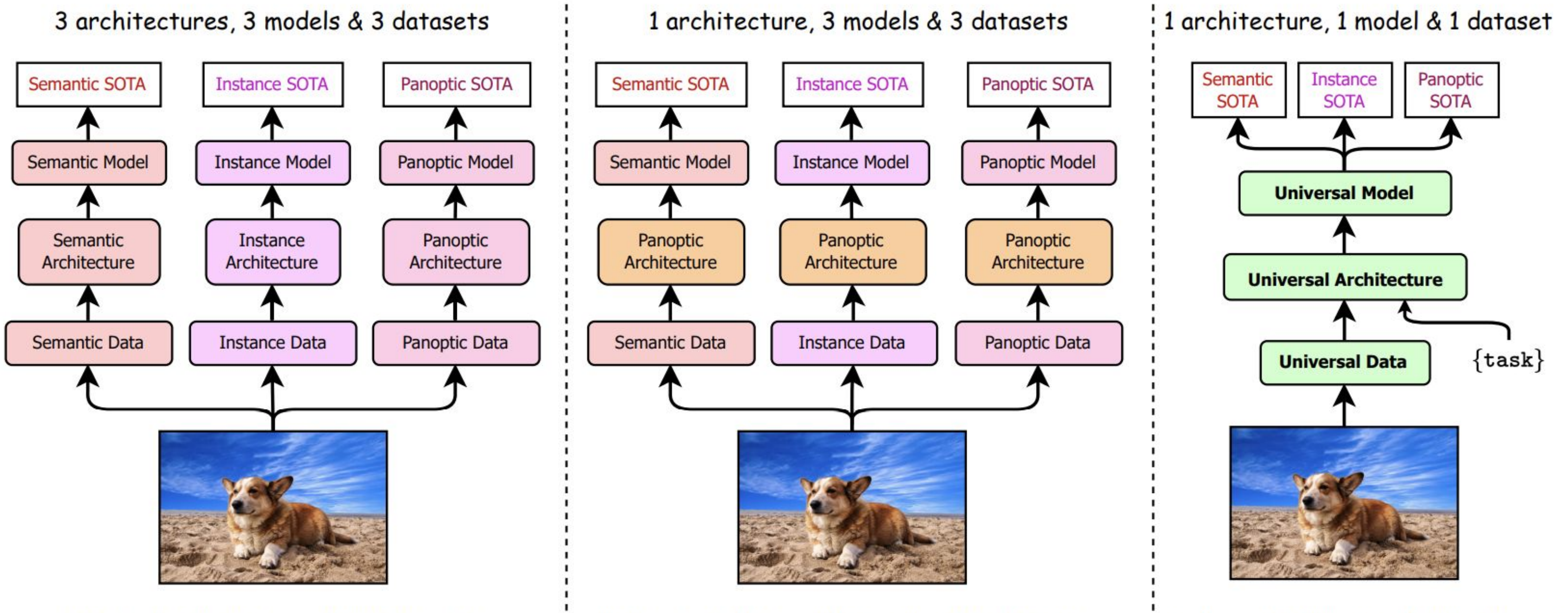
Instance Segmentation

Panoptic Segmentation

Jonathan Long *et al.*, Fully convolutional networks for semantic segmentation. CVPR 2015
Kaiming He *et al.*, Mask R-CNN. ICCV 2017
Alexander Kirillov *et al.*, Panoptic Segmentation. CVPR 2019

Goal

Develop a truly universal image segmentation framework that when trained ***only once outperforms the individually trained models*** *on all three image segmentation tasks.*



(a) Specialized Architectures, Models & Datasets

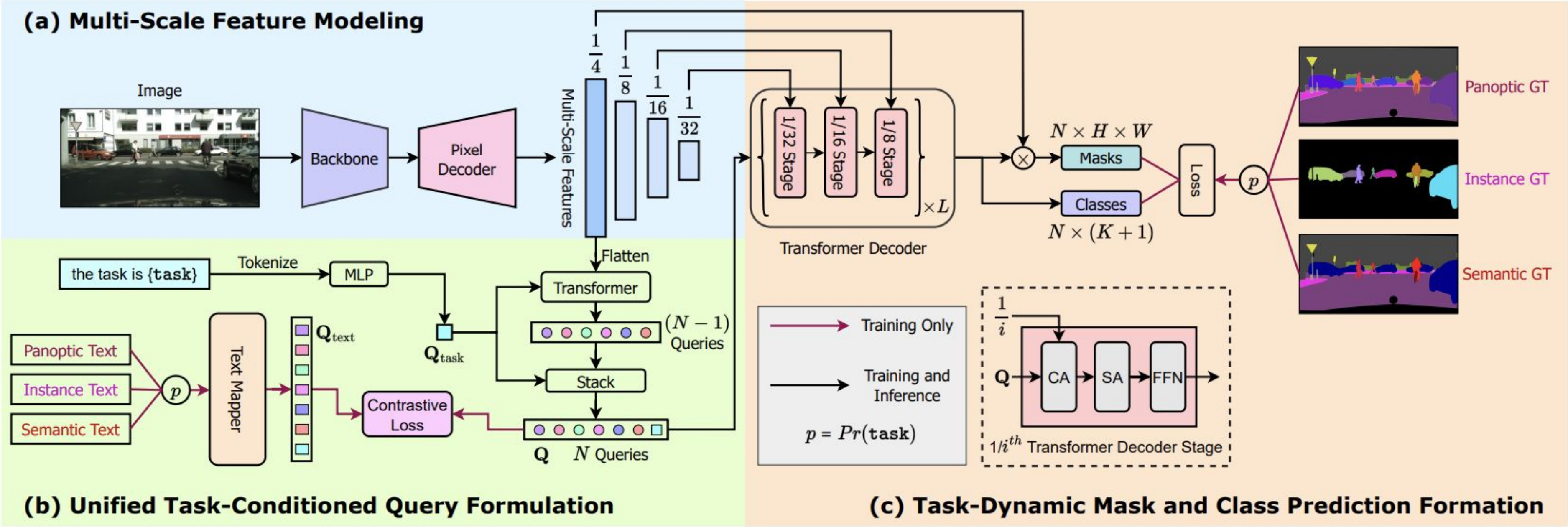
(b) Panoptic Architecture BUT Specialized Models & Datasets

(c) Universal Architecture, Model and Dataset

Methodology

OneFormer

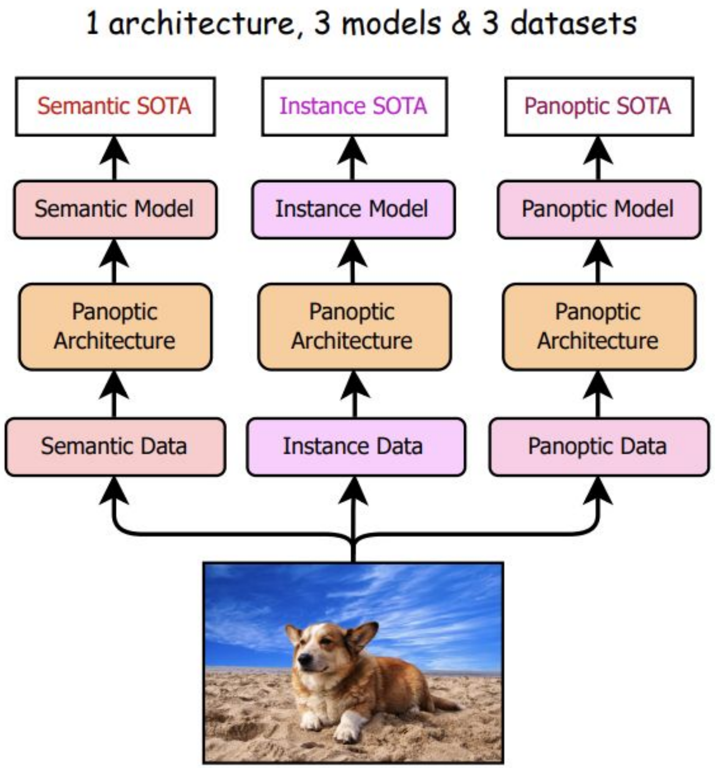
- Multi-task Model
- Task-conditioned Architecture
- Outperforms existing frameworks across semantic, instance, and panoptic segmentation tasks, despite the latter need to be trained separately on each task using multiple times of the resources.



OneFormer v/s Mask2Former

Mask2Former

- ✅Universal Architecture.
- ❌ Multiple Tasks.
- ❌ Single Set of Annotations.
- ❌ Single Model.
- ❌ Single Training Process.
- ✅SOTA Performance.

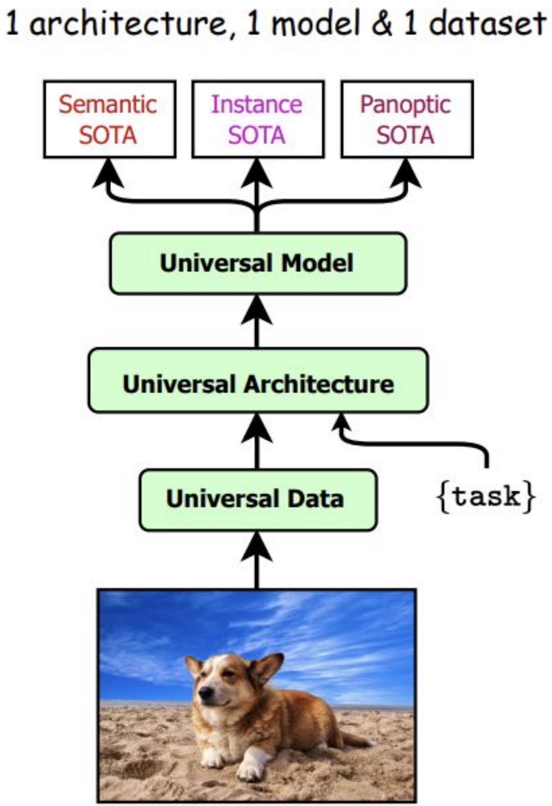


Bowen Cheng *et al.*, Masked-attention mask transformer for universal image segmentation. CVPR 2022

OneFormer v/s Mask2Former

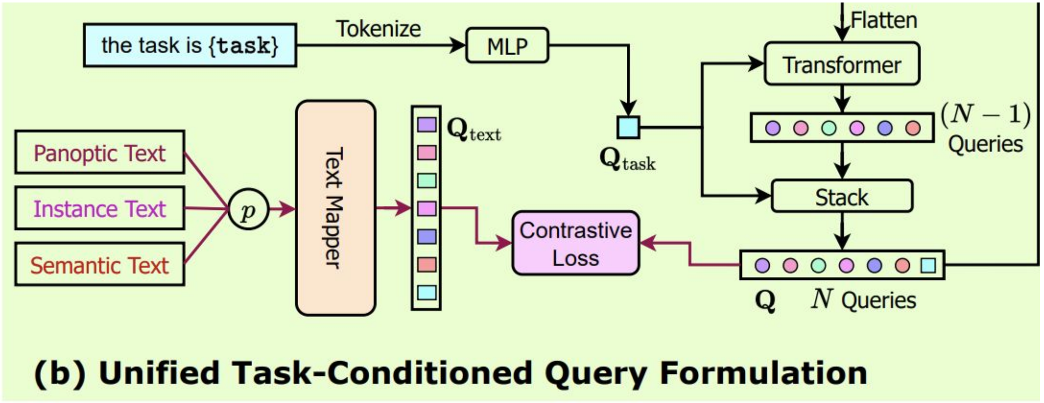
OneFormer

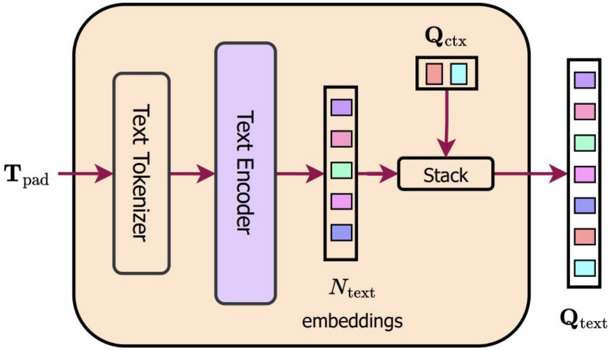
- ✅Universal Architecture.
- ✅Multiple Tasks.
- ✅Single Set of Annotations.
- ✅Single Model.
- ✅Single Training Process.
- ✅SOTA Performance.



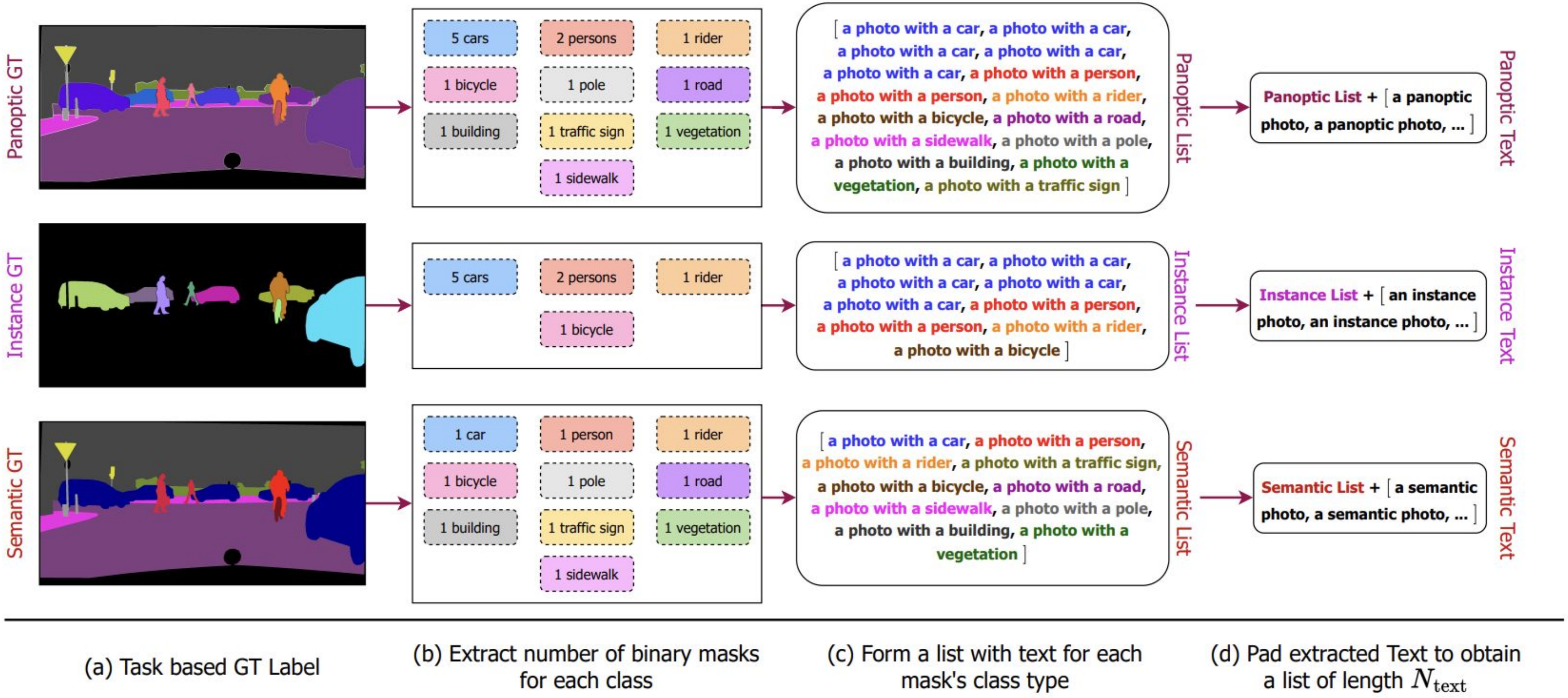
Task-Guided Joint Training

- Uniformly sample task (probability *p*) for the GT label.
- Derive all GT labels from corresponding panoptic annotations during joint training.
- Condition our architecture on the task using a “ the task is a **{task}**” input.
- The GT for a sample depends on the task domain.
  - We use a **query-text contrastive loss** for the model to learn inter-task distinctions.
- Text Mapper can be dropped during inference.





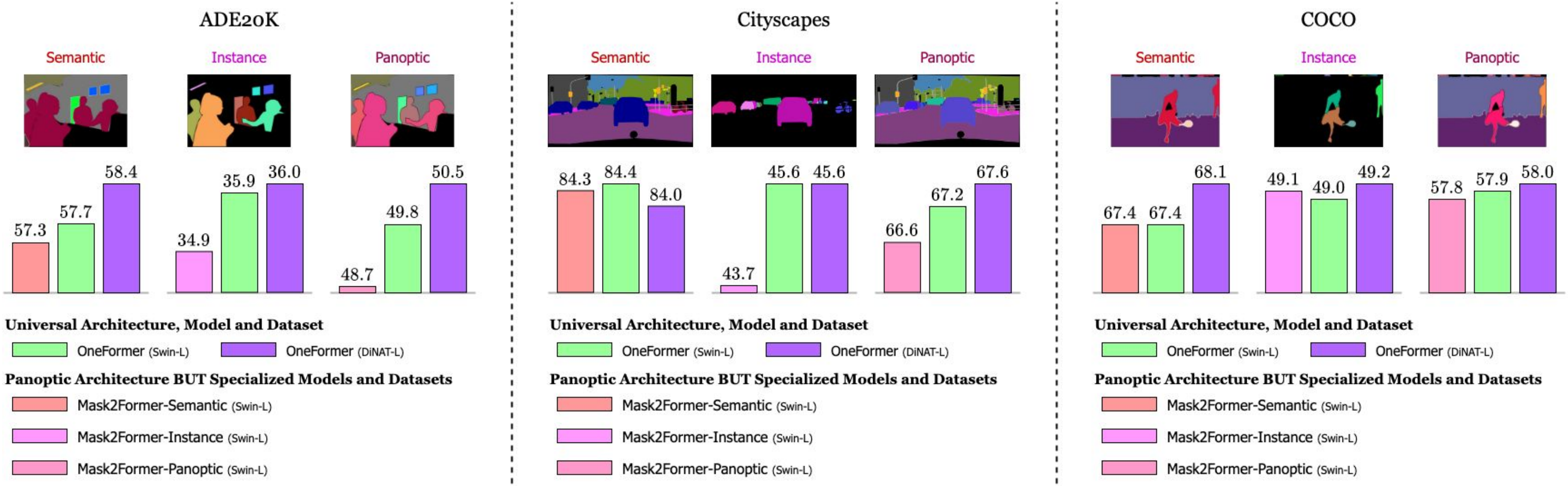
Input Text List Generation



Results

Results: *SOTA on major benchmark datasets*

OneFormer sets a new state-of-the-art performance on all three segmentation tasks compared with methods using the standard Swin-L backbone, and improves even more with new DiNAT backbone.



Ze Liu et al, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, ICCV 2021
Bowen Cheng *et al*., Masked-attention mask transformer for universal image segmentation. CVPR 2022
Hassani *et al*, . Dilated Neighborhood Attention Transformer, arXiv 2022

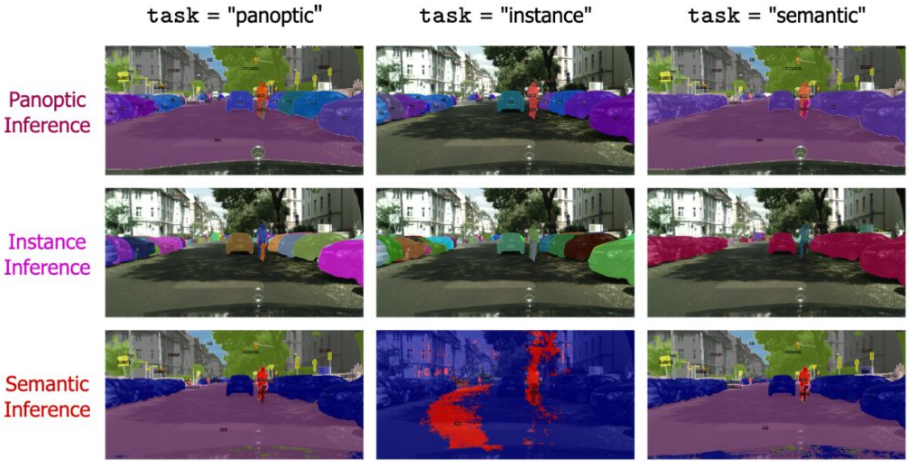
Results: *Task-Dynamic OneFormer*

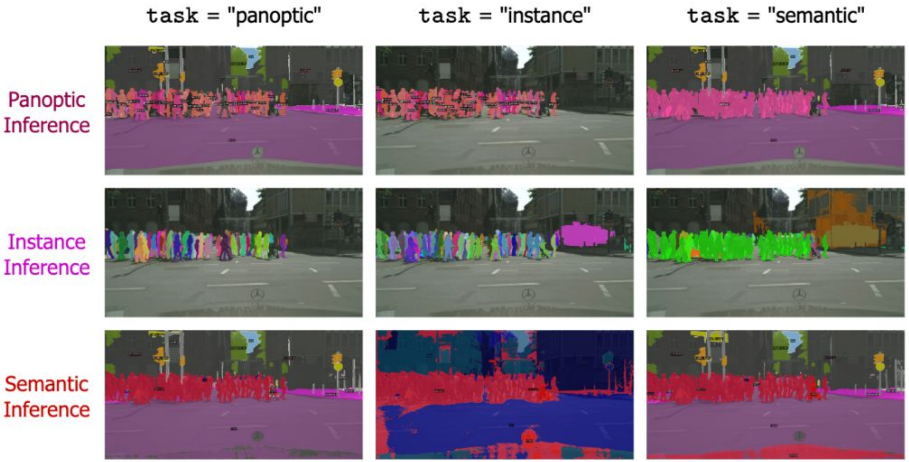
Task Token Input	PQ	PQ^{Th}	PQ^{St}	AP	mIoU
the task is `panoptic`	67.2	61.0	71.7	45.3	83.0
the task is `instance`	25.6	60.8	0.0	45.6	6.3
the task is `semantic`	56.9	36.2	71.9	27.2	83.0

Table V. **Quantitative Analysis on Task Dynamic Nature of OneFormer**. Our OneFormer is sensitive to the input task token value. We report results with Swin-L† OneFormer on the Cityscapes [14] val set. The numbers in pink denote results on secondary `task` metrics.

Task Token Input	PQ	PQ^{Th}	PQ^{St}	AP	mIoU
the task is `panoptic`	49.3	49.6	50.2	35.8	57.0
the task is `instance`	33.1	48.8	1.5	35.9	26.4
the task is `semantic`	40.4	35.5	50.2	25.3	57.0

Table 8. **Ablation on Task Token Input**. Our OneFormer is sensitive to the input task token value. We report results with Swin-L† OneFormer on the ADE20K [15] val set. The numbers in pink denote results on secondary `task` metrics.





Results: *Individual Training*

training strategy	method	PQ	AP	mIoU
Panoptic Training	Mask2Former [12]	40.7	25.2	45.6
	OneFormer (ours)	**41.4** (+0.7)	**27.0** (+1.8)	**46.1** (+0.5)
Instance Training	Mask2Former [12]	—	26.4	—
	OneFormer (ours)	—	**26.7** (+0.3)	—
Semantic Training	Mask2Former [12]	—	—	47.2
	OneFormer (ours)	—	—	**47.3** (+0.1)
Joint Training	Mask2Former† [12]	40.8	25.7	46.6
	OneFormer (ours)	**41.9** (+1.1)	**27.3** (+1.6)	**47.3** (+0.7)

Table IV. **Comparison between Individual and Joint Training.** Unlike Mask2Former [12] which shows large variance in performance among the different training strategies, OneFormer performs fairly well under all training strategies and outperforms Mask2Former [12]. We train all models with R50 [24] backbone on the ADE20K [15] dataset for 160k iterations. † We retrain our own Mask2Former [12] using the joint training strategy.

Conclusion

- Presented **OneFormer**, a new **multi-task universal image segmentation** framework with **task-guided queries**.
- Our **jointly trained single OneFormer model outperforms** the individually trained specialized Mask2Former models, the previous single-architecture state of the art, on all three segmentation tasks.
- OneFormer cuts training time, weight storage, and inference hosting requirements down to a third.
- Makes image segmentation more accessible.
- We believe OneFormer is a significant step towards making image segmentation more universal and accessible.

Scan for a demo!

https://github.com/SHI-Labs/OneFormer

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

TUE-AM-284

OneFormer: One Transformer to Rule Universal Image Segmentation

Jitesh Jain[1,2], Jiachen Li[1*], MangTik Chiu[1*], Ali Hassani[1], Nikita Orlov[3], Humphrey Shi[1,3]

[1]SHI Labs @ U of Oregon & UIUC, [2]IIT Roorkee, [3]Picsart AI Research

Thank You