Indraprastha Institute of Information Technology Delhi
CSE/ECE-344/544 - Computer Vision

EVALUATING THE TRANSFERABILITY OF COCO-TRAINED MODELS TO CITYSCAPES FOR INSTANCE SEGMENTATION

Group 29
2021522 - Avish Dahiya
2021537 - Maanas Gaur
2019196 - Sagar Keim

Fig1. Mask R-CNN on COCO

Instance Segmentation is a computer vision task that involves identifying and separating individual objects within an image, including detecting the boundaries of each object and assigning a unique label to each object. The goal of instance segmentation is to produce a pixel-wise segmentation map of the image, where each pixel is assigned to a specific object instance.

Problem Statement

Domain Shift: Current state-of-the-art instance segmentation models often train on the COCO dataset, which focuses on general object categories. However, applying these models directly to urban scenes (like Cityscapes) can lead to performance change due to domain shift

Motivation: This project aims to assess this domain shift on the effectiveness of models trained on the diverse COCO dataset when applied to the Cityscapes dataset, which focuses on urban environments.

Applications: This evaluation is crucial for real-time image analysis tasks in autonomous driving and urban planning, which rely on accurate object identification and segmentation.

COCO DATASET

The COCO dataset is renowned for its large scale and diversity, containing over 200,000 images with 80 object categories, and is widely used for training robust vision models.

CITYSCAPES DATASET

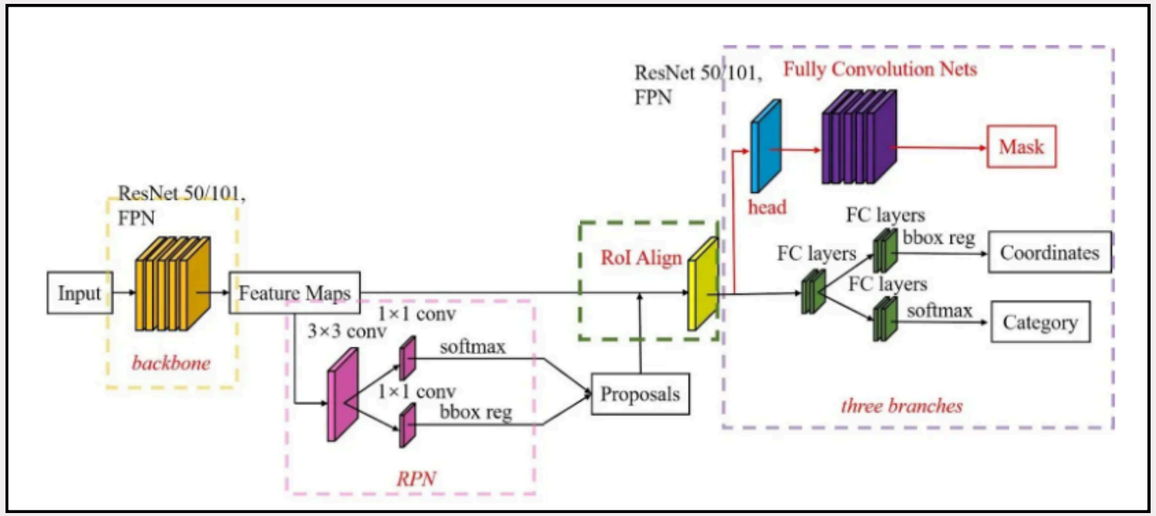
Cityscapes focuses on urban scene understanding across 50 European cities with 5,000 finely annotated images and 20,000 additional images with coarse annotations, making it ideal for testing generalization across urban scenes.

Approaches/Papers

First Model: Mask R-CNN (from Paper 1)

- **Backbone**: Mask R-CNN utilizes a pre-trained convolutional neural network (CNN) like ResNet or ResNext to extract feature maps from the input image. This backbone captures essential visual features.
- **Feature Pyramid Network (FPN)**: This network is added on top of the backbone to address the challenge of scale variations in objects. FPN combines features from different levels of the backbone, creating a multi-scale feature pyramid.
- **Region Proposal Network (RPN)**: This network operates on the feature maps from the backbone, identifying potential object regions and generating bounding box proposals.
- **RoIAlign**: Receives the proposed regions (RoIs) and feature maps. Unlike RoI pooling in Faster R-CNN, it uses bilinear interpolation for precise alignment,

https://arxiv.org/pdf/1703.06870v3.pdf



https://blog.roboflow.com/mask-rcnn/

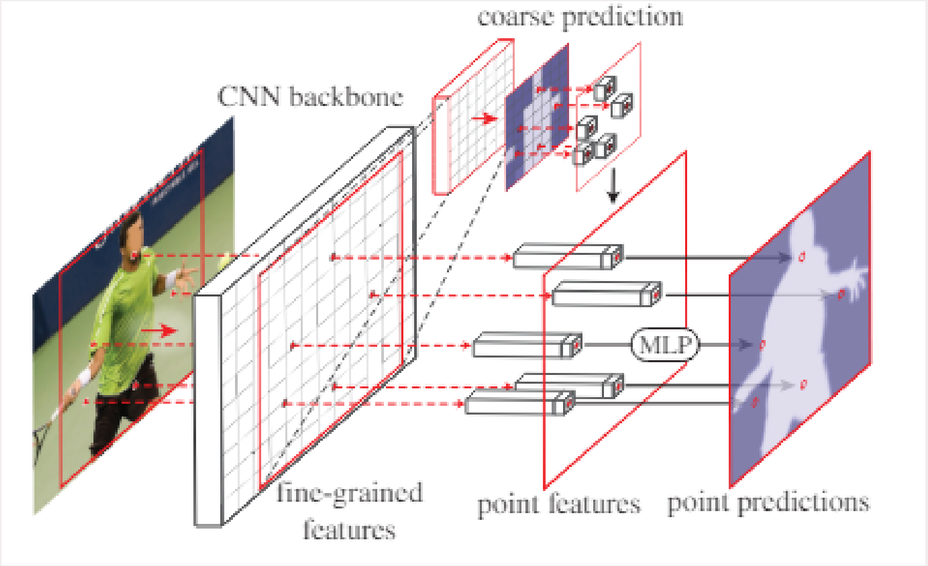
- **Classification Branch**: Predicts the class label for each proposed region.
- **Bounding Box Branch**: Refines the bounding box for each proposed region.
- **Mask Branch**: A newly introduced branch in Mask R-CNN. It operates on the same RoIAlign features as the classification and bounding box branches. This branch predicts a binary mask for each RoI, outlining the exact pixel-level segmentation of the object within the region.

Approaches/Papers

Second Model: PointRend (from Paper 2)

- PointRend is a novel module designed to **improve the efficiency and accuracy of image segmentation tasks**, including instance and semantic segmentation. It approaches segmentation as an "image rendering" problem, **aiming to create high-quality label maps with fewer computational resources**.

- PointRend offers flexibility, **seamlessly integrating into existing segmentation architectures like Mask R-CNN**. It **replaces the standard mask head with its own point-based prediction approach**. This strategy significantly reduces computational cost compared to directly predicting labels for every pixel in the image, while potentially achieving higher segmentation accuracy.



The architecture has 3 main components

- Adaptive Point Selection
- Point-wise Feature Representation:
- Point Head Subnetwork

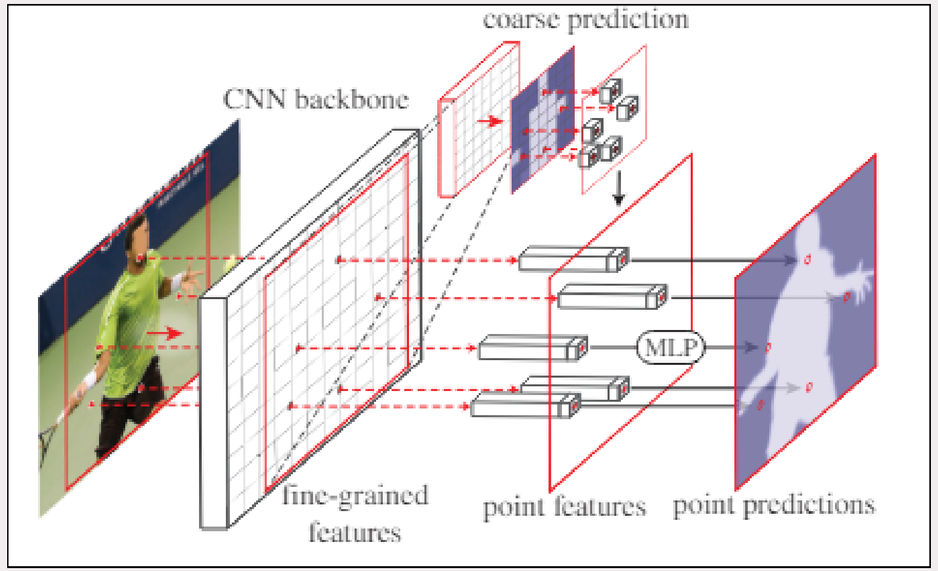
https://arxiv.org/pdf/1703.06870v3.pdf

Approaches/Papers

Three Components

- **Adaptive Point Selection:** Instead of processing every pixel in the high-resolution output grid, it selectively chooses a small number of real-value points for making predictions. It is done for both inference and training.
- During **inference**, it iteratively refines a coarse segmentation by upsampling and then identifying the most uncertain points (often on object boundaries) on the denser grid. These points are chosen for further analysis using their local features extracted from the original feature map.
- For **training**, a different random sampling approach is used, prioritizing points with uncertain coarse predictions while also maintaining some even coverage across the image.

- **Point-wise Feature Representation:** Once the points are chosen, PointRend extracts features specifically for those points. It achieves this by performing a process called bilinear interpolation. It creates a new feature vector by blending nearby grid point values which helps in gathering a unique feature description for every selected point.



https://blog.roboflow.com/mask-rcnn/

- **Point Head Subnetwork**: It's a lightweight neural network specifically designed for assigning labels to each selected point. It uses the feature vector from the previous step to predict the label for each point.

https://arxiv.org/pdf/1703.06870v3.pdf

Approaches/Papers

Loss Functions

- **Mask R-CNN** employs a multi-task loss function combining ***classification loss***, ***bounding box regression loss***, and a ***mask binary cross-entropy loss***. This encourages the model to simultaneously optimize these tasks for accurate object detection, localization, and fine-grained segmentation.

- In **PointRend** The point-wise loss represents the effectiveness in further refining the mask at selected points within the instance mask. Lower values of this loss ensure more accurate object boundaries, highlighting finer details at the selected points.

$$L_{\text{point}} = \frac{1}{N} \sum_{m} \text{smooth}_{L1}(p_m - p_{gt})$$

N is the total number of selected points

p_m is the refined prediction at point at m

p_{gt} is the ground truth label at point at m

$$L = L_{cls} + L_{box} + L_{mask} + L_{point}$$

Experiment

- **Utilize detectron2 library to load pre-trained models from the model zoo (e.g., Mask R-CNN, Mask R-CNN with PointRend).**

Experiment 1: COCO Evaluation (Source Domain)

- **Quantitative:** Average Precision (AP): Overall detection accuracy combining precision and recall. We will calculate variants like AP, AP50, AP75, APs (small objects), APm (medium objects), and APl (large objects) using Intersection over Union (IoU) thresholds.
- **Reasoning** Compare Model Performance Against Each Other
- We can analyze individual AP variants to see if the model struggles with specific object sizes (small, medium, or large).
- Identify Pre existing bias.

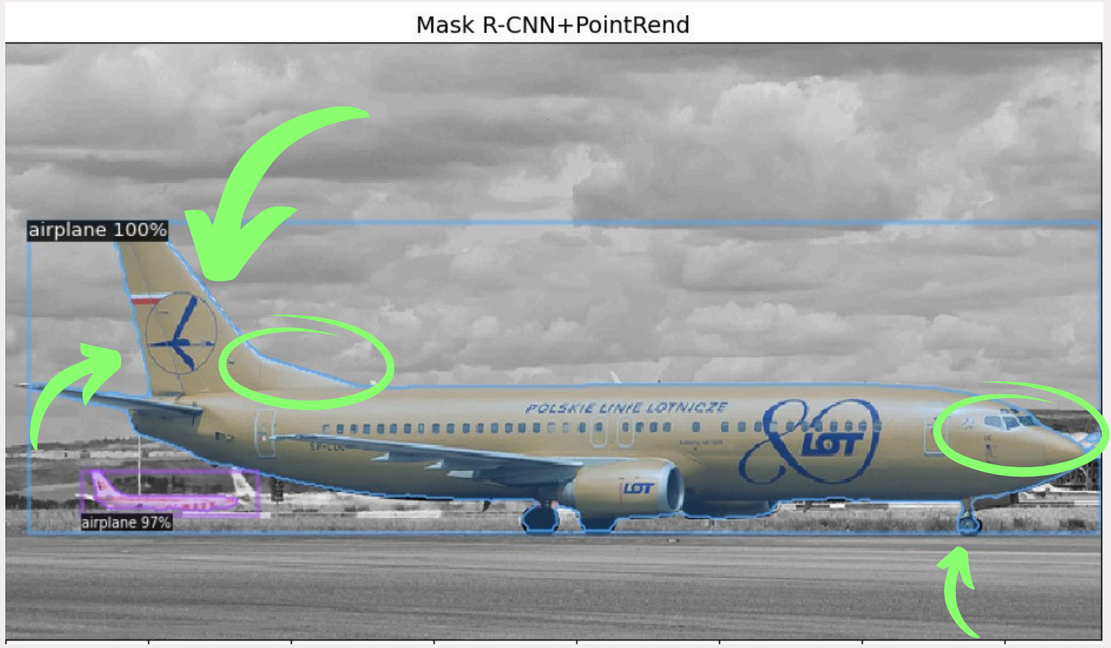
Experiment 2: Cityscapes Evaluation (Target Domain)

- **Quantitative:** Calculate AP and AP50 for each class and total AP. Compare with COCO results.
- **Reasoning:** AP helps us understand how models perform on individual object categories in Cityscapes. Comparing with COCO reveals performance drops due to domain shift (visual differences, class imbalance). Lower APs for specific classes indicate potential biases towards COCO's dominant categories

Qualitative

Reasoning: Visualizations allow us to directly see how well models localize and segment objects (bounding boxes, mask accuracy). We can identify potential issues like misidentified objects or inaccurate mask boundaries , scale of object influence. Verify Quantitative Results

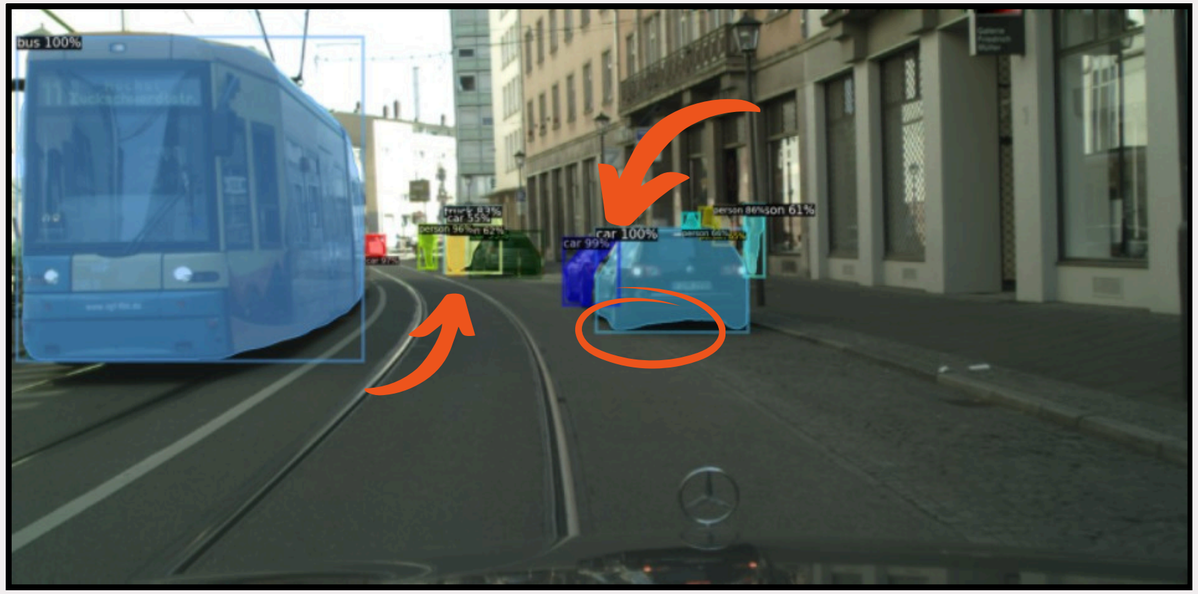
Qualitative Analysis



Mask R-CNN: In the Mask R-CNN segmentation, the mask for the airplane appears to be slightly misaligned around the edges, particularly towards the wing and the tail. This is a common issue with Mask R-CNN, as it predicts bounding boxes around objects and then refines them into masks. This process can lead to inaccuracies, especially around complex shapes.

Mask R-CNN + PointRend: The segmentation have a more precise mask that closely follows the contours of the airplane, including the wing and tail. The high-resolution masks is achieved because PointRend focuses on predicting labels at strategically selected points rather than every pixel in the image. This allows it to capture finer details along object boundaries, resulting in more accurate masks.

Qualitative Analysis



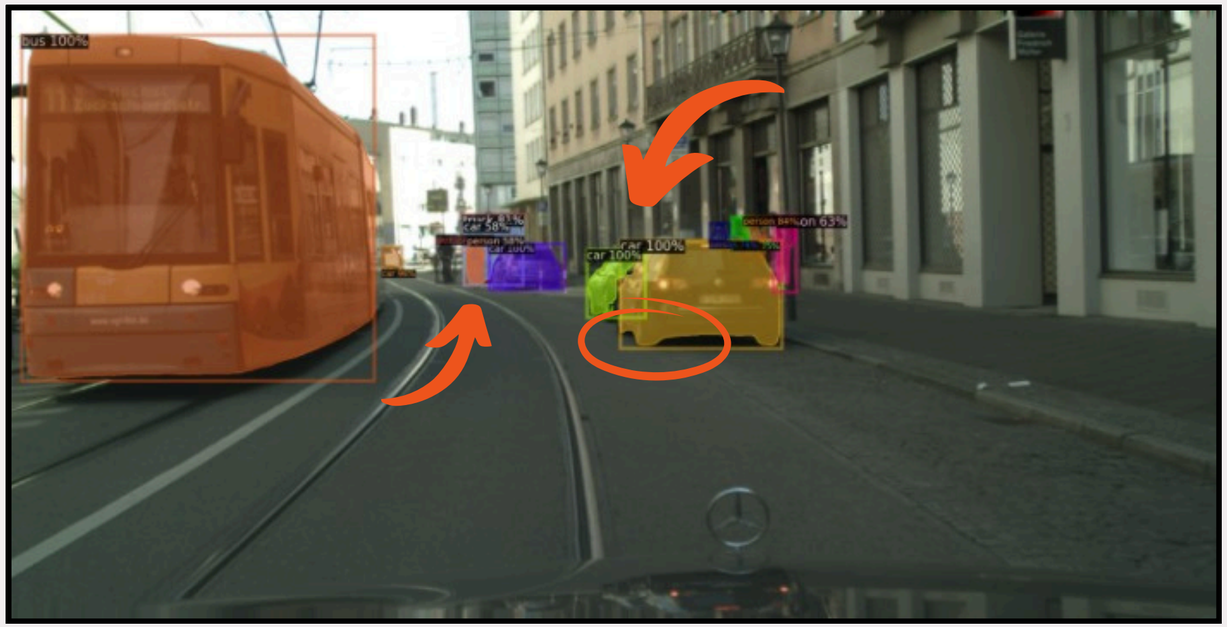
Tire Boundary(Mask R-CNN): In the *Mask R-CNN* segmentation, the mask for the car appears to include some of the background around the tire.

Tire Boundary(MaskR-CNN + PointRend): The segmentation shows some improvement around the tire boundary compared to Mask R-CNN alone. The mask follows the boundary of the tire more closely.

Overlapping

- overlapping bounding boxes can occur, particularly in smaller images. The Region Proposal Network (RPN) in Mask R-CNN generates proposals using anchor boxes, which can lead to redundancy. The additional refinement in segmentation from PointRend might increase the problem by providing a more precise delineation of object boundaries, potentially leading to even more overlapping boxes.

- Both models face challenges with domain shift, where training and testing data distributions differ. Despite PointRend's focus on segmentation quality, it doesn't directly address domain shift issues, potentially impacting performance across diverse domains.



Quantitative Analysis

TABLE III
COCO EVALUATION RESULTS

Model	Metric	AP	AP50	AP75	APs	APm	APl
Mask RCNN	bbox	37.459	54.561	41.855	19.956	40.365	50.871
	segm	34.336	52.787	37.434	15.699	36.583	50.555
Mask RCNN+PointRend	bbox	37.318	54.534	41.471	19.979	40.255	49.893
	segm	35.170	53.001	38.271	16.138	37.441	51.249

TABLE IV
PERFORMANCE COMPARISON OF MASK R-CNN AND POINTREND

Class	AP	$AP_{50\%}$
Mask R-CNN		
Person	25.500	31.874
Rider	1.270	1.587
Car	34.868	43.585
Truck	1.799	2.248
Bus	2.270	2.838
Train	9.869	12.336
Motorcycle	1.193	1.491
Bicycle	3.232	4.040
Average	10.000	12.500
PointRend		
Person	39.219	49.022
Rider	1.953	2.441
Car	53.627	67.034
Truck	2.767	3.457
Bus	3.491	4.365
Train	15.179	18.973
Motorcycle	1.835	2.293
Bicycle	4.971	6.214
Average	15.380	19.225

For the COCO evaluation:

- Mask R-CNN + PointRend achieved higher Average Precision (AP) scores across most classes compared to Mask R-CNN alone.
- PointRend's segmentation refinement led to an increase in APseg (segmentation AP) for many classes, indicating improved segmentation accuracy.
- However, the Average Precision for bounding boxes (APbbox) decreased with the addition of PointRend, suggesting potential challenges in accurately localizing objects, particularly smaller ones.

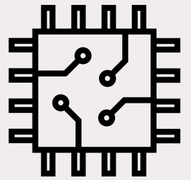
For the Cityscapes evaluation:

- Both models exhibited lower AP scores compared to the COCO evaluation, indicating challenges posed by domain shift.
- Notably, there was a significant disparity in performance between common object categories like "person" and "car" and less frequent ones, highlighting a bias towards COCO's dominant classes.
- This bias may be attributed to the models' training on COCO data, which may not adequately represent the object distribution in the Cityscapes dataset.
- The challenge with accurately segmenting smaller objects could be compounded by inherent limitations in the COCO dataset, which may does provide sufficient diversity in terms of object scales and instances.

Contribution

2021522 - Avish Dahiya [Mask RCNN+PointRend Reproduce Results and Cityscape Inference]
2021537 - Maanas Gaur [Mask RCNN Reproduce Results and Cityscape Inference].
2019196 - Sagar Keim [Mask RCNN+PointRend Reproduce Results and Cityscape Inference]

Hardware and Computational Requirements



Experiments were conducted using Kaggle P100 GPU to ensure adequate computational power for model inference. This setup was chosen to balance performance and speed,

END