Semantic Segmentation of Cityscape and Waste Images Using Large Model Support - a Comparison between IBM Power AC922, NVIDIA DGX Station and AWS p3.8xlarge *

Georgi Kostadinov and Stavri Nikolov

Imagga Technologies @IBM Think London

16 October 2019 (updated December 2019)

IDC Innovator 2016

* The choice of the three hardware architectures compared in this study was defined by the access the authors had to such hardware.

Problem Definition (1/2)

In this study we wanted to benchmark and compare the IBM Power AC922 server vs the NVIDIA DGX Station vs the Amazon Web Services p3.8xlarge instance type for state-of-the-art deep learning training.

We considered a number of common image analysis and recognition tasks, that are used in various applications and that require very intensive computing and data resources.

Initially, image classification with very large number of categories and object recognition were considered, but by state-of-the-art deep learning training infrastructure standards, these tasks are not any more so computing intensive, even for the PlantSnap plant recognition classifier (1) we trained at Imagga with over 320K plant species; we also wanted to use publicly available data sets for this benchmarking.

Other tasks and applications considered were 3D medical image processing, but a benchmarking was already done in (2).

(1) PlantSnap/Imagga: Training the world's largest plant recognition classifier, https://imagga.com/success-stories/plantsnap-case-study.html
(2) Sam Matzek, TensorFlow Large Model Support Case Study with 3D Image Segmentation, IBM Power developer portal, Published on July 27, 2018, https://developer.ibm.com/linuxonpower/2018/07/27/tensorflow-large-model-support-case-study-3d-image-segmentation/

Problem Definition (2/2)

We finally decided to focus this benchmarking study of IBM Power AC922 (3) vs NVIDIA DGX Station (4) vs Amazon Web Services p3.8xlarge instance (5) on the task of semantic segmentation of (a) cityscape images using the Cityscape data set (6, 7) and (b) of waste in the wild using the TACO data set (8) using IBM's Large Model Support.

For the choice of a neural network we were looking for a hardware demanding architecture which uses high-resolution input images for training. We found out that Gated Shape CNN (9) - a state-of-the-art CNN architecture ranking as the one of the top methodologies in the Cityscape benchmark, is the perfect candidate for benchmarking the performance of IBM Power AC922, NVIDIA DGX Station and AWS p3.8xlarge instance.

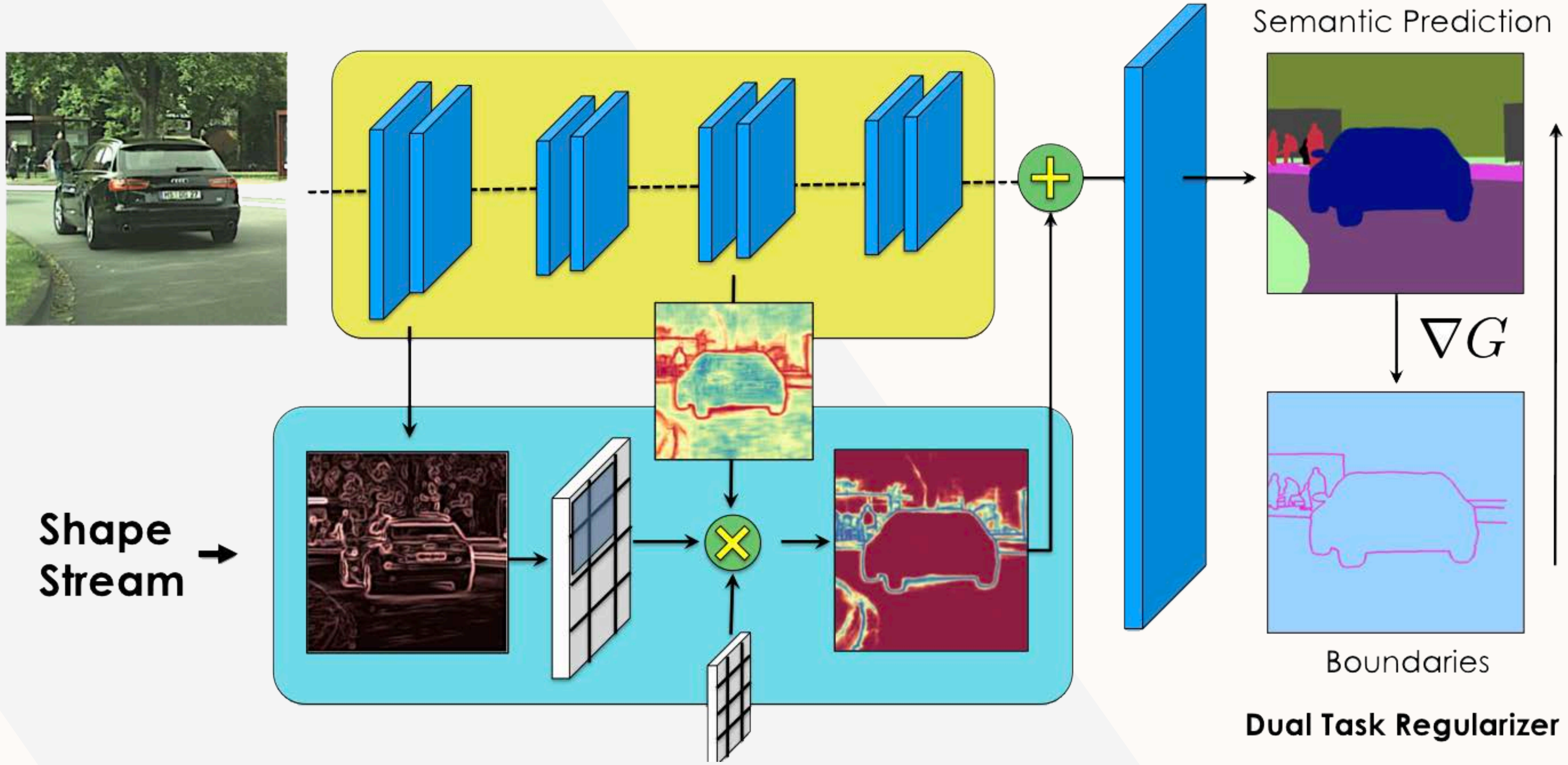
(3) IBM Power AC922 servers and processor chip, https://www.ibm.com/it-infrastructure/power/power9
(4) NVIDIA DGX Workstation, https://www.nvidia.com/en-us/data-center/dgx-station/ and https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/dgx-station/nvidia-dgx-station-datasheet.pdf
(5) AWS p3 instance type, https://aws.amazon.com/ec2/instance-types/p3/
(6) The Cityscapes data set, https://www.cityscapes-dataset.com/
(7) M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, 'The Cityscapes Dataset for Semantic Urban Scene Understanding,' in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, https://www.cityscapes-dataset.com/wordpress/wp-content/papercite-data/pdf/cordts2016cityscapes.pdf
(8) Pedro F. Proena and Pedro Simes, TACO: Trash Annotations in Context Dataset, 2019, http://tacodataset.org
(9) Takikawa, Towaki & Acuna, David & Jampani, Varun & Fidler, Sanja, Gated-SCNN: Gated Shape CNNs for Semantic Segmentation, 2019, https://arxiv.org/pdf/1907.05740.pdf

Cityscape Dataset

- Street photos from 50 cities (cityscapes)
- Several months (spring, summer, fall), daytime
- Good/medium weather conditions
- Manually selected frames
- Large number of dynamic objects
- Varying scene layout
- Varying background
- 5000 annotated images with fine annotations
- 20000 annotated images with coarse annotations
- Very challenging data set for semantic segmentation
- Various applications such as autonomous cars and driving

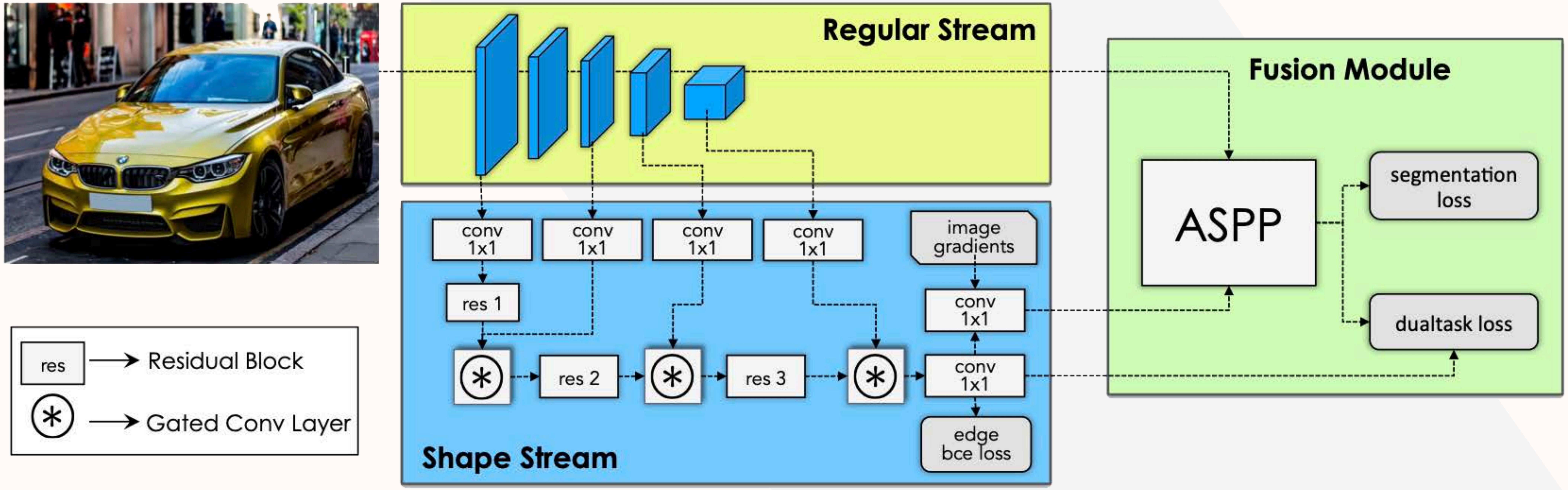
Semantic Segmentation

- Semantic image segmentation is one of the most widely studied problems in computer vision and image analysis with applications in autonomous driving, 3D reconstruction, medical imaging, image generation, etc.
- State-of-the-art approaches for semantic segmentation are predominantly based on Convolutional Neural Networks (CNN).
- Recently, dramatic improvements in performance and inference speed have been driven by new architectural designs.



G-SCNN - Architecture Overview

- State-of-the-art CNN architecture, achieving 82.8% IoU score on the Cityscapes dataset
- Originally trained on a NVIDIA DGX Station 2 with 8 NVIDIA Tesla V100
- Trained with batch size of 16 - 2 per each GPU
- Trained for 175 epochs and high-resolution input size of 800x800



Hardware (1/4) - AWS p3.8xlarge

+ **CPU:** 32-Core Intel Xeon E5-2686 v4
+ **GPU:** 4x 16GB NVIDIA Tesla V100 with NVLink
+ **System Memory:** 244GB

Hardware (2/4) - NVIDIA DGX Station

- **CPU:** 20-core Intel Xeon E5-2698 v4
- **GPU:** 4x 16GB NVIDIA Tesla V100 with NVLink, Water Cooled
- **System Memory:** 256GB

1. GPUs
4X NVIDIA Tesla® V100 16 GB/GPU
500 TFLOPS (Mixed Precision)
20,480 Total NVIDIA CUDA® Cores
2,560 Tensor Cores

2. SYSTEM MEMORY
256 GB RDIMM DDR4

3. GPU INTERCONNECT
NVIDIA NVLink™,
Fully Connected 4-Way

4. STORAGE
Data: 3 x 1.92 TB SSD RAID 0
OS: 1 x 1.92 TB SSD

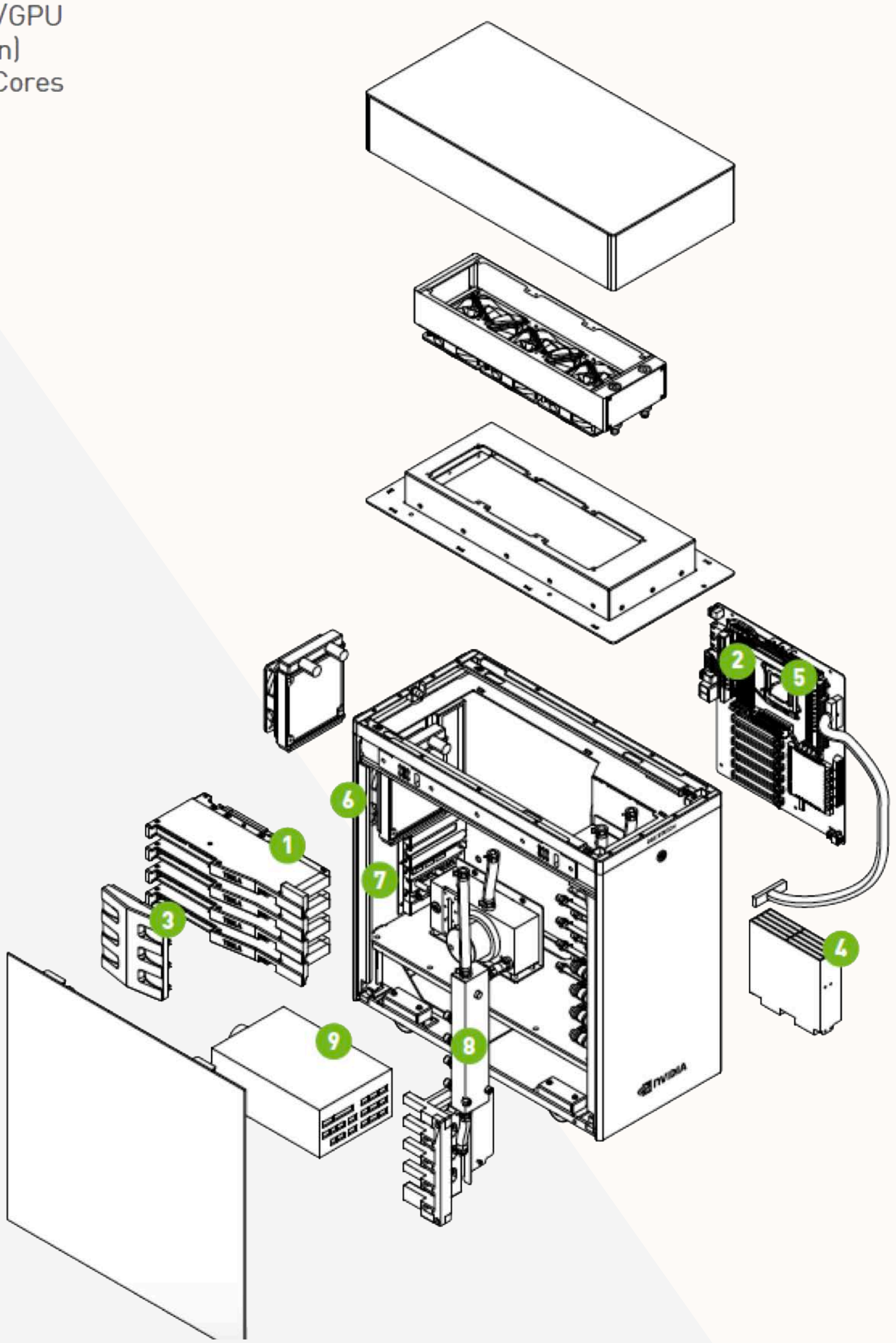
5. CPU
Intel Xeon E5-2698 v4
2.2 GHz 20-Core

6. NETWORKING
2X 10 GbE

7. DISPLAYS
3X DisplayPort,
4K Resolution

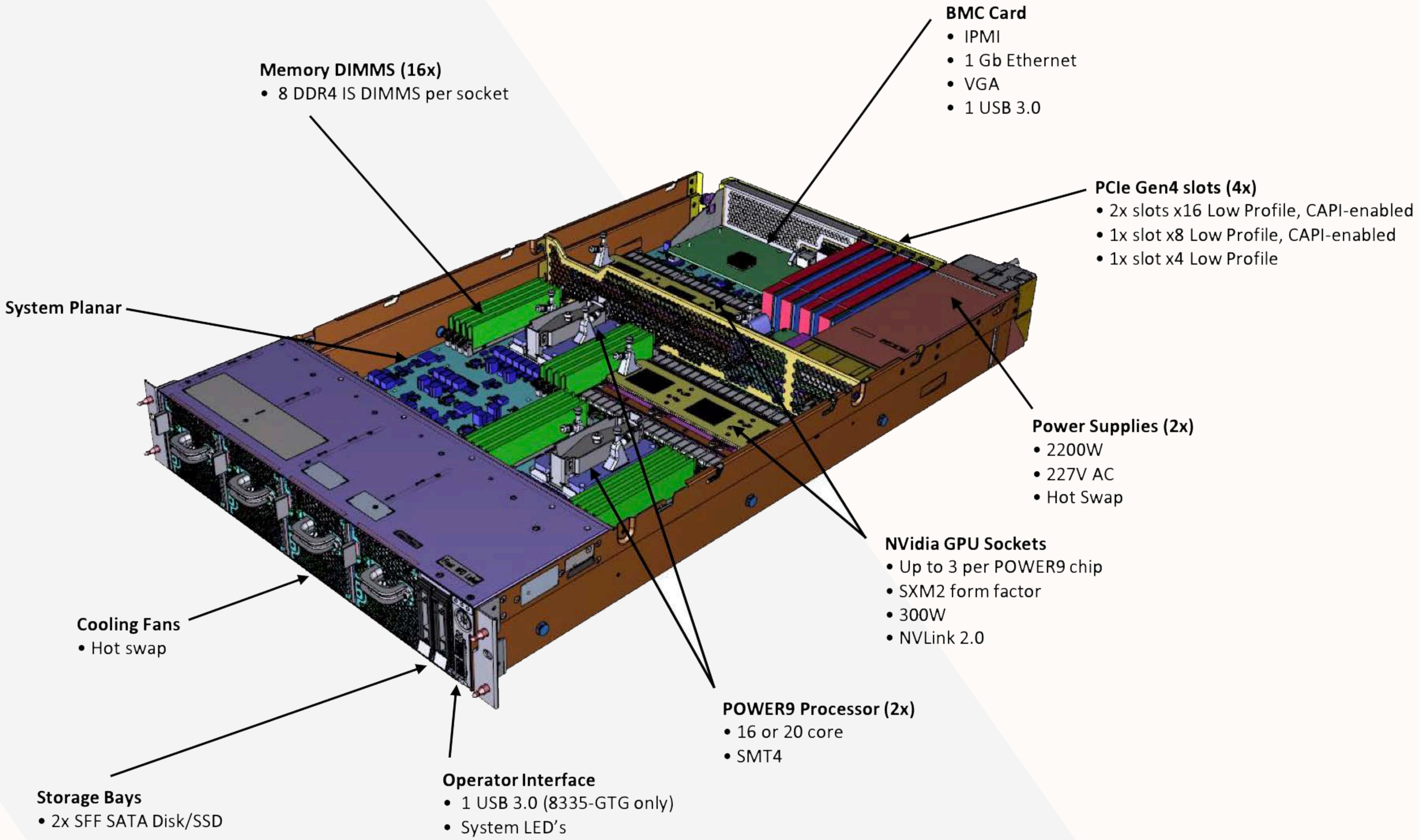
8. COOLING
Water-Cooled

9. POWER
1500 W



Hardware (3/4) - IBM Power AC922

- **CPU:** 32-Core IBM POWER9 Single Chip Module (SCM)
- **GPU:** 4x 16GB SXM2 NVIDIA Tesla V100 with NVLink, Air Cooled
- **System Memory:** 512GB



BMC Card
- IPMI
- 1 Gb Ethernet
- VGA
- 1 USB 3.0

Memory DIMMS (16x)
- 8 DDR4 IS DIMMS per socket

PCIe Gen4 slots (4x)
- 2x slots x16 Low Profile, CAPI-enabled
- 1x slot x8 Low Profile, CAPI-enabled
- 1x slot x4 Low Profile

System Planar

Power Supplies (2x)
- 2200W
- 227V AC
- Hot Swap

NVidia GPU Sockets
- Up to 3 per POWER9 chip
- SXM2 form factor
- 300W
- NVLink 2.0

Cooling Fans
- Hot swap

POWER9 Processor (2x)
- 16 or 20 core
- SMT4

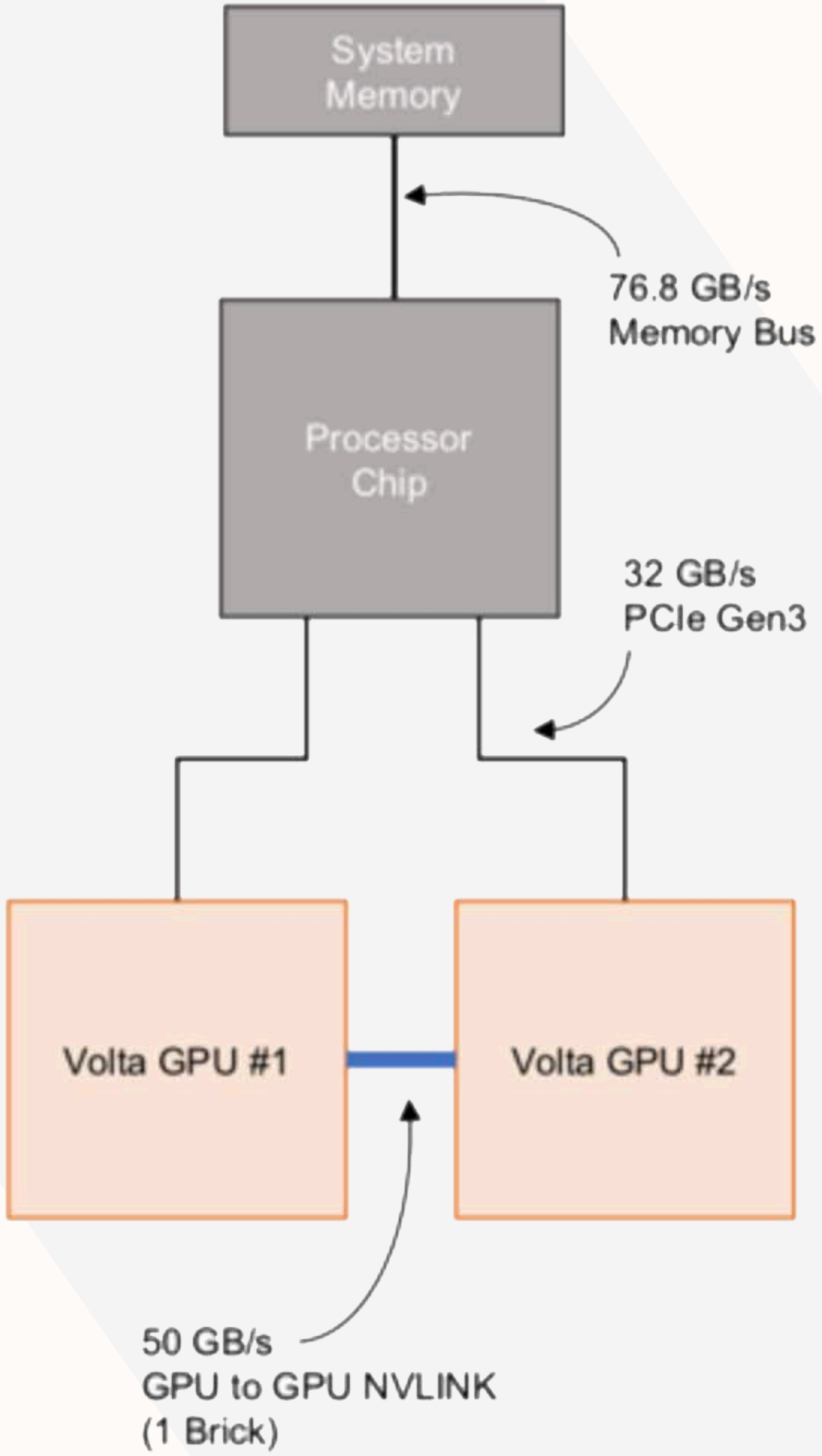
Operator Interface
- 1 USB 3.0 (8335-GTG only)
- System LED’s

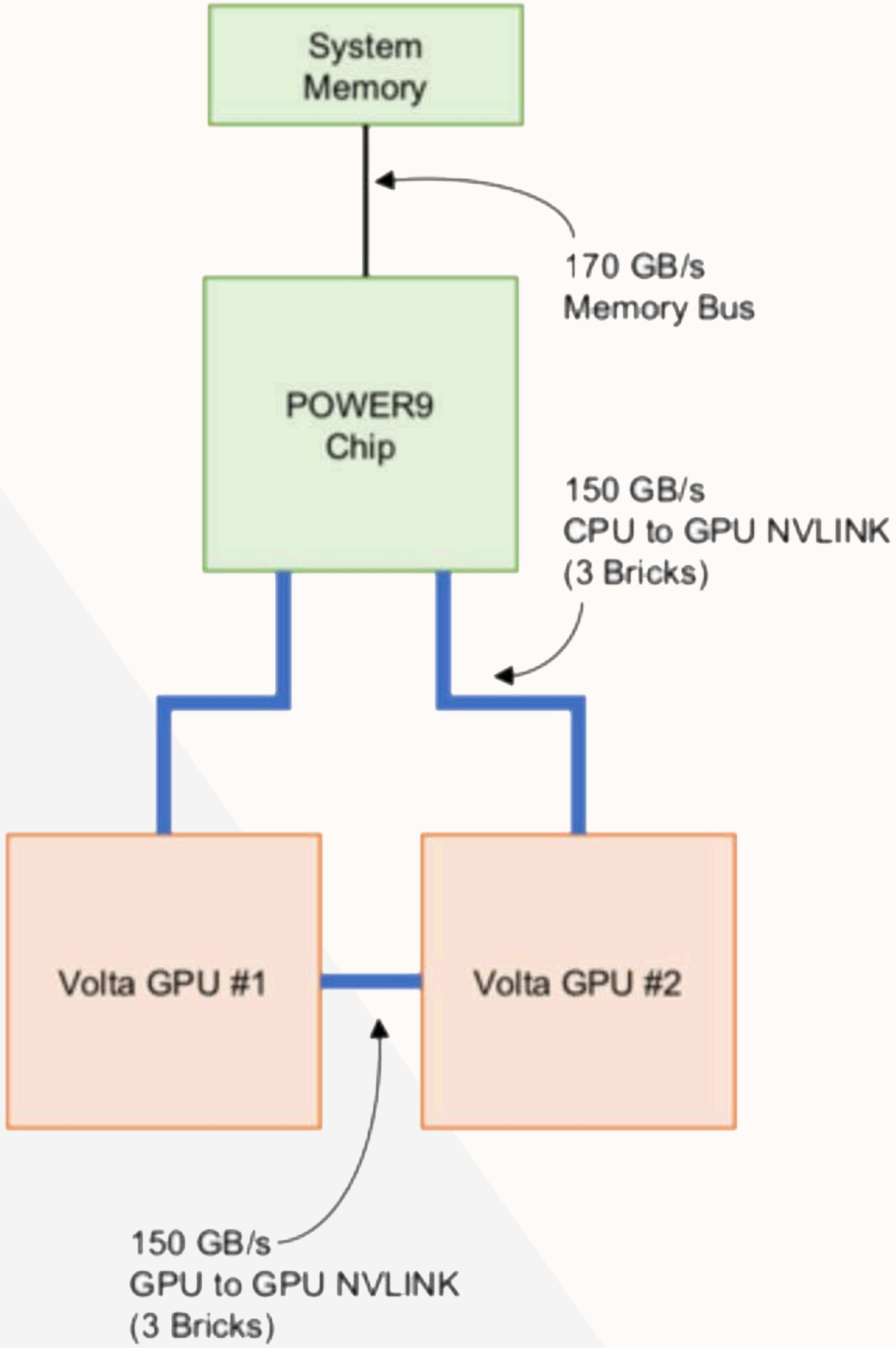
Storage Bays
- 2x SFF SATA Disk/SSD

Hardware (4/4) - NVLink comparison

- **NVIDIA DGX Station and AWS p3.8xlarge (left):** The NVIDIA Tesla V100 GPUs are each connected with a single NVLink 2.0 brick capable of 50 GB/s of bidirectional bandwidth. The CPU and GPU communication is through PCIe Gen3.

- **IBM Power AC922 (right):** The NVIDIA Tesla V100 GPUs are each connected with **three** NVLink 2.0 bricks for up to 150GB/s of bidirectional bandwidth between GPUs. **Three** NVLink 2.0 bricks also connect each GPU with the IBM Power9 CPU providing 150GB/s of bidirectional bandwidth, enabling direct system memory access.




System Memory
170 GB/s Memory Bus
POWER9 Chip
150 GB/s CPU to GPU NVLINK (3 Bricks)
Volta GPU #1
Volta GPU #2
150 GB/s GPU to GPU NVLINK (3 Bricks)

IBM Large Model Support

Large Model Support (LMS) is a feature that allows the successful training of deep learning models that would otherwise exhaust GPU memory and abort with **out of memory** errors. LMS manages this oversubscription of GPU memory by temporarily swapping tensors to host memory when they are not needed.

IBM POWER Systems with NVLink technology are especially well-suited to LMS because of their hardware topology that enables fast communication between CPU and GPUs.

One or more elements of a deep learning model can lead to GPU memory exhaustion. These include:

+ Model depth and complexity
+ Input data size (e.g. high-resolution images)
+ Batch size

With IBM LMS, deep learning models can scale significantly beyond what was previously possible and, ultimately, generate more accurate results.

Benchmarks (1/5) - Overview

To showcase the benefits of using LMS and to benchmark the performance of IBM Power AC922 vs NVIDIA DGX Station vs AWS p3.8xlarge, the following benchmarks and tests were created:

- Training time - comparing the training time for 175 epochs on the Cityscape dataset on the AC922, the DGX and the AWS p3.8xlarge instance type
- Use of LMS - showcasing what the benefits of using Large Model Support are by demonstrating "Out of Memory" situations using the G-SCNN architecture and the Cityscape dataset
- LMS overhead - overviewing the training time overhead when using Large Model Support
- GPU profiling - a detailed comparison of the two systems during training using NVIDIA profiling data

Benchmarks (2/5) - Training time

Training parameters:

- **Input size:** 800x800
- **Batch size:** 16
- **Validation batch size:** 2
- **Epochs:** 175
- **Learning rate:** 0.01
- **Learning rate policy:** polynomial

Training Time

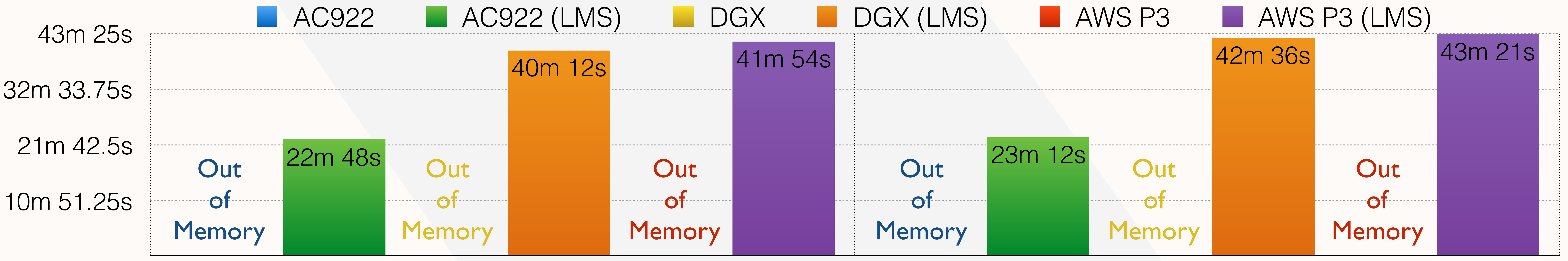
	Total Time	**Accuracy**
IBM Power AC922	3d 17h	72.6%
NVIDIA DGX Station	6d 11h	73.3%
AWS p3.8xlarge	6d 14h	70.6%

The training on IBM Power AC922 completed first - 3 days and 6 hours earlier than the one on the NVIDIA DGX Station with almost no difference in accuracy. The training on the AWS p3.8xlarge came last - 3 hours later than the DGX.

Benchmarks (3/5) - Use of LMS

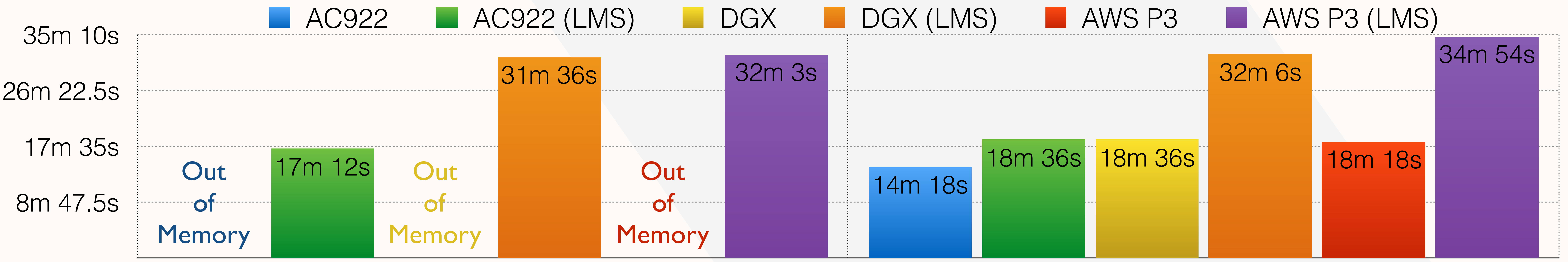
- The semantic segmentation using G-SCNN has a high-memory usage requirements due to the large input size of 800x800 and the architecture design.
- The neural network framework of our choice is the same framework used in the paper - **PyTorch**. The LMS integration was as easy as adding a single line of code:
  torch.cuda.set_enabled_lms(True)
- Without LMS activated, with batch size 16 the training couldn't fit in the 4x NVIDIA Tesla V100 GPUs on all machines, resulting in an “Out of Memory” error.
- Batch size of 8 fitted on the four GPUs but with some reduction of the input size from 800x800 to 700x700.
- Charts for epoch times are shown on the next slide.

Benchmarks (3/5) - Use of LMS



800x800, 16 batch size, epoch time

800x800, 8 batch size, epoch time



700x700, 16 batch size, epoch time

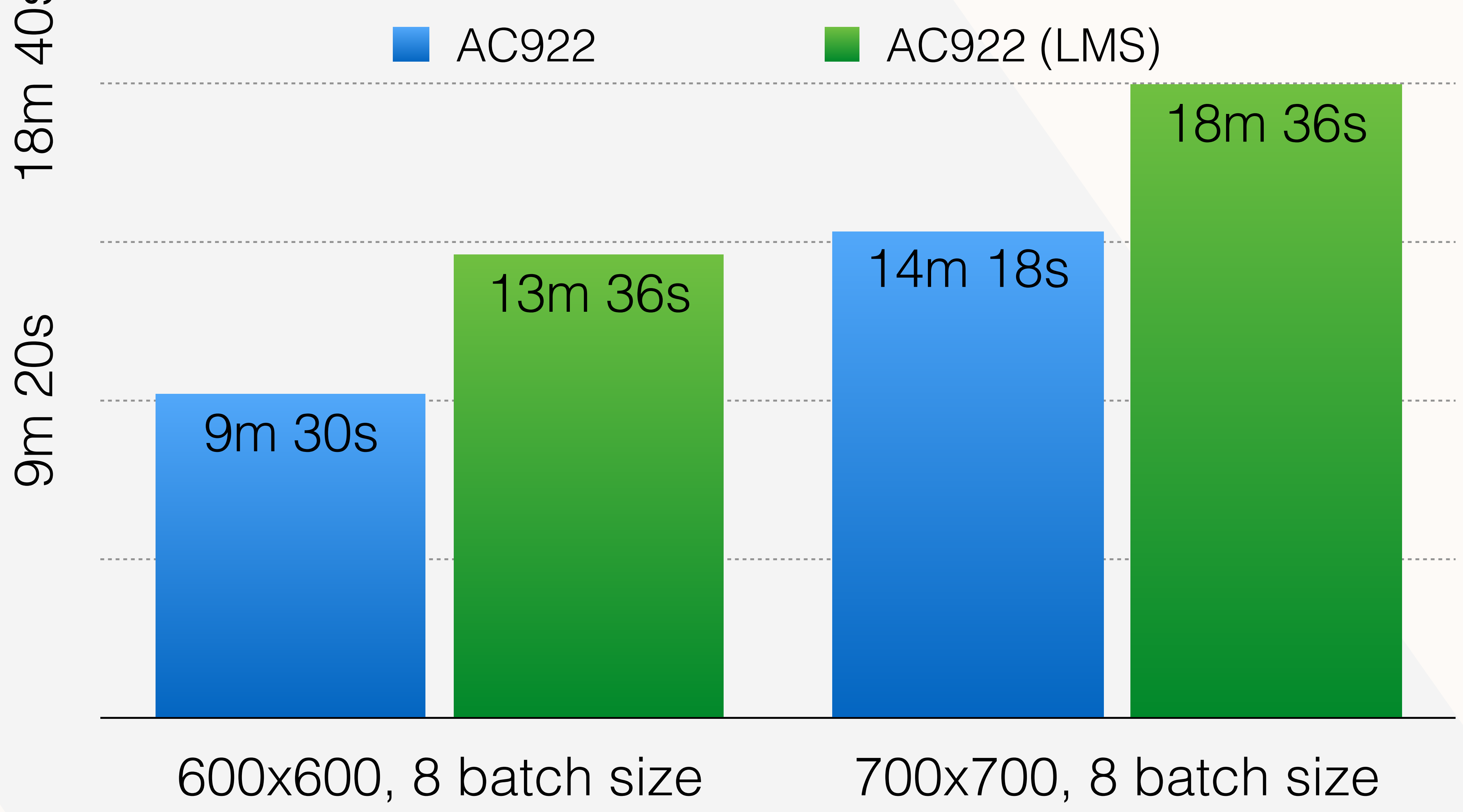
700x700, 8 batch size, epoch time

Benchmarks (4/5) - LMS overhead

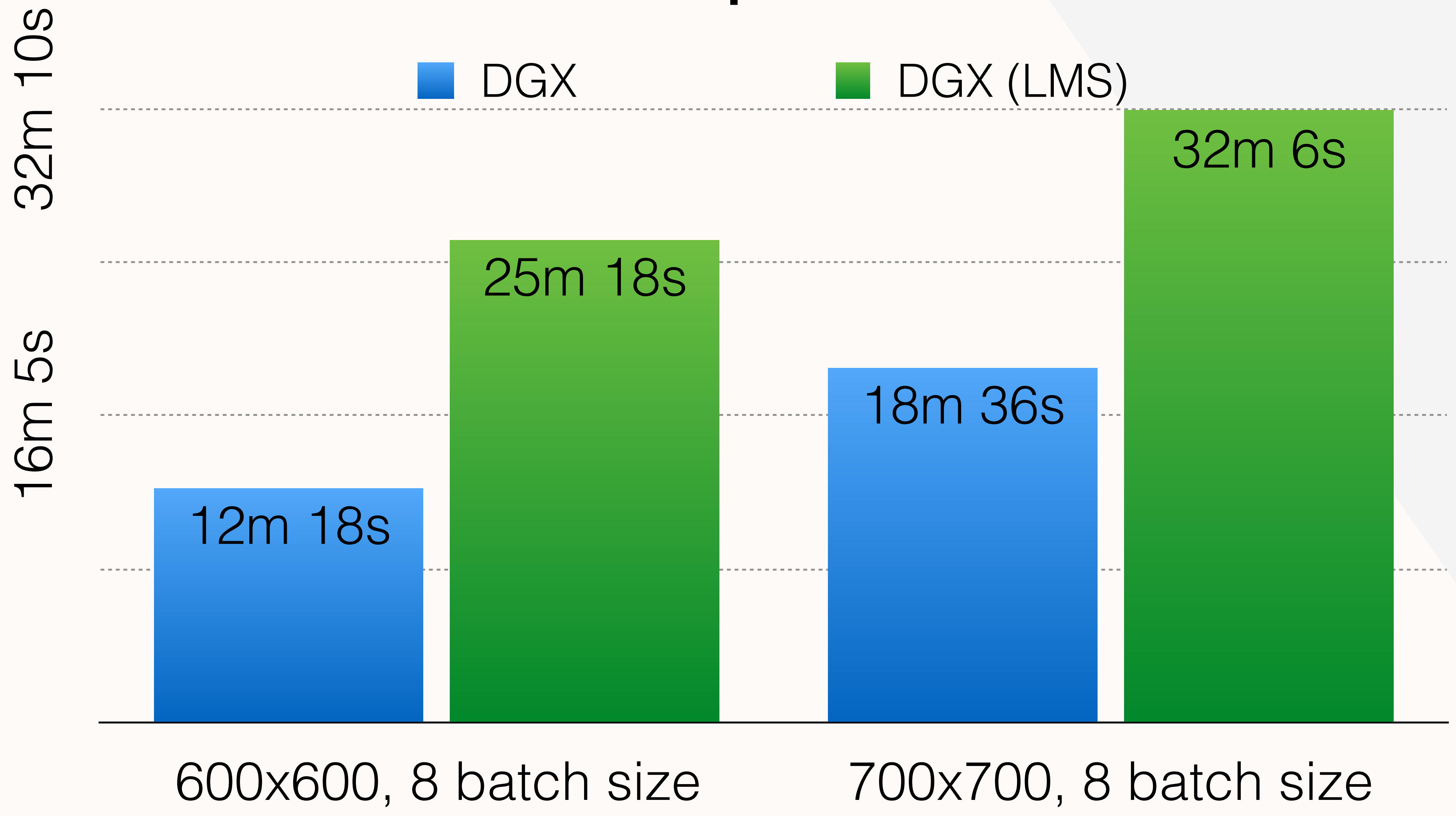
- For calculating the LMS overhead we used input sizes of 600x600 and 700x700 and a fixed batch size of 8.
- LMS overhead for 600x600 input size is 106% for the DGX , 105% for the AWS p3.8xlarge and 43% for the AC922.
- LMS overhead for 700x700 input size is 72.8% for the DGX, 91% for the AWS p3.8xlarge and 30% for the AC922.
- IBM Power AC922 shows significantly lower LMS overhead due to the NVLink connectivity between the CPU and the GPU.
- The AWS p3.8xlarge instance shows similar performance to the DGX due to the almost identical GPU and CPU system architecture.
- The next slide shows exact epoch times for each machine and train type.

Benchmarks (4/5) - LMS overhead

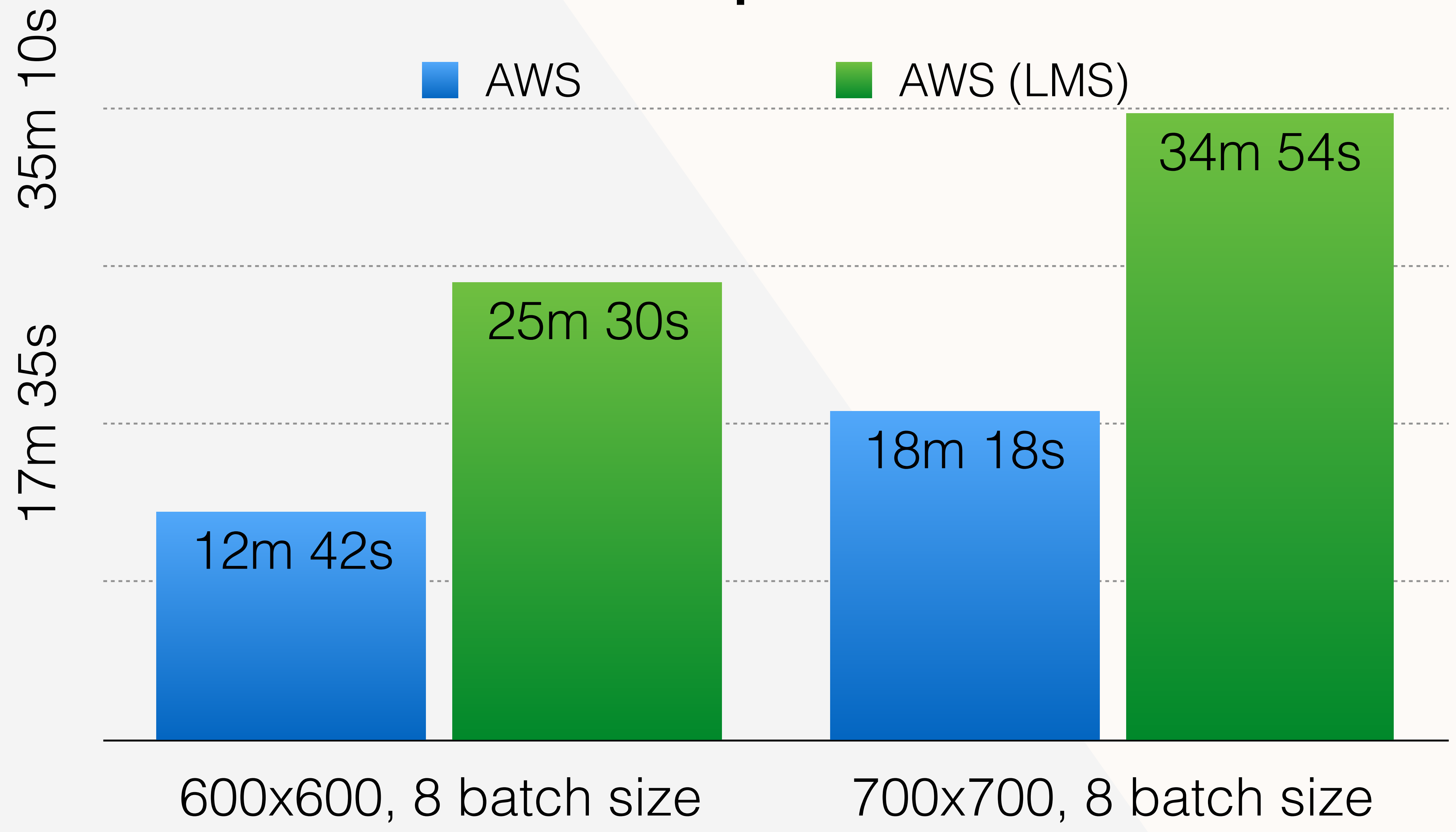
AC922 epoch time



DGX epoch time



AWS epoch time



Benchmarks (5/5) - GPU profiling

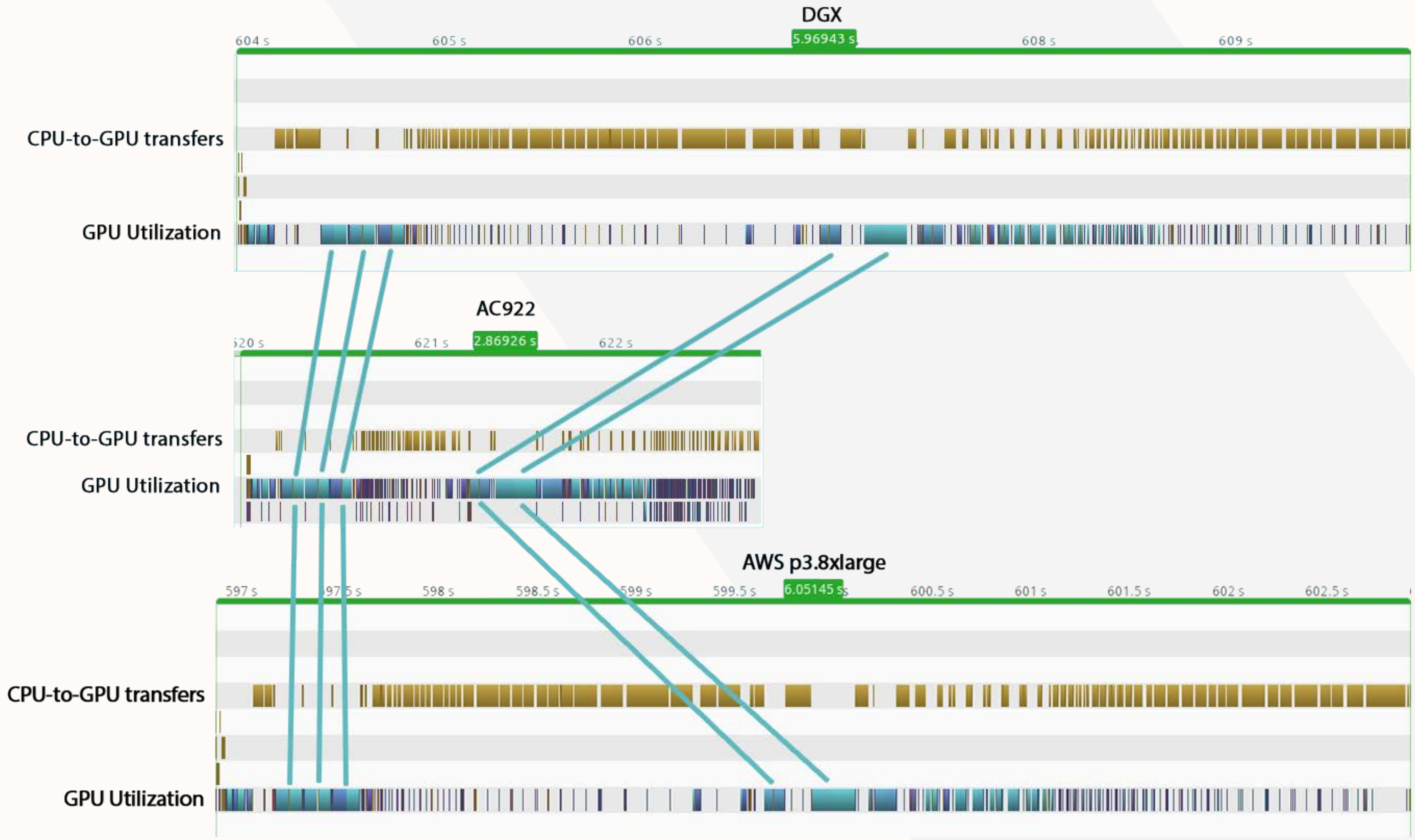
- To investigate further where the difference in the numbers between the machines come from, we used Nvprof to profile the GPU activity during epoch 2 between 40th and 60th iteration.

- Nvprof shows that the memory copies between the CPU and GPU for tensor swapping for the LMS take considerably longer on the NVIDIA DGX Station and the AWS p3.8xlarge instance type than on the IBM Power AC922 and lead to GPUs becoming idle.

- The graphic on the next slide shows GPU usage on the 50th iteration on the three machines. The blue lines relatively mark the locations of equivalent tensors on the machines.

- The gaps in the GPU utilisation for the AC922 machine are drastically smaller than both the DGX and AWS - leading to higher utilisation and faster trainings.

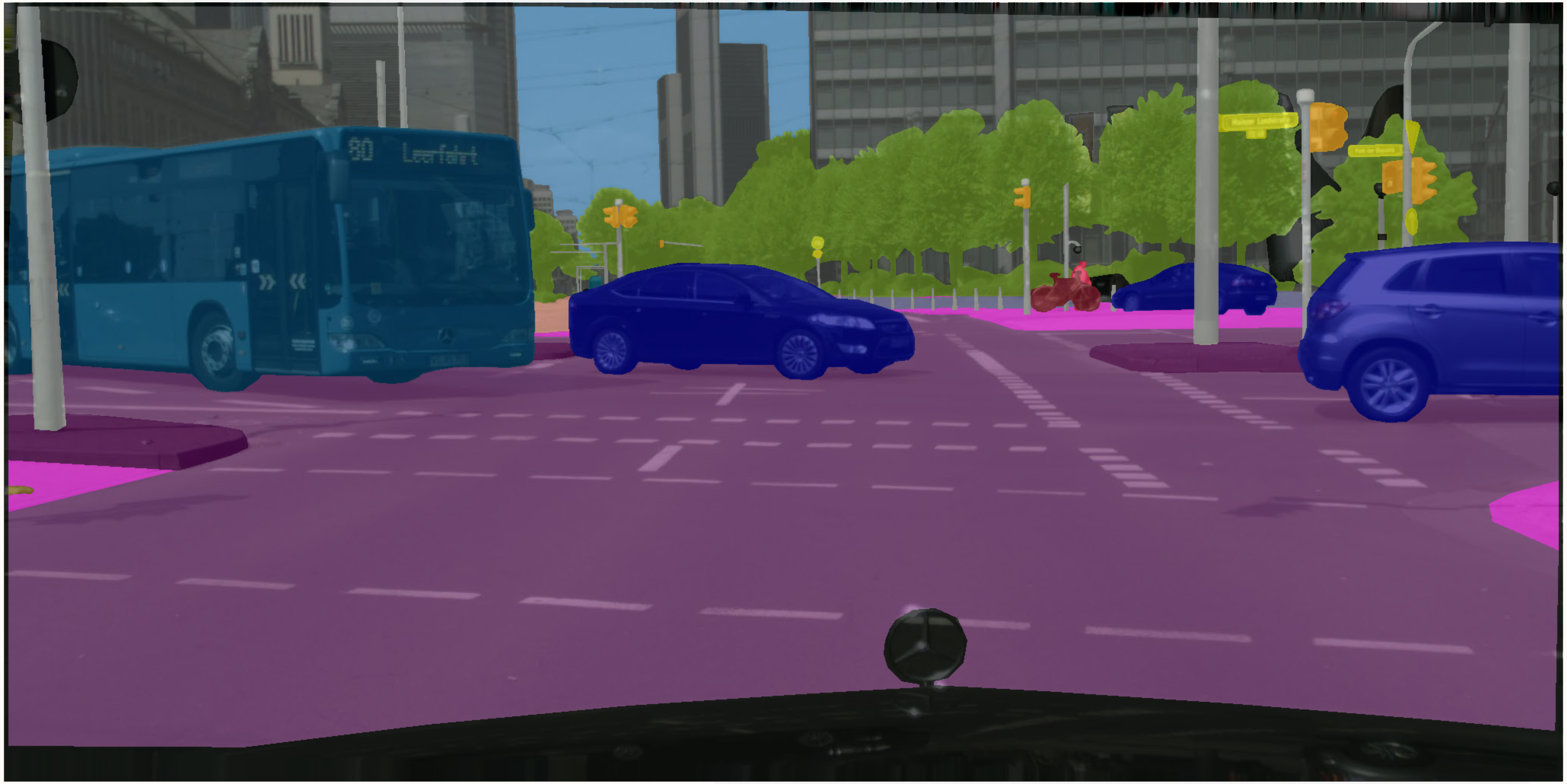
Benchmarks (5/5) - GPU profiling



	CPU-to-GPU throughput	GPU Utilization
AC922	39.25GB/s	76.5%
DGX	6.94GB/s	37%
AWS P3	6.42GB/s	29.7%

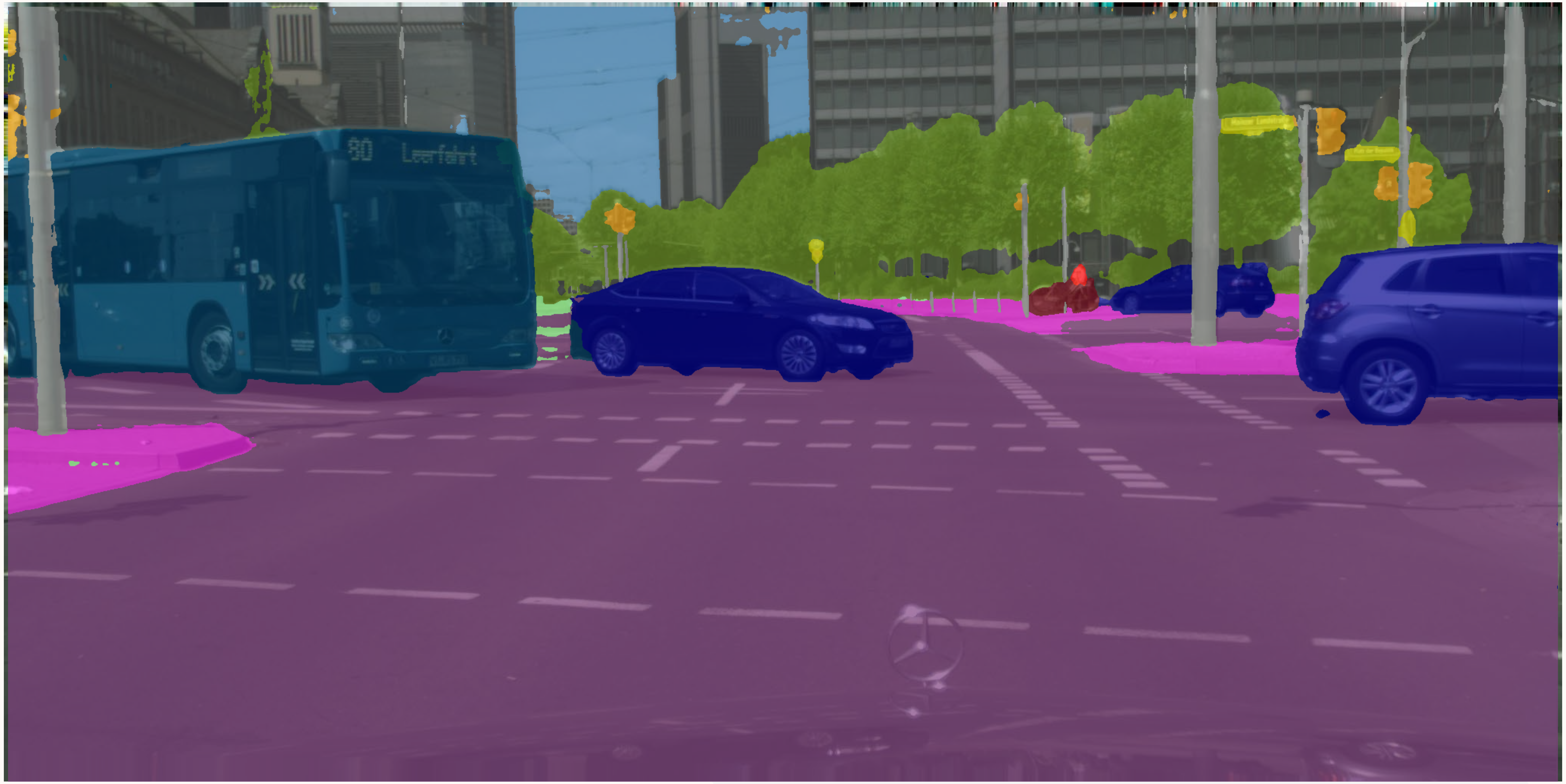
Cityscapes - Results (1/3)

Ground Truth



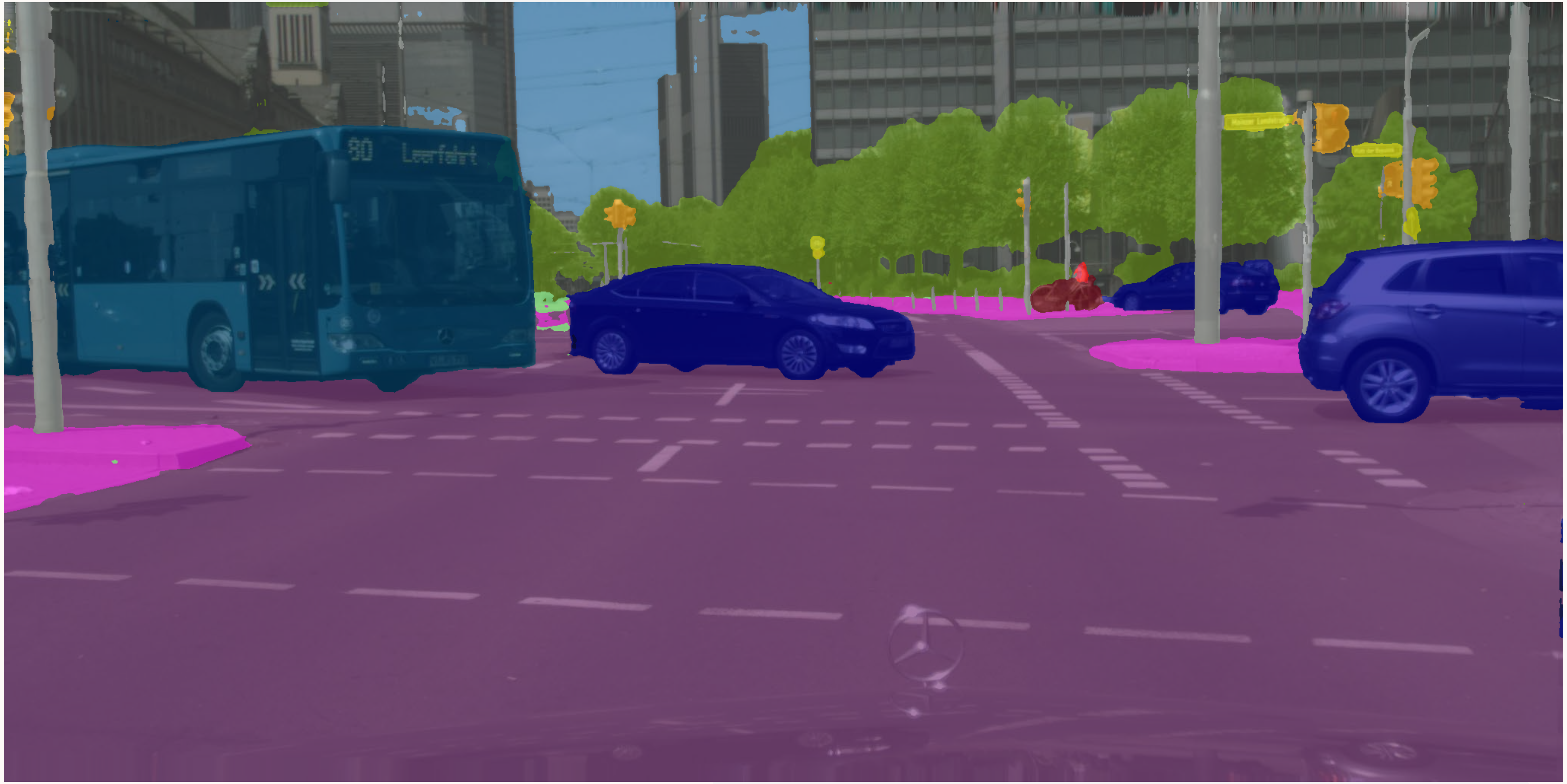
Cityscapes - Results (1/3)

Model trained on IBM Power AC922



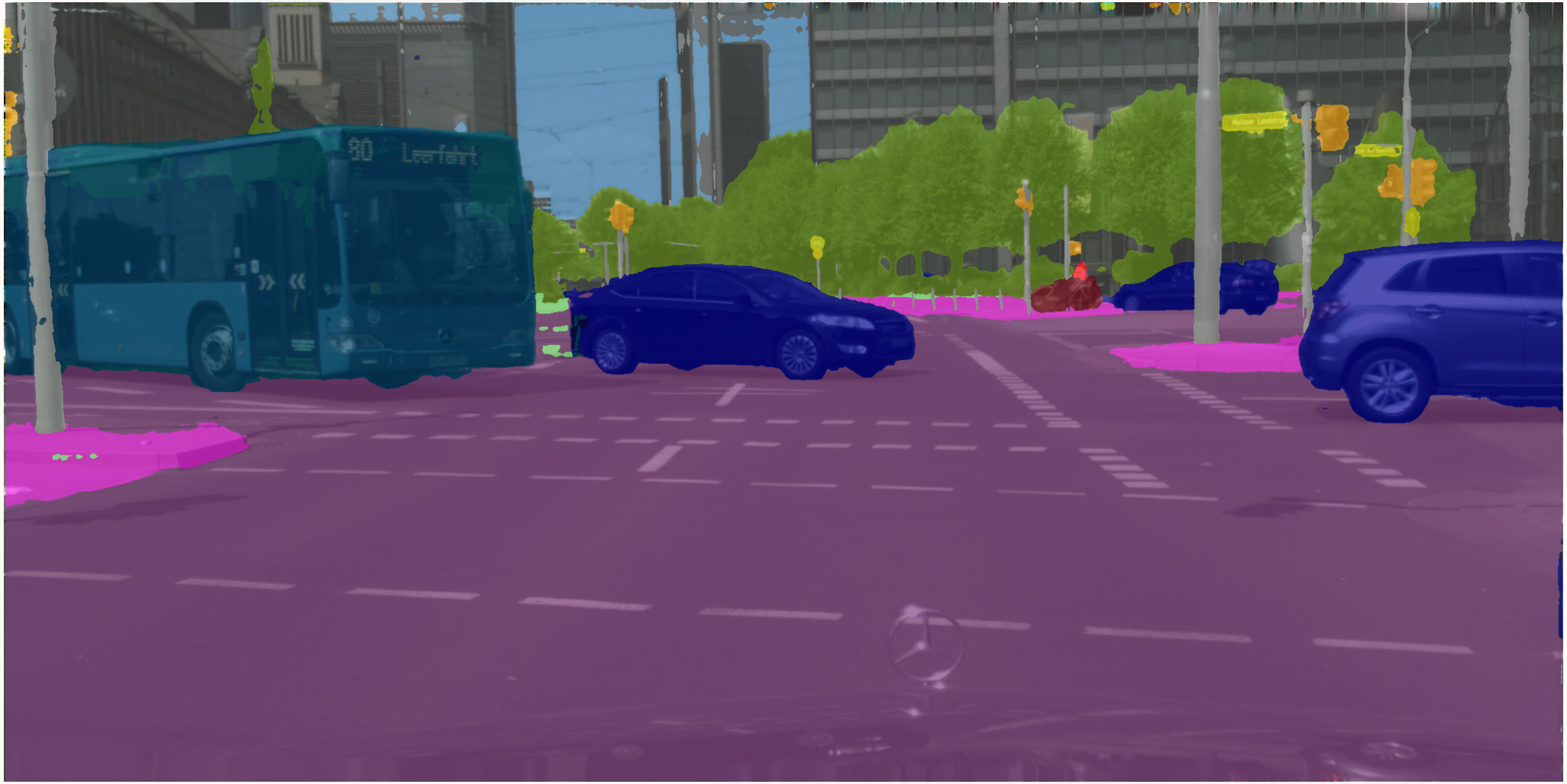
Cityscapes - Results (1/3)

Model trained on NVIDIA DGX Station



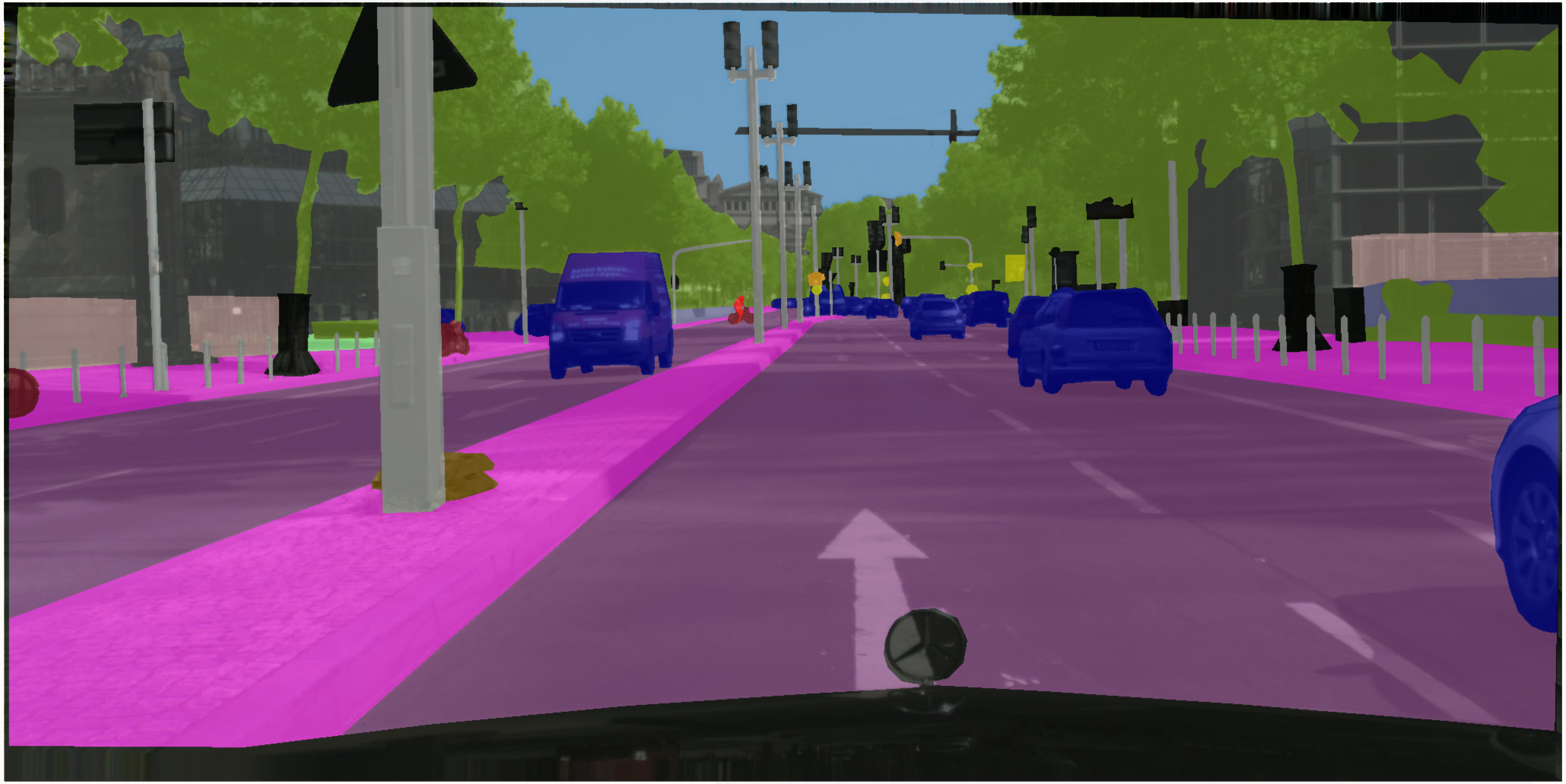
Cityscapes - Results (1/3)

Model trained on AWS p3.8xlarge



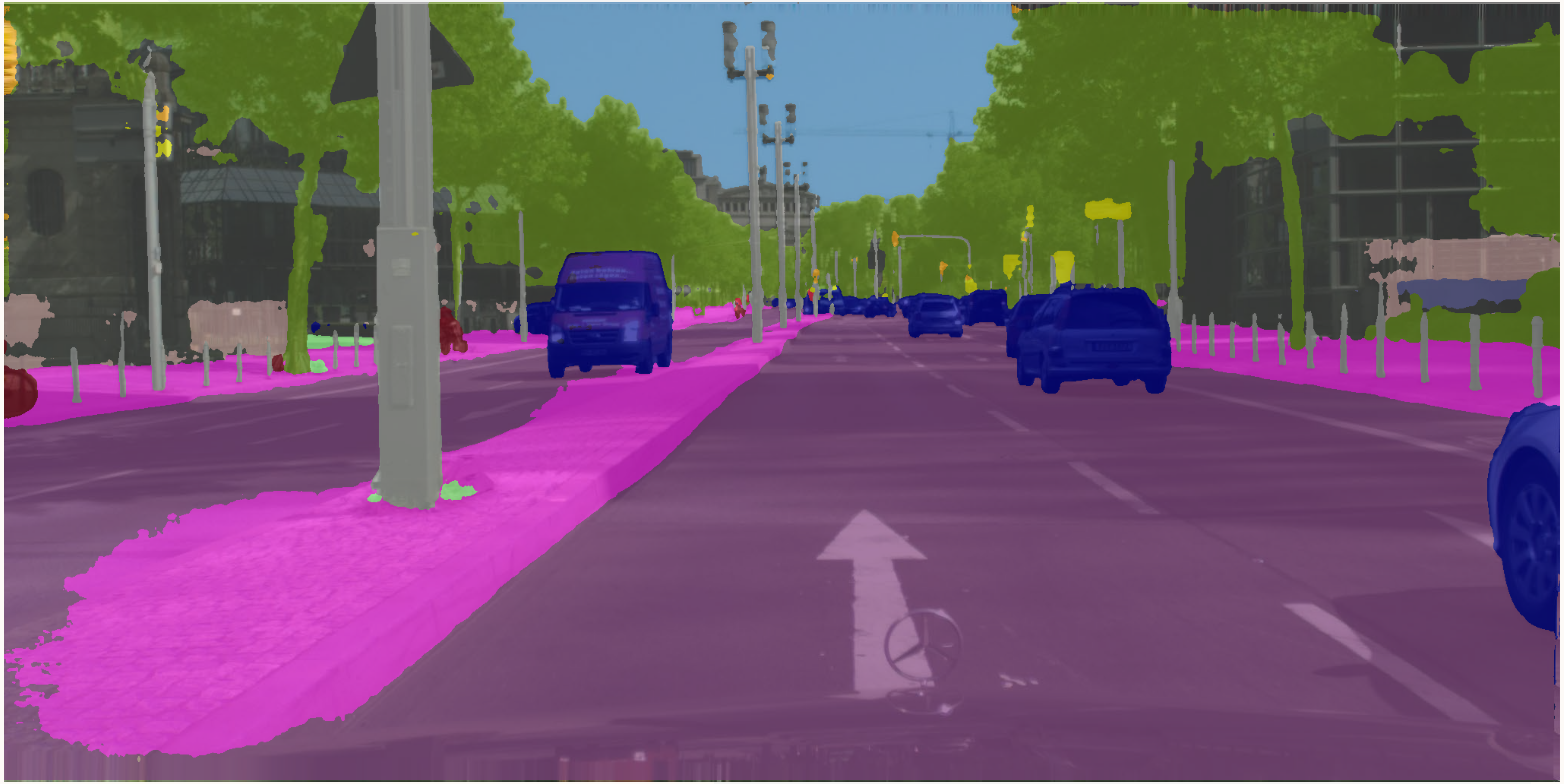
Cityscapes - Results (2/3)

Ground Truth



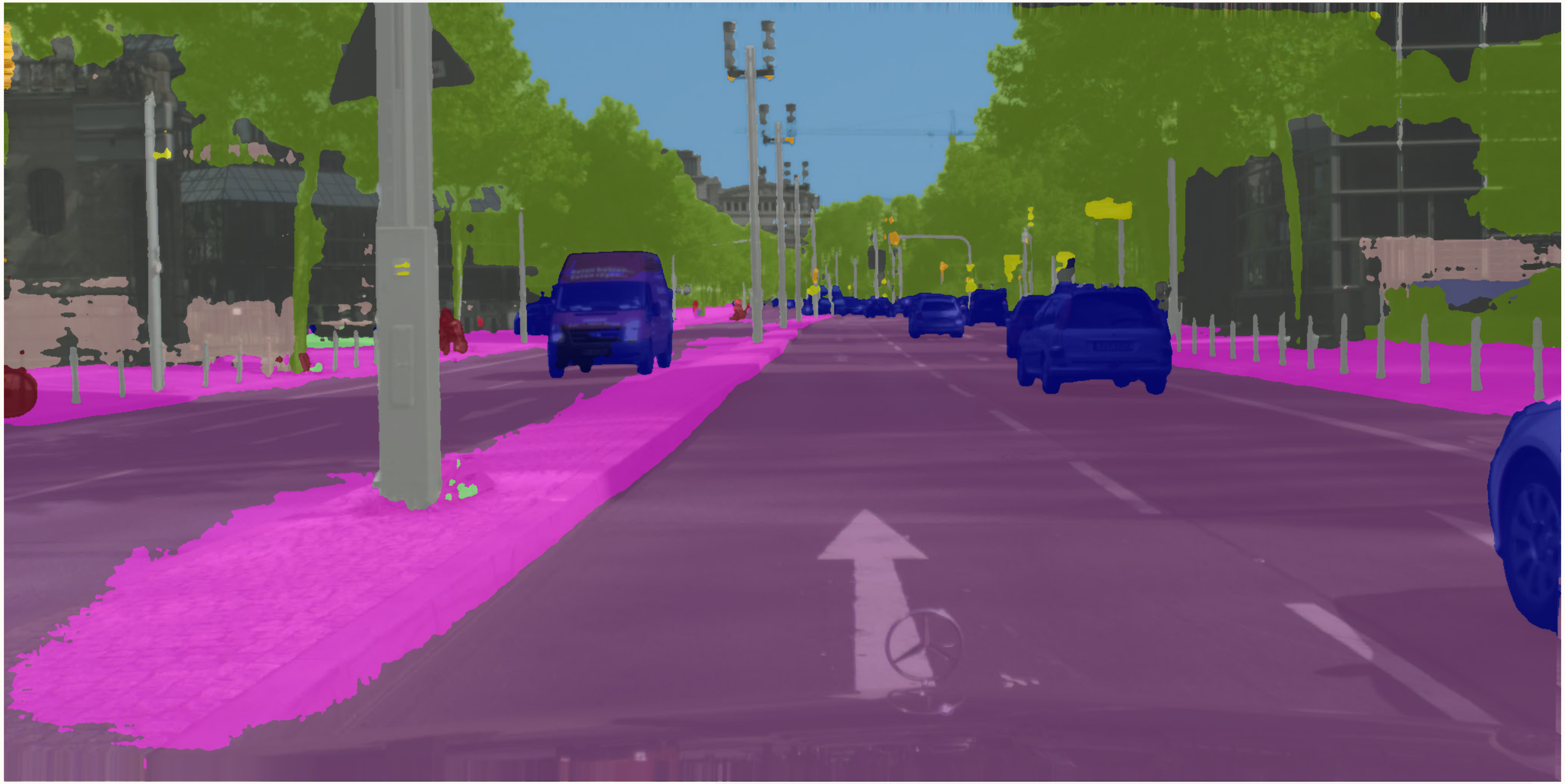
Cityscapes - Results (2/3)

Model trained on IBM Power AC922



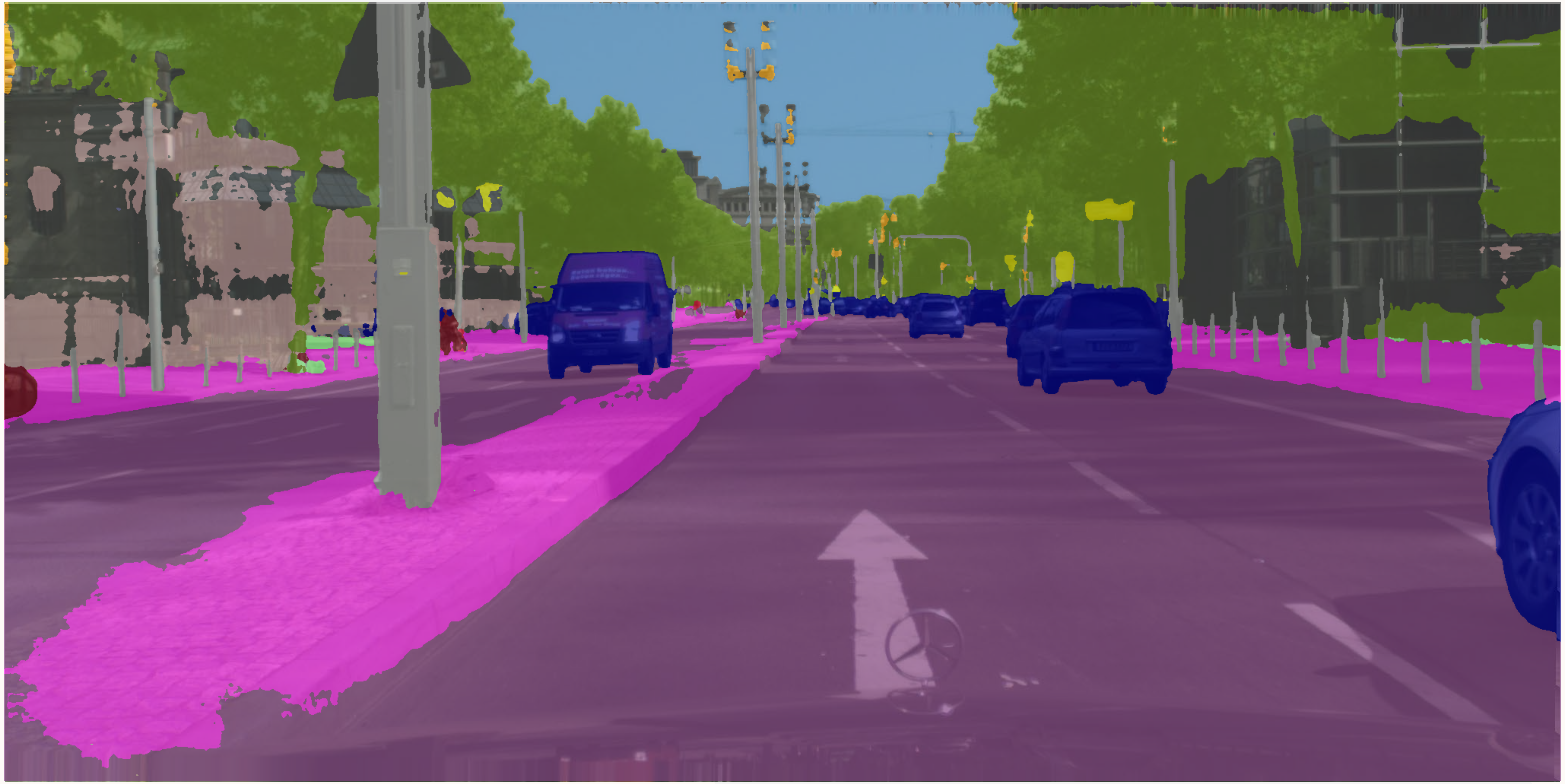
Cityscapes - Results (2/3)

Model trained on NVIDIA DGX Station



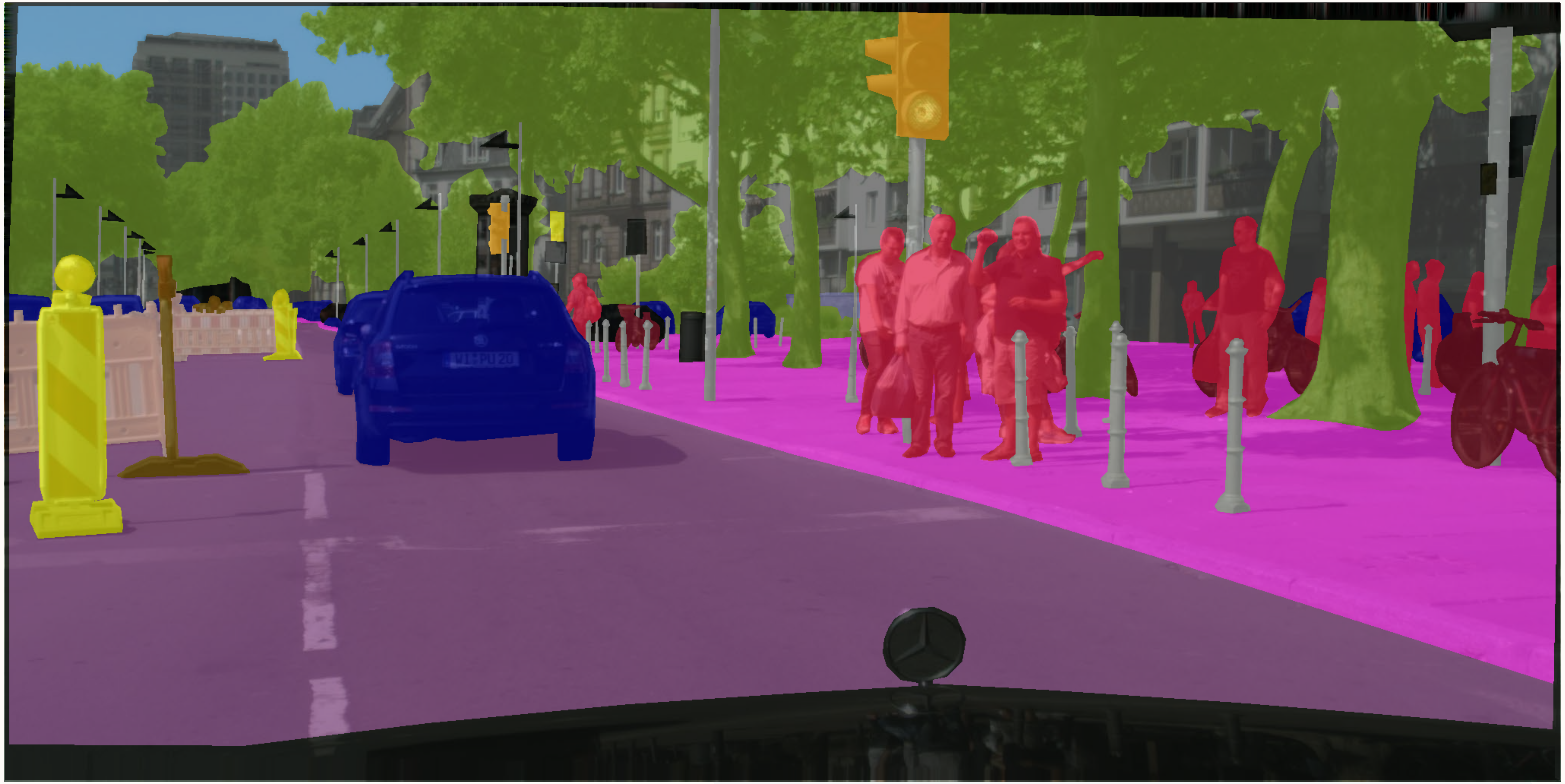
Cityscapes - Results (2/3)

Model trained on AWS p3.8xlarge



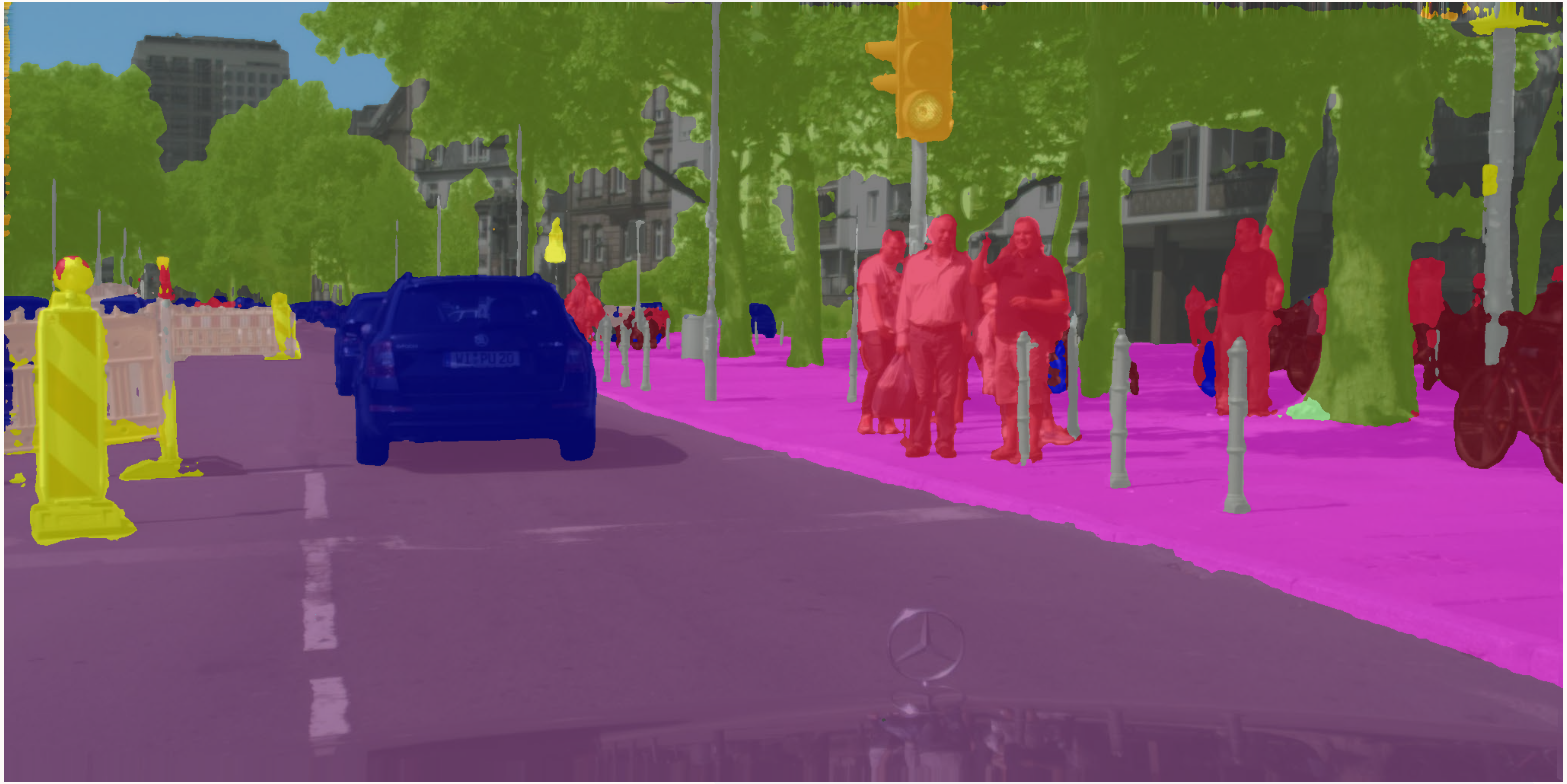
Cityscapes - Results (3/3)

Ground Truth



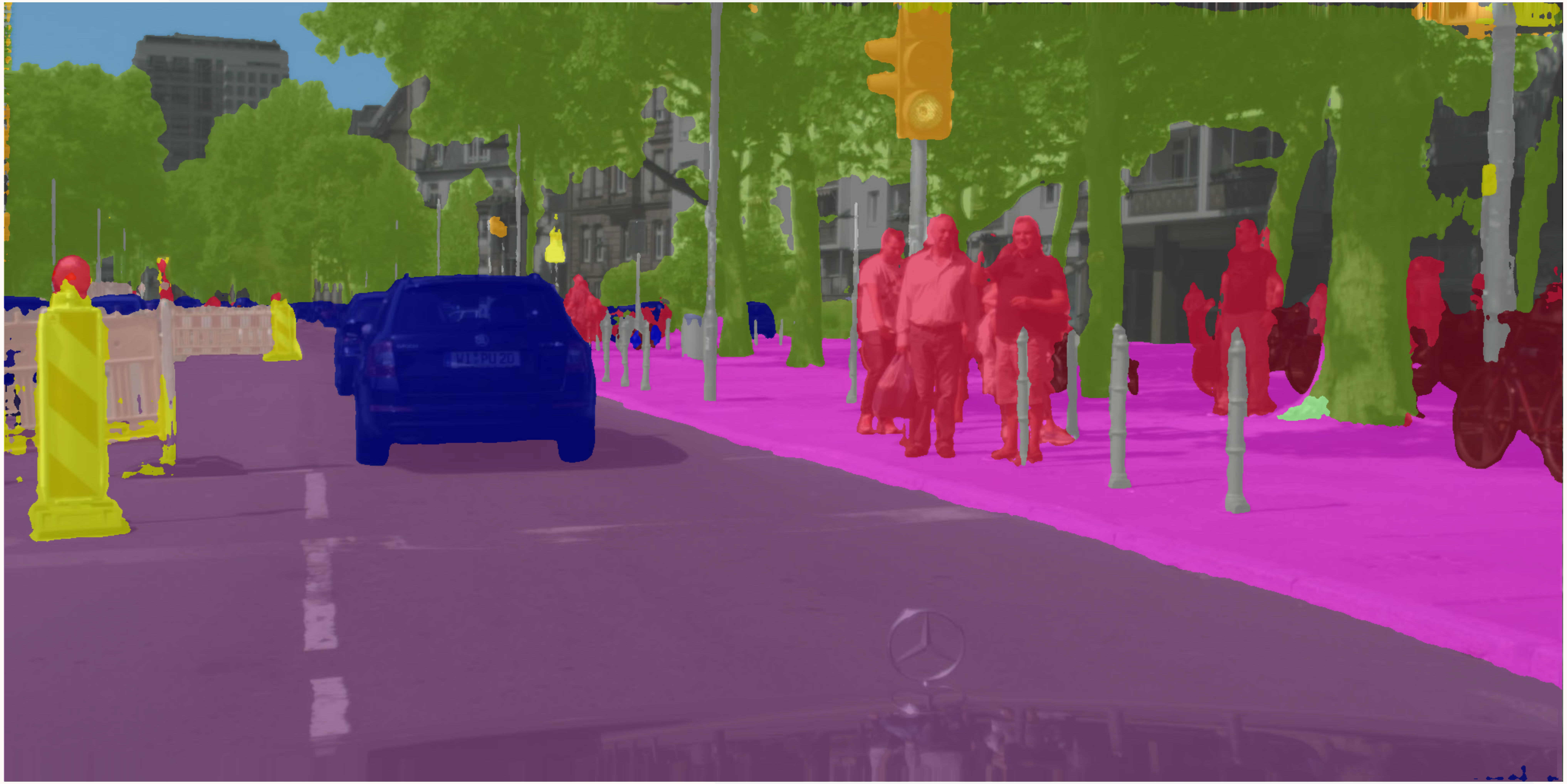
Cityscapes - Results (3/3)

Model trained on IBM Power AC922



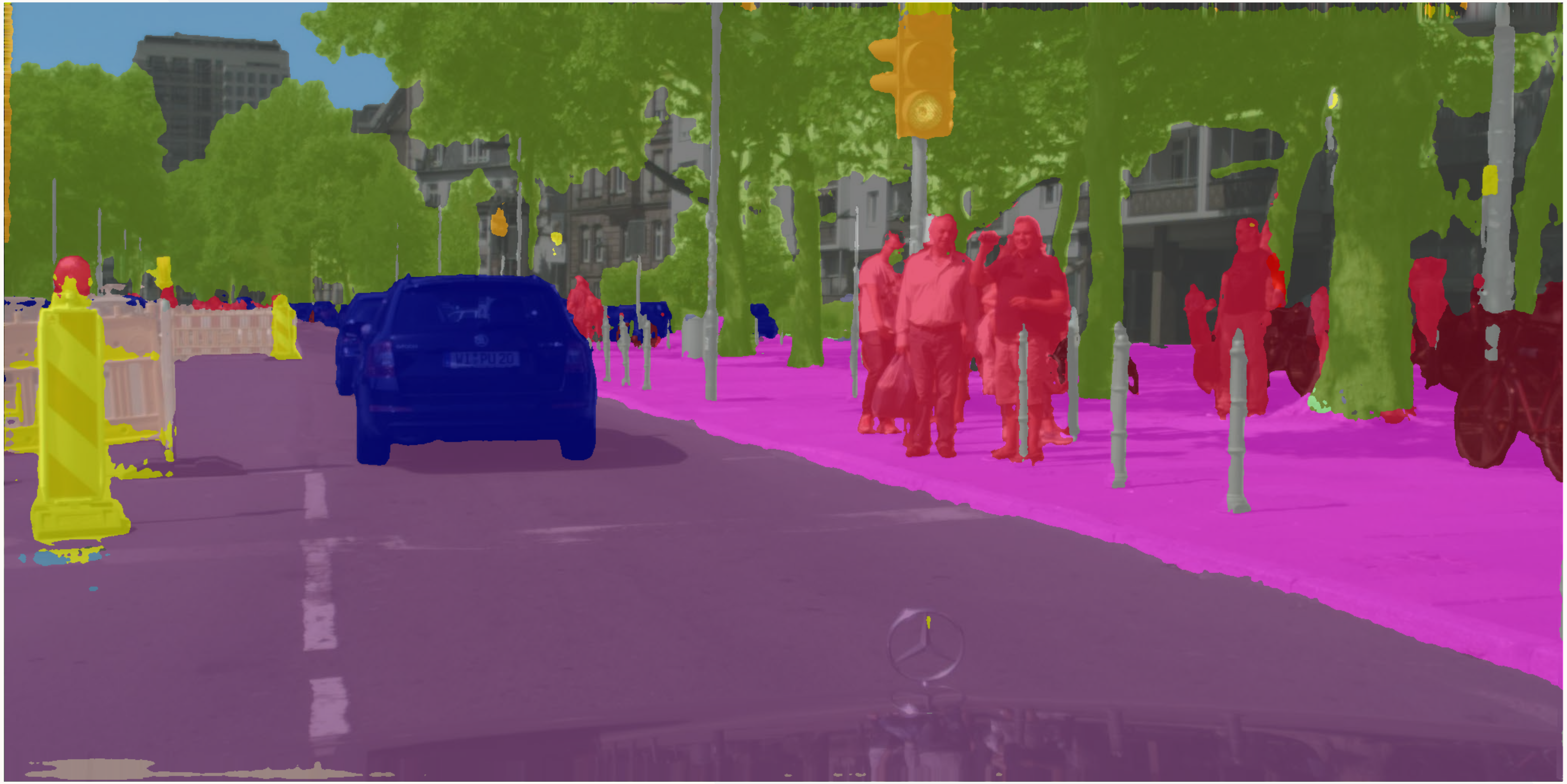
Cityscapes - Results (3/3)

Model trained on NVIDIA DGX Station



Cityscapes - Results (3/3)

Model trained on AWS p3.8xlarge

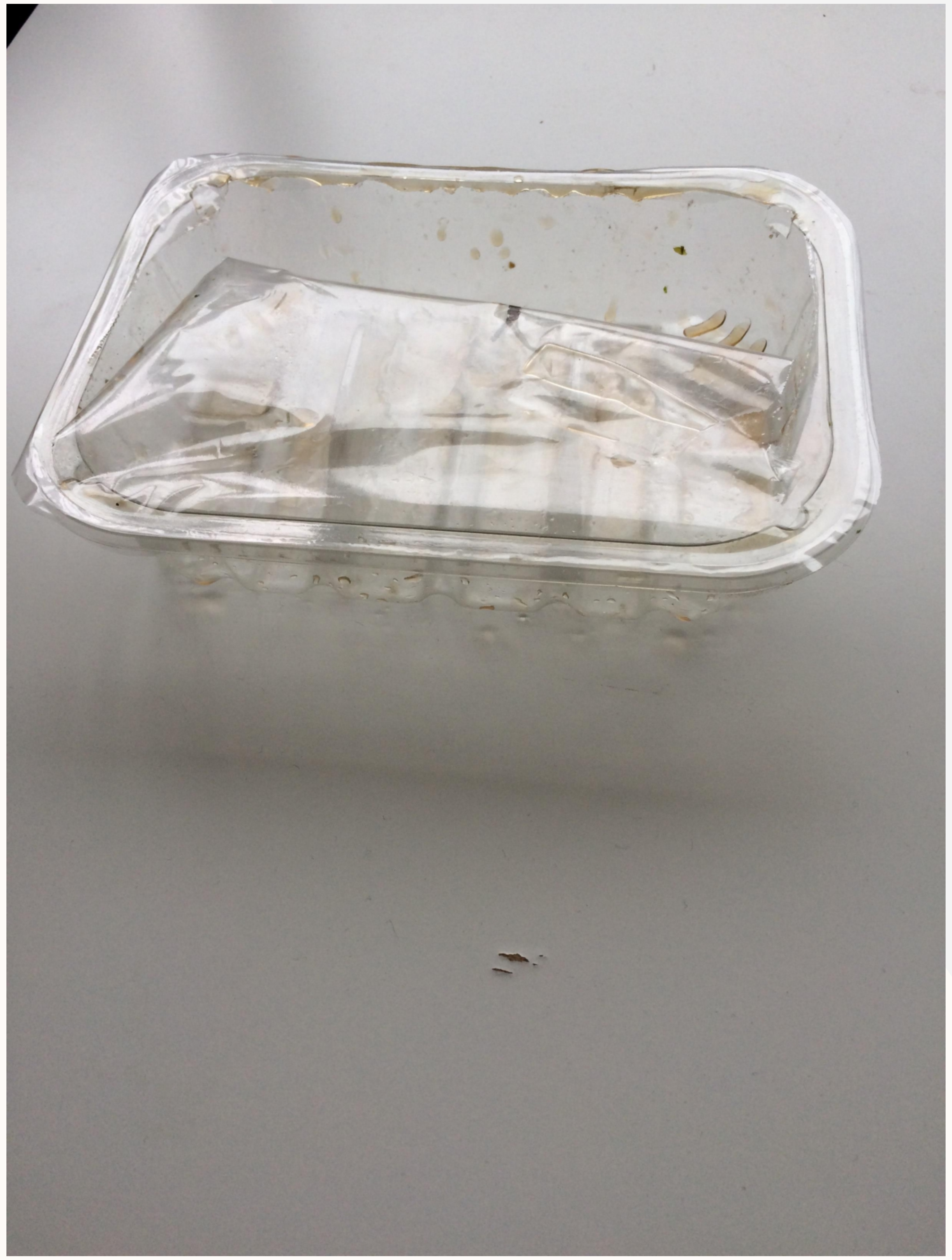


Use Case - Waste Segmentation

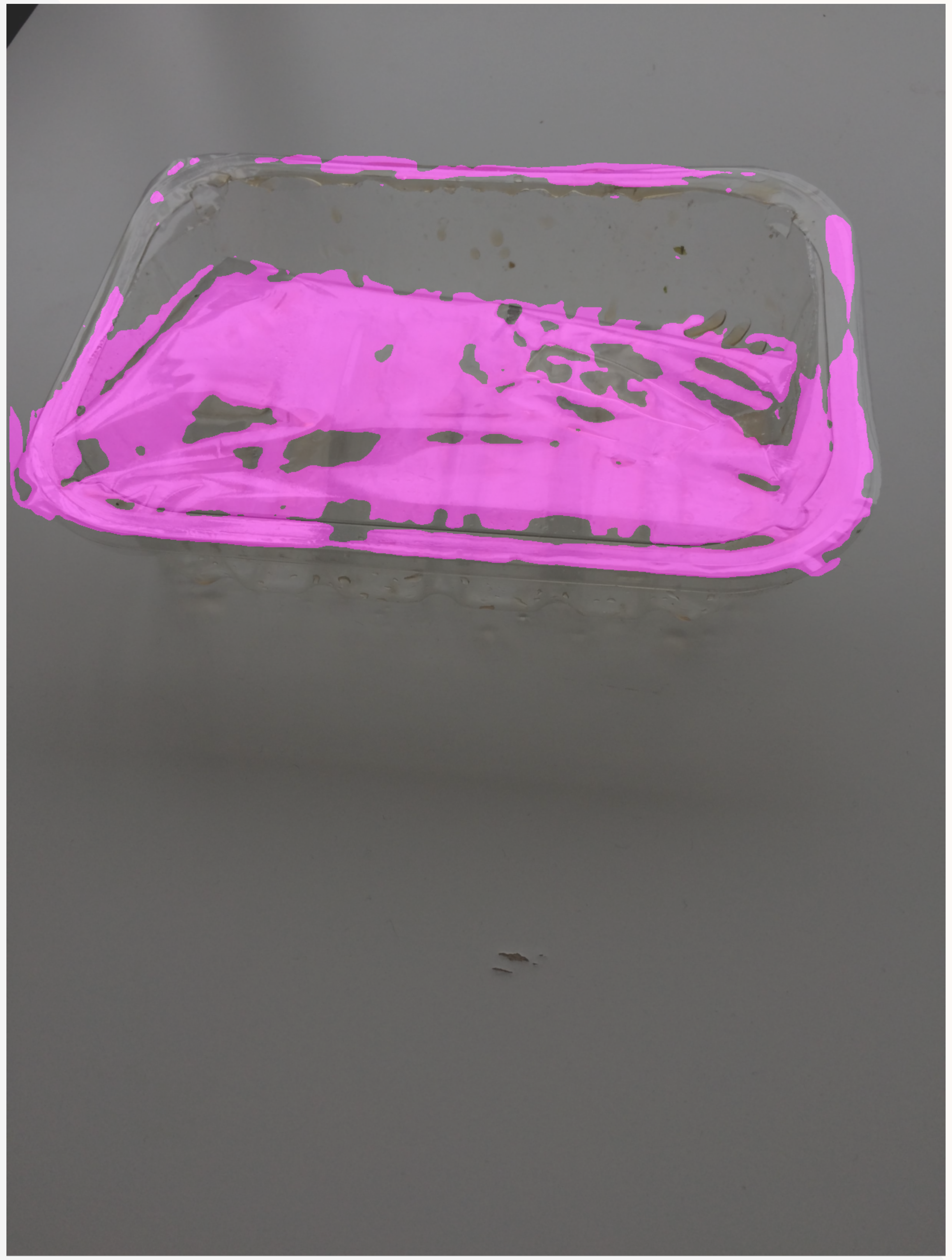
- Humans have been littering the Earth from the bottom of Mariana trench to Mount Everest. Every minute, at least 15 tonnes of plastic waste leak into the ocean, that is equivalent to the capacity of one garbage truck.
- We believe AI has an important role to play in this issue.
- One way to achieve automatic waste segmentation is using the semantic segmentation technology.
- We used the TACO dataset for training our waste segmentation model based on the G-SCNN architecture.
- The dataset consists of 715 images and 2152 annotations, labeled in 60 categories of litter.
- We trained the dataset exclusively on the IBM Power AC922 as it achieved the best performance in our benchmarks.

Waste Segmentation - Results (1/5)

Input image

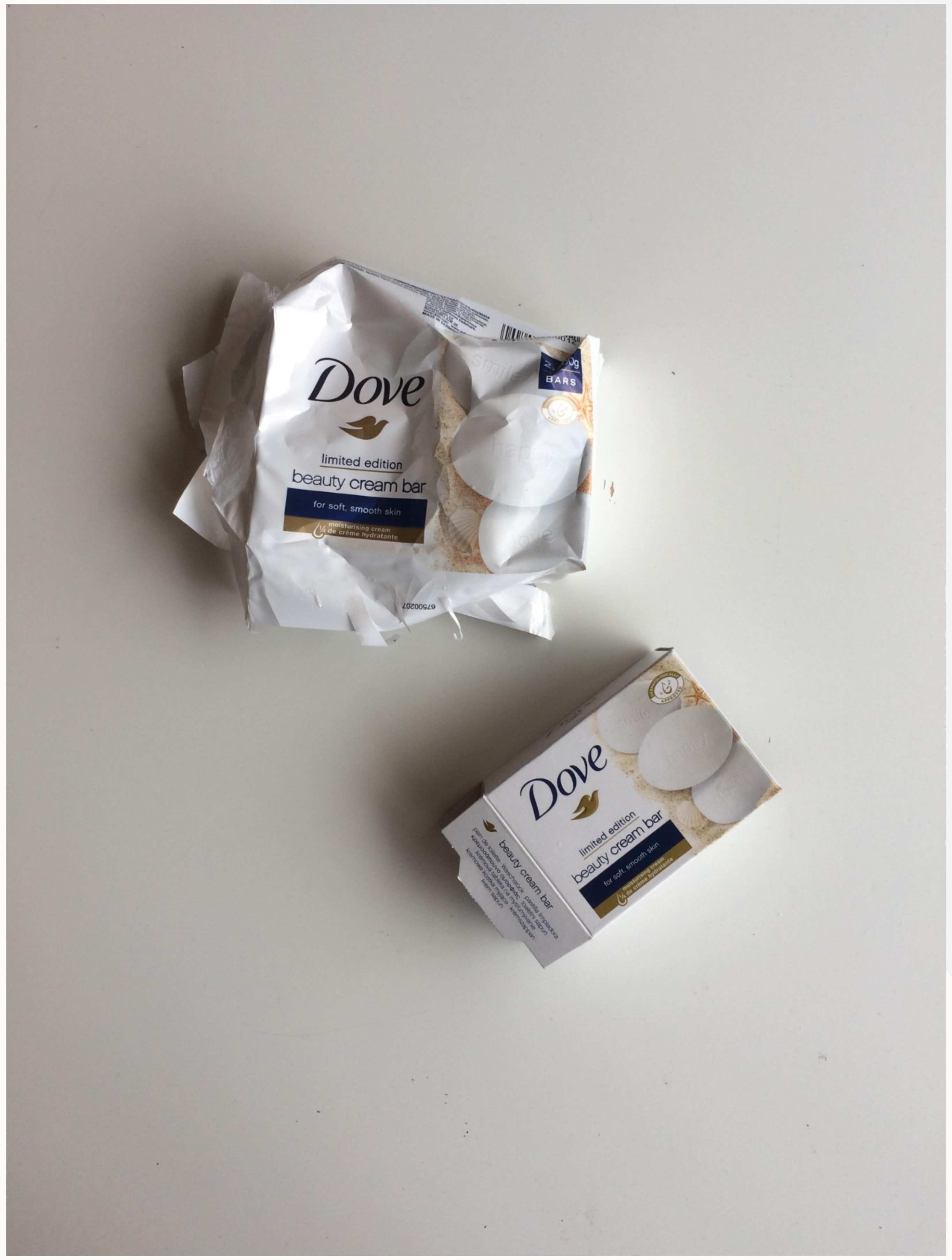


Prediction

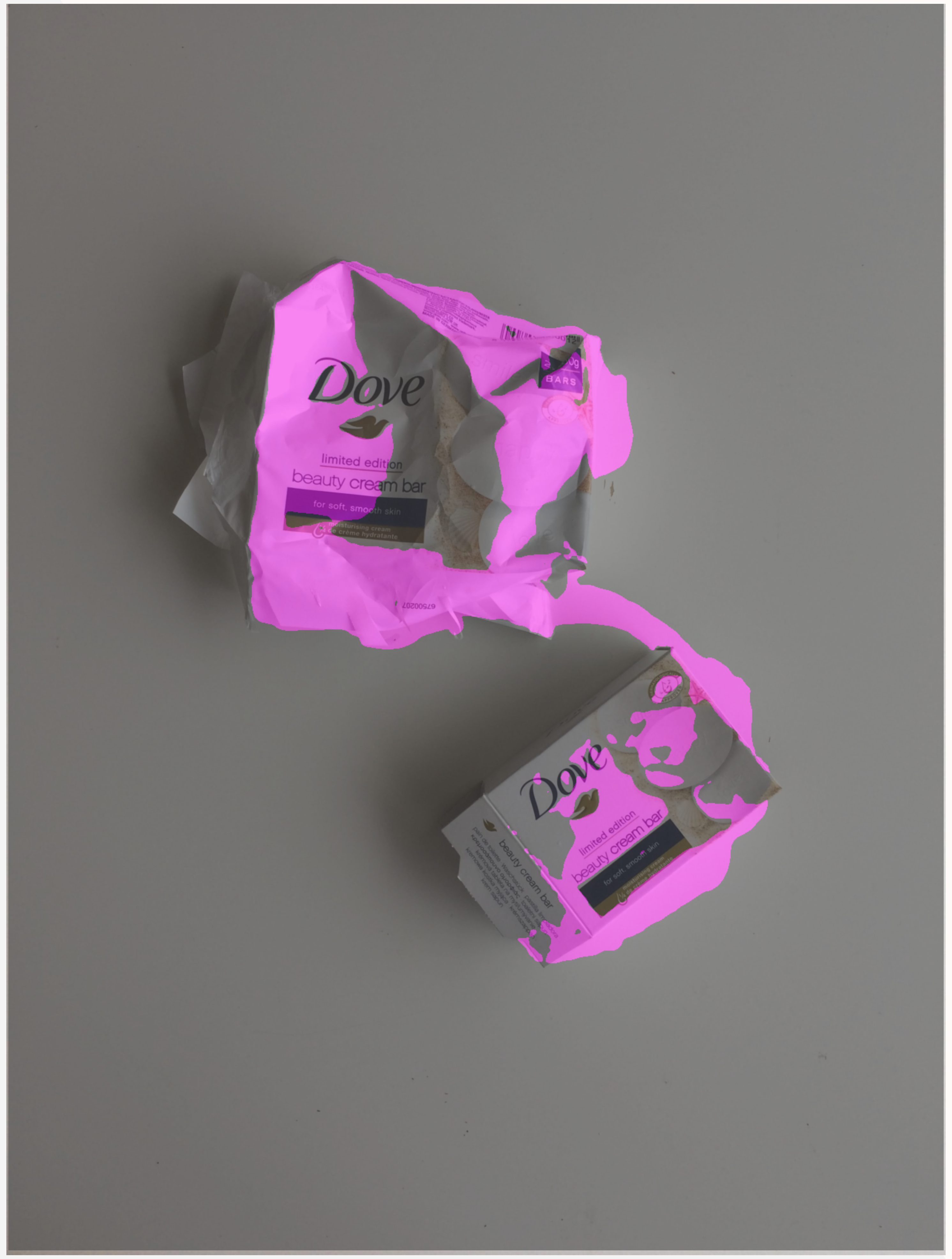


Waste Segmentation - Results (2/5)

Input image



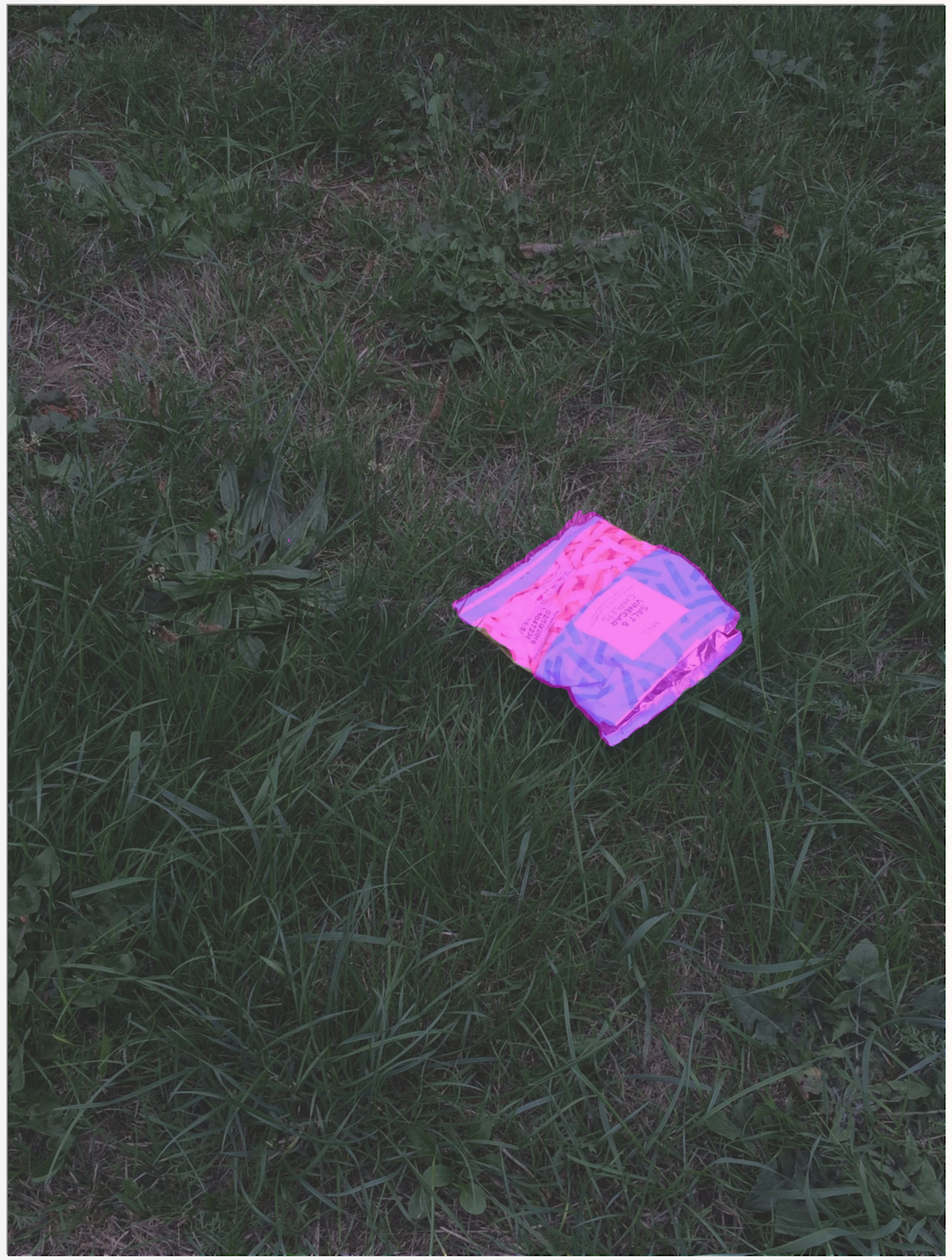
Prediction



Waste Segmentation - Results (3/5)

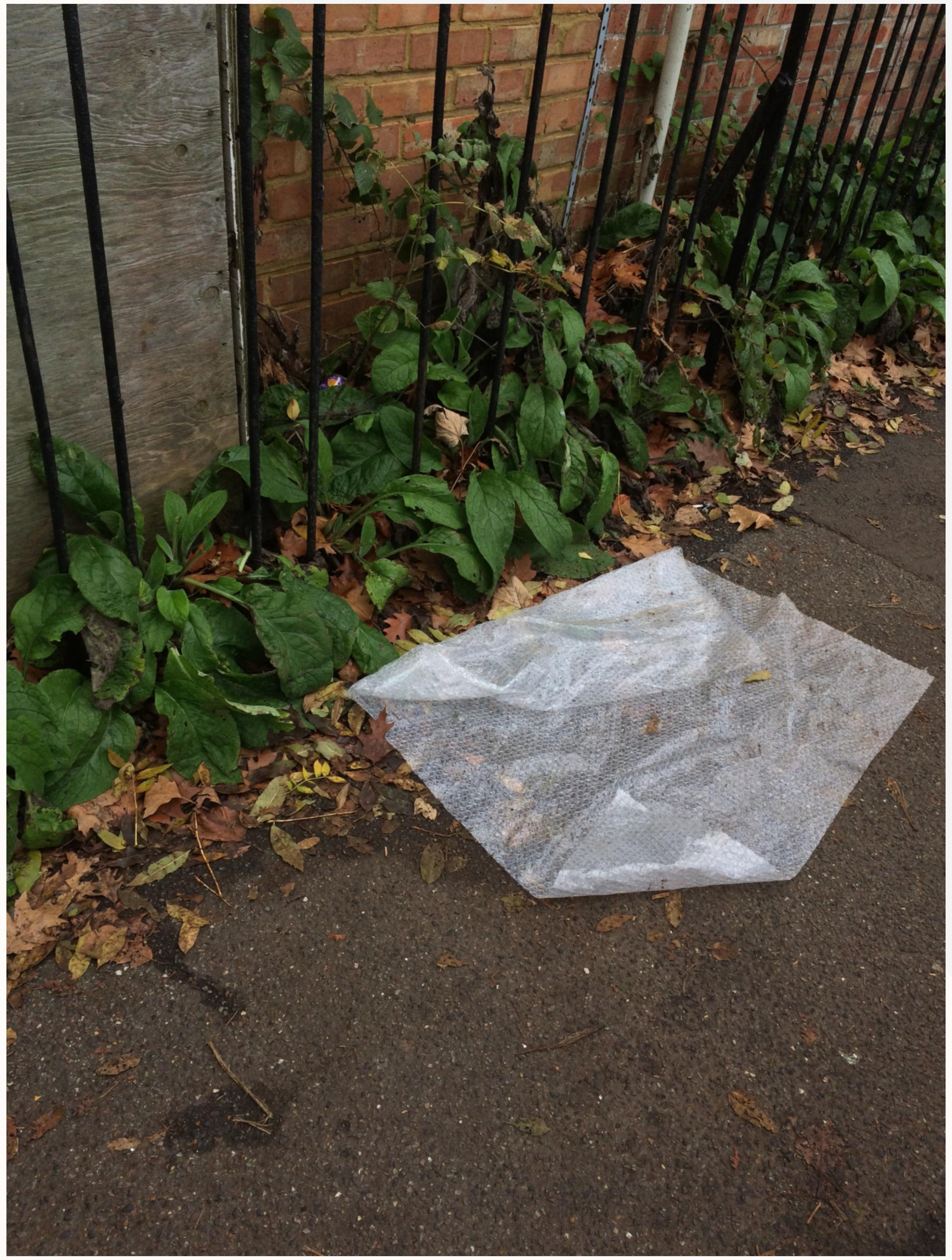
Input image

Prediction

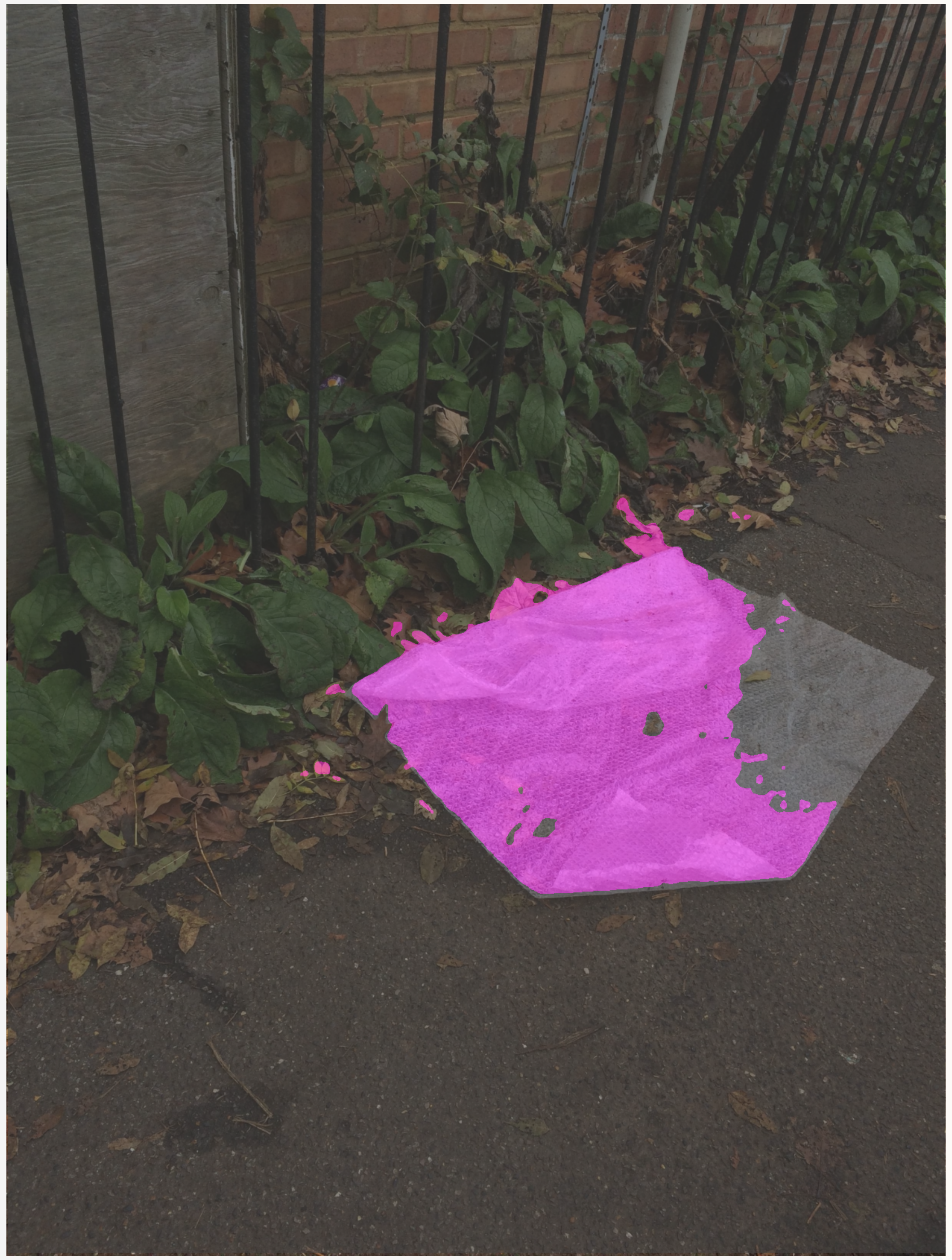


Waste Segmentation - Results (4/5)

Input image

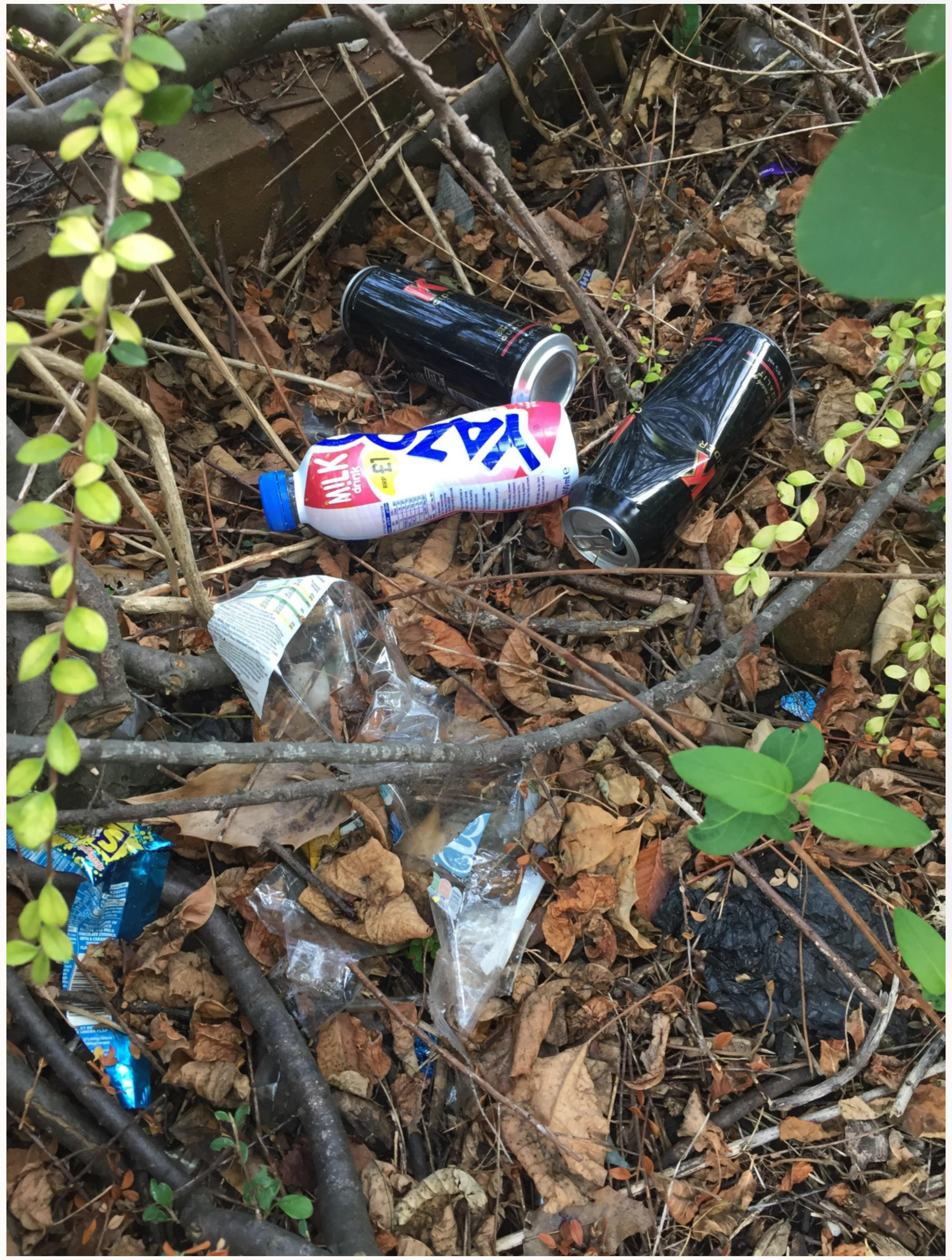


Prediction

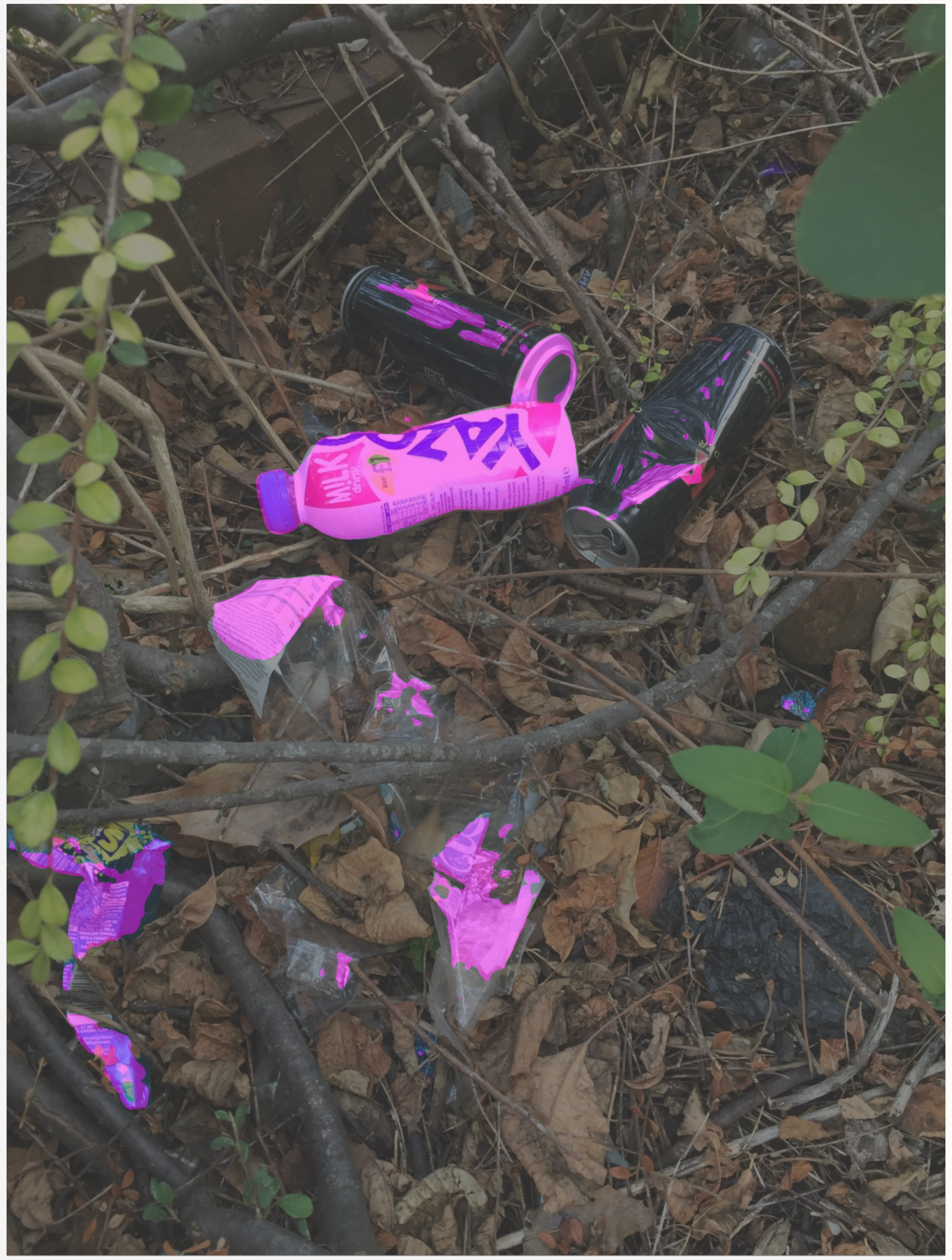


Waste Segmentation - Results (5/5)

Input image



Prediction



Conclusions

- IBM Power AC922 is significantly faster than NVIDIA DGX Station and the AWS p3.8xlarge instance type in such computationally demanding tasks as semantic segmentation.
- Large Model Support enables us to train the model with a larger batch size and input image dimensions producing better overall results.
- IBM's Large Model Support technology has less overhead when used with the IBM Power AC922 hardware, leading to more GPU utilisation and faster training time.
- IBM Power AC922 satisfies the hardware requirements for training on complex tasks such as automatic waste segmentation.

Acknowledgments

We would like to thank our partners L3C (https://l3c.cloud) for providing the IBM Power AC922 testbed used in this study, the creators of G-SCNN architecture for the semantic segmentation architecture and the code used in the benchmarks, and the creators of the Cityscapes and TACO datasets.

References

(1) PlantSnap/Imagga: Training the world's largest plant recognition classifier, https://imagga.com/success-stories/plantsnap-case-study.html

(2) Sam Matzek, TensorFlow Large Model Support Case Study with 3D Image Segmentation, IBM Power developer portal, Published on July 27, 2018, https://developer.ibm.com/linuxonpower/2018/07/27/tensorflow-large-model-support-case-study-3d-image-segmentation/

(3) IBM Power AC922 servers and processor chip, https://www.ibm.com/it-infrastructure/power/power9

(4) NVIDIA DGX Workstation, https://www.nvidia.com/en-us/data-center/dgx-station/ and https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/dgx-station/nvidia-dgx-station-datasheet.pdf

(5) AWS p3 instance type, https://aws.amazon.com/ec2/instance-types/p3/

(6) The Cityscapes data set, https://www.cityscapes-dataset.com/

(7) M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, 'The Cityscapes Dataset for Semantic Urban Scene Understanding,' in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, https://www.cityscapes-dataset.com/wordpress/wp-content/papercite-data/pdf/cordts2016cityscapes.pdf

(8) Pedro F. Proena and Pedro Simes, TACO: Trash Annotations in Context Dataset, 2019, http://tacodataset.org

(9) Takikawa, Towaki & Acuna, David & Jampani, Varun & Fidler, Sanja, Gated-SCNN: Gated Shape CNNs for Semantic Segmentation, 2019, https://arxiv.org/pdf/1907.05740.pdf

Contact Us

Bulgaria Office
47A Cherni Vrah Blvd., floor 4
1407 Sofia
Bulgaria

South Korea Office
Seocho-gu, Gangnam-dero 369,
A+ Asset Tower, 12 fl,
Seoul 06621,
South Korea

Email:

sales@imagga.com

Web:

https://imagga.com