EPFL

Pix2Seq-D: Diffusion models for Video Panoptic Segmentation

DLAV2023 - Group 30
Final presentation

Tommaso Martorella

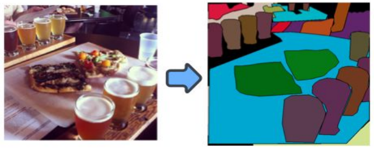
Roberto Minini

Roberto Ceraolo

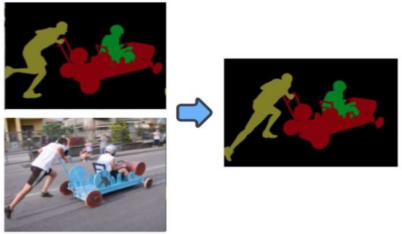
Pix2Seq-D: A Generalist Framework for Panoptic Segmentation of Images and Videos

Chen, Li, Saxena et al., Google Research, Brain Team

Denoising given previous mask and current image

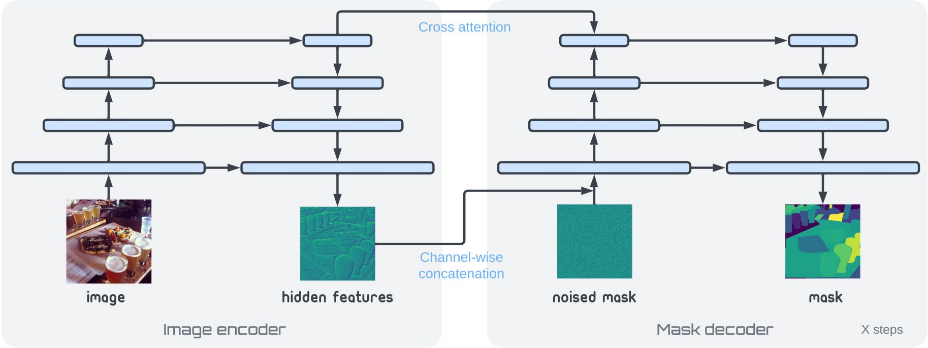


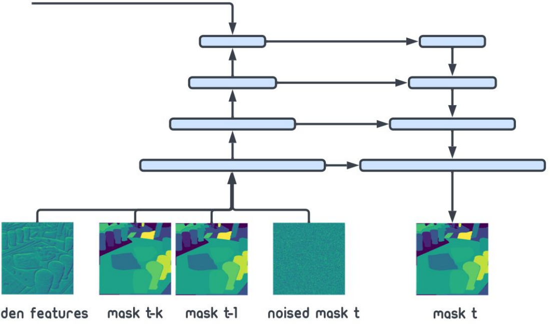
$p(\boldsymbol{m}|\boldsymbol{x})$



$p(\boldsymbol{m}_t|\boldsymbol{x}_t, \boldsymbol{m}_{t-1})$

Using Feature Pyramid Networks to extract hidden features





Extensions to videos

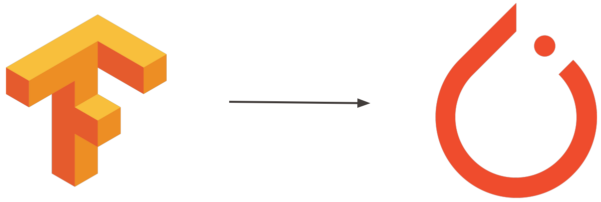
EPFL

Our contributions and results

Video Panoptic Segmentation

Conversion to PyTorch

- Re-wrote the codebase from Tensorflow to PyTorch
- We hope it can be a useful addition to the community overall



Extension to Video Panoptic Segmentation

- Built the module to extend the architecture to the task of Video Panoptic Segmentation
- Pre-trained the model on Cityscapes, trained on KITTI-STEP

Semantic segmentation

Instance segmentation

Time tracking

To have a fair comparison, we also trained the SOTA architecture on KITTI-STEP: **Video K-Net**

Video K-Net: A Simple, Strong, and Unified Baseline for Video Segmentation

Our contributions and results

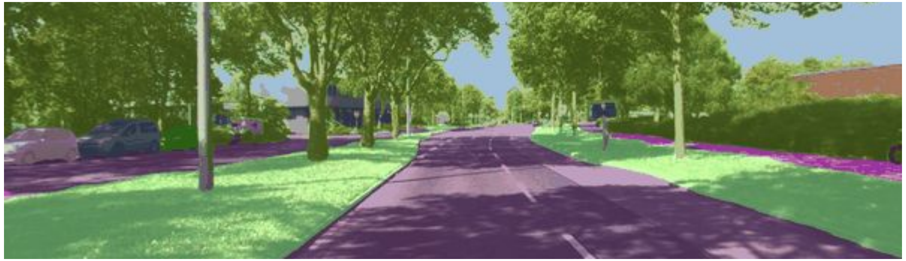
Video Panoptic Segmentation

The results are not satisfying yet, but we:

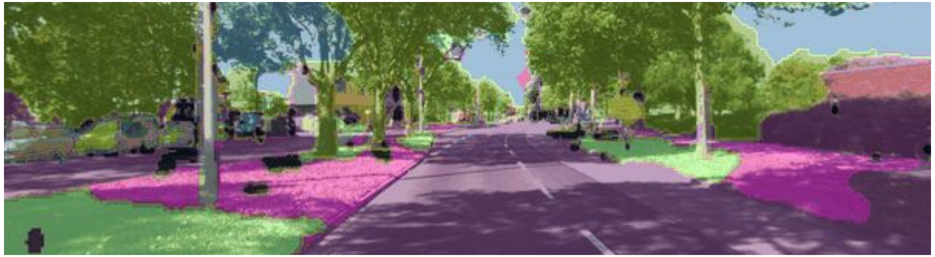
- trained for very few epochs, only on 1 GPU
- did not fine-tune hyperparameters

Overall, we believe:
Diffusion is a direction worth exploring for VPS
BUT we need more training epochs

Our inference with Video K-Net



Our inference with Pix2Seq-D



Next steps

We plan on carrying on working on this project

Training on more epochs

Pretraining on larger datasets, e.g. Mapillary
Fine-tuning on larger datasets, e.g. Cityscapes-VPS

Estimating segmentation as a variation of the previous frame

Goal: improve efficiency & performance

Thank you!