Large-scale semantic segmentation through multi-resolution processing and selective pseudolabeling

Vision for all Seasons @ CVPR2022

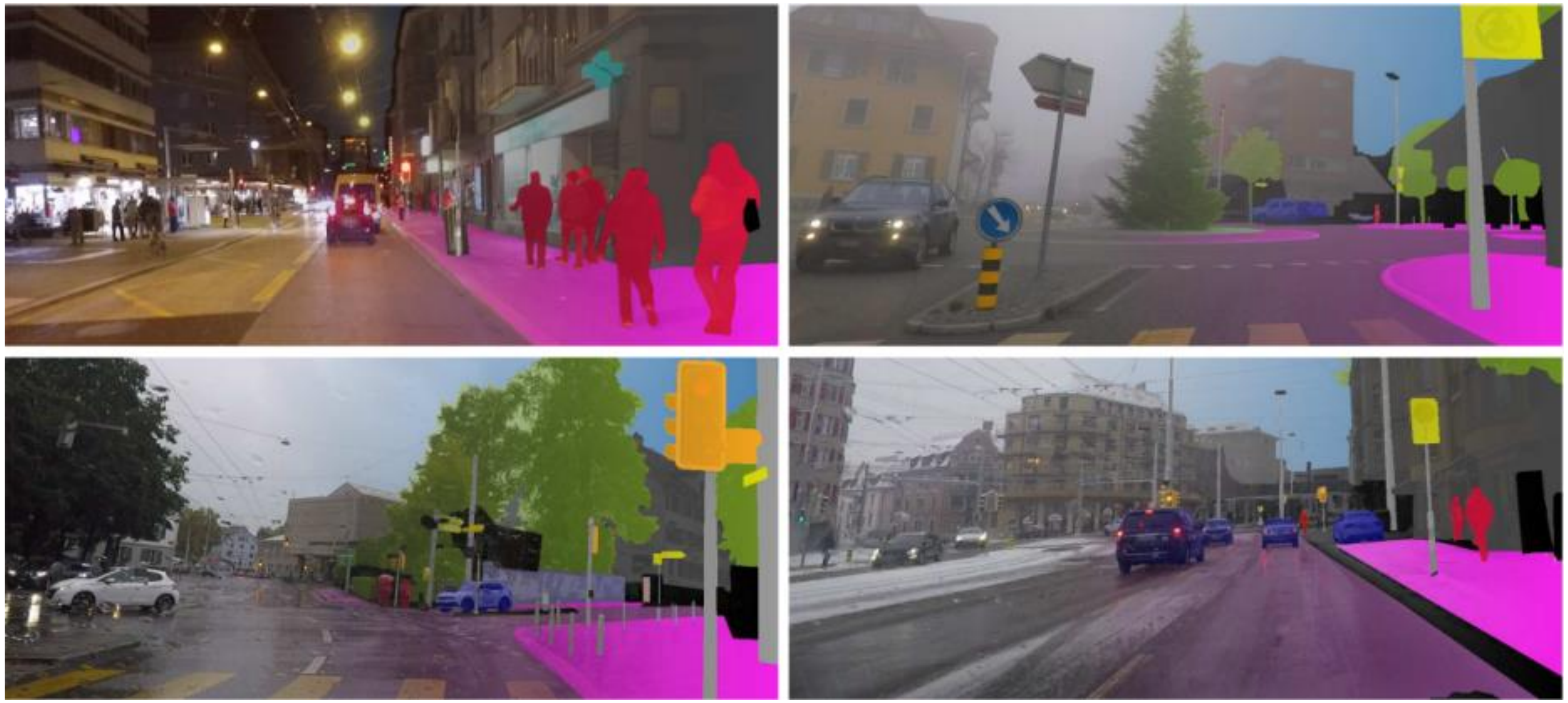
Matej Grcić
matej.grcic@fer.hr
Faculty of Electrical Engineering and Computing
University of Zagreb
Croatia

Outline

- Part I: A SwiftNet for 2020s
  - Convolutions are still competitive.
- Part II: Abundant supervision walks the walk
  - Select pseudolabels for the last mile.
- Part III: Exploiting uncertainty in semantic segmentation
  - Why can't we get better in AUIoU?

ACDC Challenge

- ACDC Dataset - 1600 train, 406 val and 2000 test images



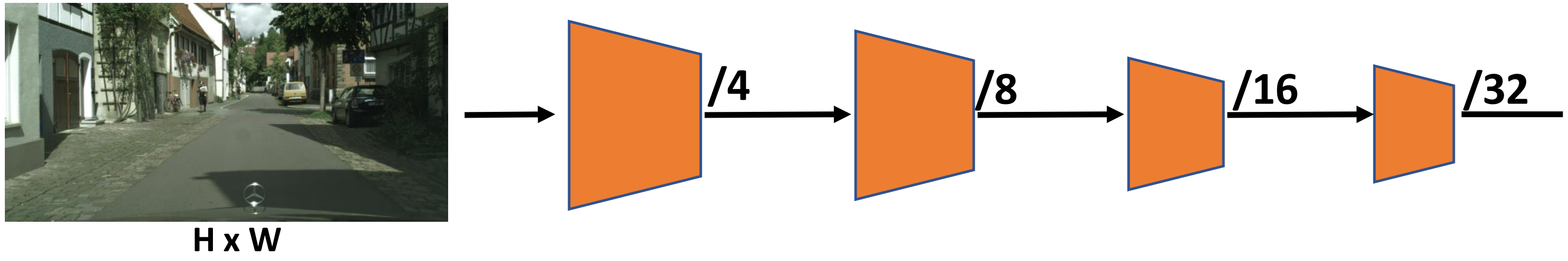
[sakaridis21iccv]

A SwiftNet for 2020s

Convolutions are still competitive on large images

Faculty of Electrical Engineering and Computing

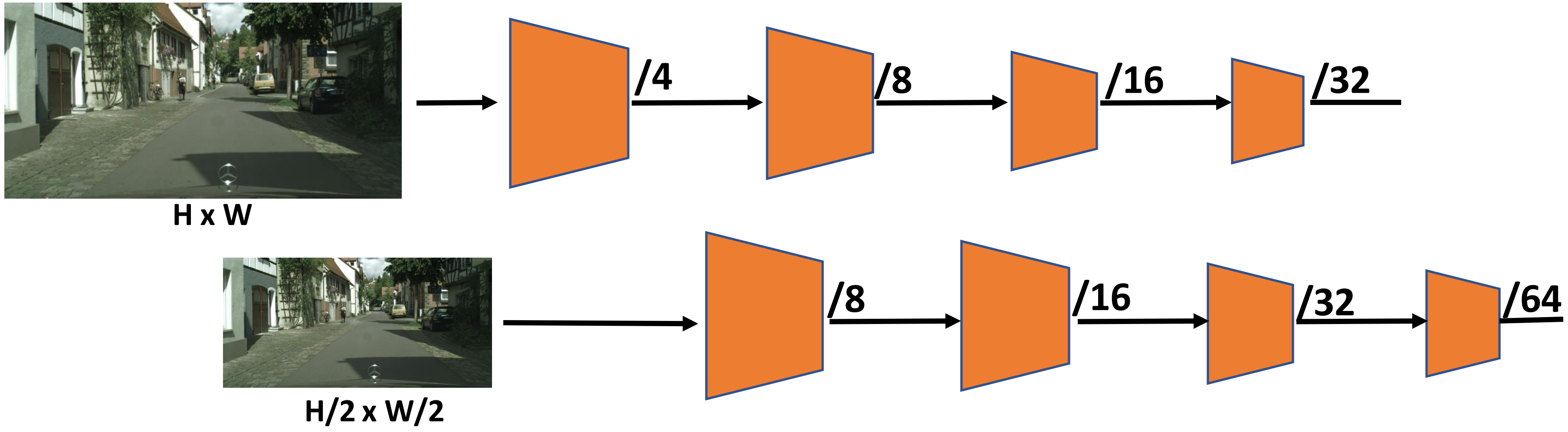
Multiple resolutions + pyramidal fusion



Convolutional backbone
Pretrained on ImageNet

[orsic21pr]

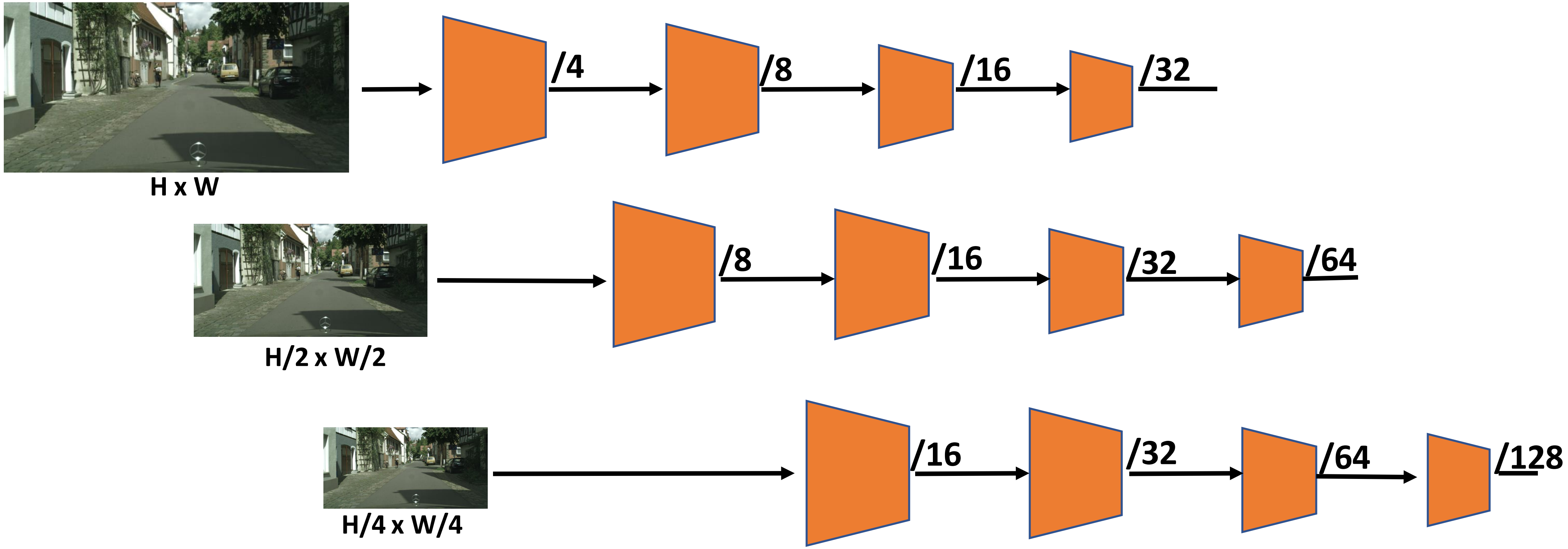
Multiple resolutions + pyramidal fusion



The same instance of the backbone

[orsic21pr]

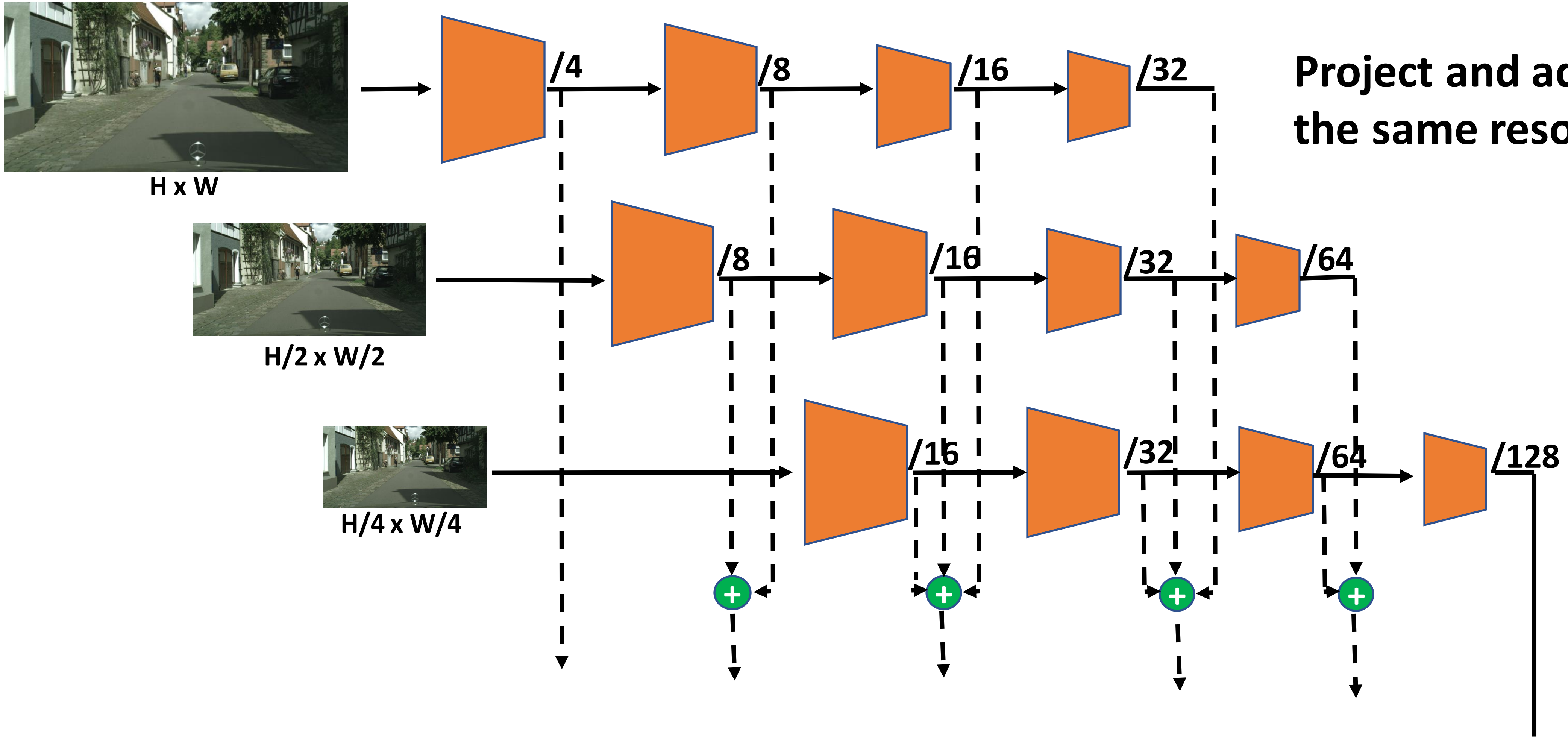
Multiple resolutions + pyramidal fusion



The same instance of the backbone

[orsic21pr]

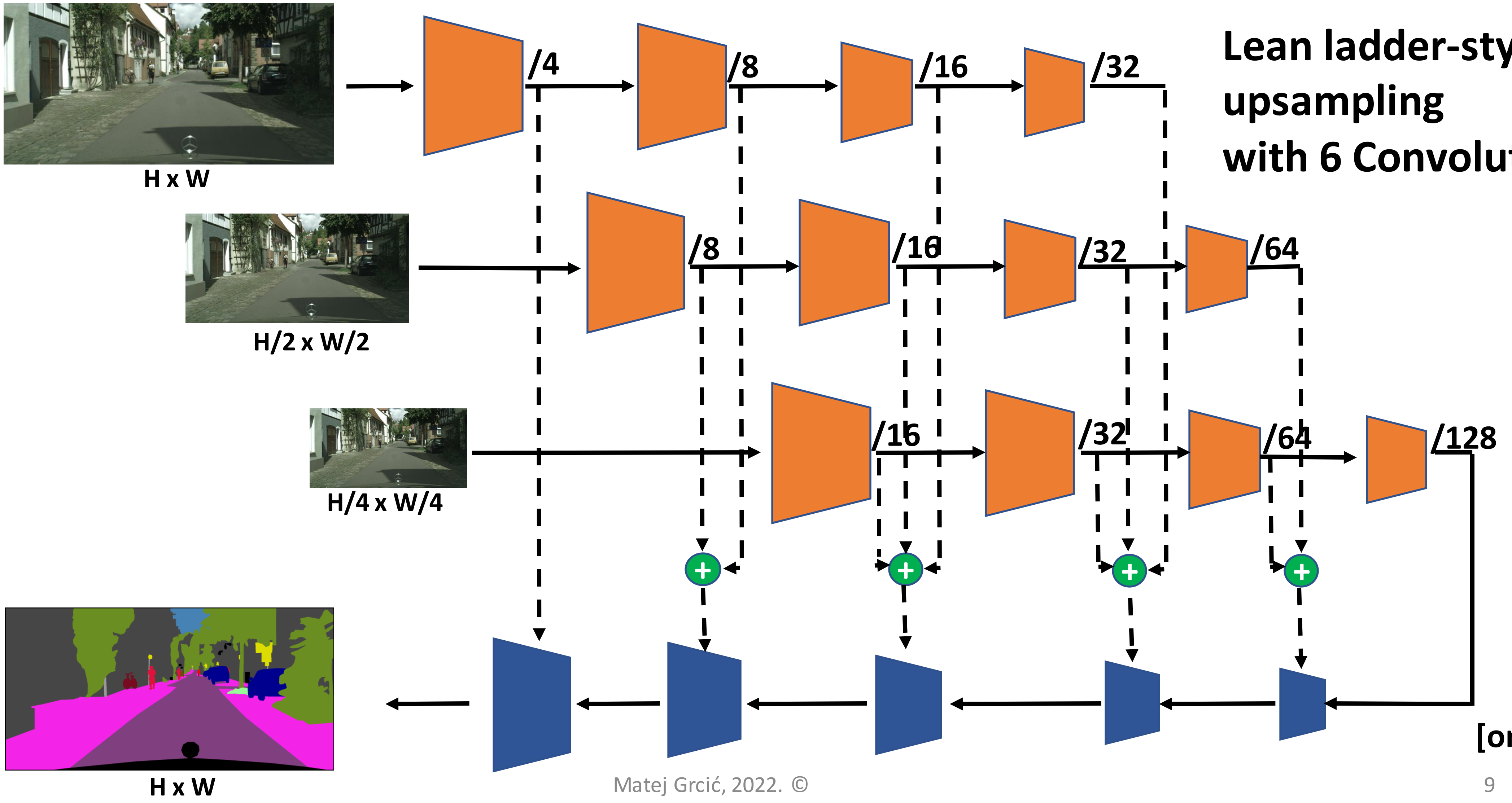
Multiple resolutions + pyramidal fusion



Project and add the same resolutions

[orsic21pr]

Lean ladder-style upsampling

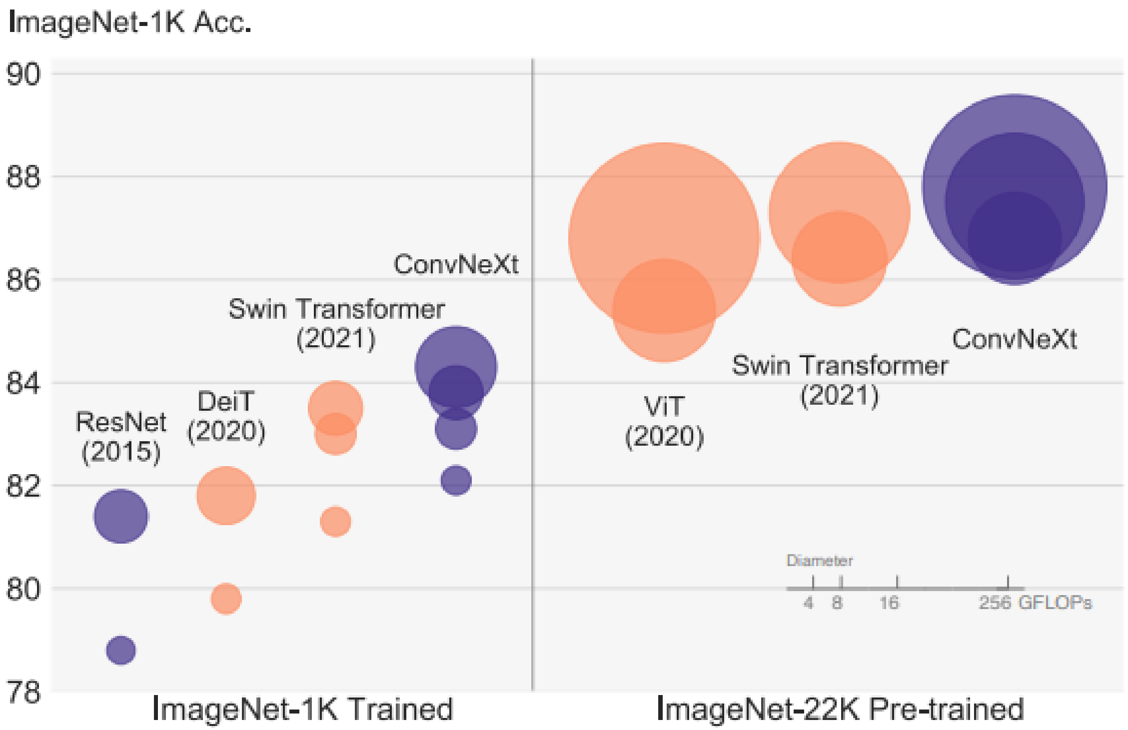


Lean ladder-style upsampling with 6 Convolutions

[orsic21pr]

Backbone: ConvNext-L [liu22cvpr]

- Fully convolutional network
- 87.5% Top1 Acc ImageNet1K
- Macro design of transformers applied to convolutional nets
- Depthwise separable 7x7 convolutions
- BatchNorm -> LayerNorm
- Inverted bottleneck
- Less activations



Mean IoU on the standard datasets

Method	Backbone	Cityscapes val	ADE20k val	GMACs
HRNetV2 + OCR [wang20tpami]	Wide ResNet48	81.6	-	-
HRViT [gu22cvpr]	-	83.2	50.2	-
Swin UperNet [liu21iccv]	SWIN-B	-	51.6	2348.1
Swin UperNet [liu21iccv]	SWIN-L	-	53.5	3183.3
Seg-B [strudel21iccv]	DeiT-B	80.5	48.5	822.6
Seg-L [strudel21iccv]	ViT-L	81.3	53.6	2697.7
SwiftNet pyramid [orsic21pr]	ConvNext-B	**83.4**	49.8	957.1
SwiftNet pyramid [orsic21pr]	ConvNext-L	83.2	51.1	2067.3

Mean IoU on the standard datasets

Method	Backbone	Cityscapes val	ADE20k val	GMACs
HRNetV2 + OCR [wang20tpami]	Wide ResNet48	81.6	-	-
HRViT [gu22cvpr]	-	83.2	50.2	-
Swin UperNet [liu21iccv]	SWIN-B	-	51.6	2348.1
Swin UperNet [liu21iccv]	SWIN-L	-	53.5	3183.3
Seg-B [strudel21iccv]	DeiT-B	80.5	48.5	822.6
Seg-L [strudel21iccv]	ViT-L	81.3	53.6	2697.7
SwiftNet pyramid [orsic21pr]	ConvNext-B	**83.4**	49.8	957.1
SwiftNet pyramid [orsic21pr]	ConvNext-L	83.2	51.1	2067.3

Faculty of Electrical Engineering and Computing

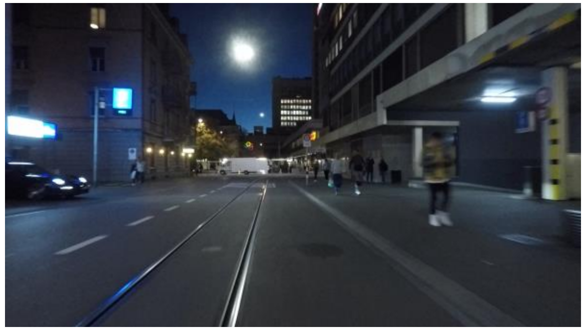
Abundant supervision walks the walk

Select pseudolabels for the last mile

Available datasets

- Datasets with Cityscapes/ACDC taxonomy
  - Cityscapes, Vistas, BDD, Wilddash2, ACDC – ≈40K fine-grained annotations

- Pool of ≈250K driving images, some with adverse conditions
  - DarkZurich, CACDC, NightOwls, NightCity, STF, BDD100k



DarkZurich

Canadian ACDC

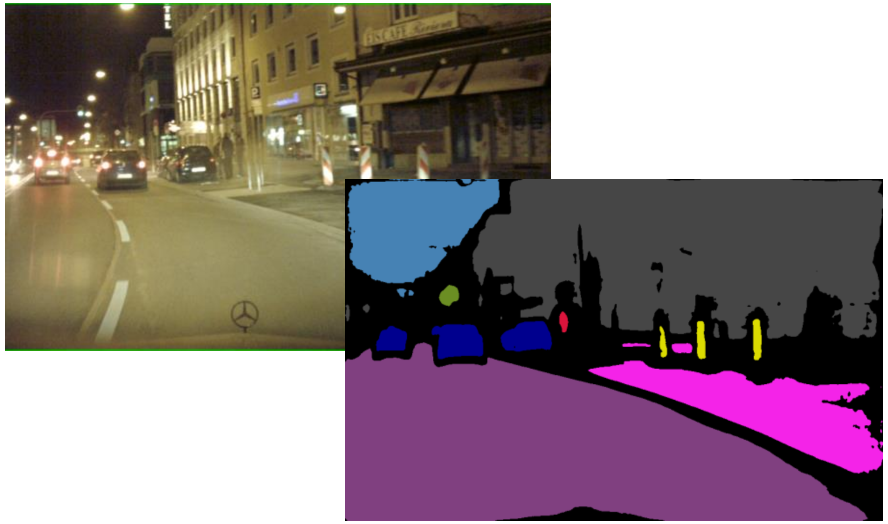
SeeingThroughFog

Semi-supervised learning procedure

Different perspective: Segmentation = recognition + boundary refinement

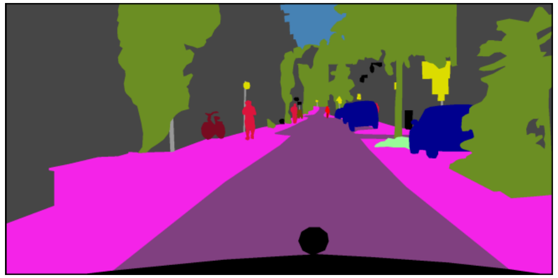
- Leverage annotated data for baseline recognition and refining semantic boundaries

- Improve recognition through pseudolabeling of highly confident predictions



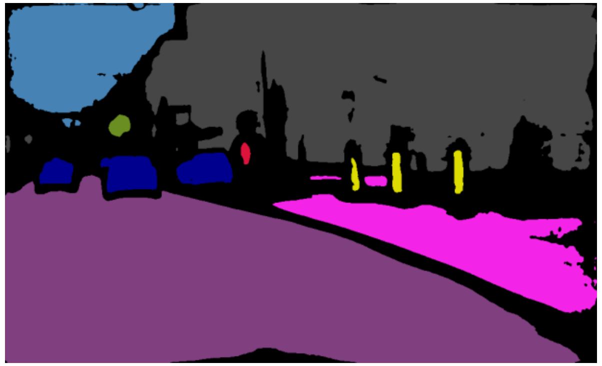
Boundary-aware focal loss [zhen19aaai]

Annotations - emphasizing boundaries ensures proper refinement



Reweighting based on boundaries

Pseudolabels - emphasizing boundaries yields higher influence of small classes



Reweighting based on boundaries

Faculty of Electrical Engineering and Computing

ACDC Challenge

Data	All conditions	Snow	Fog	Rain	Night
ACDC with DeepLabv3+	70.0	69.6	69.1	74.1	60.9
ACDC with HRNet	75.0	76.3	74.7	77.7	65.3
ACDC	77.4	77.8	77.2	80.8	69.4
+ Annotations	81.4	82.7	79.7	84.9	74.0
+ Annotations + Pseudolabels V1	82.8	85.2	80.8	85.8	75.3
+ Annotations + Pseudolabels V2	82.8	83.8	81.0	85.8	75.9

- Improvement due to pseudolabeling
- The same model achieves 84.7% on Cityscapes test

ACDC Challenge

Data	All conditions	Snow	Fog	Rain	Night
ACDC with DeepLabv3+	70.0	69.6	69.1	74.1	60.9
ACDC with HRNet	75.0	76.3	74.7	77.7	65.3
ACDC	77.4	77.8	77.2	80.8	69.4
+ Annotations	81.4	82.7	79.7	84.9	74.0
+ Annotations + Pseudolabels V1	82.8	85.2	80.8	85.8	75.3
+ Annotations + Pseudolabels V2	82.8	83.8	81.0	85.8	75.9

- Improvement due to pseudolabeling
- The same model achieves 84.7% on Cityscapes test

Faculty of Electrical Engineering and Computing

ACDC Challenge

Data	All conditions	Snow	Fog	Rain	Night
ACDC with DeepLabv3+	70.0	69.6	69.1	74.1	60.9
ACDC with HRNet	75.0	76.3	74.7	77.7	65.3
ACDC	77.4	77.8	77.2	80.8	69.4
+ Annotations	81.4	82.7	79.7	84.9	74.0
+ Annotations + Pseudolabels V1	82.8	85.2	80.8	85.8	75.3
+ Annotations + Pseudolabels V2	82.8	83.8	81.0	85.8	75.9

- Improvement due to pseudolabeling
- The same model achieves 84.7% on Cityscapes test

Faculty of Electrical Engineering and Computing

VsiTe

GIDEON

Visualizations - night

RGB Input

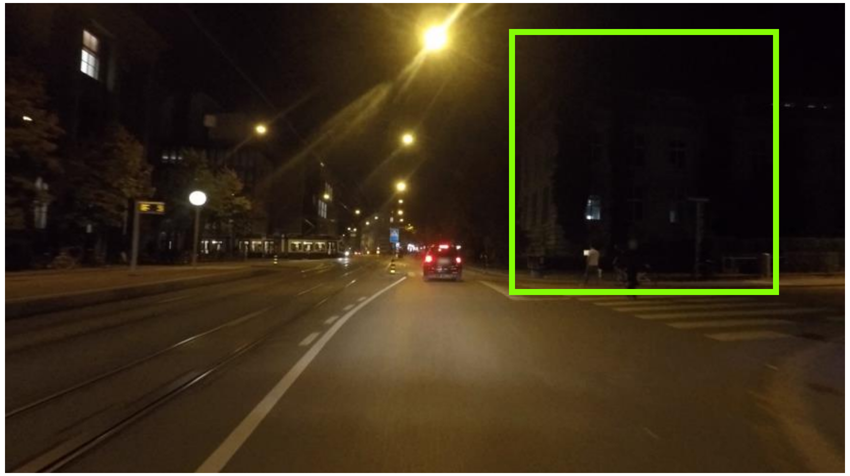
Predictions

Ground Truth

Faculty of Electrical Engineering and Computing

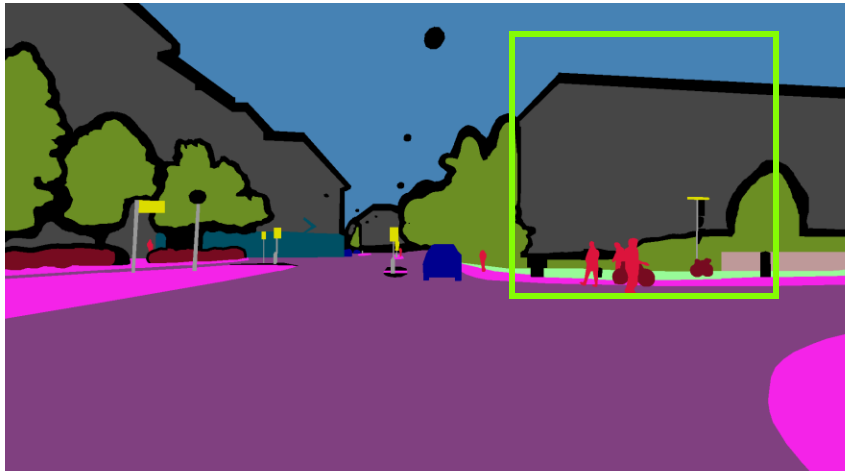
Visualizations - night

RGB Input

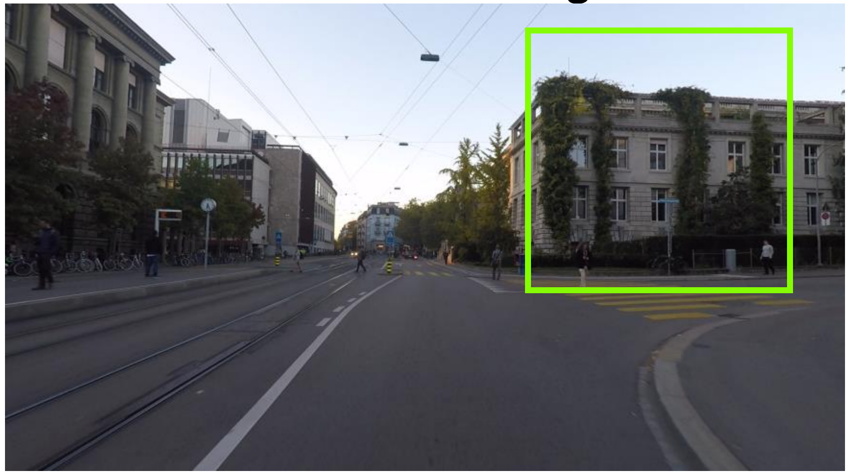


Predictions

Ground Truth



Reference Image



Exploiting uncertainty in semantic segmentation

Computers may see better than humans in dark images

Uncertainty in SemSeg: Average UIoU

- UIoU = IoU + prediction confidence
- Invalids (TI/FI) are based on confidence threshold and invalid GT

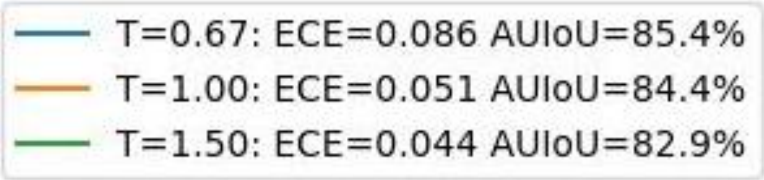
$$\text{UIoU} = \frac{|\text{TP}| + \boxed{|\text{TI}|}}{|\text{TP}| + \boxed{|\text{TI}|} + |\text{FP}| + |\text{FN}| + \boxed{|\text{FI}|}}$$

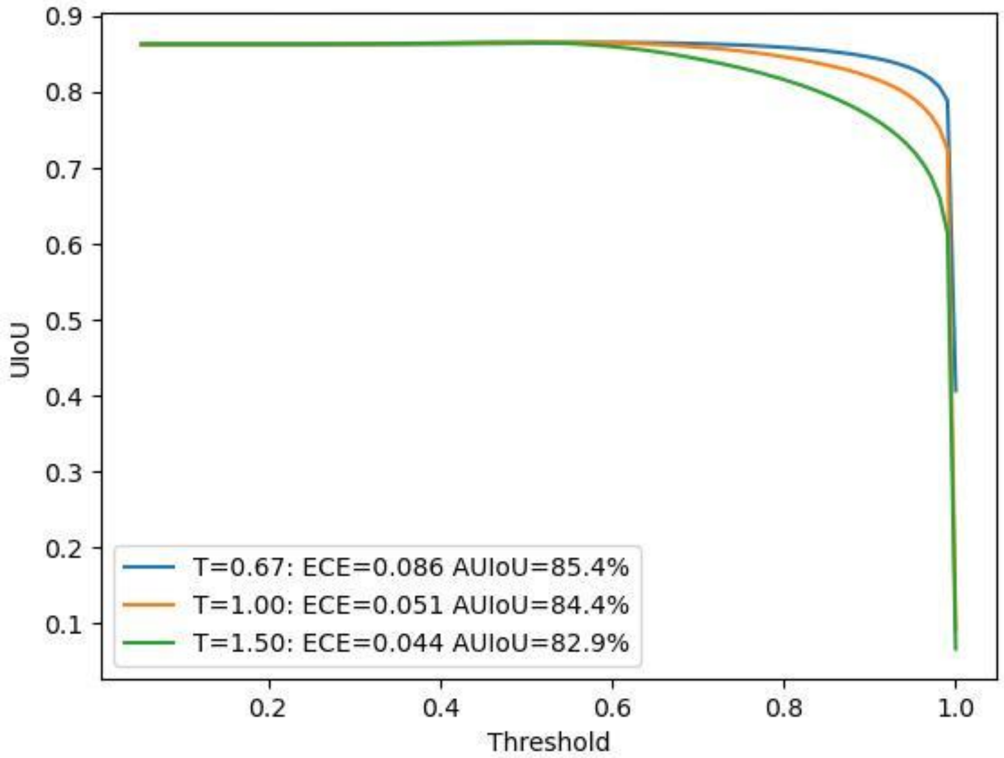
- Prediction confidence is poorly calibrated [guo17neurips]
- Better confidence calibration -> better UIoU?

Better calibration leads to worse AUIoU

- Post hoc temperature scaling of softmax improves calibration:

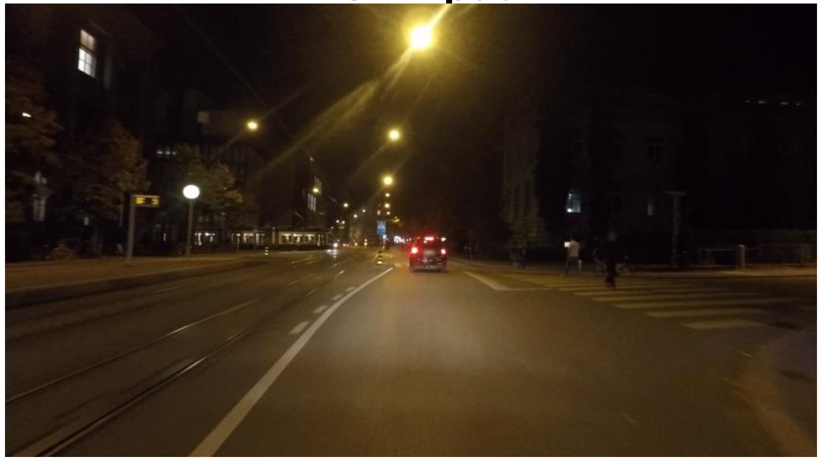
P(**y**|**x**) = softmax(logits/T)





Invalid regions: GT vs calibrated model

RGB Input



Ground Truth

Ground Truth Invalid regions

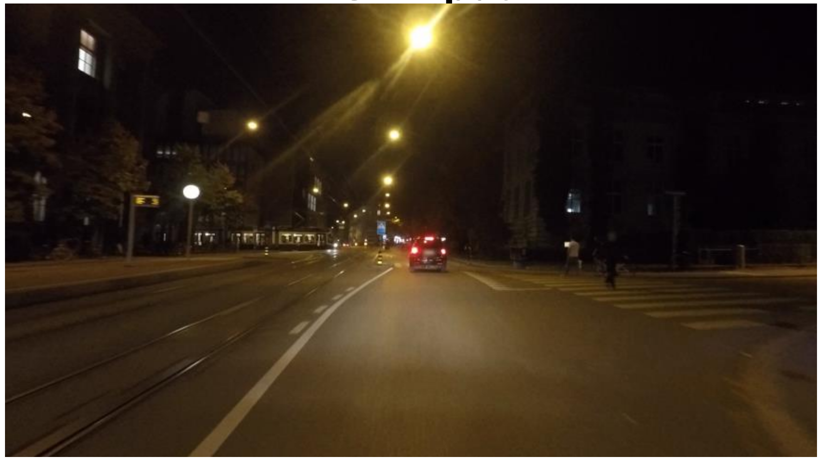
Predictions

Confidence t>75%

Confidence t>95%

Invalid regions: GT vs calibrated model

RGB Input



Ground Truth

Ground Truth Invalid regions

Predictions

Confidence t>75%

Confidence t>95%

Conclusion

- Convolutional networks deliver great performance on high resolution images
- It is easier to refine boundaries than to recognize semantics
- Unrecognizable regions for humans are different than unrecognizable regions for deep models
- Computers may see better in the dark

- Email: matej.grcic@fer.hr

This work has been supported by:

- Croatian Science Foundation (grant IP-2020-02-5851 ADEPT)
- European Regional Development Fund and Gideon ltd (KK.01.2.1.02.0119 A-Unit)
- VSITE College for Information Technologies (access to NVIDIA DGX - 8x V100 32GB)

Appendix

Training details, visualizations & more

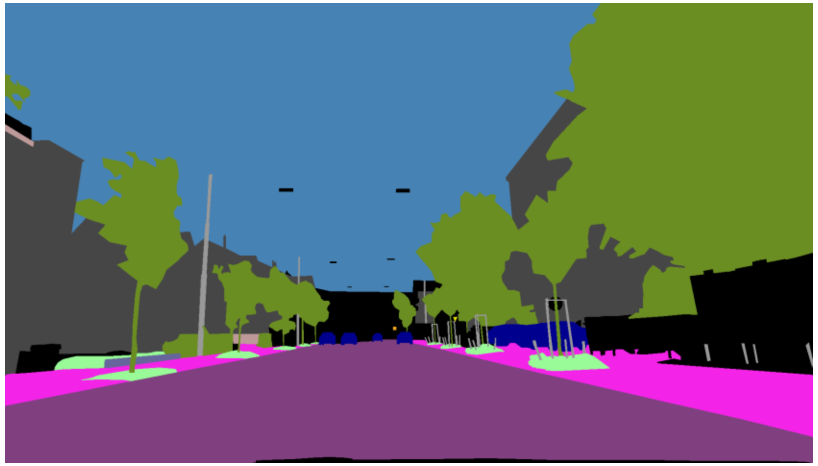
Training details

- 100K images - 40K annotated + 60K pseudolabeled
- 50 Epochs, 1.5h per epoch
- Half precision training, batch size of 3 per GPU, 8x V100 32GB
- Jittering in range [0.5, 2] followed by random crop 1024x1024
- Pytorch Lightning is a way to go!

Fog

RGB Input

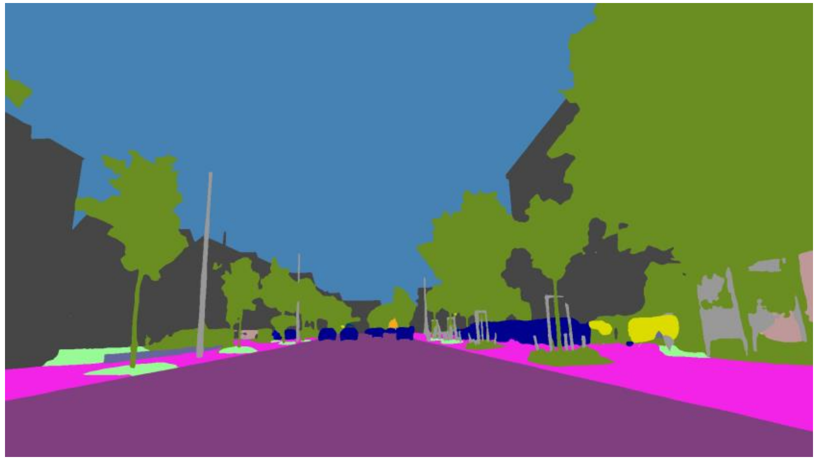
Ground Truth



Ground Truth Invalid regions

Reference Image

Predictions



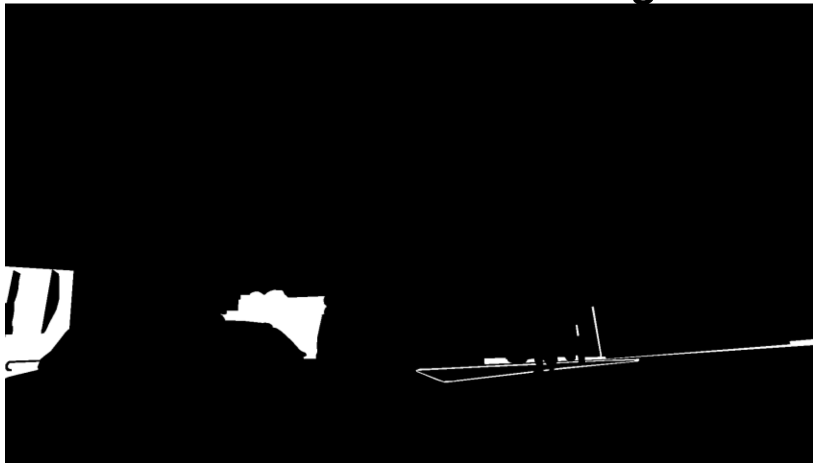
Confidence t>95%

Snow

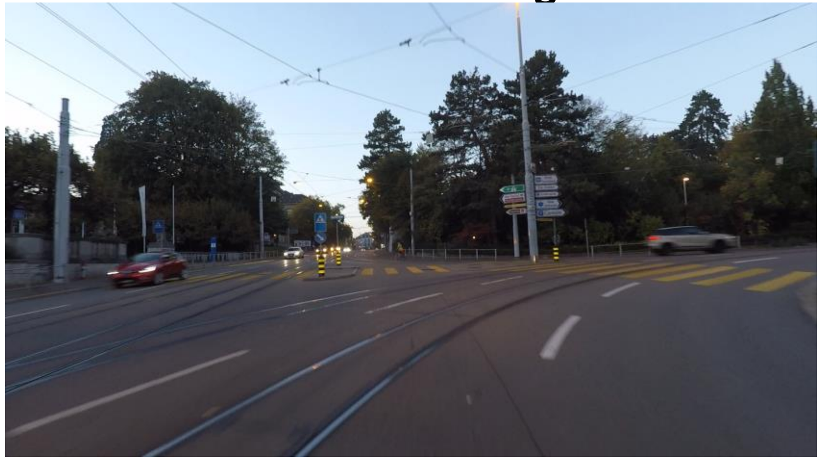
RGB Input

Ground Truth

Ground Truth Invalid regions



Reference Image

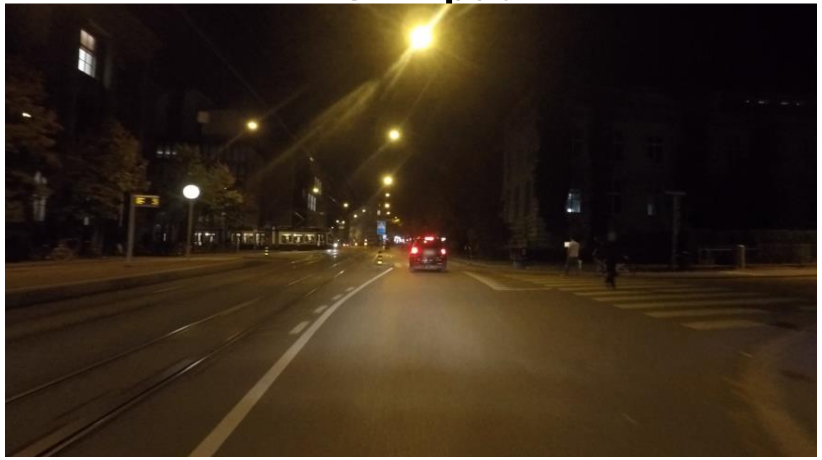


Predictions

Confidence t>95%

Night

RGB Input



Ground Truth

Ground Truth Invalid regions

Reference Image

Predictions

Confidence t>95%

Faculty of Electrical Engineering and Computing

VsiTe

GIDEON

Rain

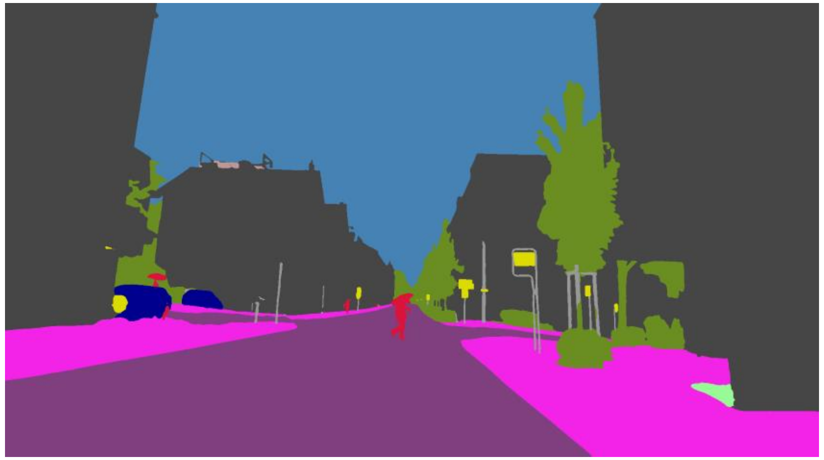
RGB Input

Ground Truth

Ground Truth Invalid regions

Reference Image

Predictions



Confidence t>95%