Reliability in Semantic Segmentation: Are we on the Right Track?

Pau de Jorge, Riccardo Volpi, Philip Torr and Grégory Rogez

West Building Exhibit Halls ABC

TUE-PM-291

Motivation of our analysis

- **Semantic segmentation** is at the core of many applications such as self-driving cars or embodied AI that need *reliable* models

- The rising popularity of transformers in computer vision has led to **many novel architectures in recent years**, but reliability metrics are often overlooked

- In our study we **compare recent segmentation models** with older baselines (i.e. ResNet) to answer the question…

Are state-of-the-art semantic segmentation models improving in terms of *reliability*?

Motivation of our analysis

- **Semantic segmentation** is at the core of many applications such as self-driving cars or embodied AI that need *reliable* models

- The rising popularity of transformers in computer vision has led to **many novel architectures in recent years**, but reliability metrics are often overlooked

- In our study we **compare recent segmentation models** with older baselines (i.e. ResNet) to answer the question…

Are state-of-the-art semantic segmentation models improving in terms of *reliability*?

Not in all metrics! Need to improve uncertainty

Reliability to “natural” domain shifts

- **Robustness**: mIoU when facing distribution shifts.
- **Calibration**: Predictive probability = expected accuracy (i.e. if the probability=0.9 then 90% of the samples should be correct)
- **Misclassification detection**: Correct samples should have a higher confidence than incorrect samples
- **Out-of-distribution detection**: Out-of-distribution (OOD) samples should have a lower confidence than in distribution

Evaluation benchmark

- We model natural domain shifts with self-driving datasets captured in diverse scenarios

- Models are trained on Cityscapes, we define the shift strength based on qualitative evaluation and ResNet baseline performance

Evaluation benchmark

Recent transformer models

- **SETR:** VIT backbone and “simple” convolution free decoders
- **Segmenter:** VIT backbone and Masked Transformer decoder (fully transformer based)
- **SegFormer:** Efficient self attention backbone + MLP decoder (speed comparable to ResNet)

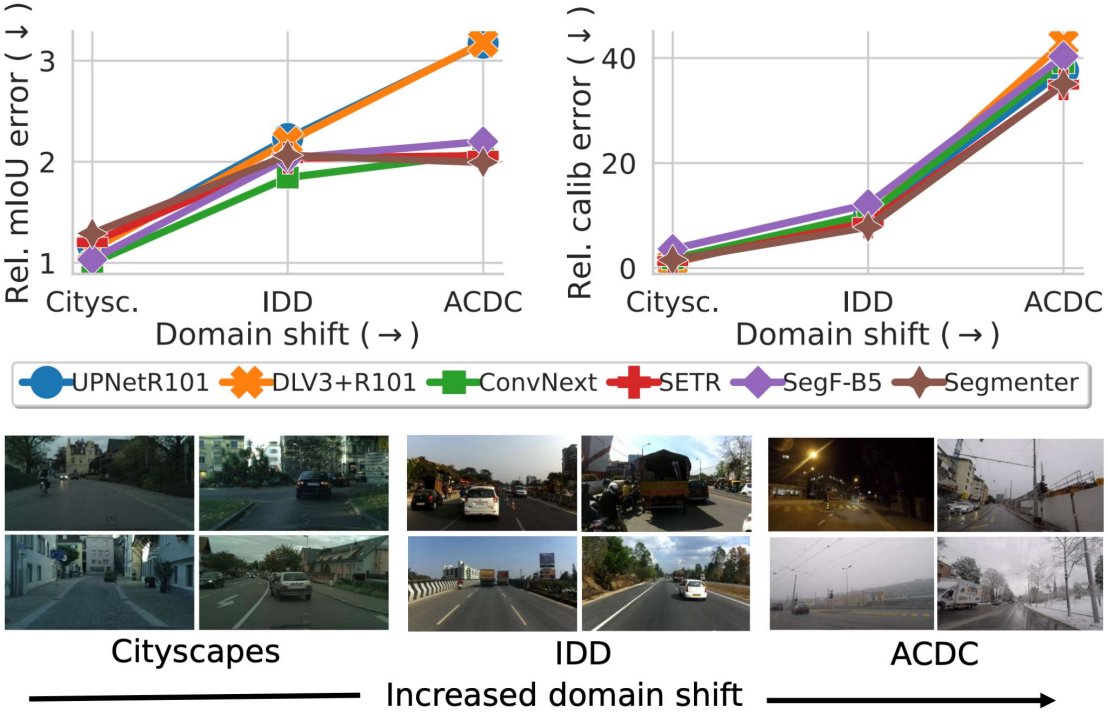
Recent convolutional models

- **ConvNext:** Inspired by transformers (uses UPerNet decoder)

Baseline

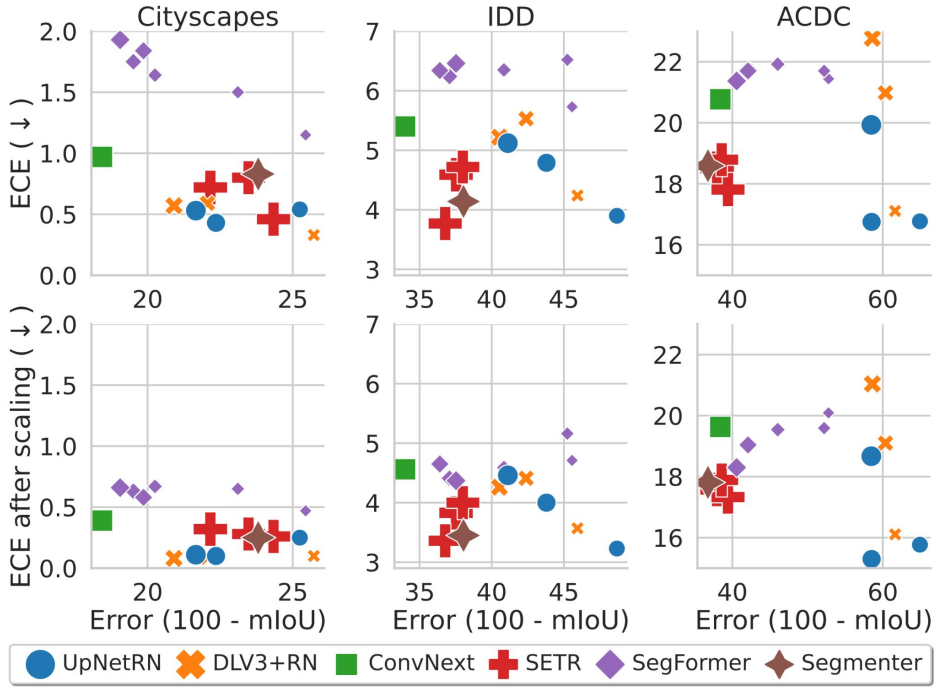
- **ResNet:** Two different decoders (DLV3+ and UPerNet)

Robustness vs Calibration



All recent models are remarkably more robust, but not much better calibrated

Robustness vs Calibration

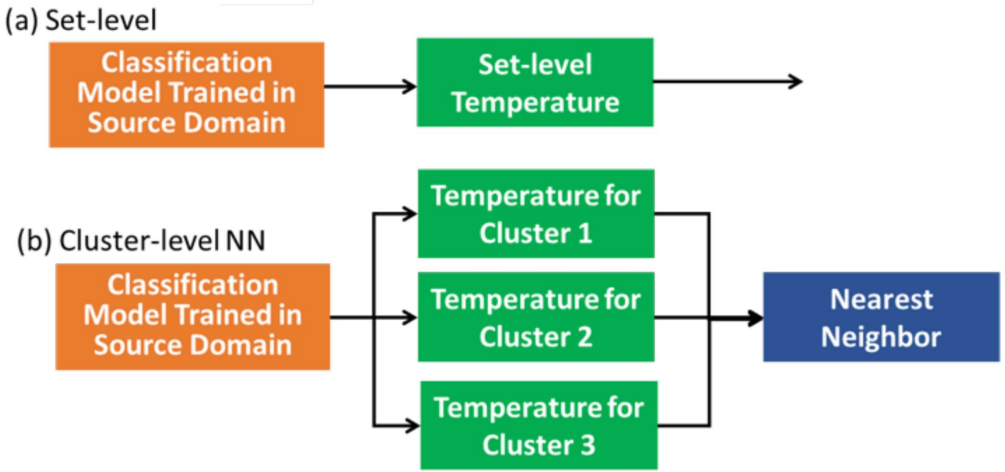


After Temp scaling (TS) calibration out of domain is still significantly higher

Can we improve on this?

Can we improve calibration out of domain?

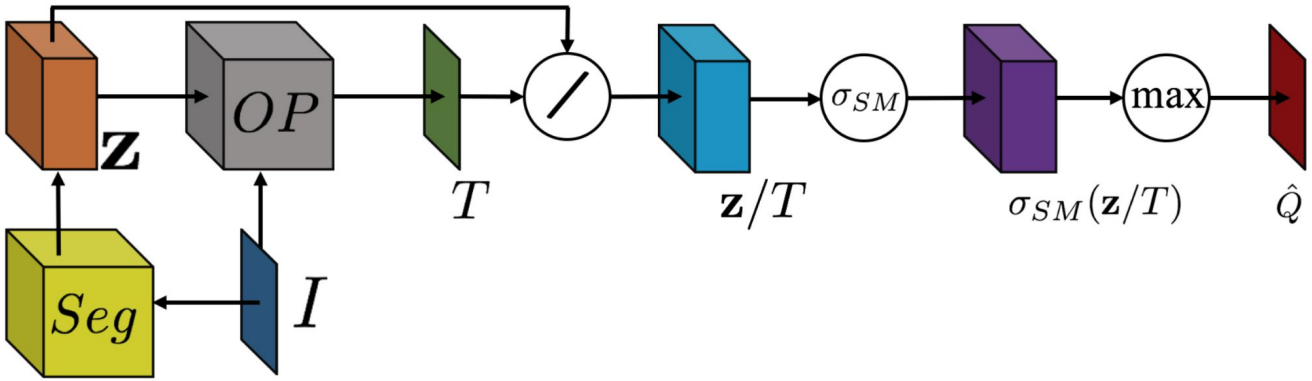
We explore existing literature to improve calibration out of domain.



Gong *et al., Confidence Calibration for Domain Generalization under Covariate Shift (ICCV 2021)*

Key idea: Clustering the calibration set to capture different domains

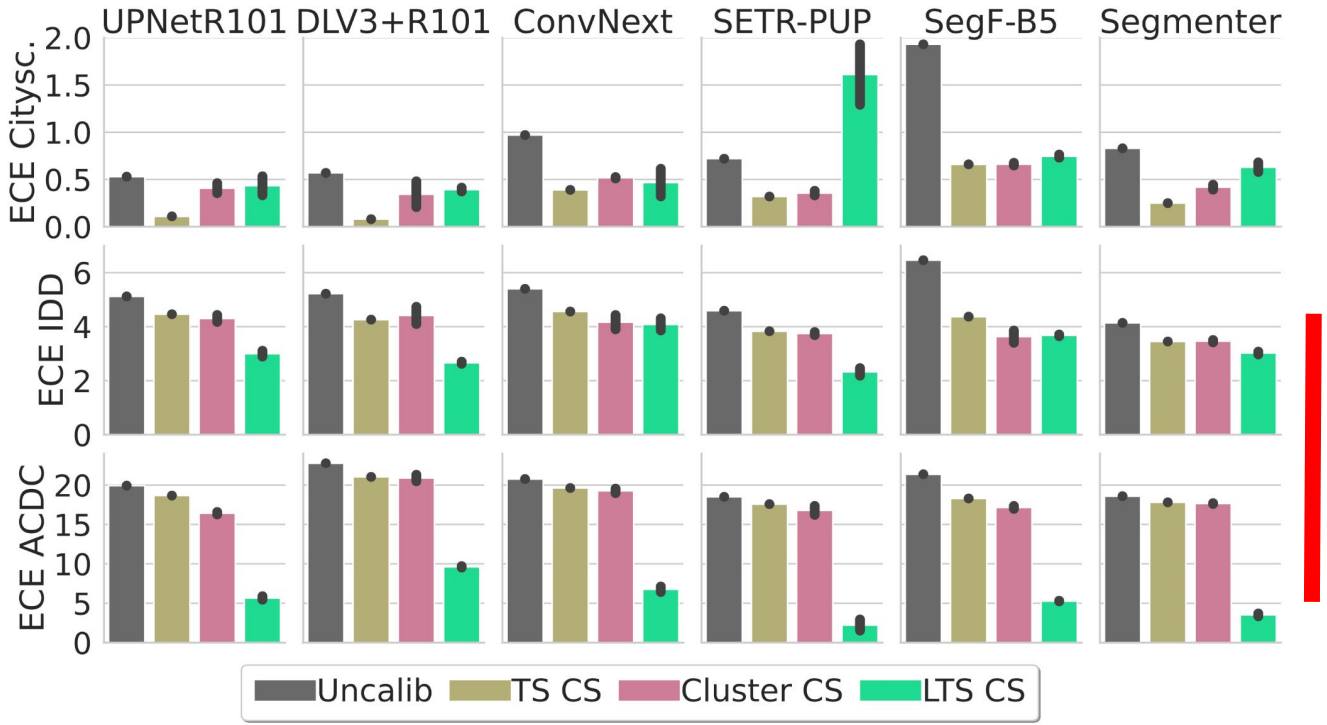
Adaptive temperature via calibration network



Ding *et al., Local Temperature Scaling for Probability Calibration (ICCV 2021)*

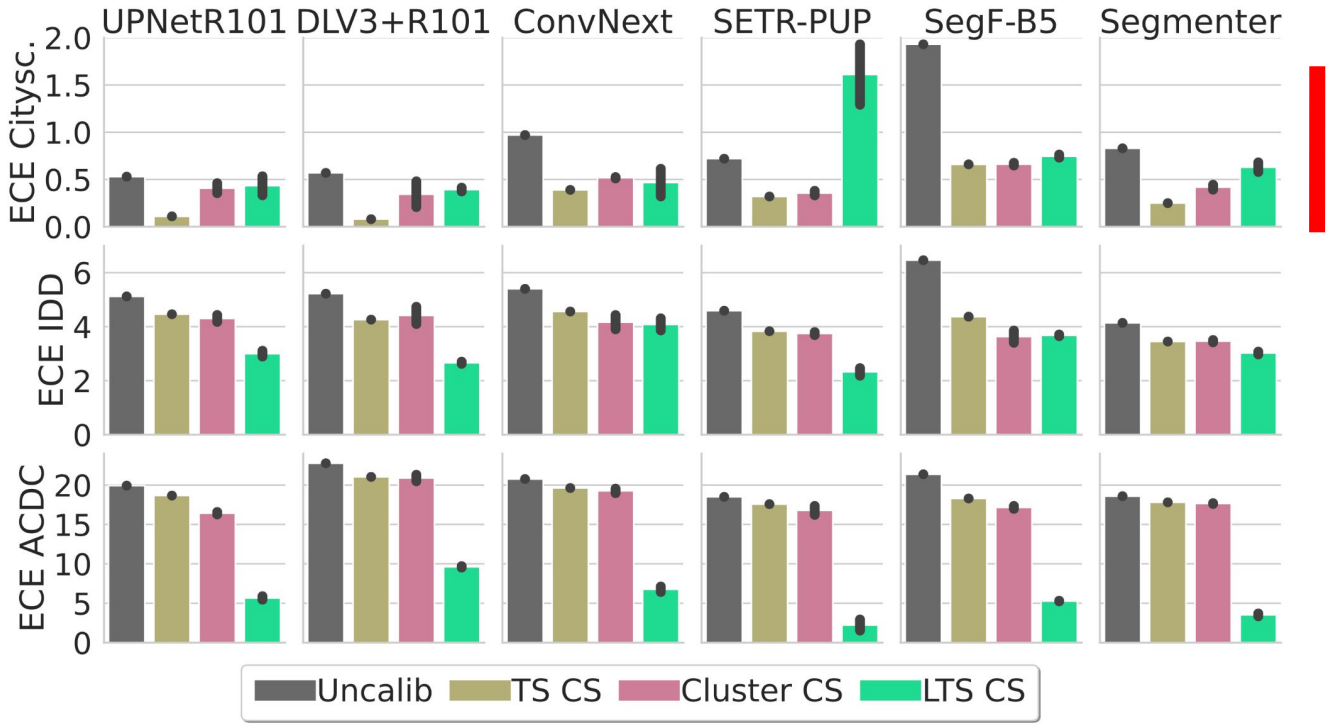
Key idea: use a small network trained on the calibration set to compute a temperature map

Adaptive temperature via calibration network



With only CS images for calibration, LTS performs best OOD

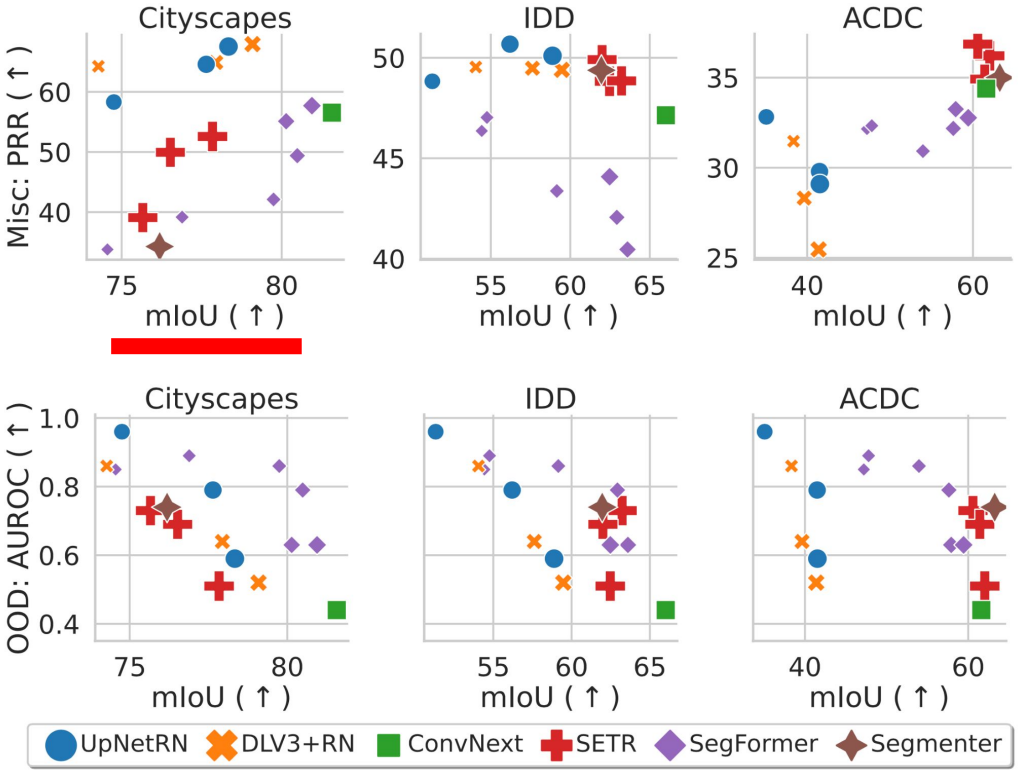
Adaptive temperature via calibration network



For CS, temperature scaling performs best

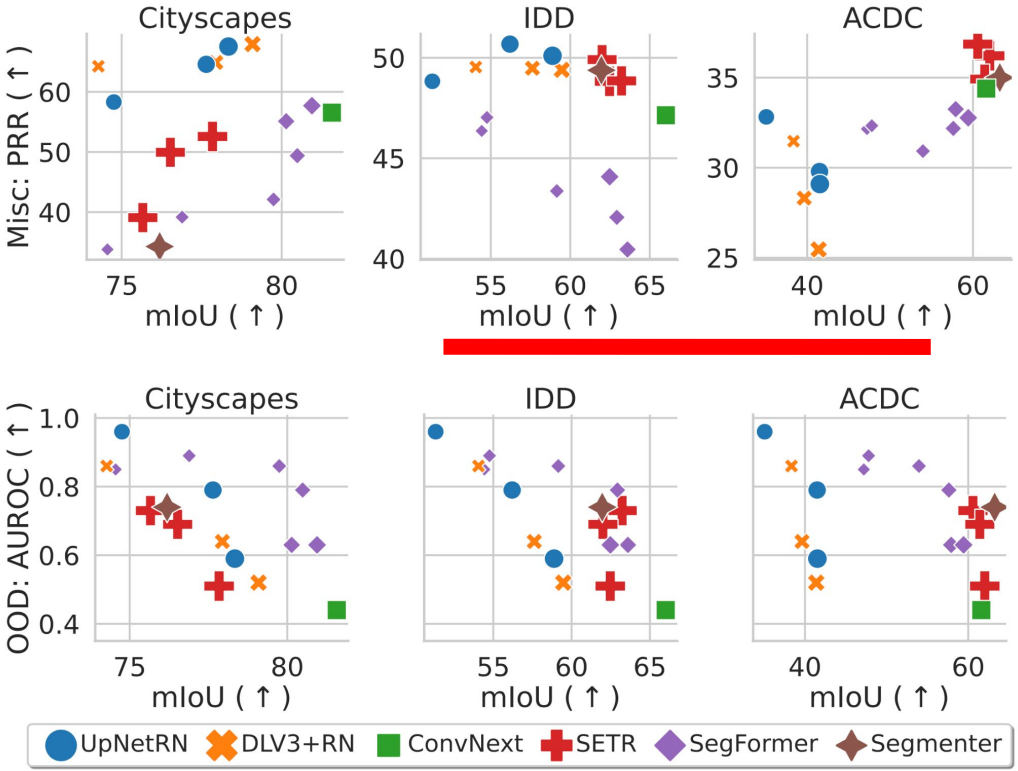
This may be due to the low variability of CS

Misclassification and OOD detection



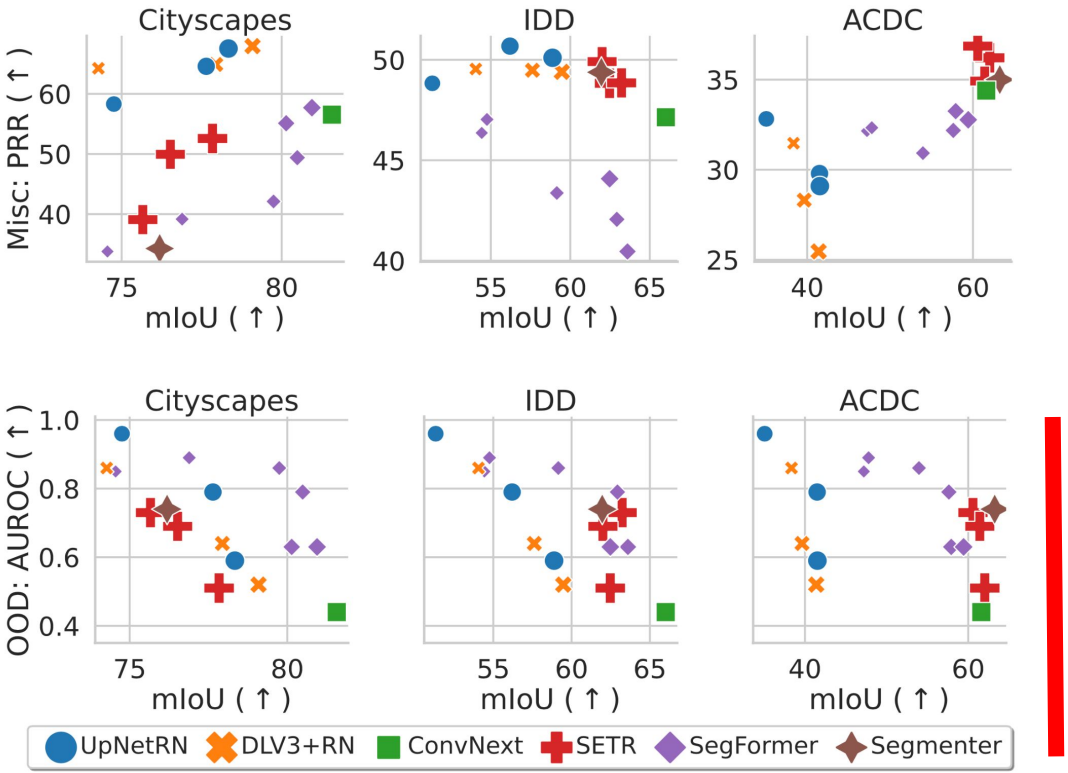
In distribution, ResNet seems to have the best Misc - mIoU trade-off

Misclassification and OOD detection



In OOD the trend changes: recent models perform best under strong shifts

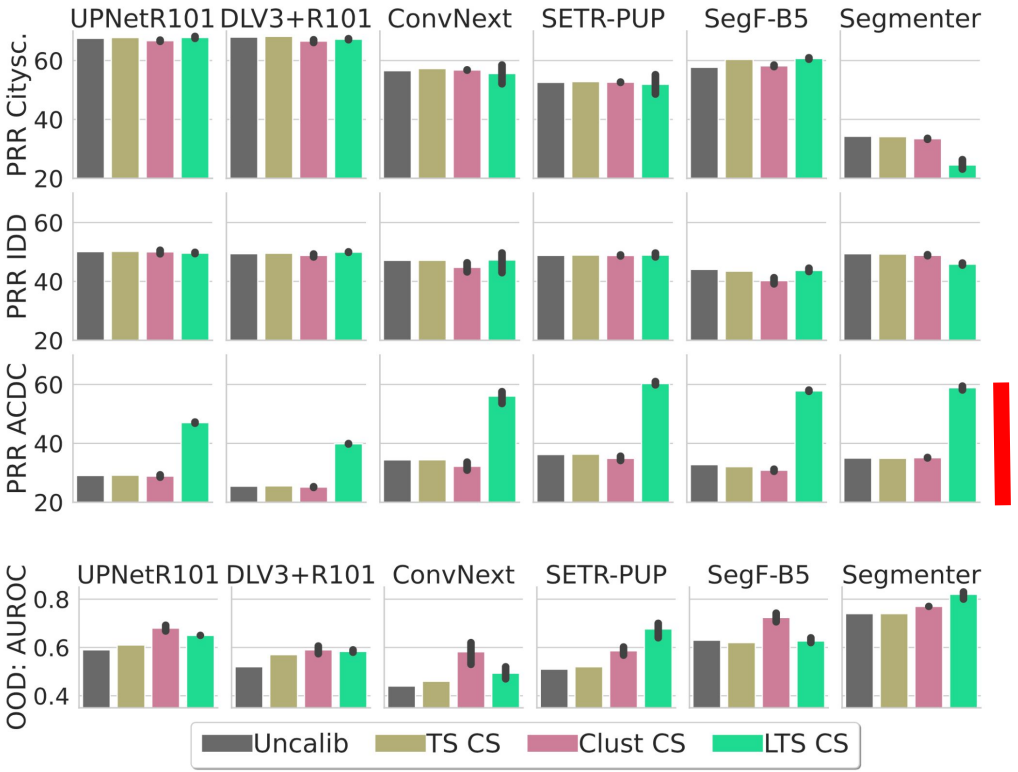
Misclassification and OOD detection



For OOD detection, there is a negative trend between mIoU and OOD detection

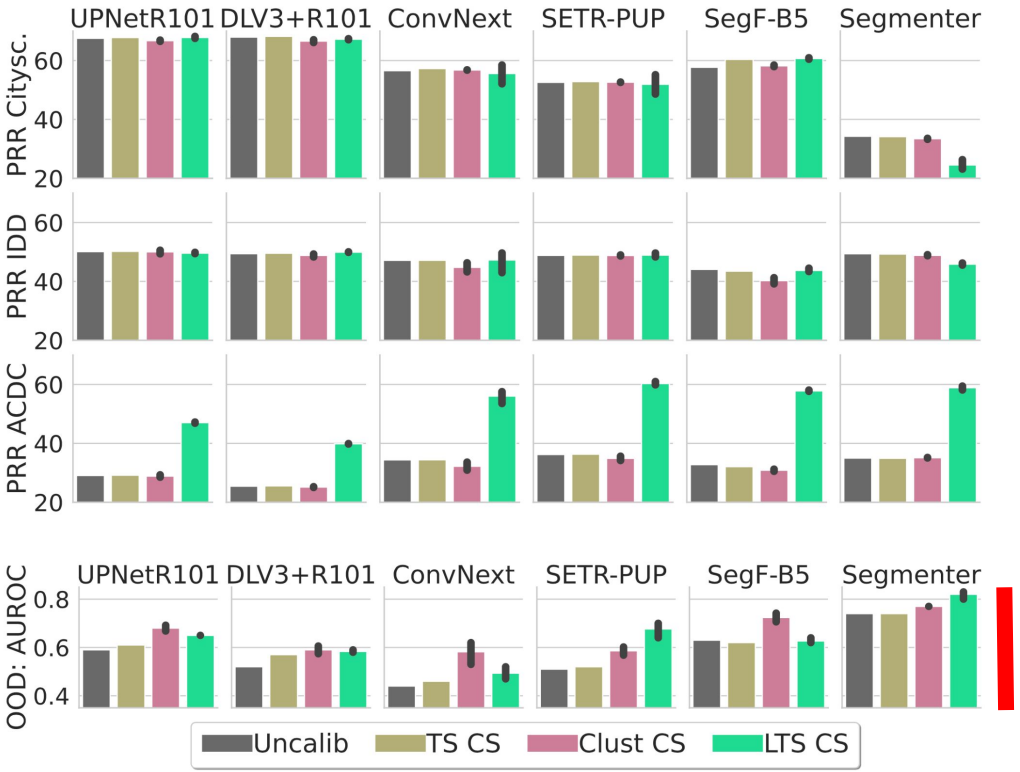
A smaller ResNet performs best for OOD

Can calibration improve other OOD metrics?



Calibration improves Misclassification detection under strong shifts.

Can calibration improve other OOD metrics?



Calibration can also improve OOD detection, although to a smaller extent

Conclusions

- Recent models present remarkable improvements in robustness but not in uncertainty estimation
- There is no single method that performs best in all scenarios
- Adaptive TS is a promising direction to improve OOD calibration
- Recent models underperform ResNets' misclassification ID but are better OOD.
- OOD detection is negatively correlated with mIoU: there is no free lunch
- Calibration can help other uncertainty metrics