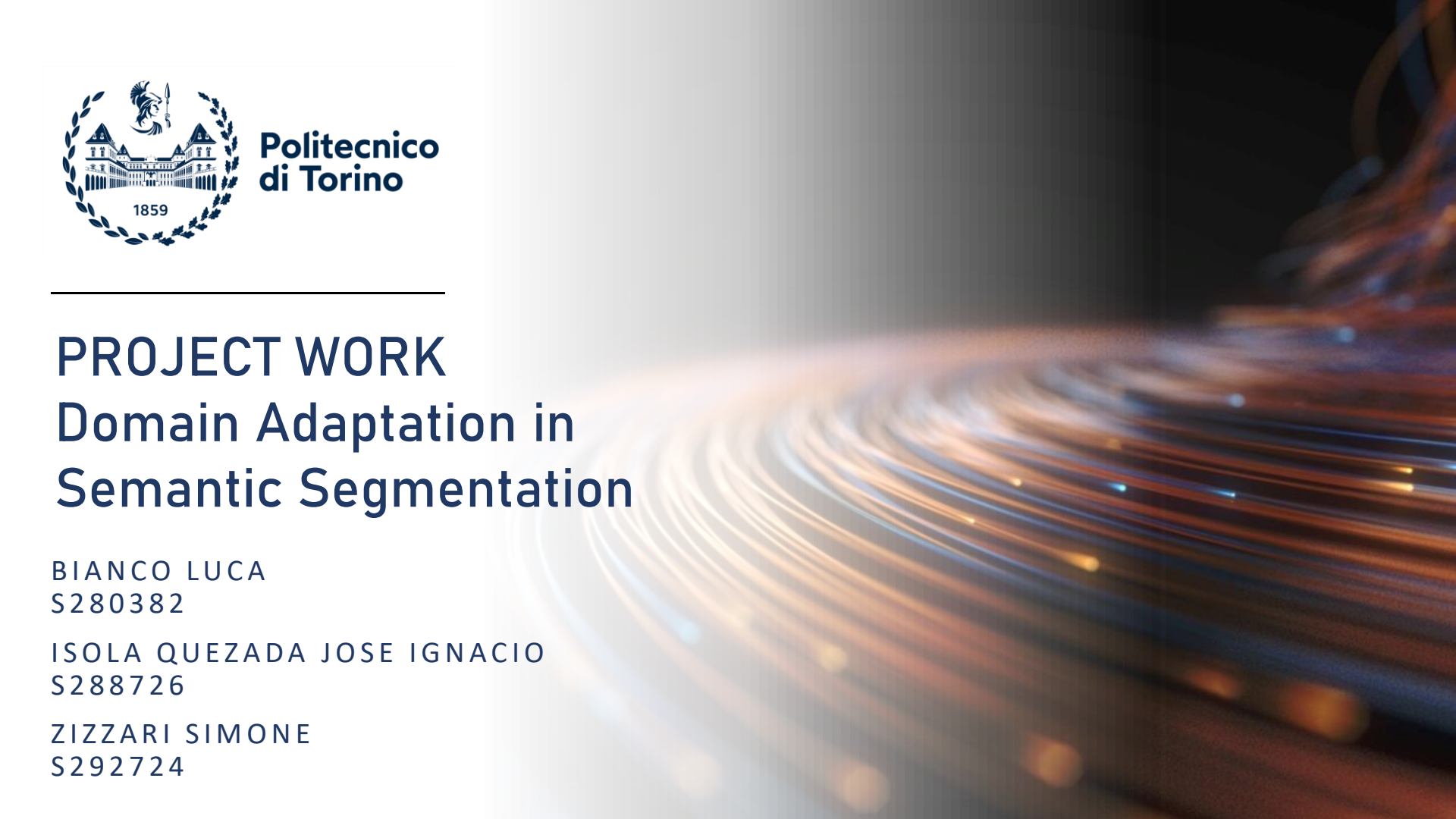


Politecnico di Torino

PROJECT WORK
Domain Adaptation in Semantic Segmentation

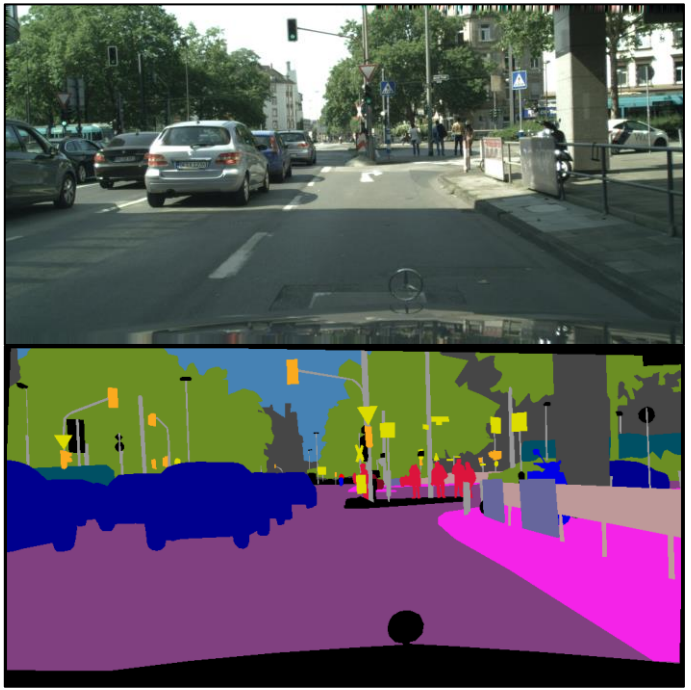
BIANCO LUCA
S280382

ISOLA QUEZADA JOSE IGNACIO
S288726

ZIZZARI SIMONE
S292724

Semantic Segmentation

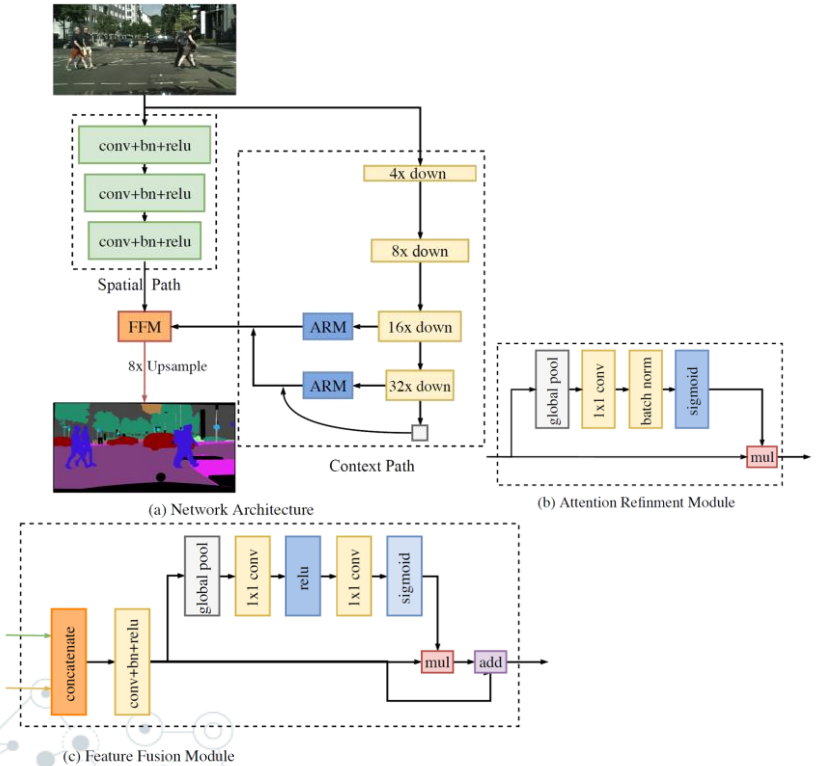
- Pixel-level labeling
- No instance detection, just labels
- Fully convolutional
- Problems:
  - Large receptive fields
  - Computationally expensive on high resolution images



Real-Time Application

- Good for computer-vision, especially in AVs, as the vehicle needs to understand where it can go
- Needs to be lightweight to achieve good performance
- Needs considerable amounts of training for real-word applications

An Example: BiSeNet

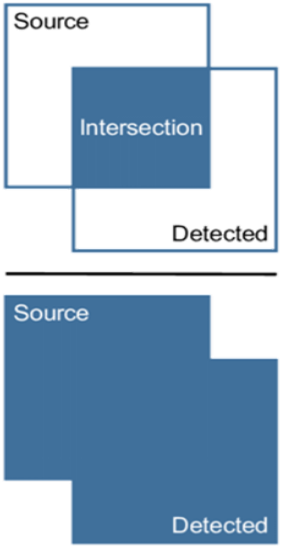


- Two paths: Spatial path and context path
- **Spatial path** retains spatial information (exact positioning of the labels)
- **Context path** presents large receptive fields and fast downsampling for quick and accurate labeling
- **Attention Refinement Module** adds global context information to improve feature recognition done in the context path. It does not need upsampling, computational cost is low
- **Feature Fusion Module** matches the low level (detailed) description of the spatial path with the high-level semantics of the context path

The Metrics

- **Mean Intersection over Union (MIoU):** Shows how much the labels overlap with the ground truth mask, averaged over all classes.
- **Pixel Accuracy:** Ratio between correctly classified pixels and all pixel predictions. Drawback: misleading results for uncommon classes.

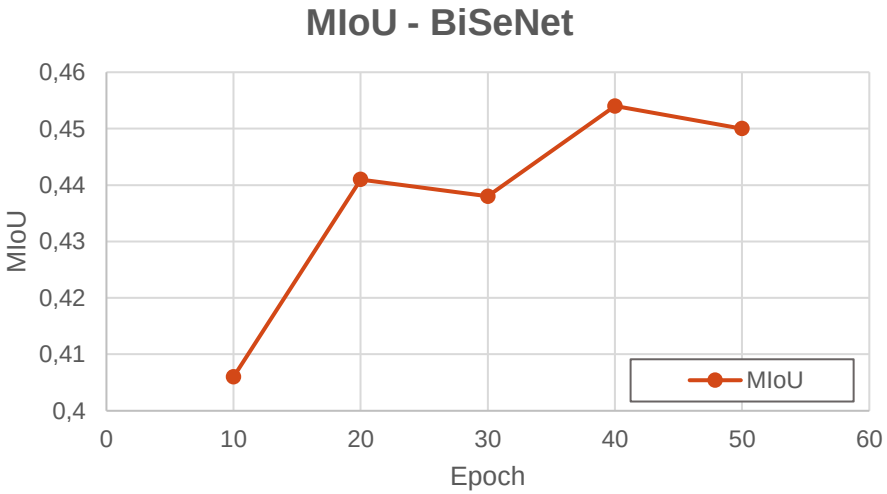
$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

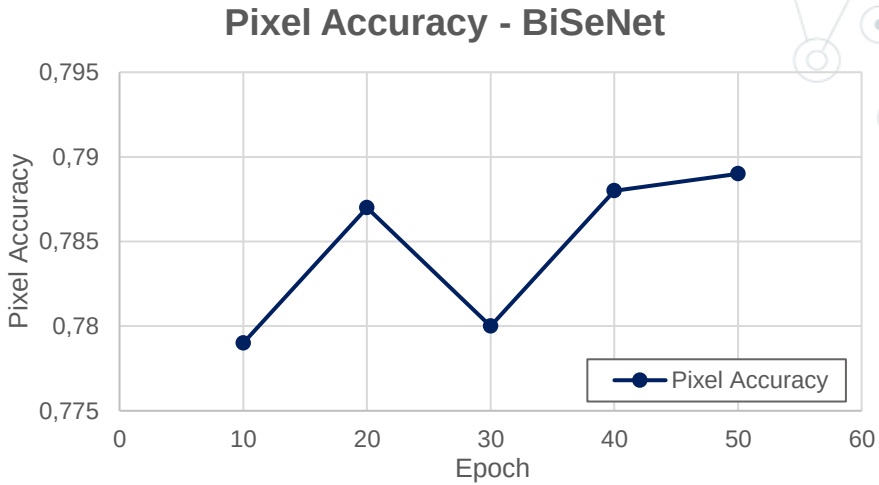


BiSeNet Implementation

- ResNet-18 backbone (Context Path)
- 50 epochs
- Polynomial learning rate: "*poly*" $initial\_lr \cdot \left(1 - \frac{iter}{max\_iter}\right)^{power}$
- Cross-entropy loss
- SGD optimizer
- Cityscapes dataset: subset with fully labeled images, taken from real car videos
- Good performance, after 40 epochs the MIoU slightly worsens (overfitting)

BiSeNet Accuracy and MIoU





The Need for Labeled Data and Domain Shift

GTA 5 synthetic dataset

- Realism
- Curse of dataset annotation
- Detouring:
  - Fast and semi-automated annotation process

Domain Shift

- Different visual appearance between image datasets
- Semantic Segmentation performance degradation
- Need for Domain Adaptation

Domain Adaptation

Possible solution to domain shift issues

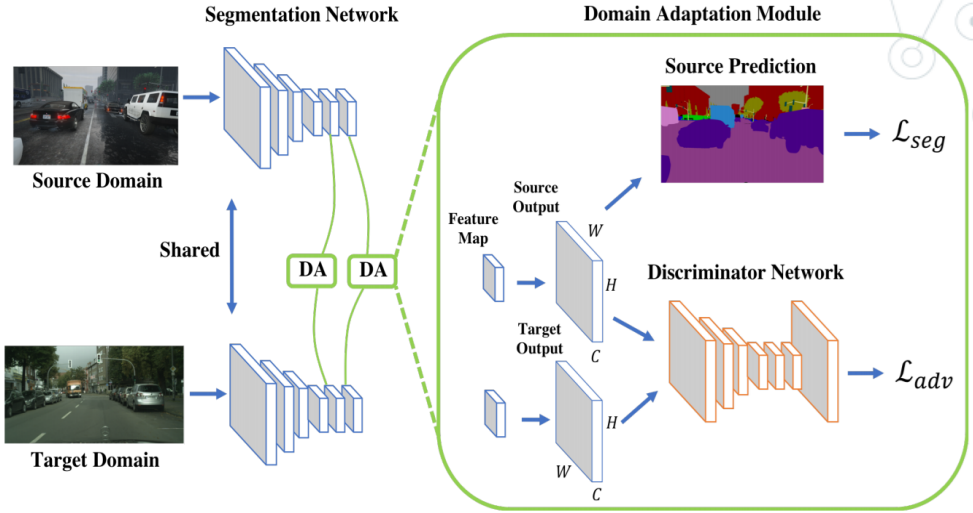
- Can be supervised, semi-supervised or unsupervised, depending on the number of labeled target samples. In this case, we try an unsupervised approach with a single source dataset (GTA5).
- Various techniques available, with different pros and cons. Semantic segmentation networks normally use Adversarial Discriminative Methods.

COMPARISON OF DIFFERENT SINGLE-SOURCE DUDA CATEGORIES. (THE MORE STARS THE METHOD HAS, THE BETTER IT IS.)

	Theory guarantee	Efficiency	Task scalability	Data scalability	Data dependency	Optimizability	Performance
Discrepancy-based methods	★★★	★★★	★	★★	★★★	★★★	★★
Adversarial discriminative methods	★★	★★	★★★	★★★	★	★	★★★
Adversarial generative methods	★	★	★★	★	★	★	★★★
Self-supervision methods	★	★★	★★★	★★★	★★	★★★	★★

An Example – Adversarial Discriminative Domain Adaptation

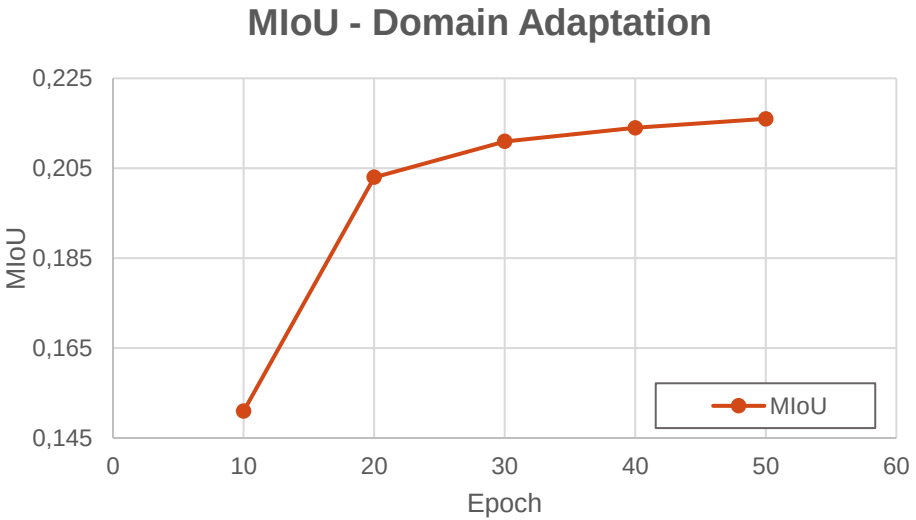
- Idea: apart from the semantic segmentation network, we introduce a discriminator.
- The generator must fool the discriminator, meaning its segmentation is "just as good" on the target domain as it is on source domain.
- The objective is to minimize the probability of the discriminator correctly labeling images coming from the target domain.

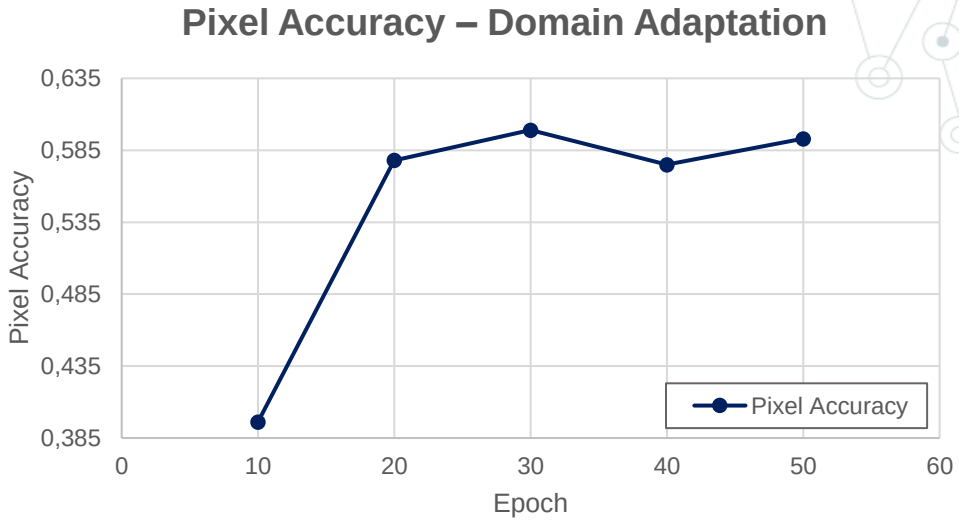


An Example – Adversarial Discriminative Domain Adaptation

- We introduce a discriminator with a fully convolutional architecture: 5 convolution layers with kernel 4 × 4 and stride of 2, each ending in a leaky ReLU (except the last one)
- Training is performed by loading one source (GTA5) image, its corresponding ground truth labels and one target (Cityscapes) image (without labels)
- Generator creates labels for both source and target images
- Both generated labels train the discriminator

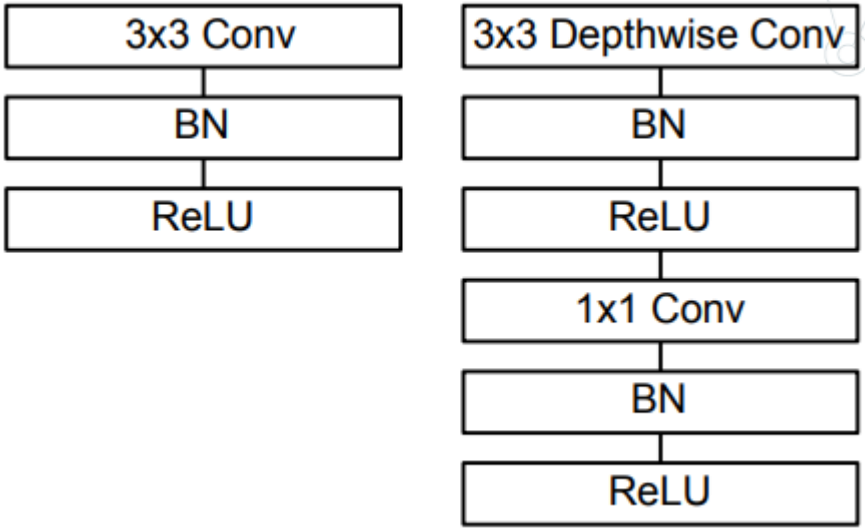
Domain Adaptation Accuracy and MIoU





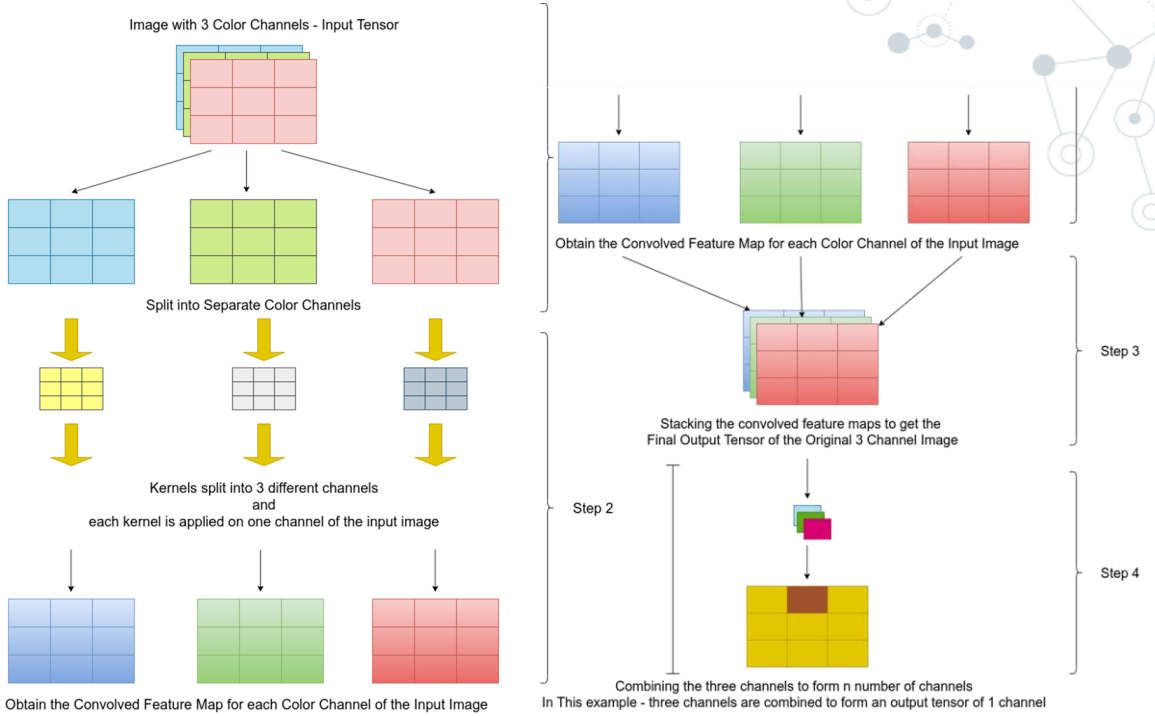
Going Lightweight
Depthwise-Separable Convolutions

- Less operations required
- The full convolutions of the discriminator can be changed for depthwise-separable convolutions

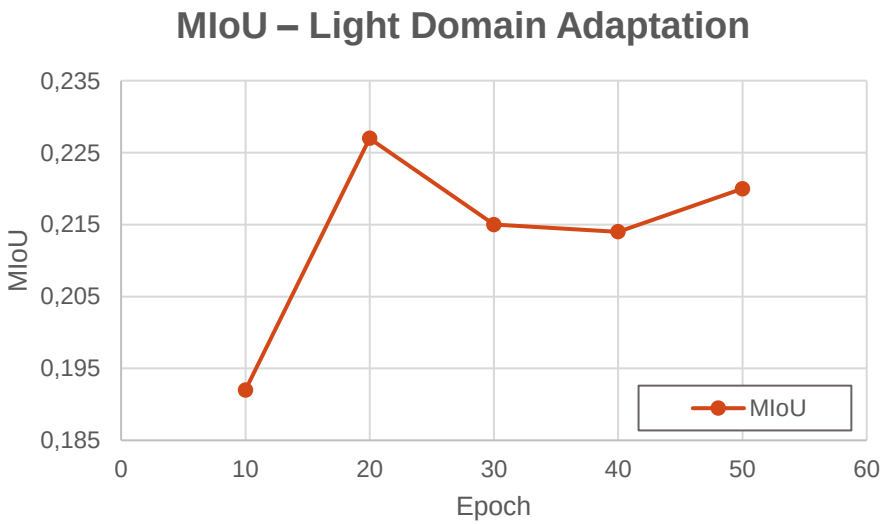


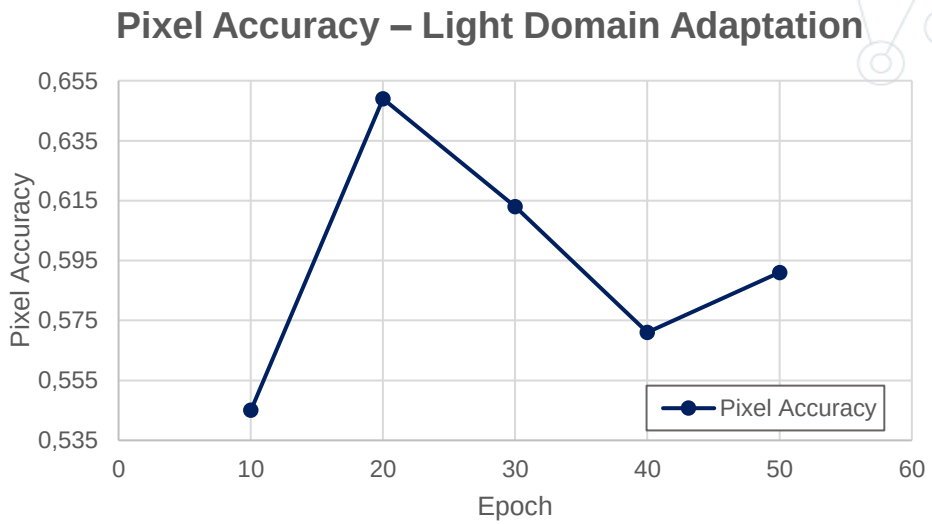
Going Lightweight Depthwise-Separable Convolutions

- In contrast to fully convolutional layers, channels are kept separate when running 2D convolutions (depthwise convolutions)
- The final output is made combining the separate single-channel outputs into as many channels as desired (depthwise separable convolutions)



Lightweight Domain Adaptation Accuracy and MIoU



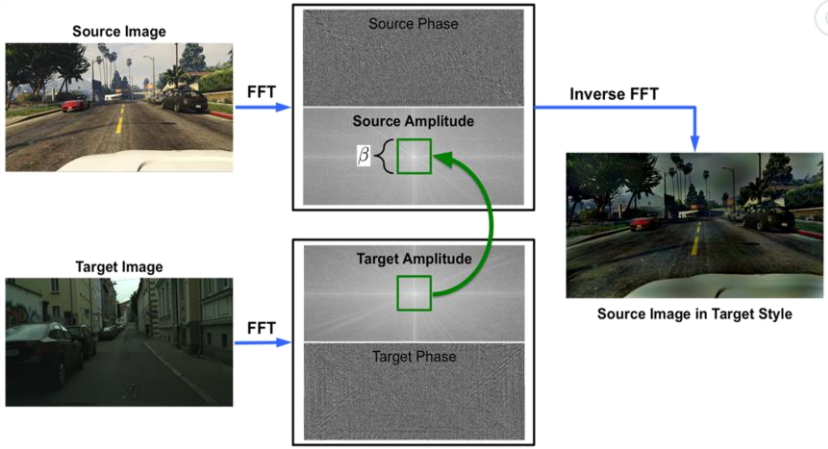


Limitations of Domain Adaptation

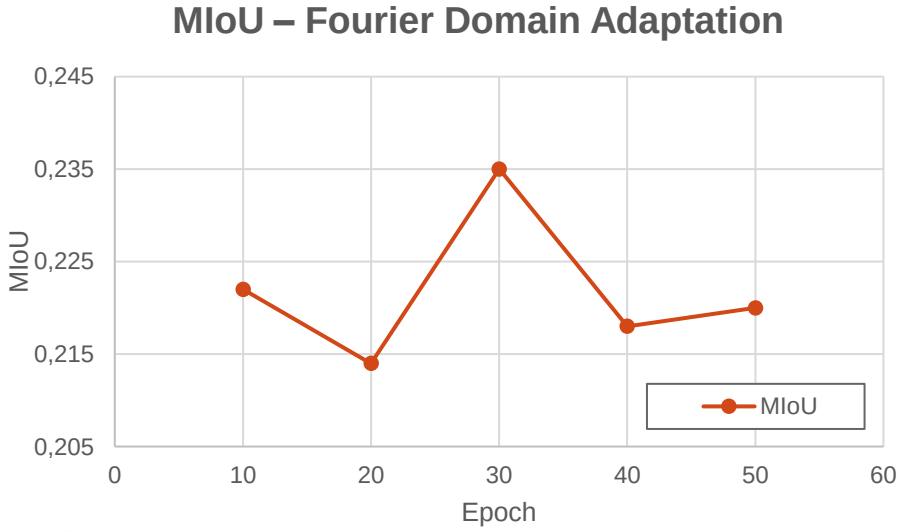
- Performance is worse than directly training on a labeled target dataset
- Categories might not match perfectly
- Further optimization is needed (higher complexity)
- Models are heavier in training (adversarial D.A.)
- Different D.A. methods involve different tradeoffs

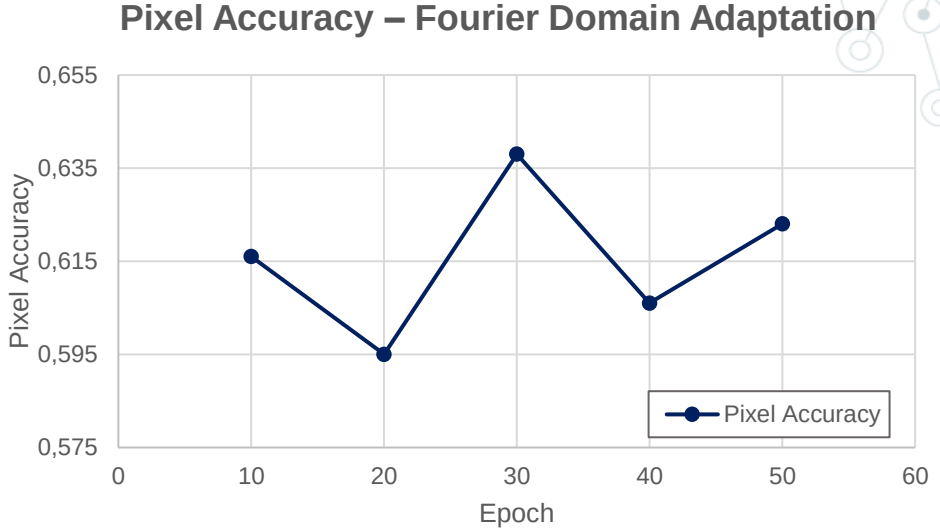
Extension: Fourier Domain Adaptation (FDA)

- Alternative approach to DA: no discriminator needed.
- Uses fast Fourier transform to extract the frequency spectra of both source and target images.
- The amplitudes of the low frequency components of the target image replace those of the source, which is re-converted to an image by means of an inverse Fourier transform.
- The new images to be fed to the network keep the features of the source, but have the appearance of the target.



Fourier Domain Adaptation Accuracy and MIoU

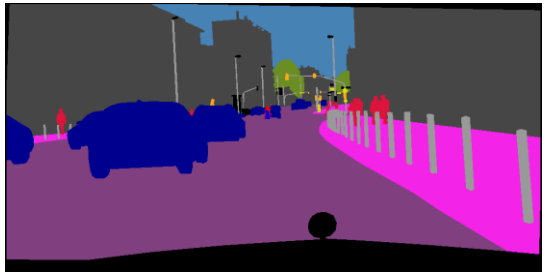




Results comparison – Table

Experiment	Accuracy (%)	mIoU (%)	Total Parameters	GFLOPs
BiseNet (ResNet-18)	0.788	0.454	12.6 M	51.4
Adversarial Domain Adaptation	0.593	0.216	12.6 M + 2.78 M	51.4 + 30.9
Lightweight Adversarial Domain Adaptation	0.649	0.227	12.6 M + 191k	51.4 + 2.17
Fourier Domain Adaptation	0.638	0.235	12.6 M	51.4

Results comparison – Images



Ground Truth

BiSeNet Benchmark

Domain Adaptation

Lightweight Domain Adaptation

Fourier Domain Adaptation

Final Considerations

- Semantic segmentation can be very powerful for computer vision applications, but it needs large and fully labeled image datasets
- A possible approach is offered by synthetic images, which are more easily labeled, but domain adaptation is needed
- Domain adaptation methods, based on adversarial networks or other algorithms, are promising but further research is needed to reach performances close to the ones of directly trained networks

Thanks for your attention