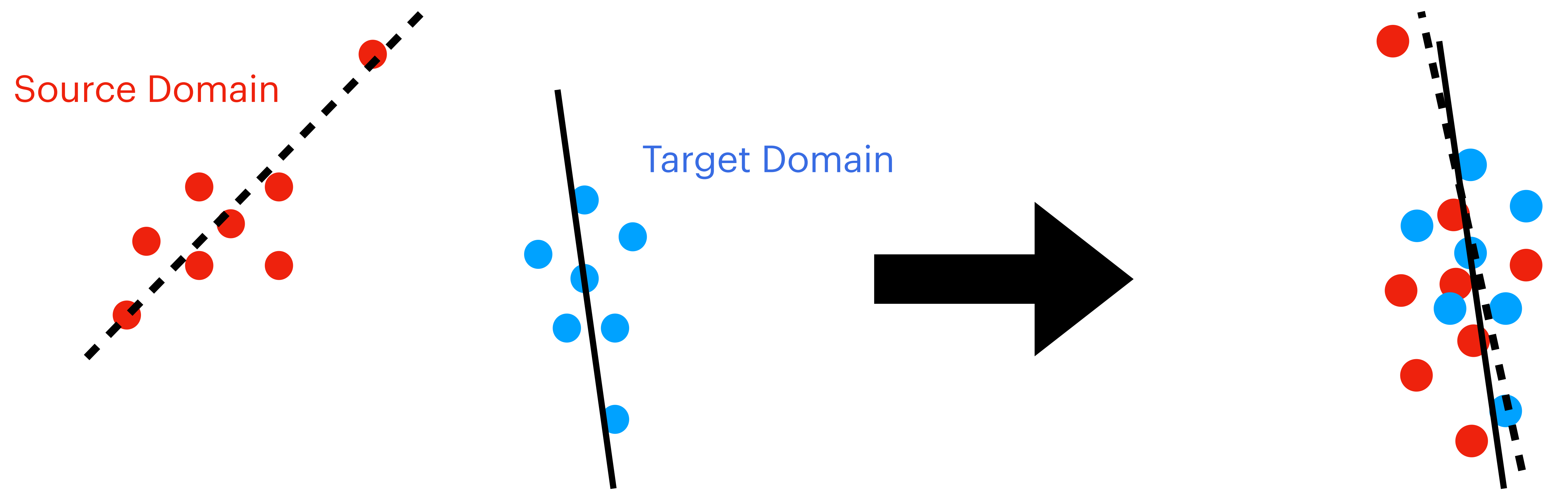
Robust Domain Adaptation

8 ECTS Semester Project - CSE

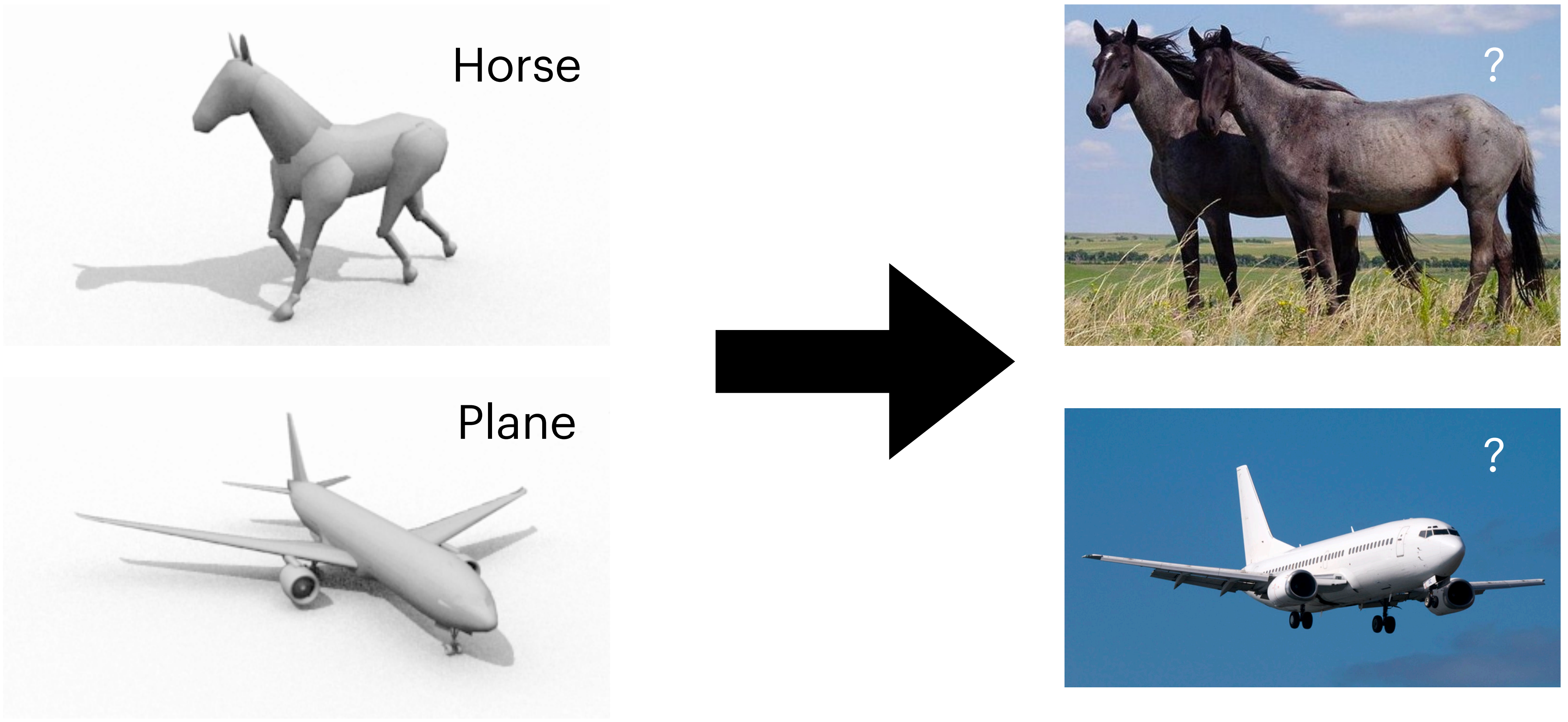
Eelis Mielonen

Motivation

- In many practical ML applications, there is an inevitable distribution shift between training data and real world data. This results in performance degradation.
- You can try to retrain continuously on the target domain, but you can't always obtain labels.
- Domain Adaptation and transfer learning methods have been developed to address this problem.
- We're interested in the case where we have no labels in the target domain, ie. **Unsupervised Domain Adaptation (UDA)**
- An outstanding question is whether existing UDA techniques are robust to labelling errors, adversarial attacks, etc. in the target domain.



Examples - Visda2017 and PACS Datasets

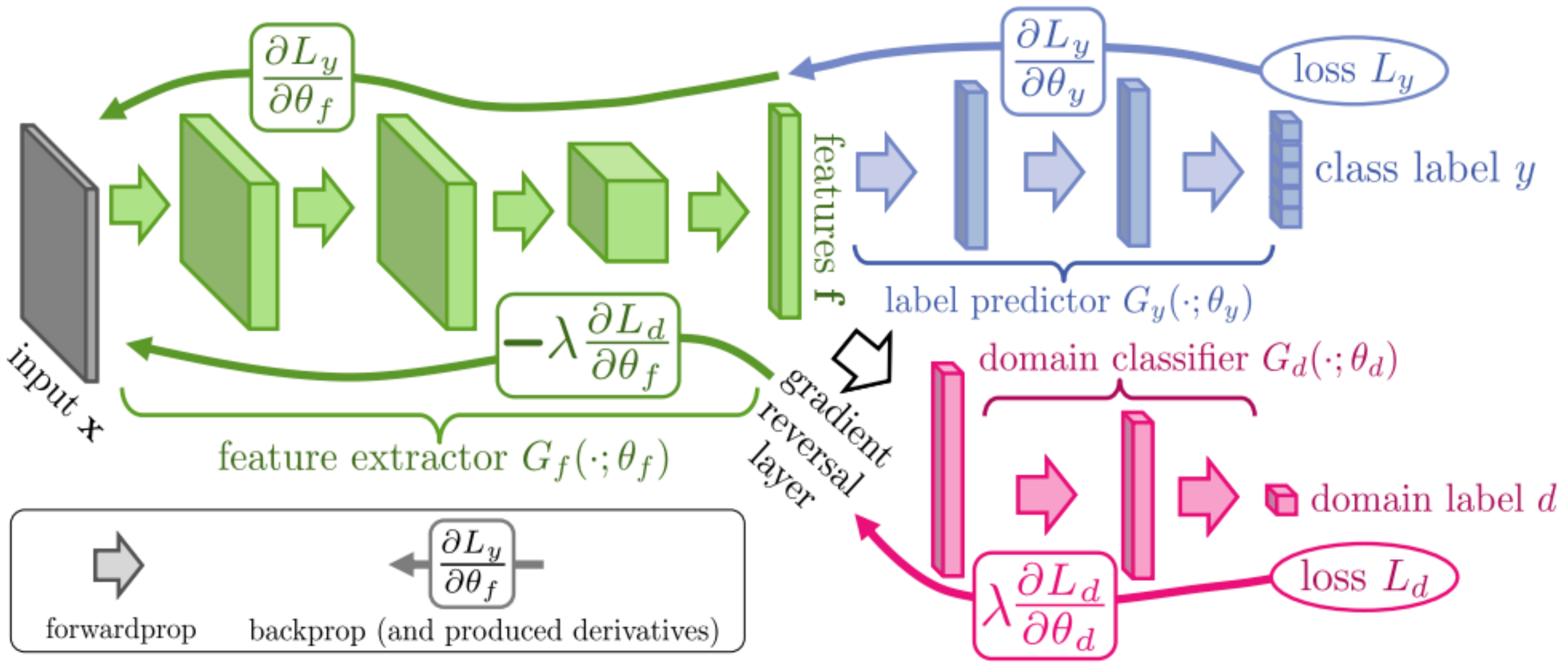


(a) Photo (b) Art painting (c) Cartoon (d) Sketch

Related Work

Most current UDA methods rely on balancing two objectives in a min-max game:

1. **Maximising** domain confusion / minimising domain discrepancy
2. **Minimising** model classification loss



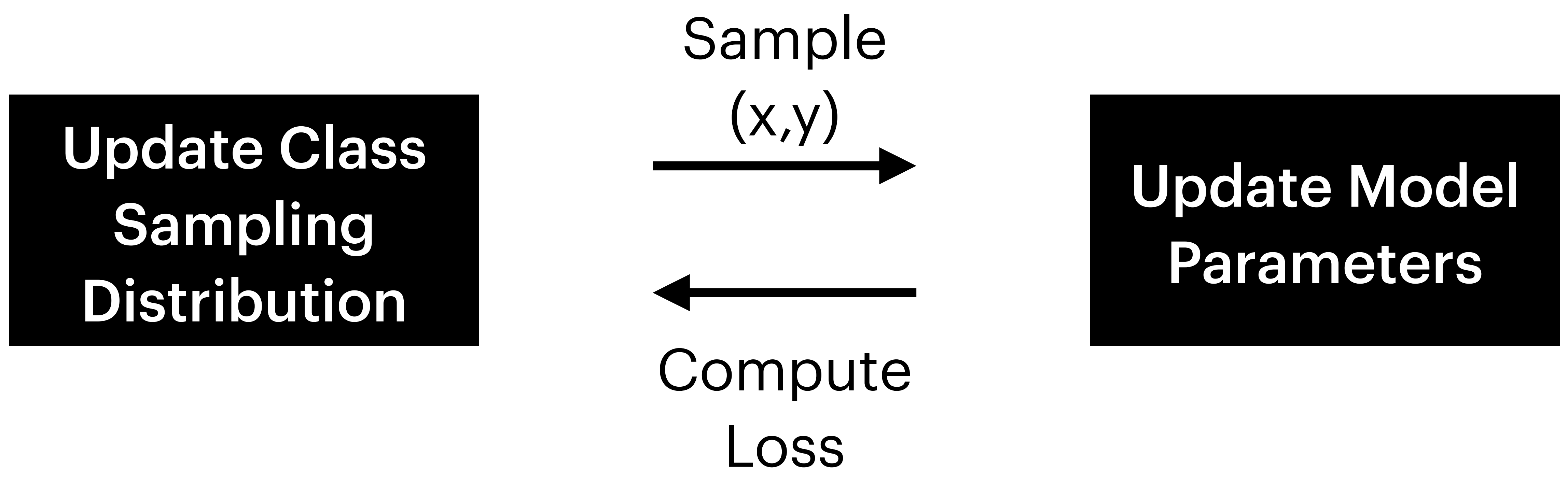
- Domain-Adversarial Training of Neural Networks (2015 - ICML)
- Bridging Theory and Algorithm for Domain Adaptation (2019 - ICML)
- Reusing the Task-specific Classifier as a Discriminator: Discriminator-free Adversarial Domain Adaptation (2022 - CVPR)
- Revisiting adversarial training for the worst-performing class (2022 - ?)

Methods

- In adversarial training, even when the average accuracy of the model is acceptable, some classes can be noticeably worse than others. **Not ok** in some applications.
- **Class Focused Online Learning:** Instead of minimising the average risk, we minimise maximum **class conditioned** risk (ie. the weakest link):

$$\min_{\theta} \mathbb{E}_{(x,y)\sim P(x,y)}[L(\theta, x, y)] \quad \Rightarrow \quad \min_{\theta} \max_{y\in[k]} \mathbb{E}_{x\sim P(\cdot|y)}[L(\theta, x, y)]$$

- This objective can be minimised by cycling between two steps:



Initial Experiments: Diagnosing Problems with UDA

- Selected a few competitive baselines (MCC, MDD, DALN), and common benchmarks for unsupervised domain adaptation (Visda2017, MNIST-MNIST-M, Office-31).
- We measured the accuracies across classes for the source domains and the target domains.
- We also measured the effect of adversarial attacks on both source and target domains.

Hard to Classify Examples

- Observation: the weakest classes in the source domain are likely to be the weakest classes in the target domain.
- As expected, relative differences in per-class accuracies are amplified in the target domain.
- The idea is to use Class Focused Online Learning to try to close the performance gap between classes in the target domain as well.

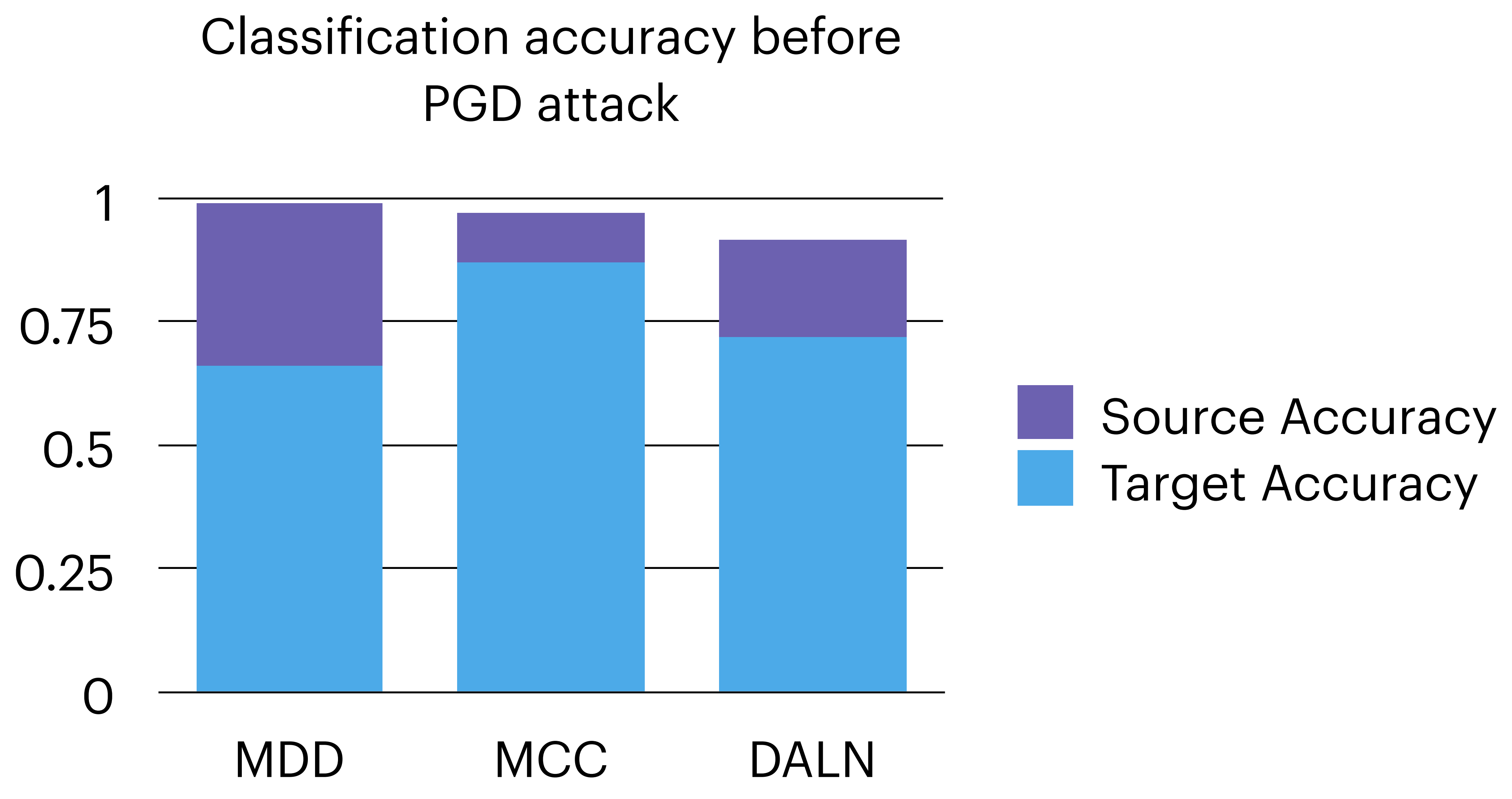
	MNIST (Source)	**MNIST-M (Target)**
Weakest Classes	Nine, Eight	Eight, Nine
Performance Gap	1.2%	2.5%

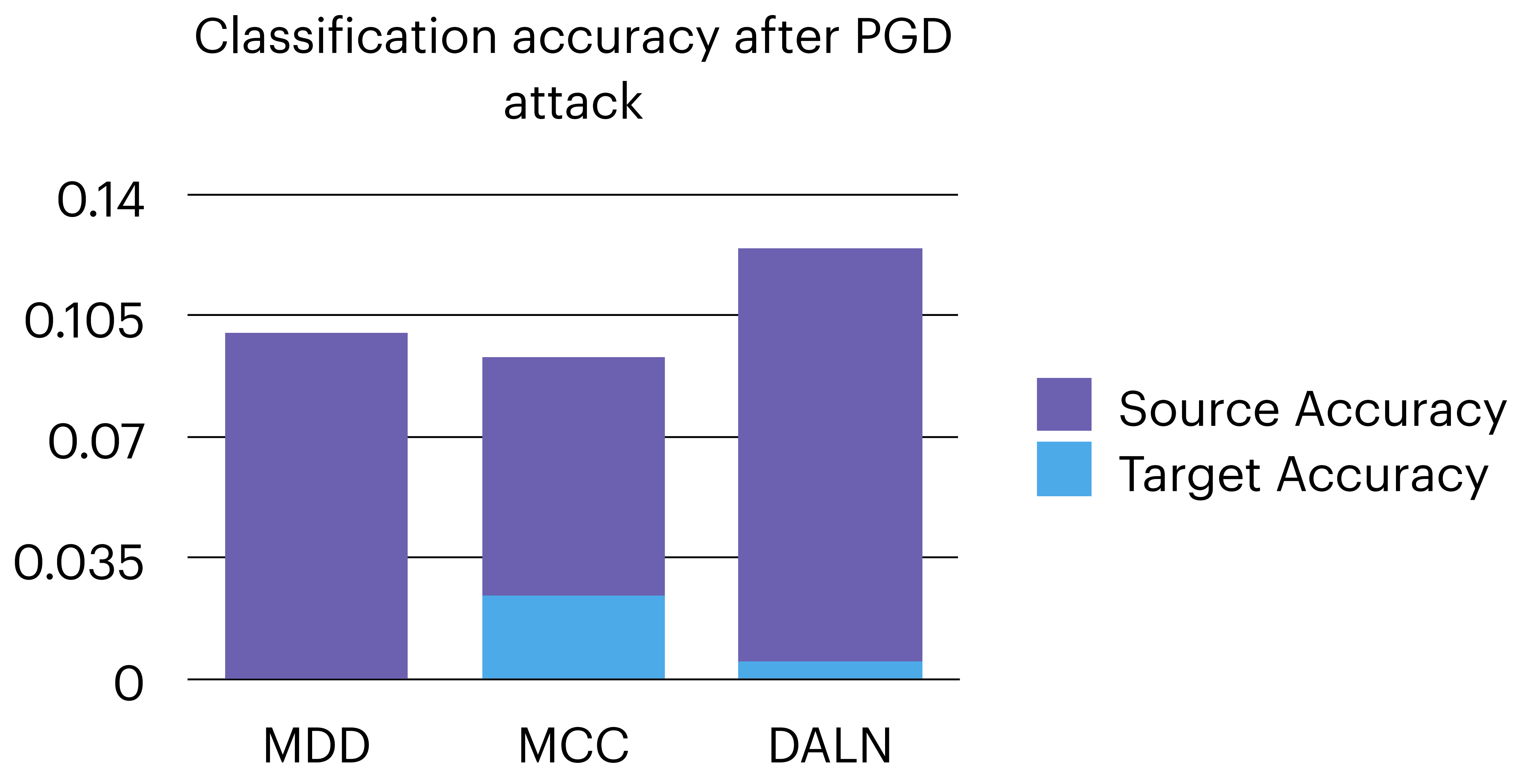
	Visda Synthetic (Source)	**Real (Target)**
Weakest Classes	Car, Bus, Truck	Bus, Car, Truck
Performance Gap	3%	43%

Performance Gap = Best Class Accuracy - Worst Class Accuracy

Adversarial Attacks

- As expected, adversarial attacks are also more powerful on the target domain.





Idea:

- Apply adversarial training with CFOL in the source domain, observe whether improves robustness in both domains.

Next Steps (open to discussion)

The next steps would be to try to fix the diagnosed issues

- Measure per-class accuracies with CFOL.
- Measure robust per-class accuracies with and without CFOL.
- Write the report.