# Deep Learning Algorithms for Galaxy Merger Identification

PHYS 449 Final Presentation

Gabriel Carvalho, Urja Nandivada,
Agniya Pobereshnikova, Brian Ruan

Image Credit: ESA/Hubble

# The Paper

## DeepMerge II: Building Robust Deep Learning Algorithms for Merging Galaxy Identification Across Domains

Ciprijanovic A., Kafkes D., Downey K., Jenkins S., Perdue G.~N., Madireddy S., Johnston T., et al., 2021, MNRAS, 506, 677

| | |
|---|---|
| Paper | arxiv.org/pdf/2103.01373v1.pdf |
| GitHub | github.com/AleksCipri/DeepMergeDomainAdaptation |
| Data | zenodo.org/record/4507941 |

# Review

Image Credit: ESA/Hubble

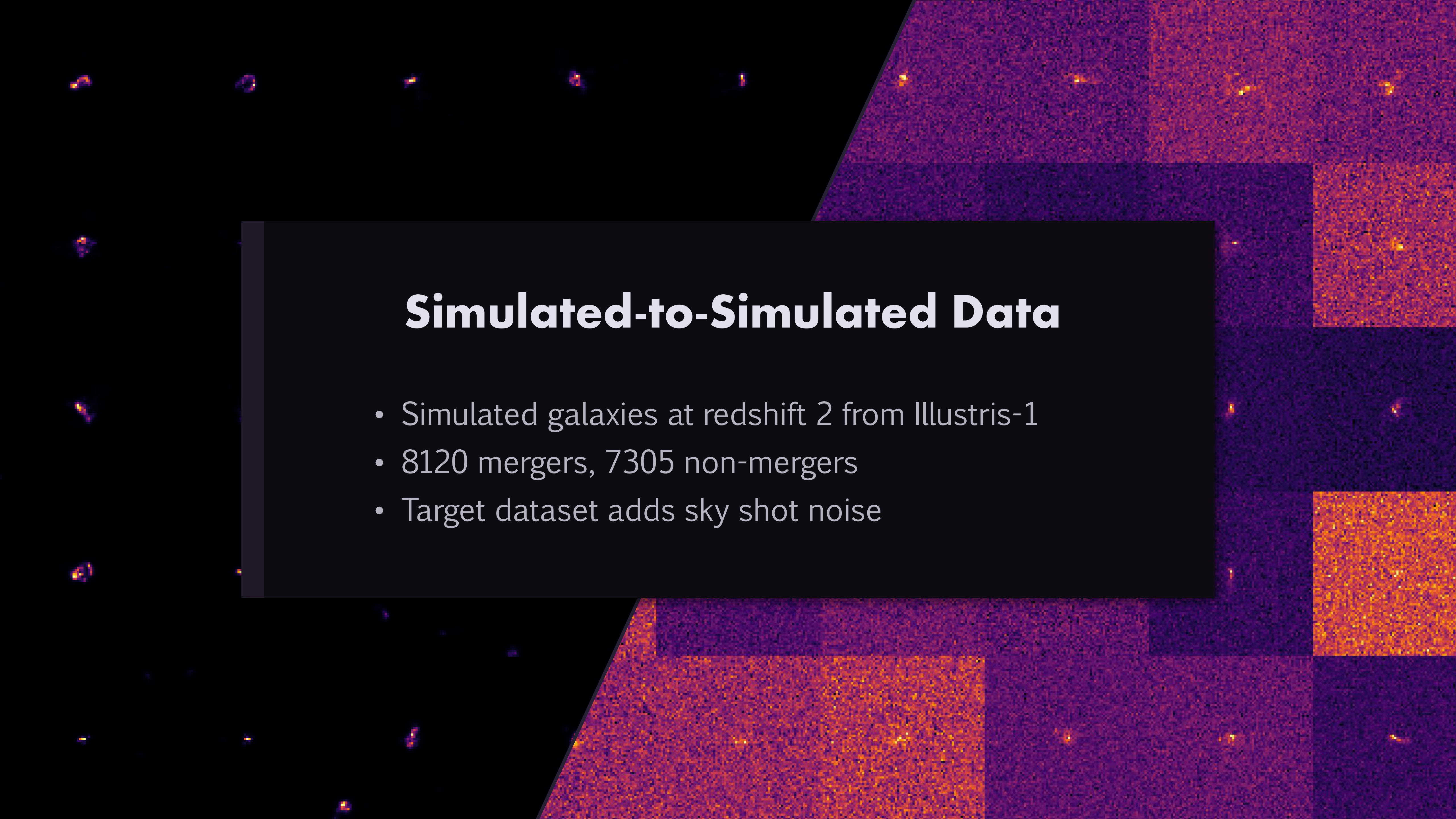
# Goals

1. Binary classification: label images as either **non-interacting** galaxies or **mergers** of multiple galaxies

2. Use **domain adaptation** techniques to train a classifier that works on multiple datasets

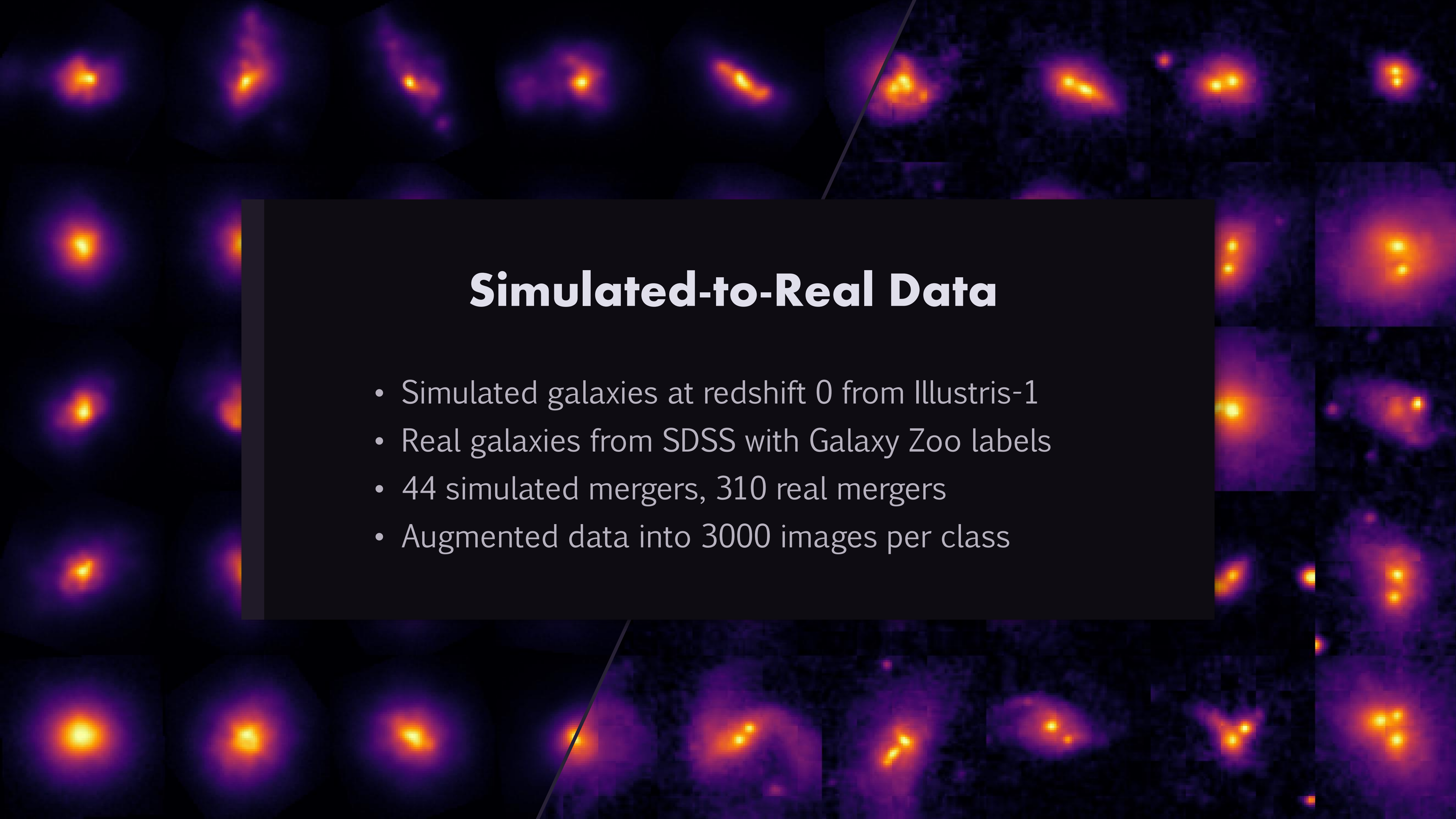
# Domain Adaptation

- Train on data from **two separate domains** (e.g., clipart images of items, and images of the real items)
- The **source** domain is **labeled**, and the **target** domain is **unlabeled**
- DA techniques should help the network find **domain-invariant features**



# Simulated-to-Simulated Data

- Simulated galaxies at redshift 2 from Illustris-1
- 8120 mergers, 7305 non-mergers
- Target dataset adds sky shot noise



# Simulated-to-Real Data

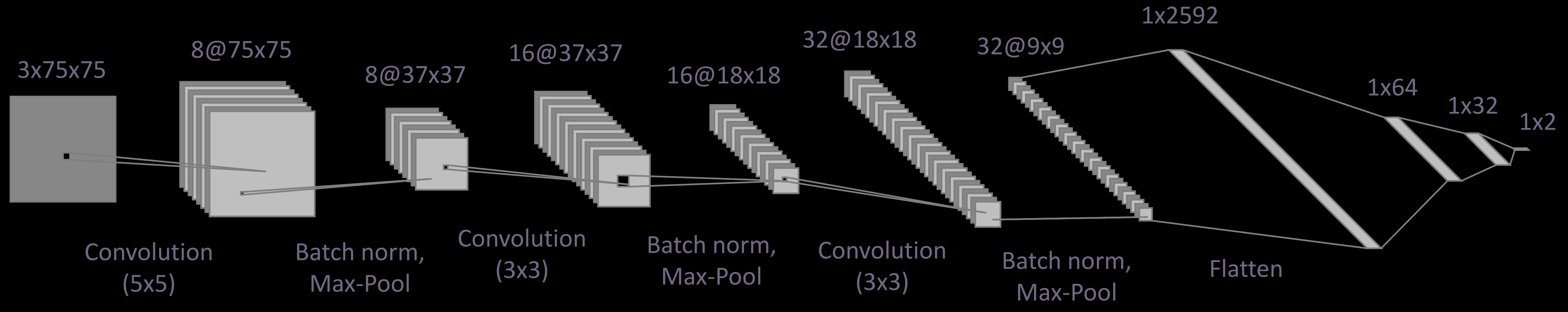
- Simulated galaxies at redshift 0 from Illustris-1
- Real galaxies from SDSS with Galaxy Zoo labels
- 44 simulated mergers, 310 real mergers
- Augmented data into 3000 images per class

# Architecture

DeepMerge

The convolutional feature extractor computes a set of 2592 **features**.

The fully-connected label predictor ends in a 2-class **logit**.



## Architecture

ResNet18

The paper also ran each experiment with the popular ResNet18 architecture, which is over 100x larger than DeepMerge.

The results from ResNet18 were not as good as DeepMerge, so we have focused on reproducing the DeepMerge results only.

# Paper Results

## Simulated-to-Simulated Experiments

| Loss | Target Domain Accuracy |
|---|---|
| No Domain Adaptation | 0.58 |
| Maximum Mean Discrepancy | 0.77 |
| MMD + Fisher + Entropy | 0.77 |
| Adversarial | 0.79 |
| Adv. + Fisher + Entropy | 0.74 |

# Paper Results

Simulated-to-Real Experiments

| Loss | Target Domain Accuracy |
|---|---|
| No Domain Adaptation | 0.50 |
| MMD | 0.53 |
| Transfer Learning + MMD | 0.69 |

Image Credit: ESA/Hubble

# Our Results

## Hyperparameters and Training

We used the hyperparameters stated in the paper, except for cycle length.

We opted to use a fixed learning rate instead of the “one-cycle” learning rate scheduler used in the paper.

Use early stop in some case to prevent overfitting

# No Domain Adaptation



| Metric | Source (Paper) | Source (Us) |
|---|---|---|
| Accuracy | 0.85 | 0.84 |

| Metric | Target (Paper) | Target (Us) |
|---|---|---|
| AUC | 0.74 | 0.74 |
| Accuracy | 0.58 | 0.63 |
| Precision | 0.99 | 0.84 |
| Recall | 0.08 | 0.39 |
| F1 score | 0.14 | 0.54 |
| Brier score | 0.47 | 0.37 |

The baseline experiment. Trained on source using cross-entropy loss.

- Over all results very similar to paper
- The network classified almost all images as mergers, resulting in a low accuracy

# Maximum Mean Discrepancy



Calculates the distance between mean embeddings of the source and target probability distribution from the L infinity norm

## Training Procedure

1. Concatenate batches from source and target datasets
2. Do forward pass to get features (output of convolutional layers) and logits (predicted labels)
3. MMD loss is calculated by comparing the features from the source and target

# Maximum Mean Discrepancy



| Metric | Source (Paper) | Source (Us) |
|---|---|---|
| Accuracy | 0.87 | 0.80 |

| Metric | Target (Paper) | Target (Us) |
|---|---|---|
| AUC | 0.85 | 0.83 |
| Accuracy | 0.77 | 0.76 |
| Precision | 0.81 | 0.76 |
| Recall | 0.72 | 0.80 |
| F1 score | 0.76 | 0.78 |
| Brier score | 0.17 | 0.24 |

- Used early stopping, best result after 30 epochs
- Our source domain accuracy was lower than the paper, but target results were good

# MMD + Fisher + Entropy



## Fisher Loss

Minimizing this increases class compactness and between-class separability in feature space for labelled source classes. Improves source domain accuracy.

## Entropy Minimization

Loss calculated from the entropy of the predicted labels only. Ensure better generalization of decision boundry between source domain classes to the target domain

$$L_{\mathrm{EM}} = -\sum_{j}^{n}\sum_{m}^{M} p(y_m|\boldsymbol{h}_{mj}) \log p(y_m|\boldsymbol{h}_{mj})$$

## Training Procedure

1. Initialize a Fisher class with random centroids before training
2. Concatenate batches from source and target datasets
3. Do forward pass to get features and logits
4. Entropy loss is calculated using all logits
5. Fisher loss is calculated with the source features and true labels

# MMD + Fisher + Entropy



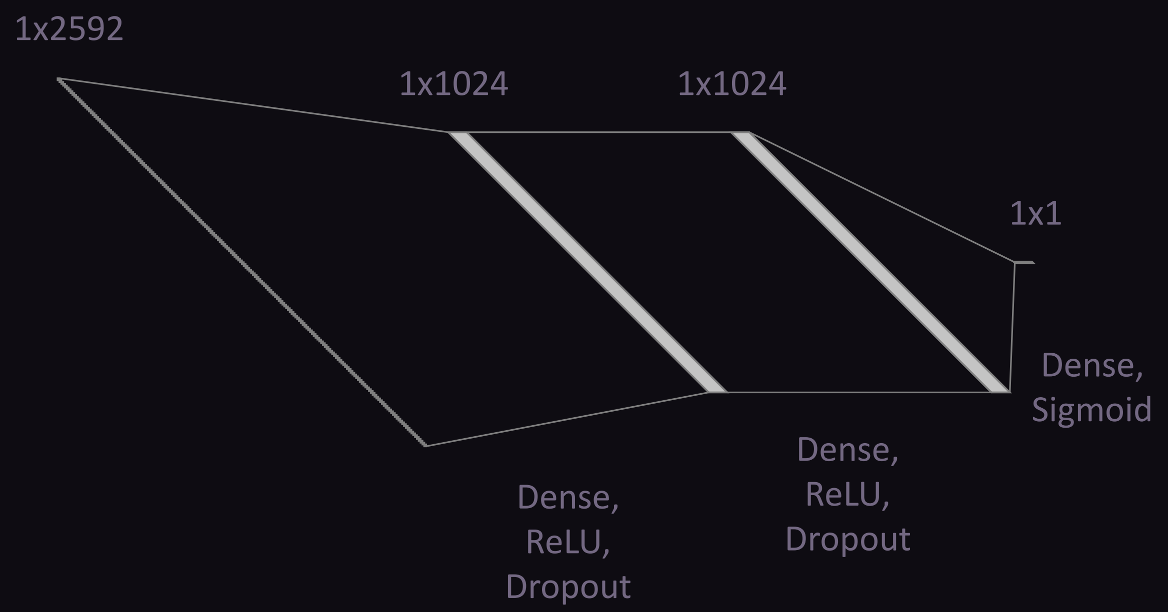
| Metric | Source (Paper) | Source (Us) |
|---|---|---|
| Accuracy | 0.84 | 0.78 |

| Metric | Target (Paper) | Target (Us) |
|---|---|---|
| AUC | 0.86 | 0.81 |
| Accuracy | 0.77 | 0.72 |
| Precision | 0.79 | 0.78 |
| Recall | 0.75 | 0.69 |
| F1 score | 0.77 | 0.73 |
| Brier score | 0.16 | 0.28 |

- Did not produce better target domain results
- Fisher loss resulted in a small decrease in source accuracy due to finding better domain-invariant features

# Adversarial

Adds a domain classifier network built from dense layers. A gradient reversal layer is used to increase loss when the classifier can distinguish between domains.




## Training Procedure

1. Concatenate batches from source and target datasets
2. Do forward pass on base network to get features and logits
3. All features are fed through the gradient reversal layer, then domain classifier to get domain labels
4. Calculate binary cross-entropy loss between predicted and true domains.

# Adversarial



| Metric | Source (Paper) | Source (Us) |
|---|---|---|
| Accuracy | 0.87 | 0.80 |

| Metric | Target (Paper) | Target (Us) |
|---|---|---|
| AUC | 0.87 | 0.82 |
| Accuracy | 0.79 | 0.74 |
| Precision | 0.79 | 0.78 |
| Recall | 0.81 | 0.75 |
| F1 score | 0.80 | 0.76 |
| Brier score | 0.16 | 0.26 |

- Possibly turning off early stop or setting it to a larger early stop tolerance, to improve results
- Results may have been better with longer training

# Adv. + Fisher + Entropy



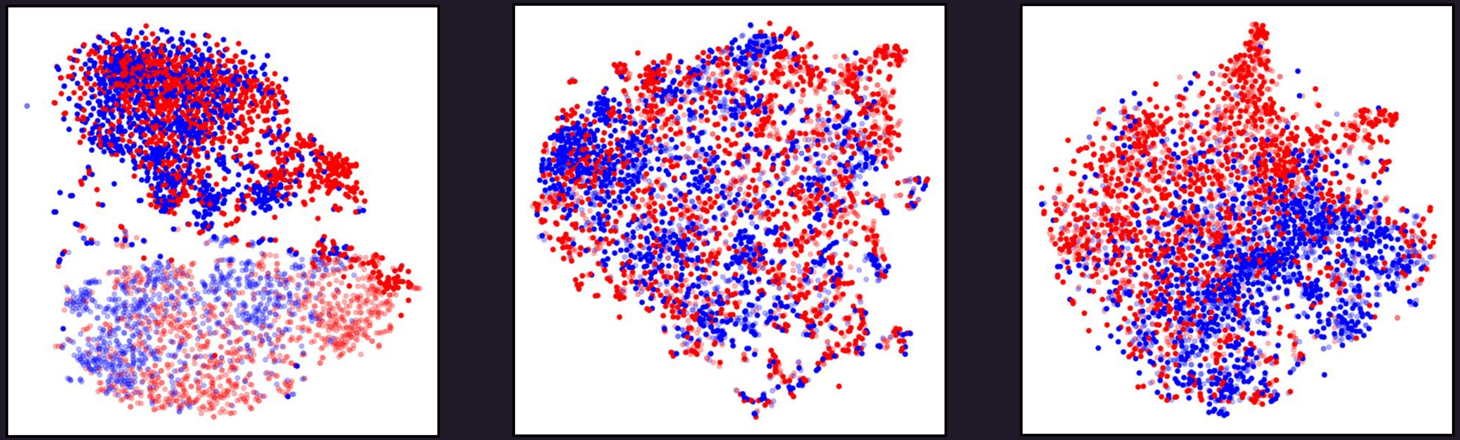
| Metric | Source (Paper) | Source (Us) |
|---|---|---|
| Accuracy | 0.87 | 0.77 |

| Metric | Target (Paper) | Target (Us) |
|---|---|---|
| AUC | 0.82 | 0.83 |
| Accuracy | 0.74 | 0.75 |
| Precision | 0.78 | 0.79 |
| Recall | 0.68 | 0.73 |
| F1 score | 0.73 | 0.76 |
| Brier score | 0.21 | 0.25 |

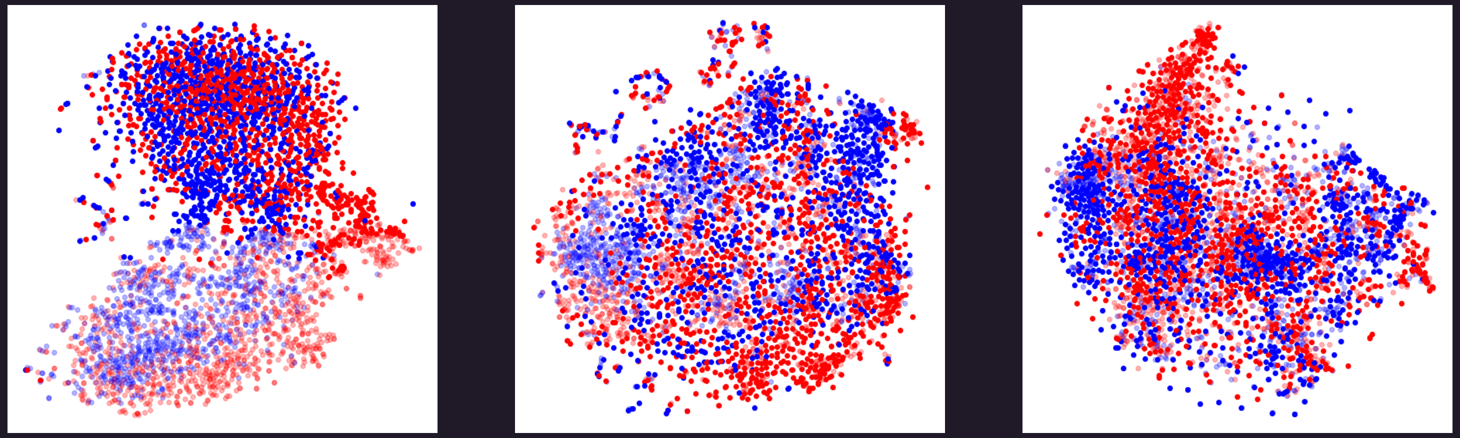
Note: we did not use the domain classifier learning rate multiplier from the paper for this experiment

- Like with simulated-simulated, source accuracy was lower due to finding domain-invariant features
- Results may have been better with longer training

## From the paper



No DA | MMD | MMD + F + EM



## From our experiments

### t-SNE Plots

These plots visualize the distance between features extracted by the network in different domains.

Red dots are mergers

Blue dots are non-mergers

Opaque dots are the source domain

Transparent dots are the target domain

Moving on to the **Simulated-to-Real** experiments

# No Domain Adaptation



| Metric | Source (Paper) | Source (Us) |
|---|---|---|
| Accuracy | 0.92 | 0.83 |

| Metric | Target (Paper) | Target (Us) |
|---|---|---|
| AUC | 0.58 | 0.65 |
| Accuracy | 0.50 | 0.62 |
| Precision | 0.50 | 0.56 |
| Recall | 0.80 | 0.82 |
| F1 score | 0.62 | 0.67 |
| Brier score | 0.49 | 0.39 |

The baseline experiment. Trained on source using cross-entropy loss

- Different results from sim-sim
  - Precision low, accuracy high
  - Classified more as non-mergers

# MMD



| Metric | Source (Paper) | Source (Us) |
|---|---|---|
| Accuracy | 0.94 | 0.79 |

| Metric | Target (Paper) | Target (Us) |
|---|---|---|
| AUC | 0.60 | 0.72 |
| Accuracy | 0.53 | 0.68 |
| Precision | 0.53 | 0.68 |
| Recall | 0.63 | 0.61 |
| F1 score | 0.58 | 0.64 |
| Brier score | 0.40 | 0.32 |

- Trade off between source and target.
- Not sure how paper decided on stopping parameters
- Source would most likely overfit if we trained for longer

# Transfer Learning + MMD



| Metric | Source (Paper) | Source (Us) |
| --- | --- | --- |
| Accuracy | 0.83 | 0.64 |

| Metric | Target (Paper) | Target (Us) |
| --- | --- | --- |
| AUC | 0.76 | 0.50 |
| Accuracy | 0.69 | 0.47 |
| Precision | 0.68 | 0.44 |
| Recall | 0.74 | 0.49 |
| F1 score | 0.71 | 0.47 |
| Brier score | 0.23 | 0.53 |

This experiment started from our best MMD + Fisher + Entropy network from our Simulated-to-Simulated experiments.

- Clearly not same results as paper
- Training on more epochs might have helped out
  - Took a while to stabilize

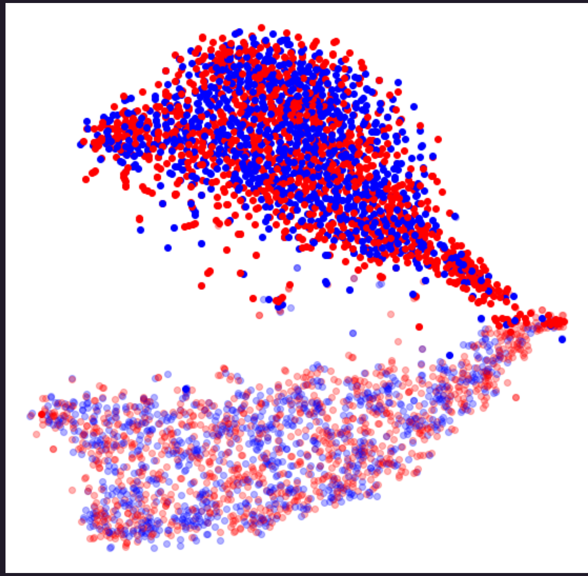
## Note

Adversarial

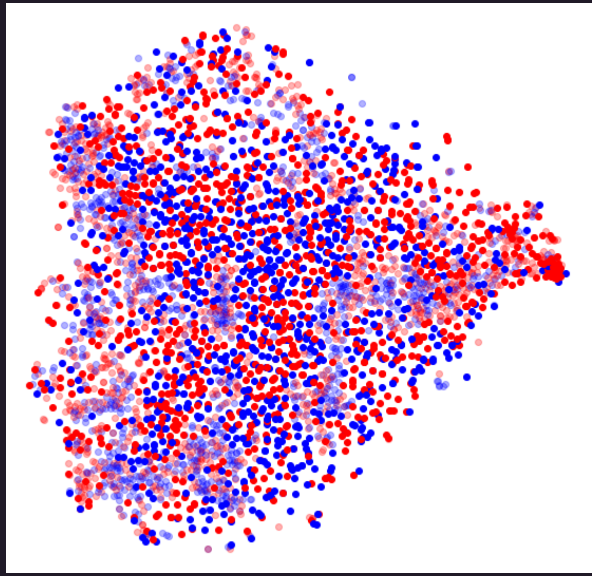
The paper stated that simulated-to-real experiments using adversarial methods were not successful.

# t-SNE Plots for Simulated-to-Real

No DA



MMD



Transfer + MMD

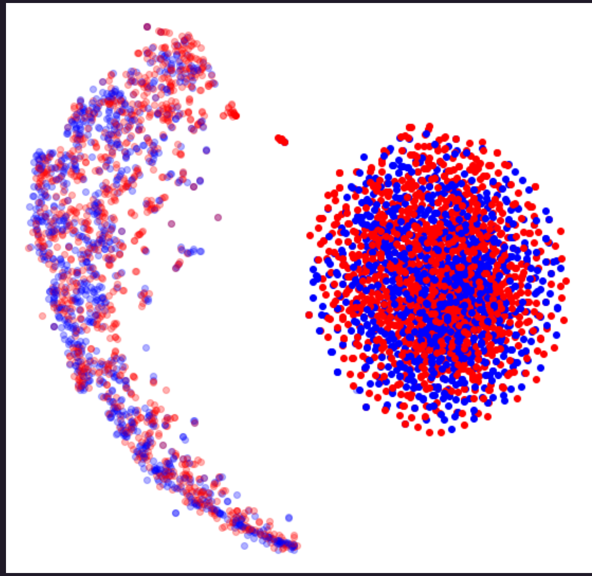


Image Credit: ESA/Hubble

# Conclusions

# Key Take-Aways

1. Domain adaptation techniques **can** be used to train a merger classifier on an unlabeled dataset.
2. Fisher loss and entropy minimization are good techniques for finding domain-invariant features
3. DA models are highly sensitive to changes in hyperparameters when there is high discrepancy between domains.

# Possible Next Steps

1. Update the dataset to be larger, decrease domain discrepancy (new simulations?)
2. More complex morphological tasks

# Thank You

# References


- Gretton et al., 2012a, Advances in Neural Information Processing Systems 25. Curran Associates, Inc. Pp. 1205-1213. http://papers.nips.cc/paper/4727-optimal-kernel-choice-for-large-scale-two-sample-tests.pdf
- Gretton et al., 2012b, Journal of Machine Learning Research, 13, 723.
- Ganin et al., 2016, Journal of Machine Learning Research, 17, 1.
- Zhang et al., 2020a, arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2020arXiv200305636Z/abstract
- Grandvalet Y., Bengio Y., 2005, in Saul L. K., Weiss Y., Bottou L., eds, Advances in Neural Information Processing Systems 17. MIT Press, pp. 529-536 http://papers.nips.cc/paper/2740-semi-supervised-learning-by-entropy-minimization.Pdf
- Somla et al., 2007, in Algorithmic Learning Theory, Letcure Notes in Computer Science 4754. Springer, Berlin, Germany, pp. 13-31
- Ciprijanovic et al., 2020, Astronomy and Computing, 32, 100390
- He et al., 2015, arXiv e-prints, https://ui.adsabs.harvard.edu/abs/2015arXiv151203385H/abstract

# Appendix

Simulated-to-Simulated Experiments

# No Domain Adaptation

| Metric | Source (Paper) | Source (Us) | Target (Paper) | Target (Us) |
|---|---|---|---|---|
| AUC | 0.92 | 0.91 | 0.74 | 0.74 |
| Accuracy | 0.85 | 0.83 | 0.58 | 0.63 |
| Precision | 0.88 | 0.88 | 0.99 | 0.84 |
| Recall | 0.83 | 0.81 | 0.08 | 0.39 |
| F1 score | 0.86 | 0.84 | 0.14 | 0.54 |
| Brier score | 0.11 | 0.16 | 0.47 | 0.37 |

Simulated-to-Simulated Experiments

# MMD

| Metric | Source (Paper) | Source (Us) | Target (Paper) | Target (Us) |
|---|---|---|---|---|
| AUC | 0.93 | 0.88 | 0.85 | 0.83 |
| Accuracy | 0.87 | 0.80 | 0.77 | 0.76 |
| Precision | 0.88 | 0.80 | 0.81 | 0.76 |
| Recall | 0.87 | 0.86 | 0.72 | 0.80 |
| F1 score | 0.88 | 0.83 | 0.76 | 0.78 |
| Brier score | 0.10 | 0.20 | 0.17 | 0.24 |

Simulated-to-Simulated Experiments

# MMD + Fisher + Entropy

| Metric | Source (Paper) | Source (Us) | Target (Paper) | Target (Us) |
|---|---|---|---|---|
| AUC | 0.92 | 0.87 | 0.86 | 0.81 |
| Accuracy | 0.84 | 0.77 | 0.77 | 0.72 |
| Precision | 0.87 | 0.83 | 0.79 | 0.78 |
| Recall | 0.84 | 0.75 | 0.75 | 0.69 |
| F1 score | 0.86 | 0.79 | 0.77 | 0.73 |
| Brier score | 0.11 | 0.22 | 0.16 | 0.28 |

Simulated-to-Simulated Experiments

# Adversarial

| Metric | Source (Paper) | Source (Us) | Target (Paper) | Target (Us) |
|---|---|---|---|---|
| AUC | 0.94 | 0.86 | 0.87 | 0.82 |
| Accuracy | 0.87 | 0.80 | 0.79 | 0.74 |
| Precision | 0.88 | 0.83 | 0.79 | 0.78 |
| Recall | 0.89 | 0.79 | 0.81 | 0.75 |
| F1 score | 0.87 | 0.81 | 0.80 | 0.76 |
| Brier score | 0.09 | 0.20 | 0.16 | 0.26 |

Simulated-to-Simulated Experiments

# Adv. + Fisher + Entropy

| Metric | Source (Paper) | Source (Us) | Target (Paper) | Target (Us) |
|---|---|---|---|---|
| AUC | 0.94 | 0.85 | 0.82 | 0.83 |
| Accuracy | 0.87 | 0.77 | 0.74 | 0.75 |
| Precision | 0.90 | 0.82 | 0.78 | 0.79 |
| Recall | 0.86 | 0.76 | 0.68 | 0.73 |
| F1 score | 0.88 | 0.79 | 0.73 | 0.76 |
| Brier score | 0.10 | 0.23 | 0.21 | 0.25 |

Simulated-to-Real Experiments

# No Domain Adaptation

| Metric | Source (Paper) | Source (Us) | Target (Paper) | Target (Us) |
|---|---|---|---|---|
| AUC | 0.97 | 0.87 | 0.58 | 0.65 |
| Accuracy | 0.92 | 0.83 | 0.50 | 0.62 |
| Precision | 0.91 | 0.92 | 0.50 | 0.56 |
| Recall | 0.92 | 0.70 | 0.80 | 0.82 |
| F1 score | 0.92 | 0.79 | 0.62 | 0.67 |
| Brier score | 0.06 | 0.17 | 0.49 | 0.39 |

Simulated-to-Real Experiments

# MMD

| Metric | Source (Paper) | Source (Us) | Target (Paper) | Target (Us) |
|---|---|---|---|---|
| AUC | 0.98 | 0.86 | 0.60 | 0.72 |
| Accuracy | 0.94 | 0.79 | 0.53 | 0.68 |
| Precision | 0.92 | 0.86 | 0.53 | 0.68 |
| Recall | 0.95 | 0.68 | 0.63 | 0.61 |
| F1 score | 0.94 | 0.75 | 0.58 | 0.64 |
| Brier score | 0.05 | 0.21 | 0.40 | 0.32 |

Simulated-to-Real Experiments

# Transfer Learning + MMD

| Metric | Source (Paper) | Source (Us) | Target (Paper) | Target (Us) |
|---|---|---|---|---|
| AUC | 0.90 | 0.64 | 0.76 | 0.50 |
| Accuracy | 0.83 | 0.64 | 0.69 | 0.47 |
| Precision | 0.89 | 0.82 | 0.68 | 0.44 |
| Recall | 0.74 | 0.32 | 0.74 | 0.49 |
| F1 score | 0.80 | 0.46 | 0.71 | 0.47 |
| Brier score | 0.13 | 0.36 | 0.23 | 0.53 |

# Comparisons

Simulated – Simulated data

Average time per epoch using CUDA:

MMD : 8.4 sec

MMD + Fisher + EM : 19.5 sec

Adversarial : 4.5 sec

Adversarial + Fisher + EM : 13.2 sec