



Unsupervised Domain Adaptation

Final Project Presentation

Michele Yin, Roberto Mazzaro, Andrea Bonora, Filippo Daniotti, Giovanni Ambrosi



Outline

1. Introduction
2. Works on DANN
 - DANN + Discrepancy loss
 - DANN + improvement for Adversarial methods and combinations
 - Incremental DANN
3. Gradual self training
 - Comparison with other datasets
 - Ablation study
4. Conclusions
 - Future works
 - Our opinion

Intro - Domain Adaptation

Labelled dataset: when we train networks

- **Pros:** training results in good performances
- **Cons:** expensive -> we can't have a label dataset for each application

Unlabelled data: usually for real applications we only have unlabeled data

- **Pros:** Almost free -> we want to exploit them
- **Cons:** domain shift

$\{X_s, Y_s\}$



$\{X_t\}$



Intro - Our work

Previously:

- General introduction to **Domain Adaptation**
- More specifically on **UDA** for **classification**
- Overview of different UDA techniques
- Test of some standard methods

Today:

- Work over DANN: **improve DANN with other methods and combination** of them
- Gradual self-training: in **depth ablation analysis** of a single method
- All our work can be found here: <https://github.com/filippodaniotti/TACV-DA-project>

Experiment Setting

Ideas:

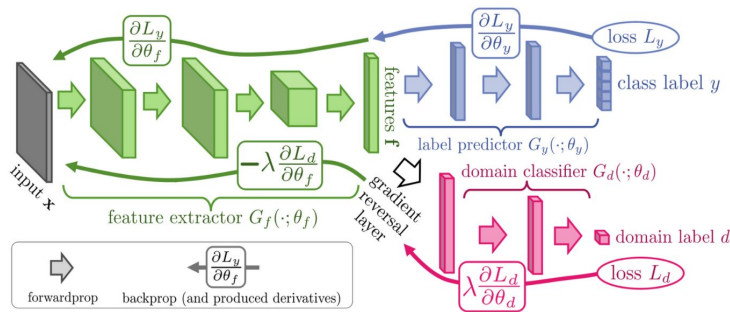
- Start with **DANN**
- **Improve** it with newer methods
- Try to **combine** more than one method together

Why DANN:

- **Simple** -> fast to implement and test
- Investigate if the concept make sense
- Don't look for the greatest accuracy

Dataset: Office31

- **Pros:** more complex than digits, lighter than OfficeHome
- **Cons:** unbalanced in the 2 domains



Office31

WEBCAM (795 images)



AMAZON (2817 images)





DANN + Discrepancy loss

- Starting from DANN
 - Add a loss at the output of the feature extractor
 - Simple and effective
 - Align the features before classification and discrimination

- Maximum Mean Discrepancy Loss

$$\begin{aligned} MMD^2(P, Q) &= \|\mu_P - \mu_Q\|_{\mathcal{F}}^2 \\ &= \mathbb{E}_{x \sim P} [k(x, x')] + \mathbb{E}_{y \sim Q} [k(y, y')] - 2\mathbb{E}_{x, y \sim P, Q} [k(x, y)] \end{aligned}$$

- Coral Loss

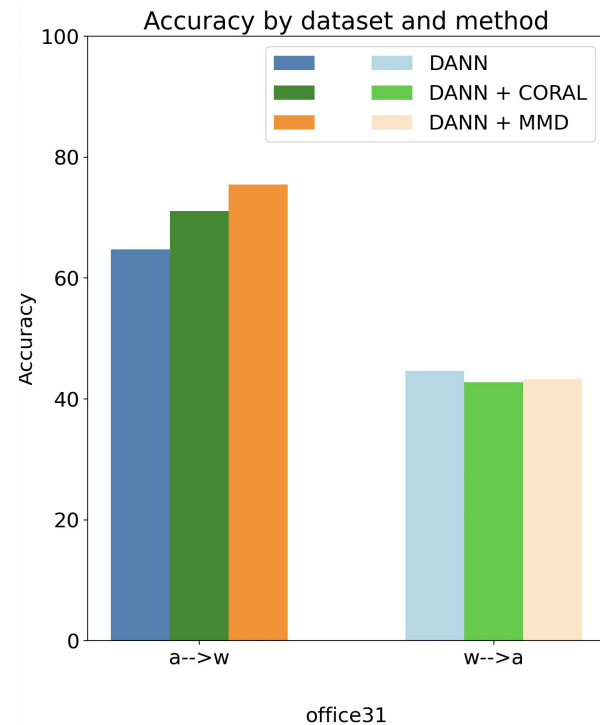
$$\ell_{CORAL} = \frac{1}{4d^2} \|C_S - C_T\|_F^2$$



Obtained Results

- MMD loss best on Amazon -> Webcam domain, gain over 10%
- Coral loss improves of 6%
- Both losses don't provide improvements on Webcam -> Amazon direction, **unbalanced dataset**

Model	A->W A	A->W W	Gain	W->A W	W->A W	Gain
DANN	86.35	64.78	-	93.71	44.68	-
DANN + MMD	83.16	75.47	+10.69	95.60	42.02	-2.66
DANN + Coral	87.06	71.07	+6.29	94.34	43.25	-1.43



Experiment Setting

Ideas:

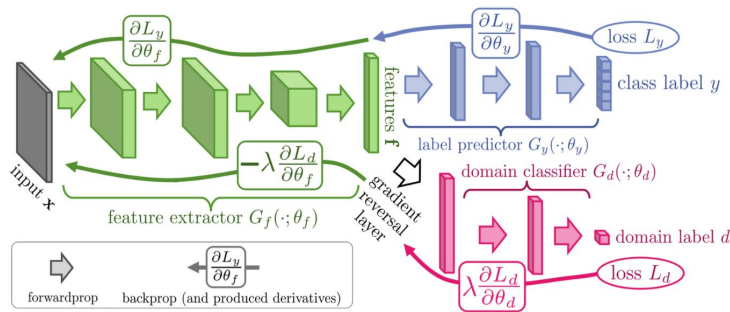
- Start with **DANN**
- **Improve** it with newer methods
- Try to **combine** more than one method together

Why DANN:

- **Simple** -> fast to implement and test
- Investigate if the concept make sense
- Don't look for the greatest accuracy

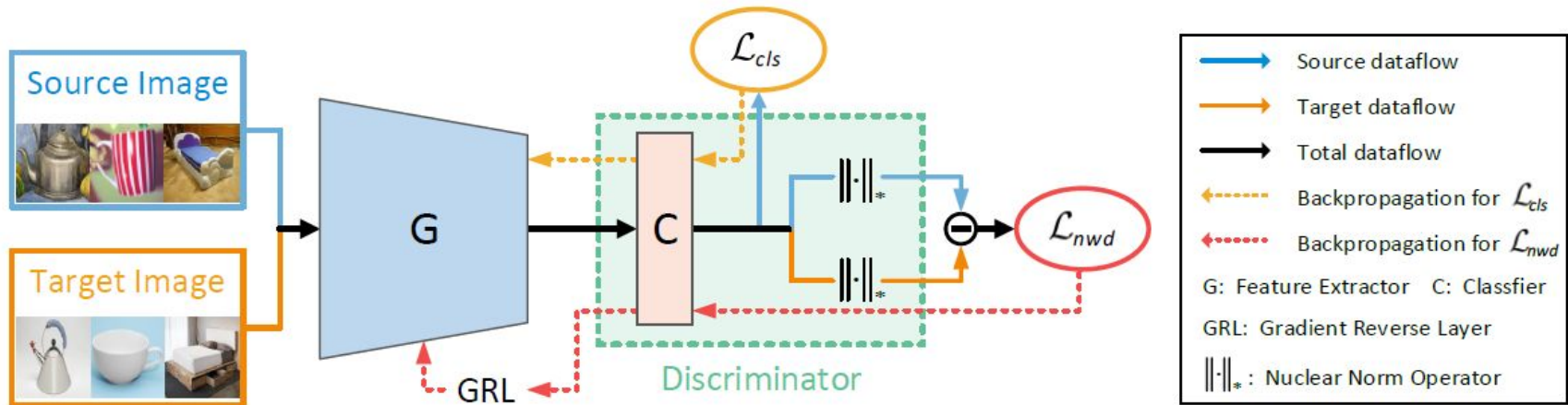
Dataset: Office31

- **Pros:** more complex than digits, lighter than OfficeHome
- **Cons:** unbalanced in the 2 domains



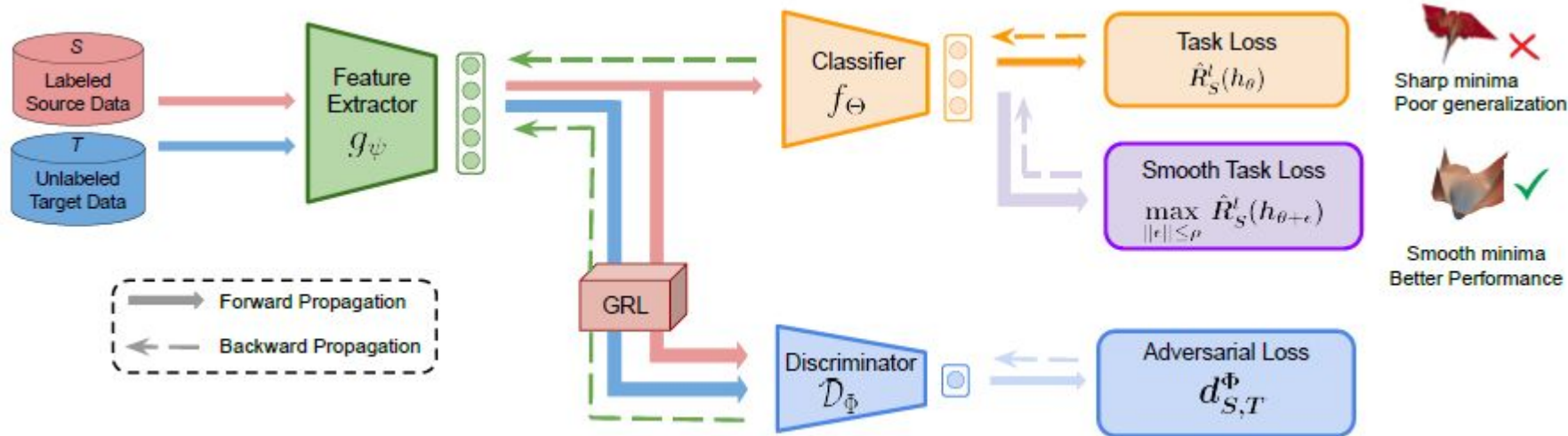
Methods Recap - DALN

- Remove the discriminator
- Use classifier + NWD module to discriminate the domain



Methods Recap - SDAT

- Find **smoother minima** for classification loss
- New optimizer** with additional gradient computation steps

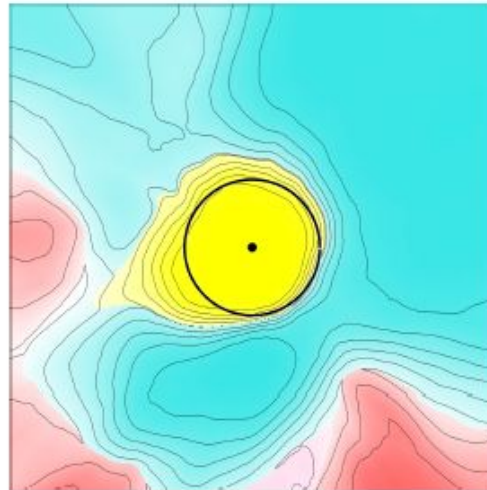


Methods Recap - JREG

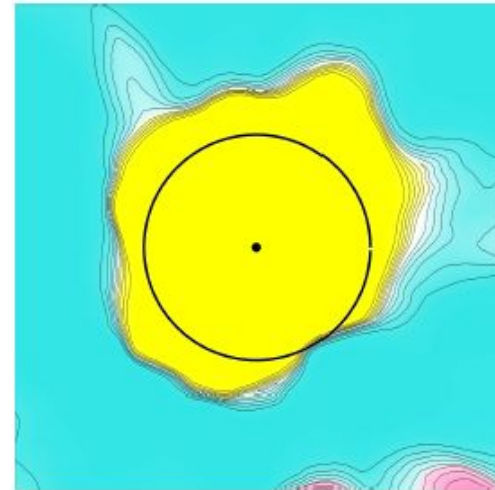
- Regularization method
- Push decision boundaries further away
- Used inside FGDA



(a) Without regularization



(b) With L^2 regularization

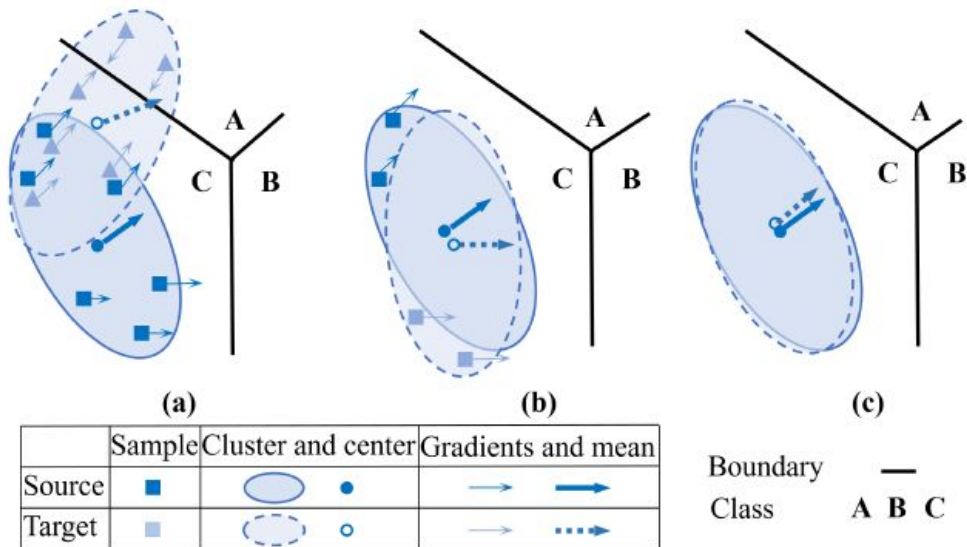


(c) With Jacobian regularization



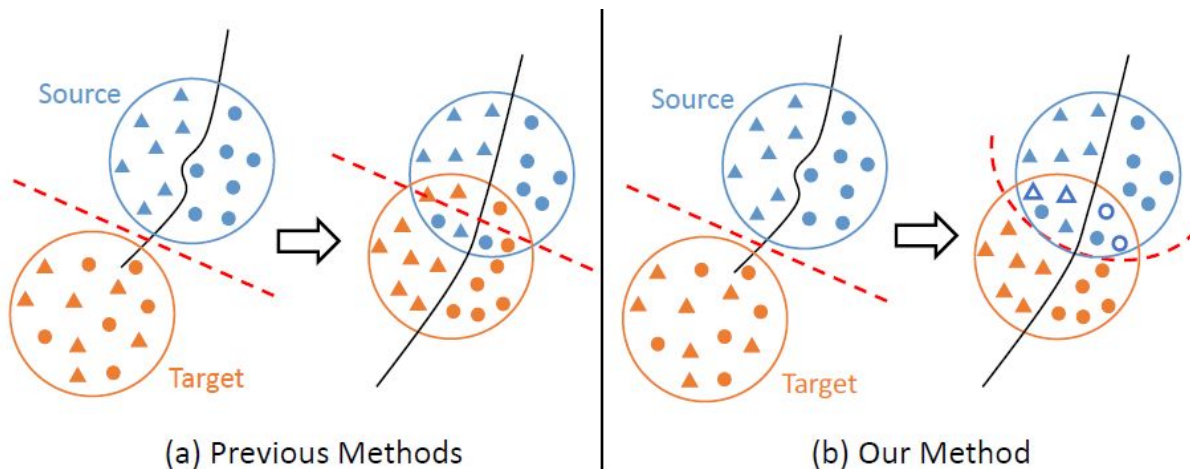
Methods Recap - FGDA

- **Constrain** feature **gradients** of two domains to have **similar distributions**
- Pseudo labels computed to obtain target loss
- Jacobian Regularization used inside



Methods Recap - RADA

- **Relabel** well aligned target **samples** as source domain
 - Well aligned samples -> domain discriminator entropy higher than a threshold
 - **Mixup** at feature level used with relabel samples to softly mix features
 - Domain relabeling doesn't influence classification
-
- **No official implementation available**





Combining methods - How

FGDA + DALN: **no conflicts and lighter model**

- FGDA use an additional grad_discriminator to align gradient distributions
- Adversarial discriminator can be substituted by DALN

RADA + FGDA: **no conflicts**

- RADA change domain labels but doesn't influence classification task
- Just add FGDA

Any + SDAT: SDAT is a different optimizer so can be applied to any method

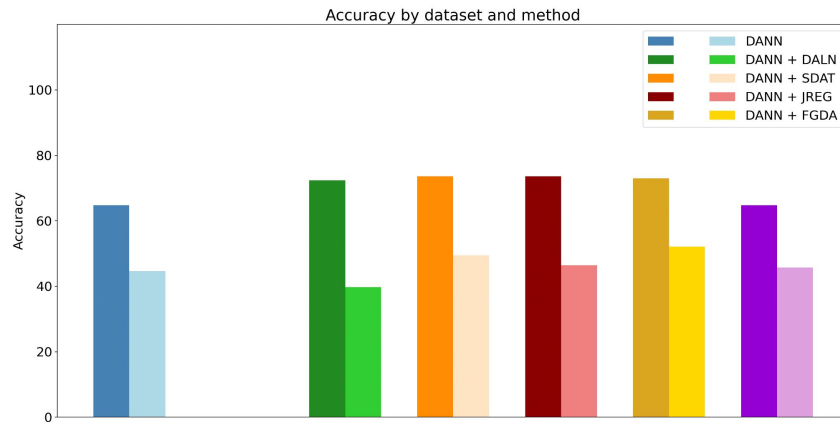
RADA + DALN: **creates conflicts**

- RADA use domain discriminator entropy as policy to re-align samples



Obtained Results

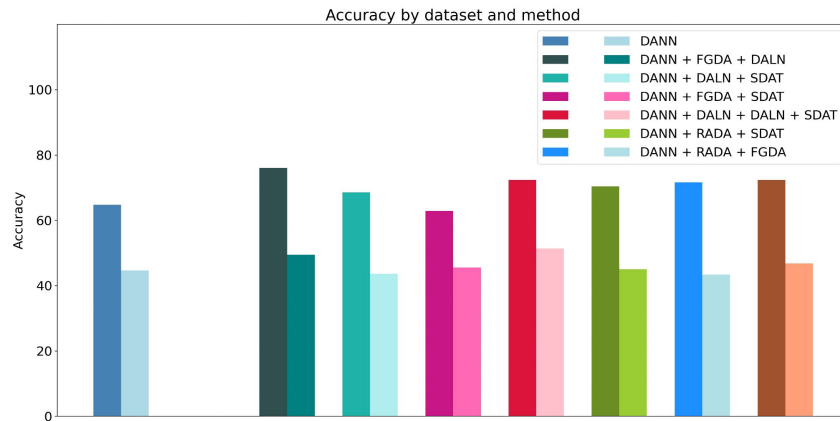
- Almost all methods improve DANN
- RADA in A->W test doesn't improve
 - Neither worsen and relabeling started at epoch 17
 - No official code and training parameters available
- DALN in W->A is suffering the dataset imbalance
- JREG very effective but FGDA improve it a lot in W->A



Model	A->W A	A->W W	Gain	W->A W	W->A W	Gain
DANN	86.35	64.78	-	93.71	44.68	-
DANN + DALN	83.16	72.33	+7.55	95.60	39.72	-4.96
DANN + SDAT	87.06	73.58	+8.80	94.34	49.47	+4.79
DANN + JREG	85.82	73.58	+8.80	94.97	46.45	+1.77
DANN + FGDA	86.35	72.96	+8.18	96.86	52.13	+7.45
DANN + RADA	85.99	64.78	0	93.08	45.74	+1.06

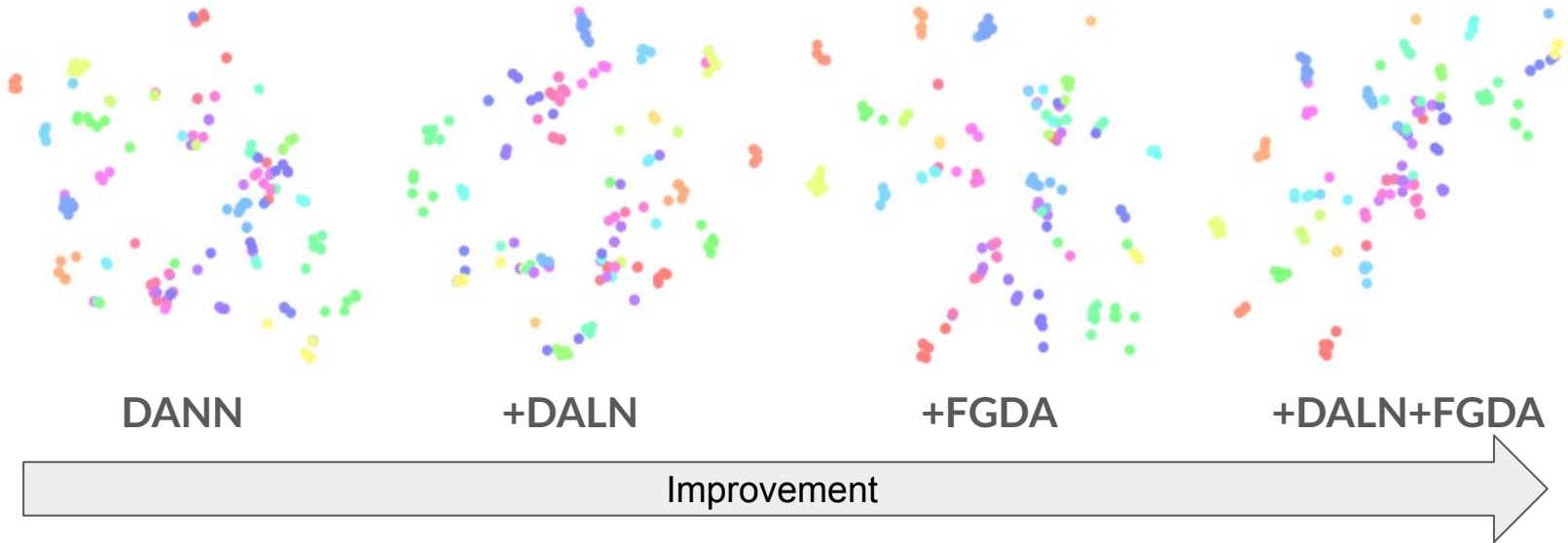
Obtained Results

- FGDA + DALN seems a good idea
 - Best method in A->W test with gain of +11.32
 - In W->A test suffer the poor performances of DALN in this direction
- RADA + FGDA might be a good idea
 - Increase RADA performances
 - Problem are RADA poor performances due to non optimal training params.
- SDAT very sensitive to training params. -> if not well selected decrease performances



Model	A->W A	A->W W	Gain	W->A W	W->A W	Gain
DANN + FGDA + DALN	84.75	76.10	+11.32	93.08	49.47	+4.79
DANN + DALN + SDAT	85.64	68.55	+3.77	91.82	43.62	-1.06
DANN + FGDA + SDAT	81.21	62.89	-1.89	93.71	45.57	+0.89
DANN + DALN + DALN + SDAT	85.82	72.33	+7.55	94.97	51.42	+6.74
DANN + RADA + SDAT	83.87	70.44	+5.66	94.34	45.04	+0.36
DANN + RADA + FGDA	83.87	71.70	+6.92	94.34	43.44	-1.24
DANN + RADA + FGDA + SDAT	84.57	72.33	+7.55	93.71	46.81	+2.13

TSNE Analisy

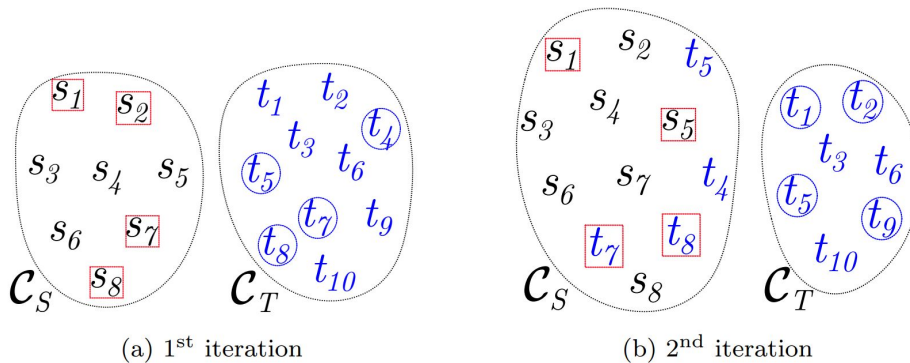


- TSNE plots for A->W test of predicted target domain labels
- Better inter class separation
- Better intra class compactness

Incremental Method

Idea:

- Start from a **trained model**
- Assign a pseudo label to **k samples**
- At each iteration train the model
 - **First**, train the model as usual
 - **Next**, only on the new pseudo labeled samples
- At the end, a model from scratch **only** with the target data

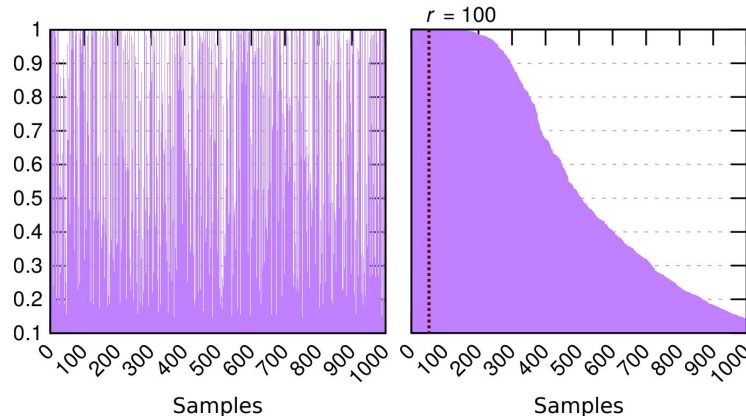




Incremental Method

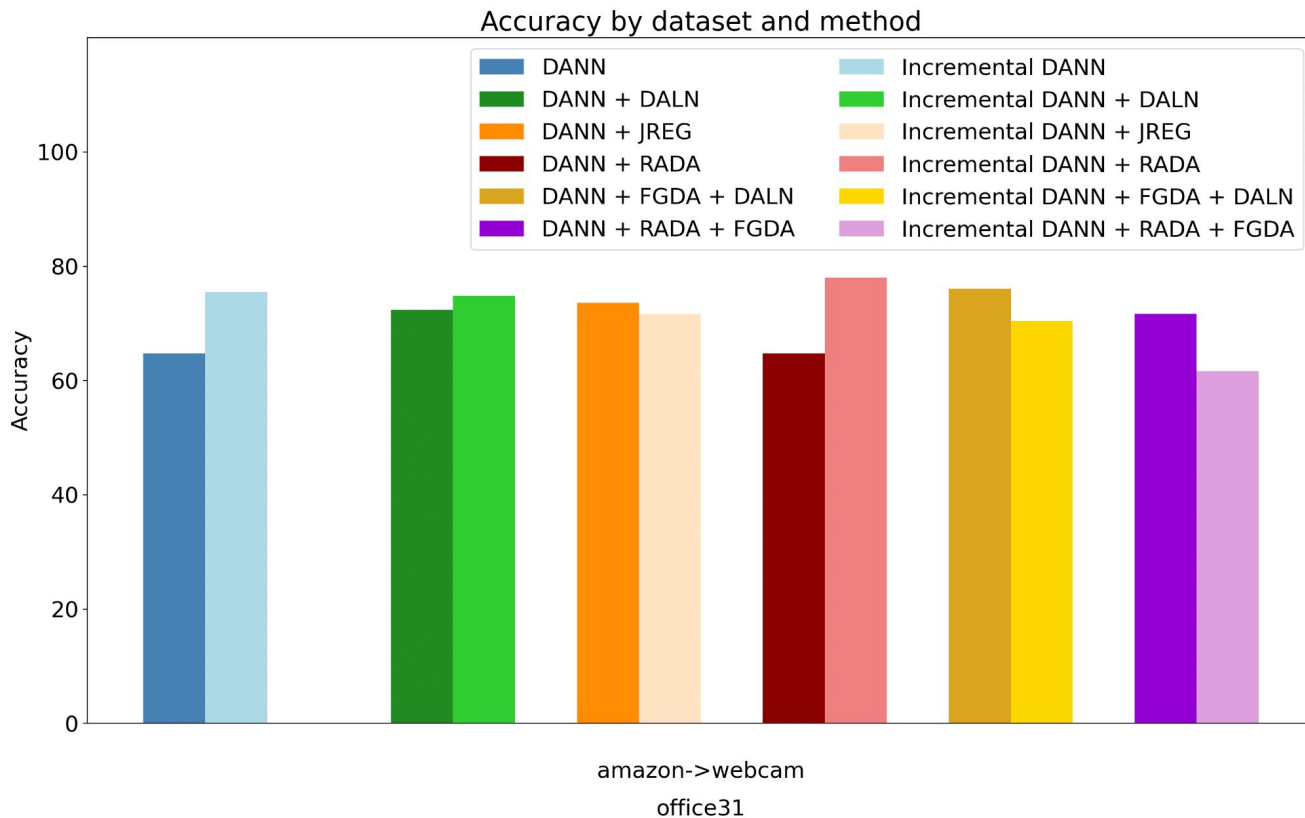
How to assign a label to the data?

- **Confidence policy:** select the samples with the **highest confidence** in the classifier predictions
- **Possible issue:** samples with a very low confidence will distort the training of the model
- **Possible solution:** when the confidence is lower than a certain threshold assign all the remaining data to a label without training the model anymore





Results





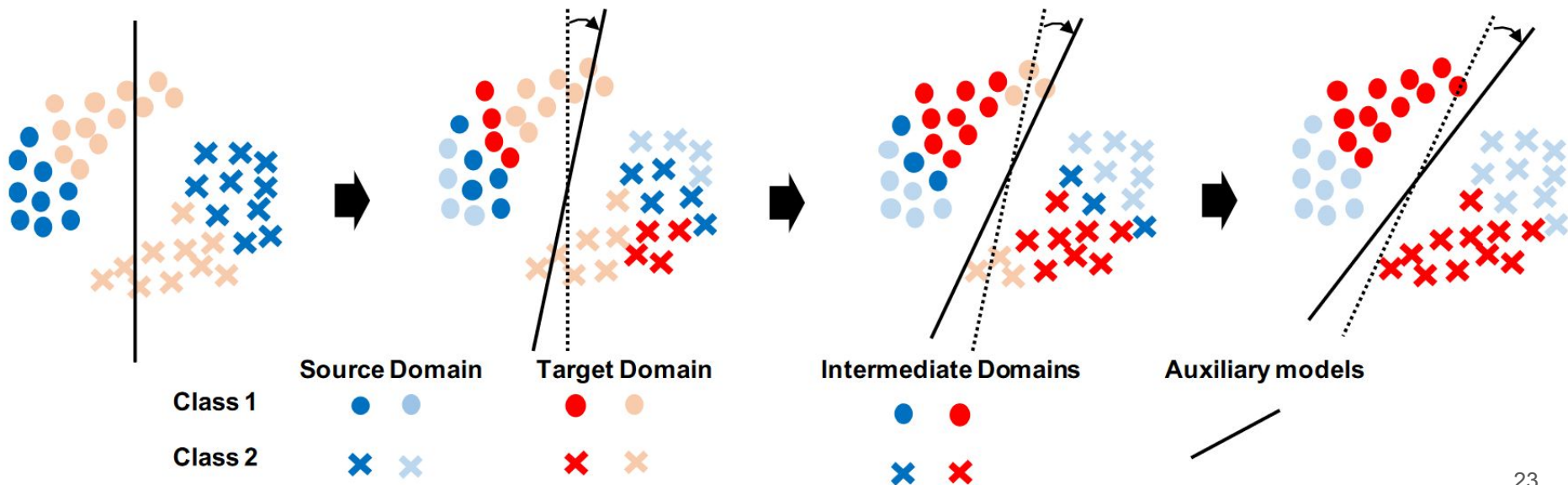
Adversarial Future Works

- Try different training parameters to possibly obtain better results in particular with RADA and SDAT
- Try different starting architectures (e.g. CDAN)
- Test on different datasets (e.g. MNIST or OfficeHome) to have a better understanding
- Try a different alignment measure for RADA not based on discriminator output allowing to fuse RADA with DALN
- Try different policies for selecting samples in the incremental method (e.g. k-NN)
- Test the incremental method with different hyperparameters setting

AuxSelfTrain

Key Idea:

- gradually replace source samples with target samples
- assign pseudo-labels through self-training



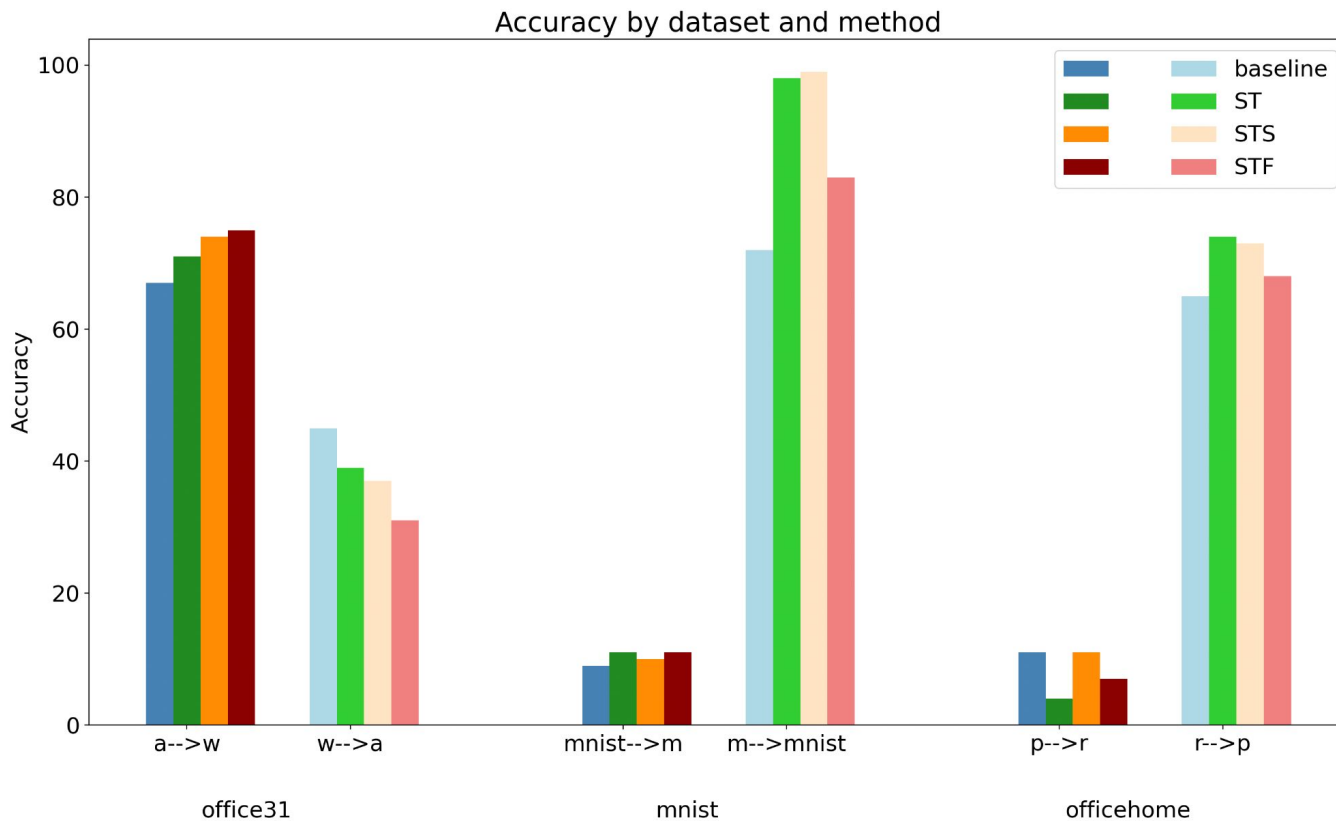


Experiments

- AuxSelfTrain samples selection
 - *target* - highest **confidence** pseudo-label
 - *source* - **closest** to target distribution
- Ablation studies:
 - *ST* - Full approach
 - *STS* - **source** samples are **randomly** selected
 - *STF* - both **source** and **target** are **randomly** selected

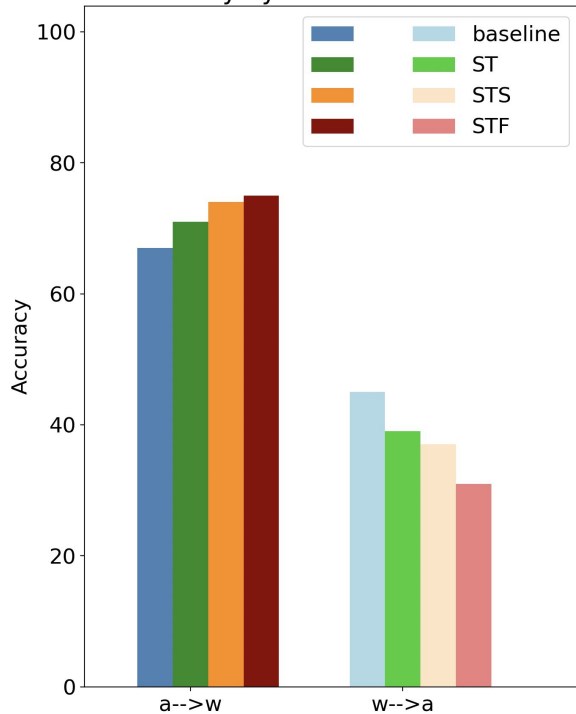


Results



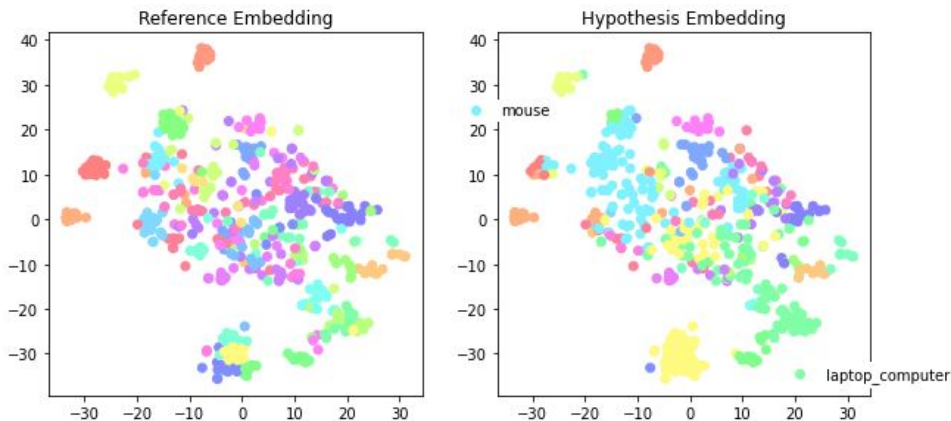
Experiments: Office31

Accuracy by dataset and method



office31

- Improvements in A -> W...
- ... But drop in W -> A
- Two problems:
 - Clustering fails
 - Some classes are over-represented



STS experiment

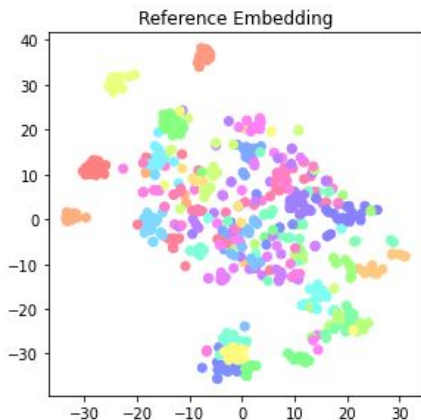


Experiments: Office31

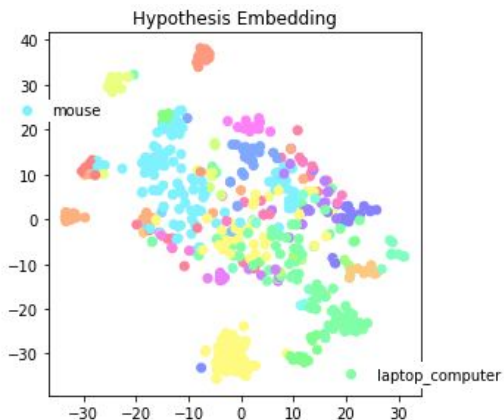
Target Confusion Matrix

	back_pack	bike	bike_helmet	bookcase	bottle	calculator	desk_chair	desk_lamp	desktop_computer	file_cabinet	headphones	keyboard	laptop_computer	letter_tray	mobile_phone	monitor	mouse	mug	paper_notebook	pen	phone	printer	projector	punchers	ring_binder	ruler	scissors	speaker	stapler	tape_dispenser	trash_can
back_pack	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bike	0	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bike_helmet	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bookcase	0	0	0	9	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
bottle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
calculator	0	0	0	0	0	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
desk_chair	0	0	0	0	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
desk_lamp	0	0	0	0	0	0	3	3	0	0	0	0	3	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
desktop_computer	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
file_cabinet	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0
headphones	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0	0	0
keyboard	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
laptop_computer	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
letter_tray	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mobile_phone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
monitor	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
mouse	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0	0
mug	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
paper_notebook	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
pen	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0
phone	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
printer	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	0	0	0	0	0	0	0	0	0	0
projector	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
punchers	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ring_binder	0	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0
ruler	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0
scissors	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0
speaker	0	0	0	0	0	0	0	0	0	0	0	6	0	0	0	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
stapler	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
tape_dispenser	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
trash_can	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- Hypothesis: unbalanced dataset
 - A: ~3000 samples
 - W: ~800 samples
- Perform experiments balanced dataset



STS experiment



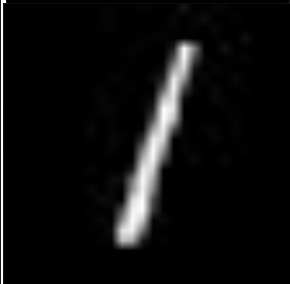


Additional Datasets

MNIST

mnist-m

mnist



~4000

~4000

OfficeHome

Product

Real World



~2000

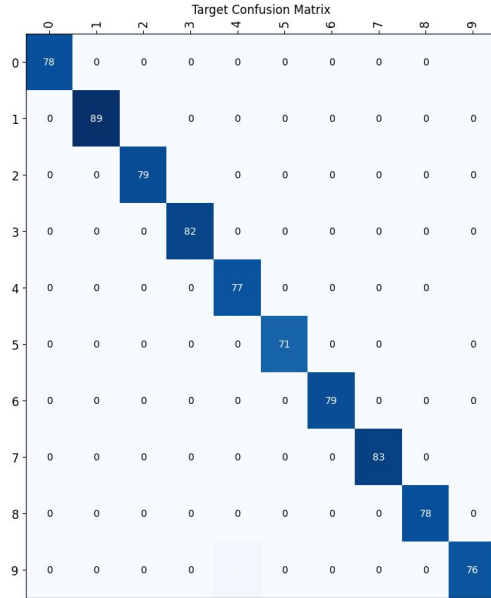
~2000



MNIST-M \rightarrow MNIST

Great!

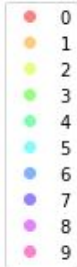
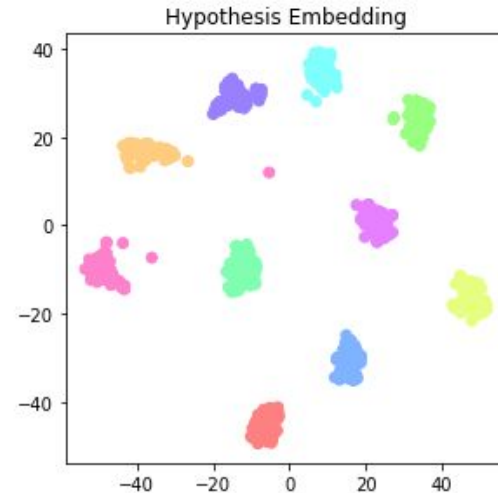
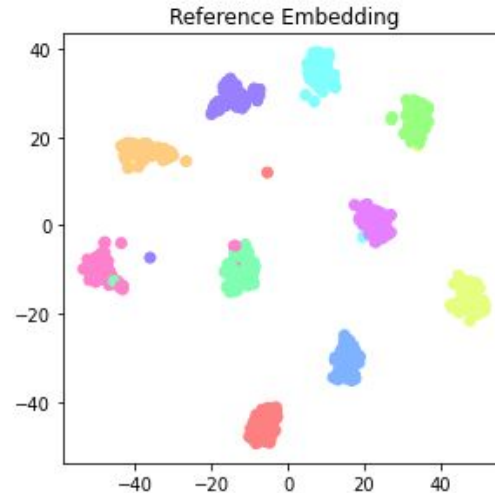
Baseline is ~72%



Target

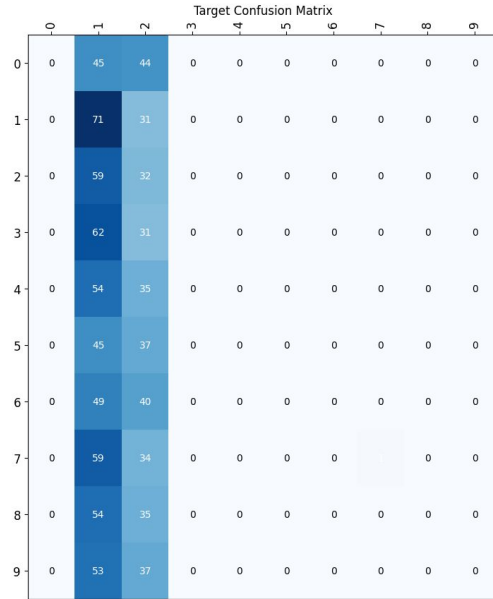
Accuracy: 98.88

Loss: 0.00





MNIST -> MNIST-M

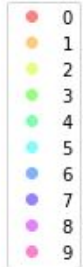
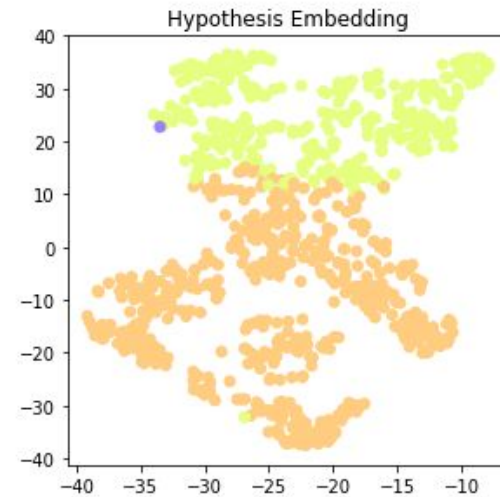
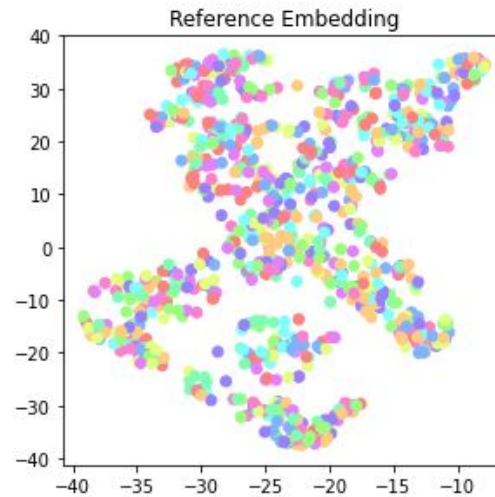


Target

Accuracy: 11.45

Loss: 0.10

Yikes!



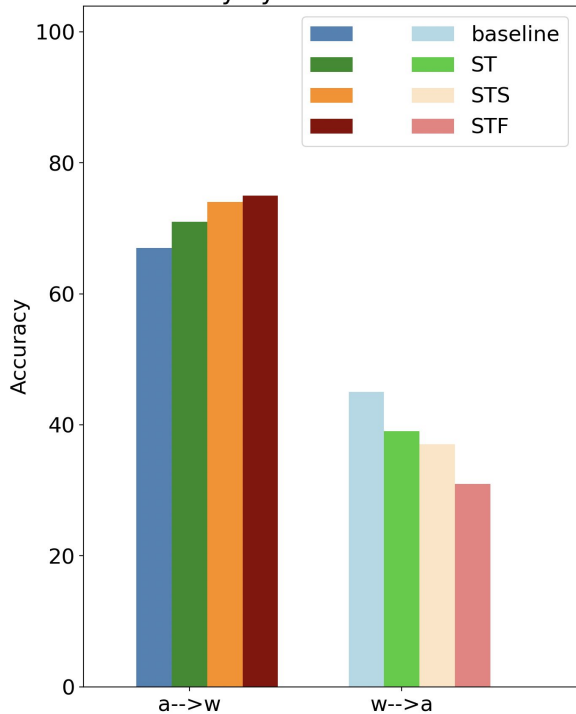


Asymmetric domains

- Results show similar patterns in OfficeHome
 - $R \rightarrow P$ - good results
 - $P \rightarrow R$ - no results
- Probably due to an **asymmetric** domain shift
 - MNIST-M is MNIST but with **more** information
 - Same for R and P
- The model only work with **small** domain shift
 - Otherwise, it is **over-confident** on one single label

Back to Office31

Accuracy by dataset and method



office31

It performs “suspiciously well” on A -> W

- Amazon has some webcam-style images



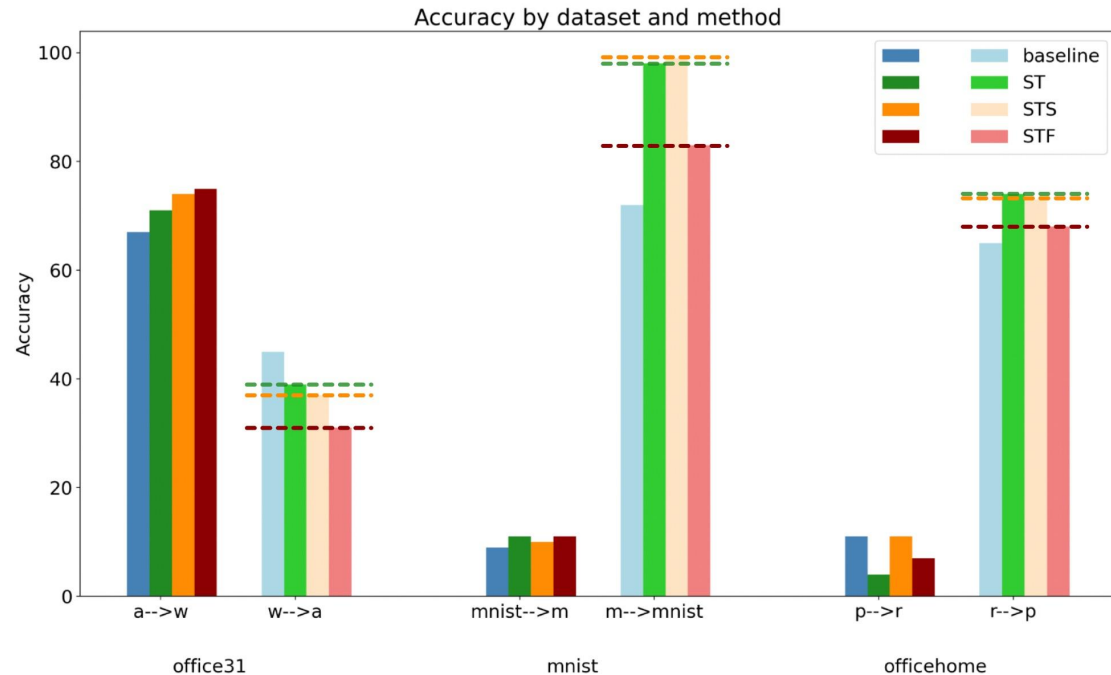
It performs “suspiciously bad” on W -> A

- The domains are not balanced!



Ablation study

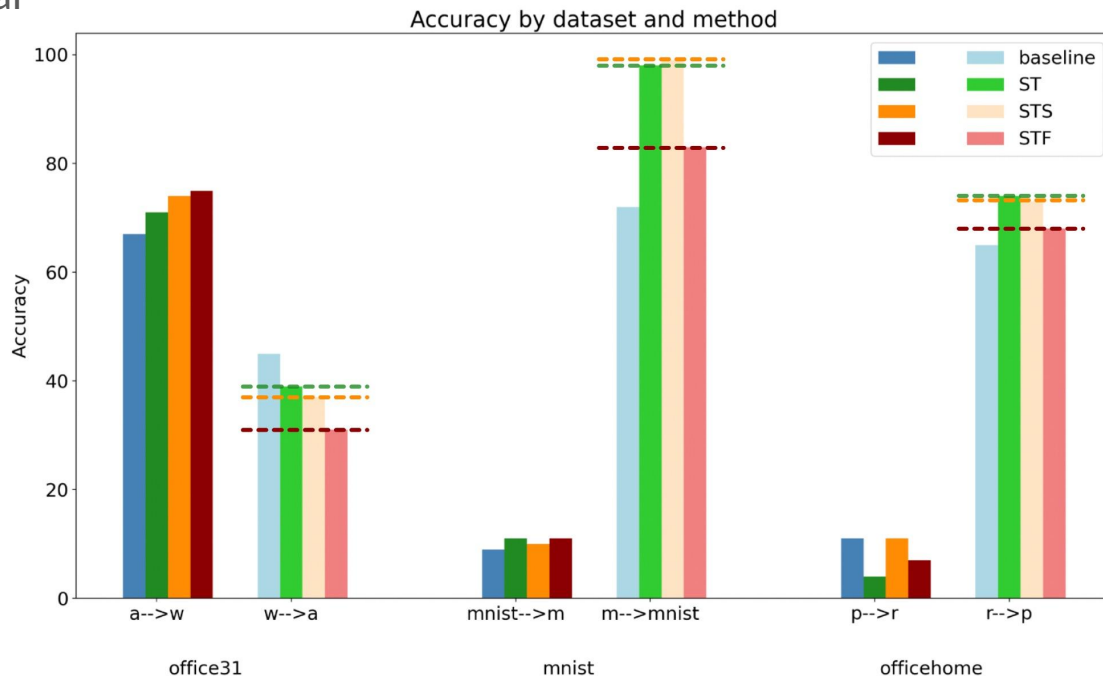
- ST, STS \gg STF
 - Target sample selection works





Ablation study

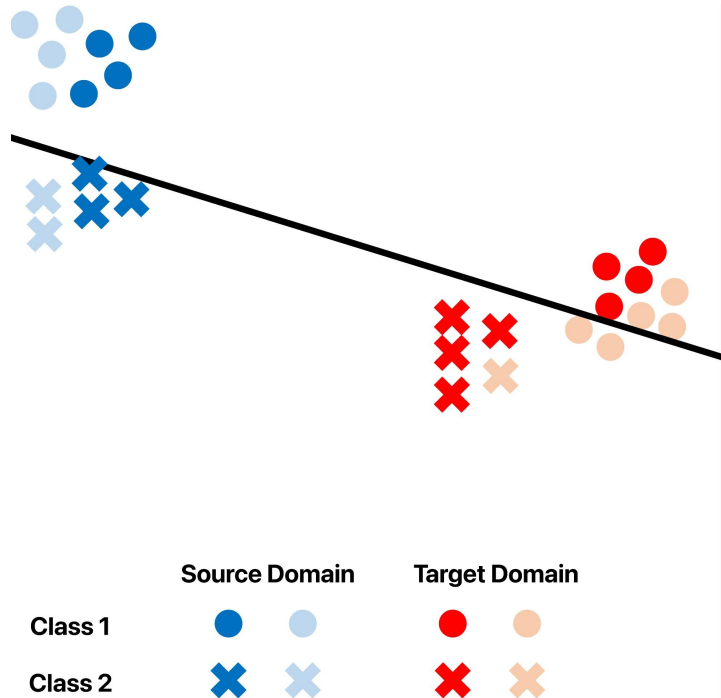
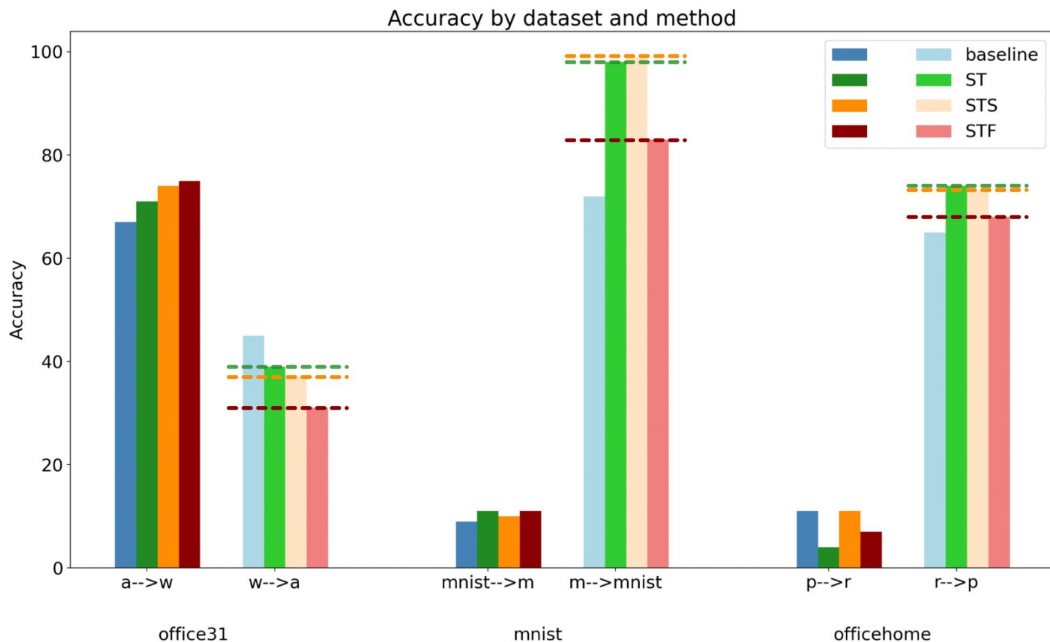
- $ST \cong STS$
 - Source sample selection does not
 - Distributions are too far





Ablation study

- $ST \cong STS$
 - Source sample selection does not
 - Distributions are too far





Future works

- Implement **model ensemble** on the target source samples selection strategy
 - Requires significant computational power
- Behaviour on MNIST -> MNIST-M resembles the **mode collapse** problem of GANs
 - Use toolchains from GAN literature to further explore
 - e.g. batch discrimination
 - Add penalty/threshold when few classes are over-represented
- Test our DANN methods on other datasets
 - Provide insights on your model



Conclusion

- **Avoid** to **over complicate** the model
- **Keep** the model **simple** but exploit it better (change losses and/or optimizer)
- Some models require careful **fine-tuning** of hyperparameters
- There is **no panacea** model
 - A thorough dataset exploration is crucial
 - Pick the best DA approach given the dataset



Thank you for your attention



**Thank you for your attention
...and merry Christmas!**