ISTITUTO ITALIANO DI TECNOLOGIA
PATTERN ANALYSIS AND COMPUTER VISION

UNIVERSITÀ di VERONA
Dipartimento di INFORMATICA

Domain Adaptation and Generalization

Vittorio Murino, Pietro Morerio

April 8, 2022

Session 2
Recent Methods
(Deep learning)

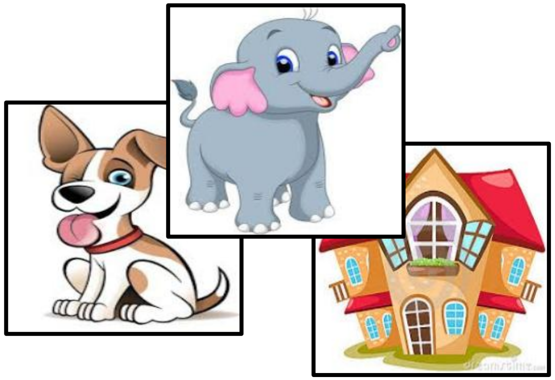
Outline

Session 2 - Recent Methods (Deep learning) (1h)

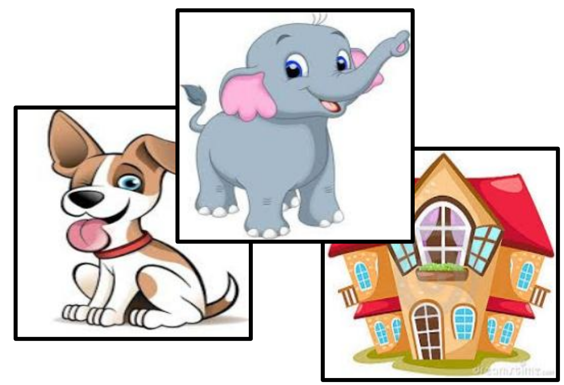
- Adversarial DA
- Image translation methods
- Feature alignment/confusion
- Batchnorm-based methods
- Pseudo-labeling

Adversarial Domain Adaptation

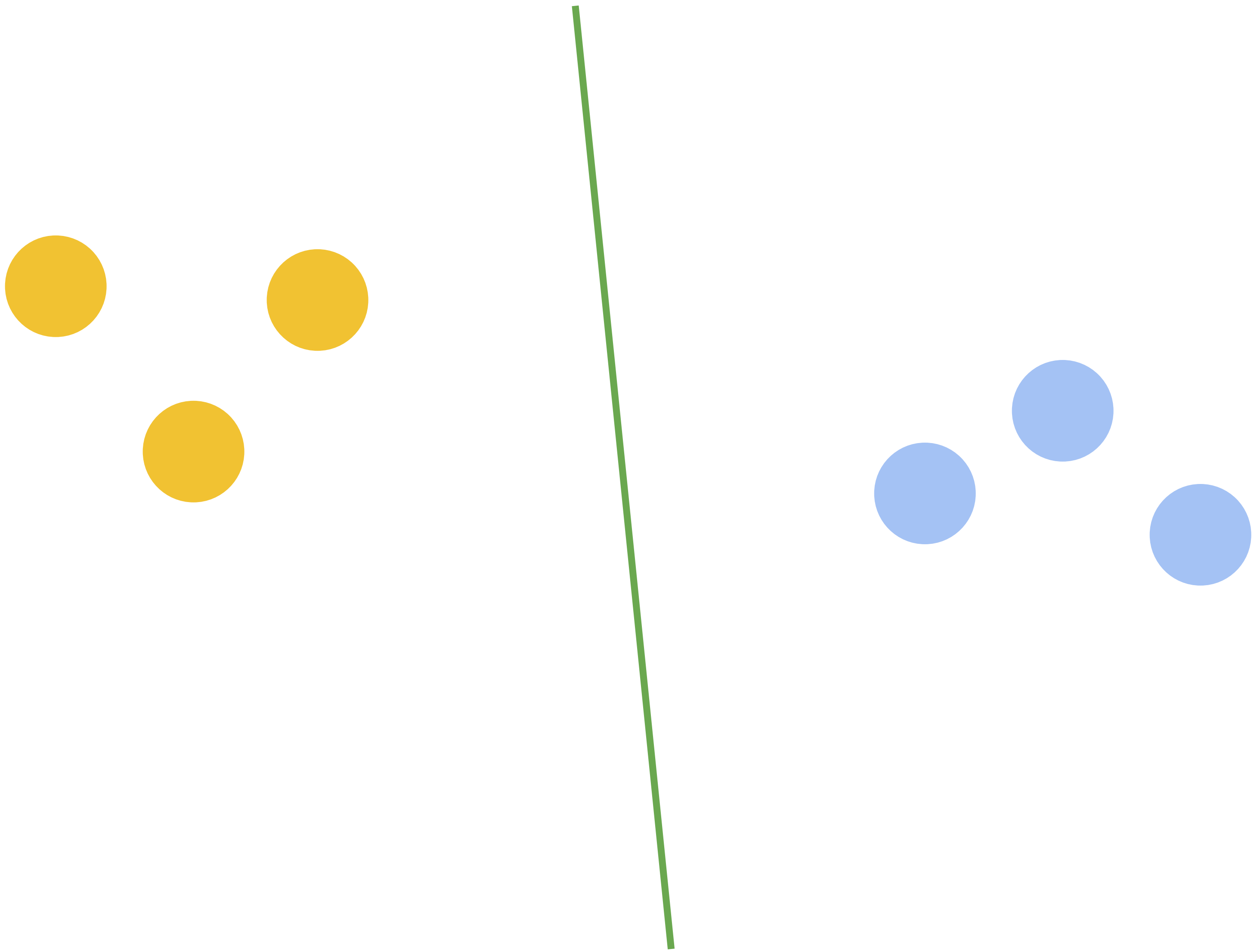
Domain Classification



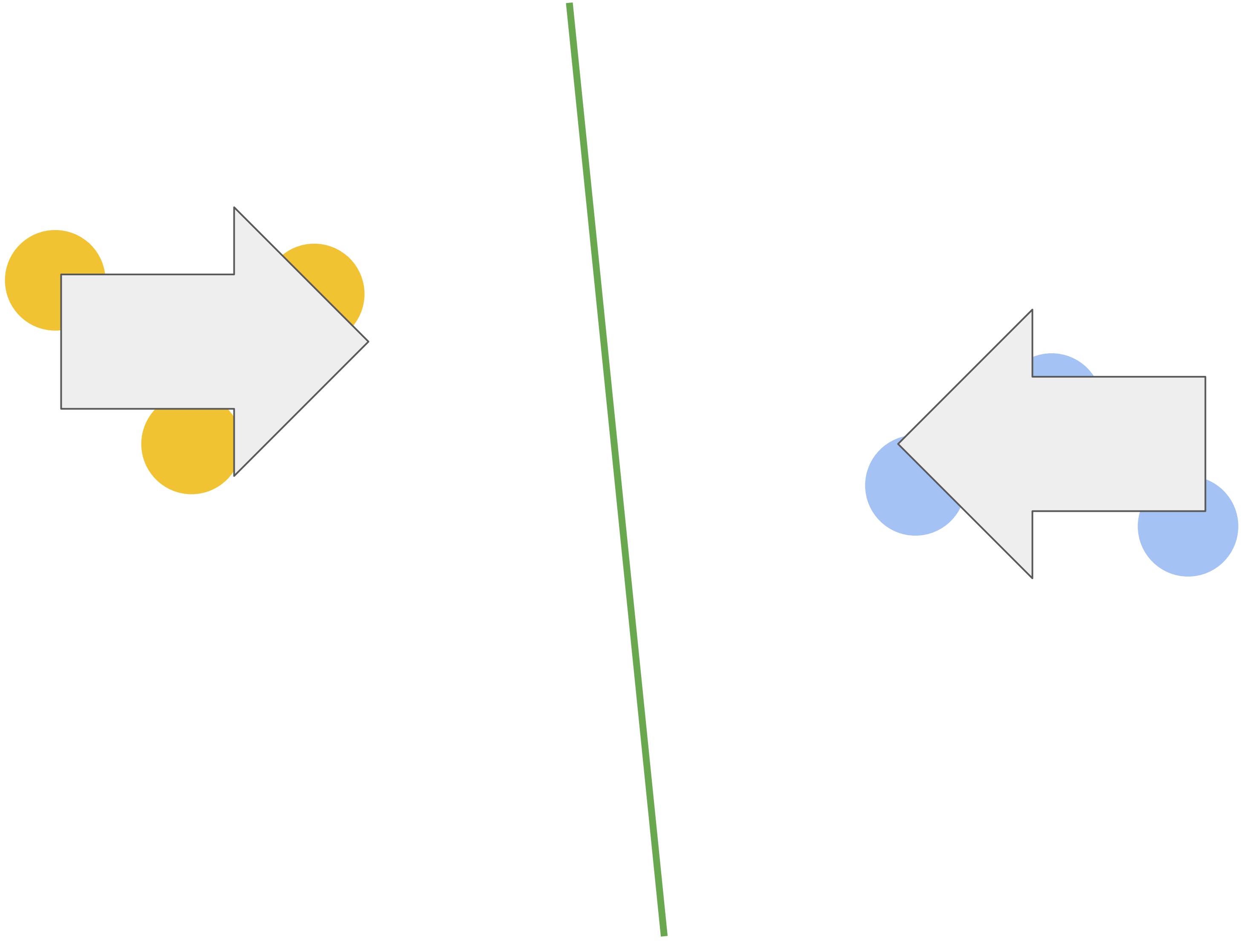
Domain Classification



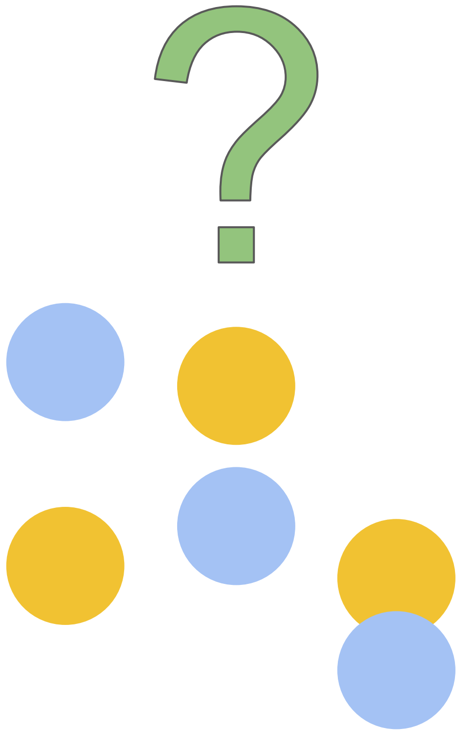
Domain Classification



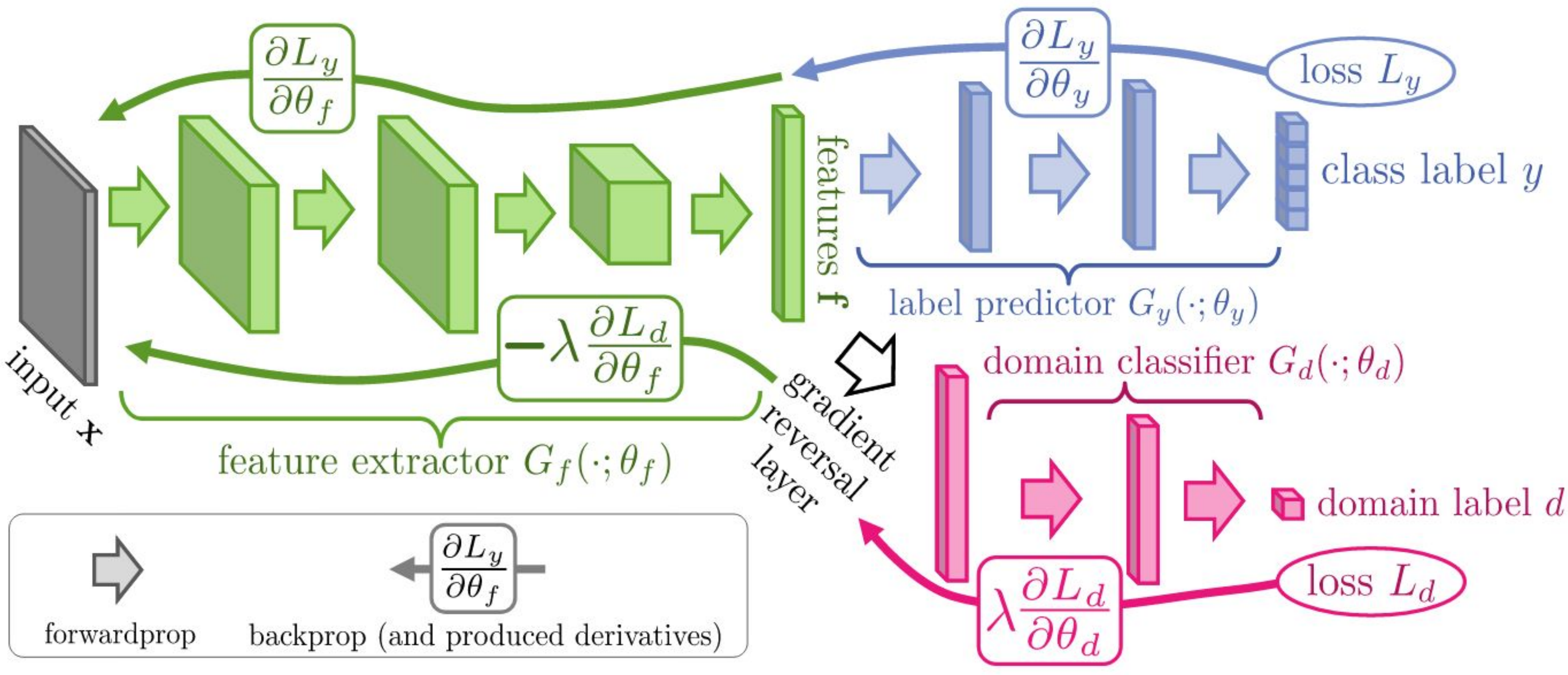
Domain Classification



Domain Classification



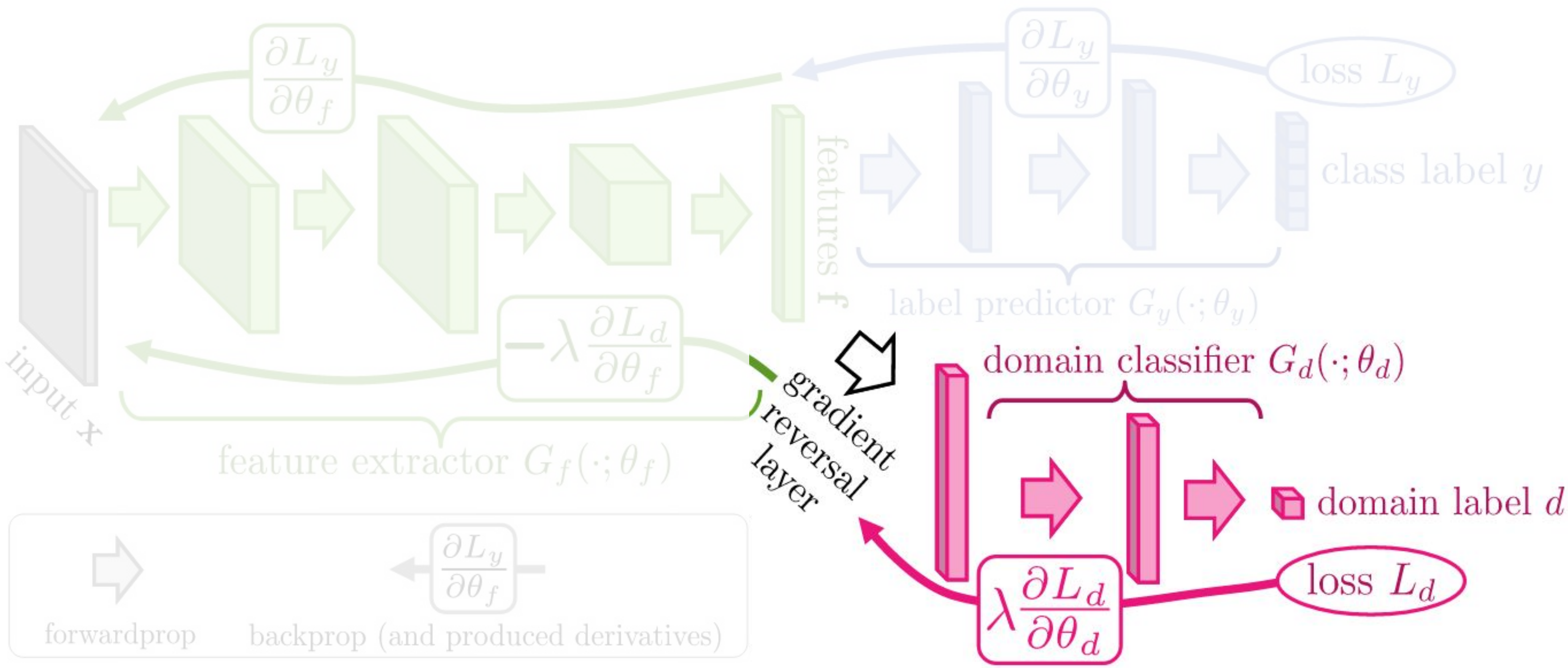
Domain Adversarial Training



$$R_\lambda(\mathbf{x}) = \mathbf{x}, \quad \frac{dR_\lambda}{d\mathbf{x}} = -\lambda \mathbf{I}, \quad \lambda > 0$$

Ganin, Y., et al. "Domain-adversarial training of neural networks". The Journal of Machine Learning Research, 17(1), 2096-2030, 2016..

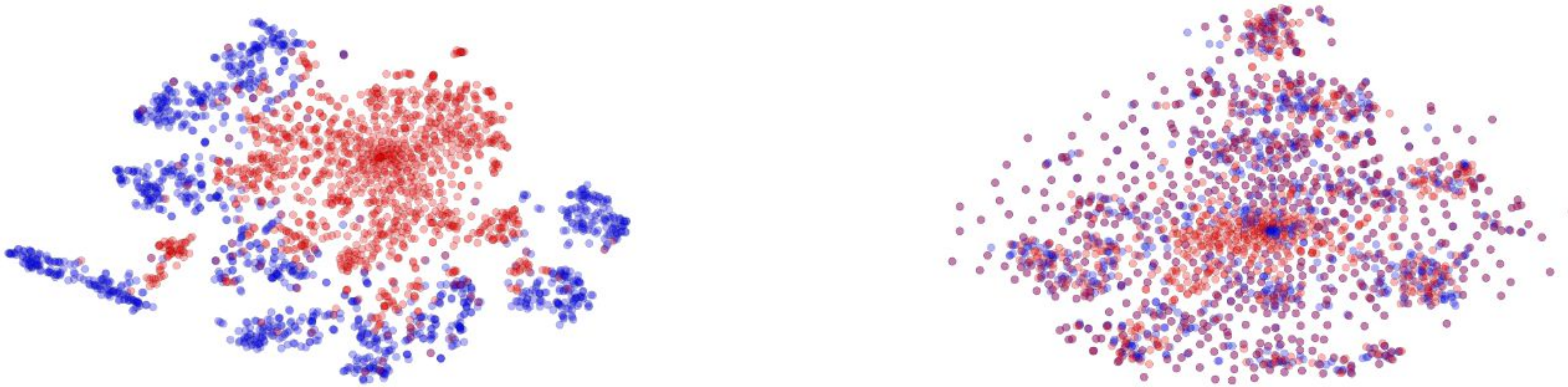
Domain Adversarial Training



Ganin, Y., et al. "Domain-adversarial training of neural networks". The Journal of Machine Learning Research, 17(1), 2096-2030, 2016..

Domain Adversarial Training

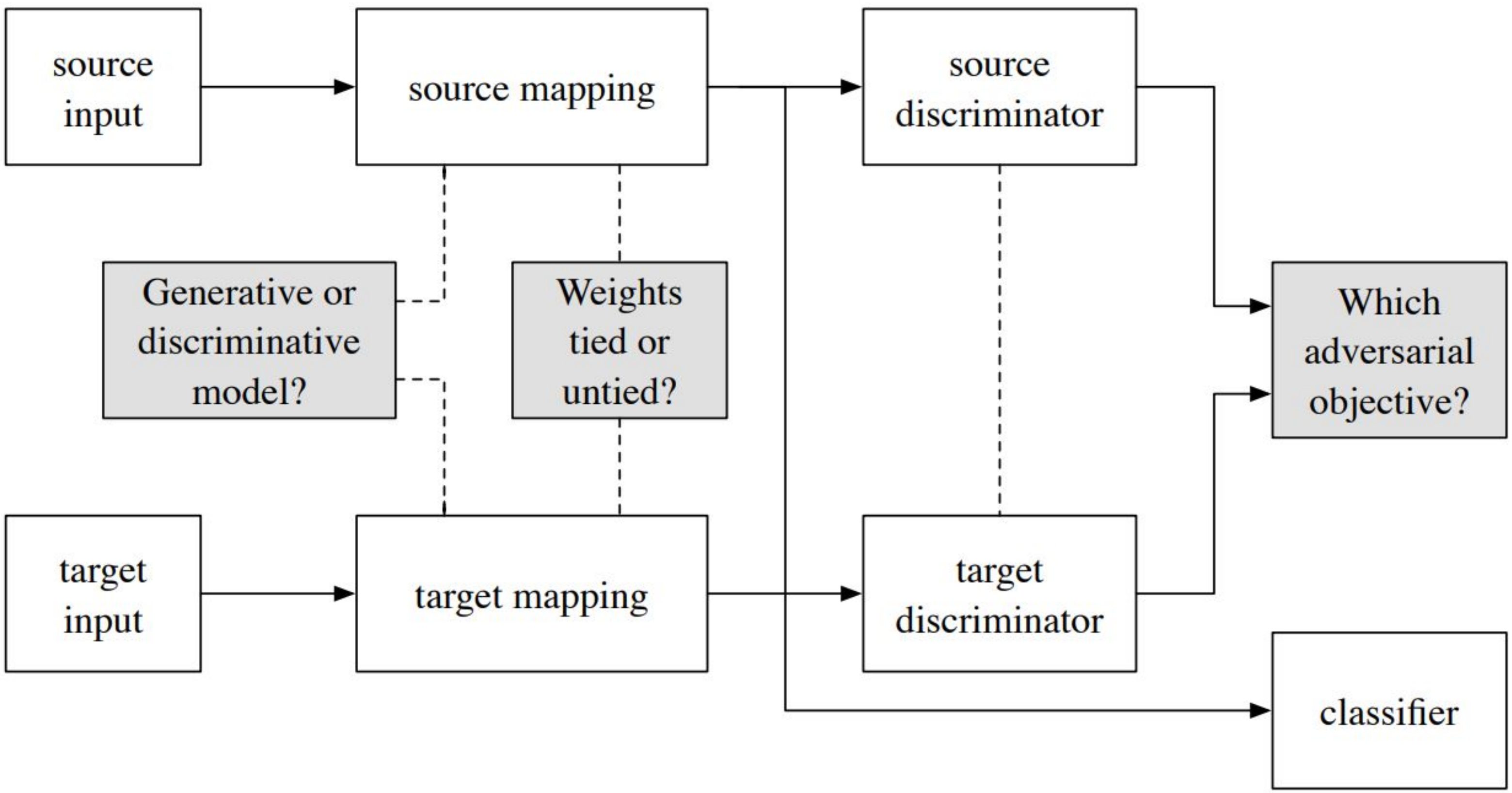
MNIST → MNIST-M: top feature extractor layer



(a) Non-adapted (b) Adapted

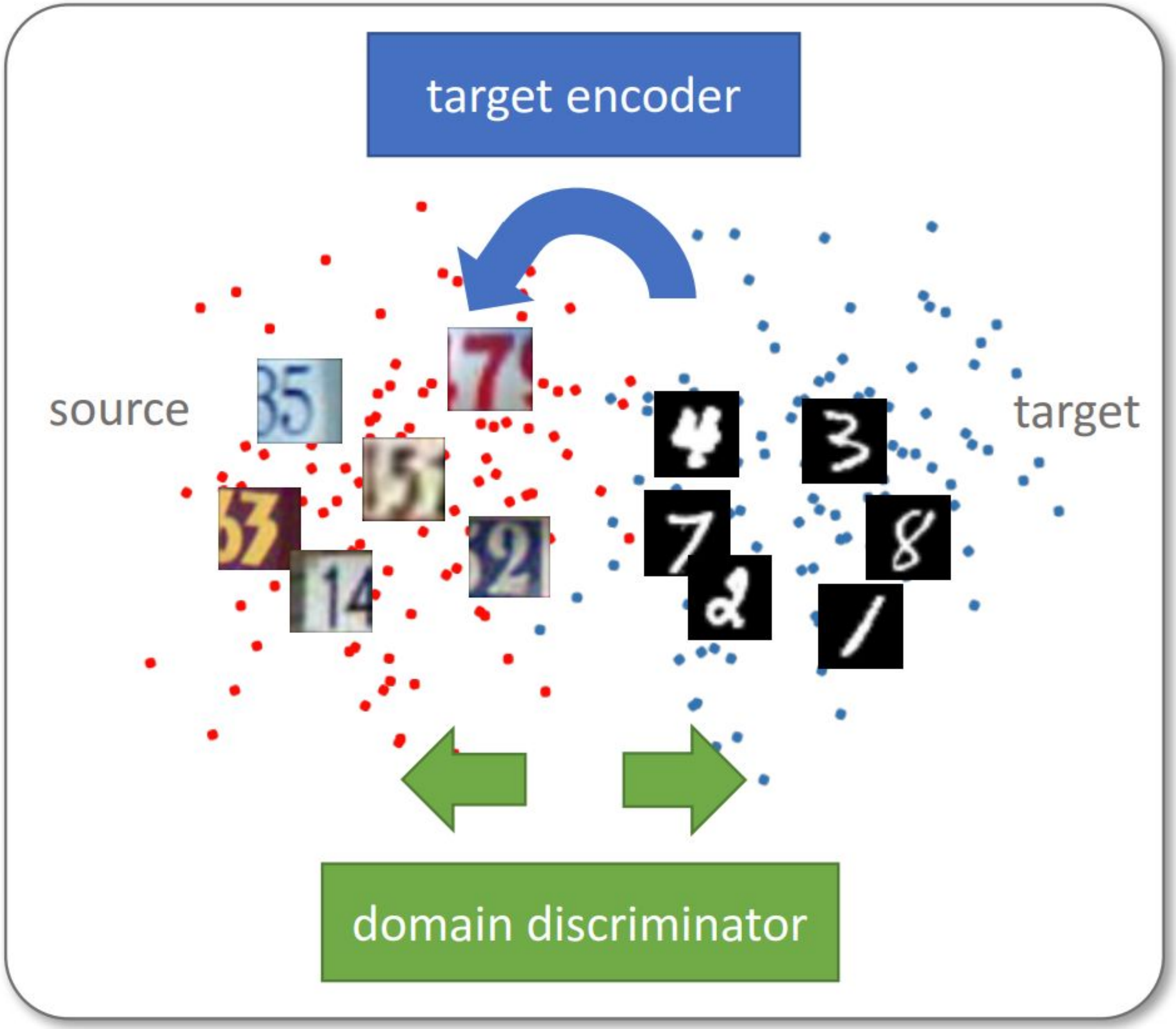
Ganin, Y., et al. "Domain-adversarial training of neural networks". The Journal of Machine Learning Research, 17(1), 2096-2030, 2016..

ADDA



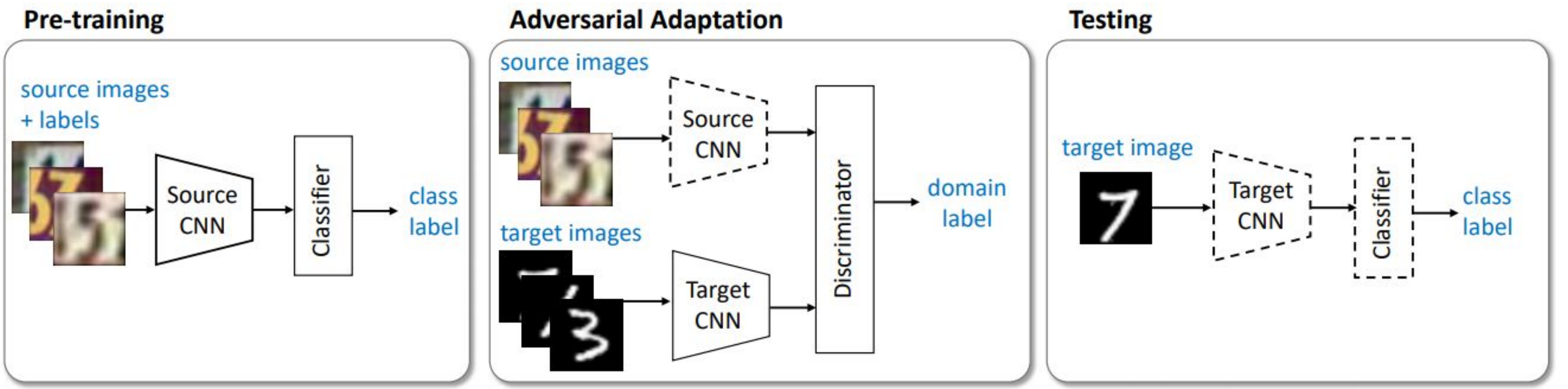
Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T., " Adversarial discriminative domain adaptation". In CVPR 2017.

ADDA



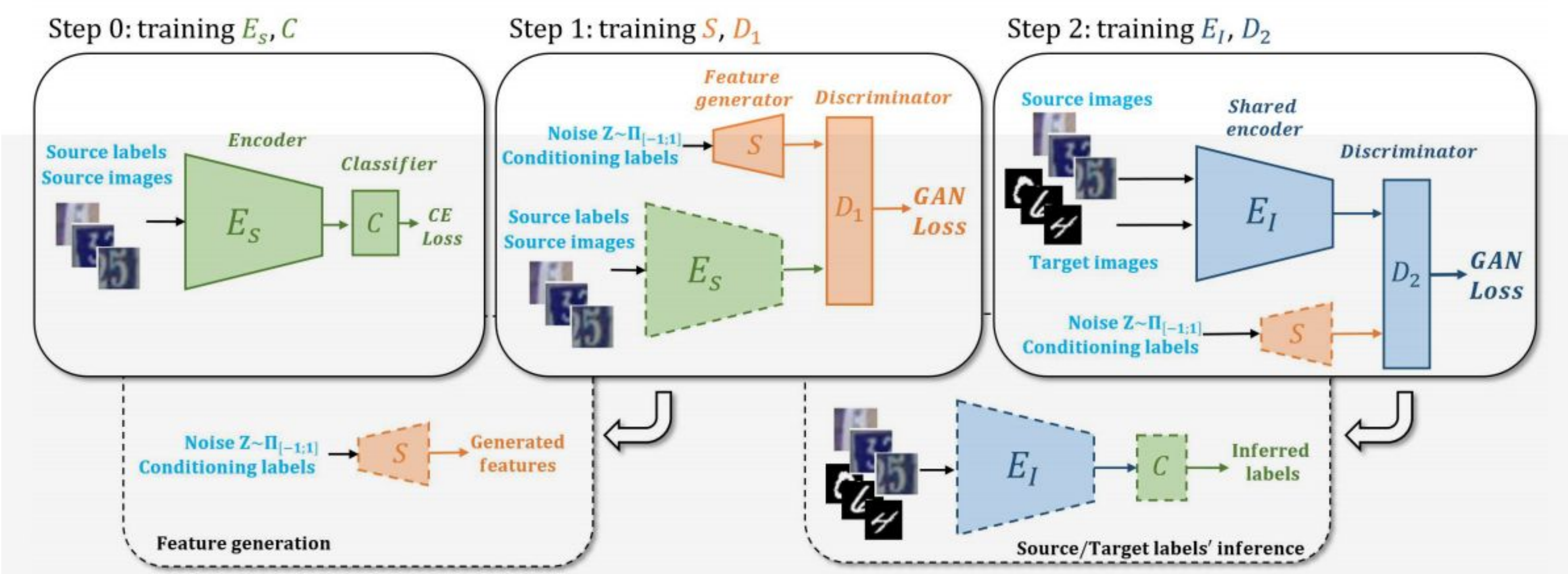
Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T., " Adversarial discriminative domain adaptation". In CVPR 2017.

ADDA - methodology



Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T., " Adversarial discriminative domain adaptation". In CVPR 2017.

Domain invariant Feature Augmentation (DIFA)



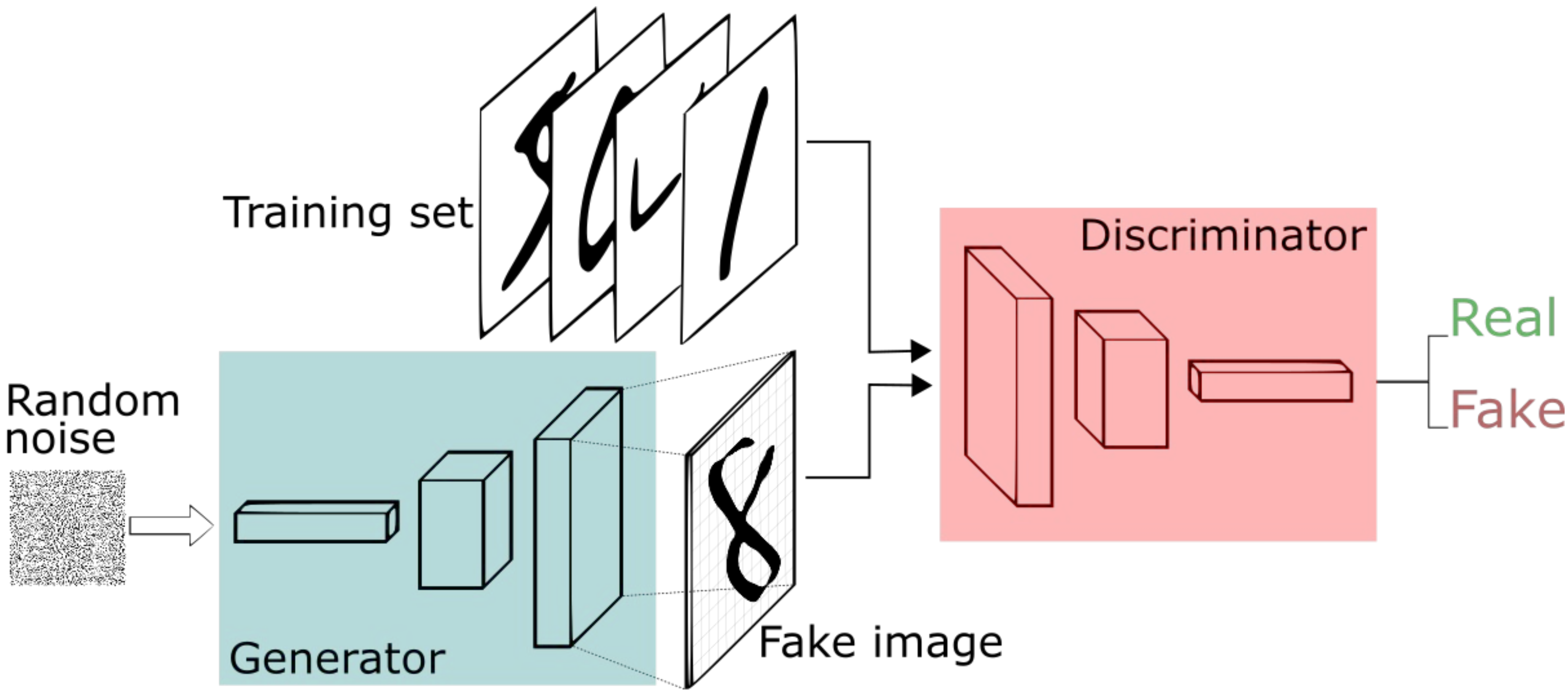
Improves ADDA by

1. Sampling source feature to perform augmentation
2. Making the encoder domain invariant (anchoring to the source)

Volpi, R., Morerio, Murino, V. "Adversarial Feature Augmentation for Unsupervised Domain Adaptation", CVPR 2018

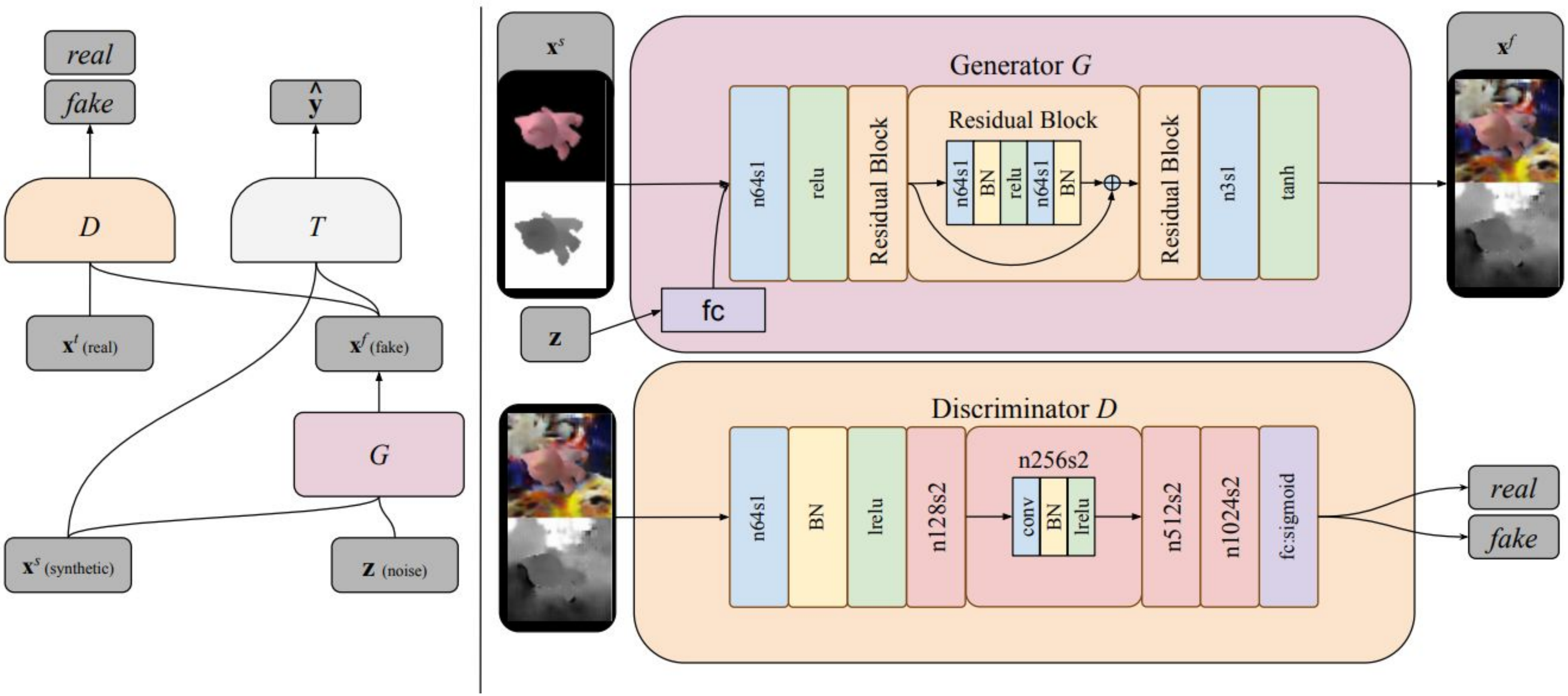
Domain Adaptation through Image Translation

Generative Adversarial Networks (GAN)



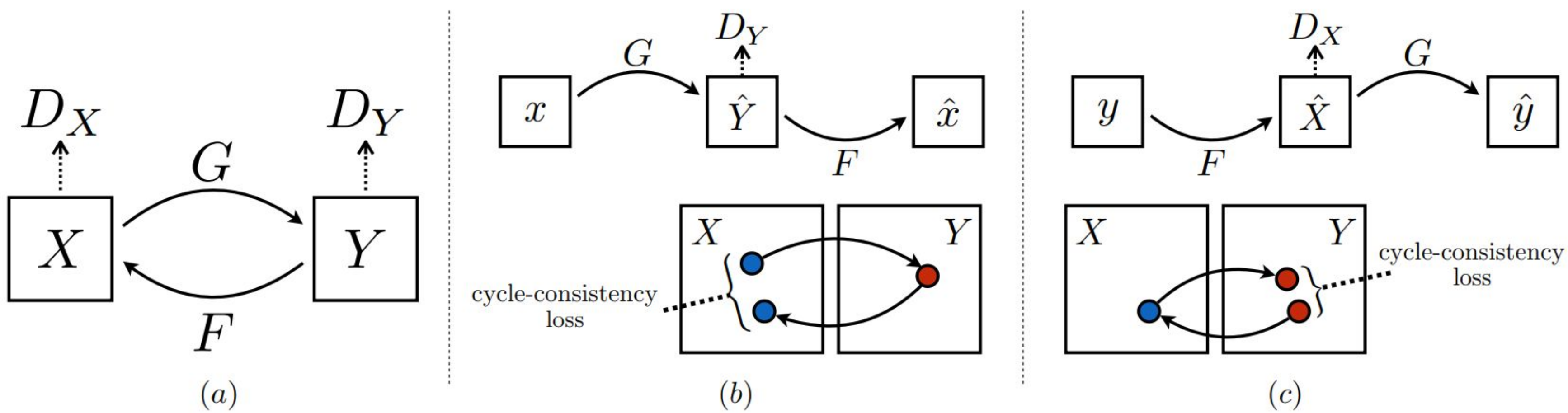
https://skymind.ai/wiki/generative-adversarial-network-gan

Pixel-Level DA



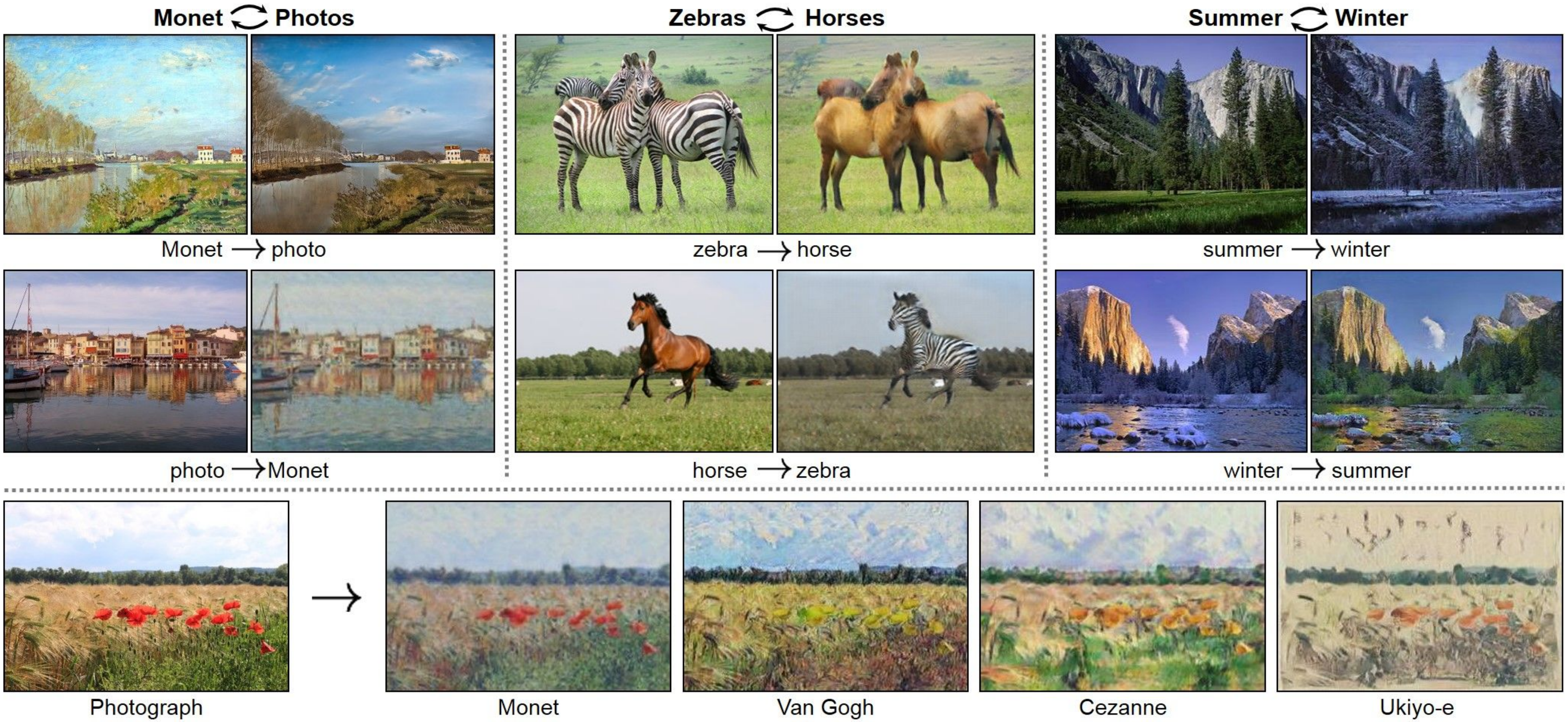
Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., & Krishnan, D., "Unsupervised pixel-level domain adaptation with generative adversarial networks". In CVPR 2017.

Cycle GAN



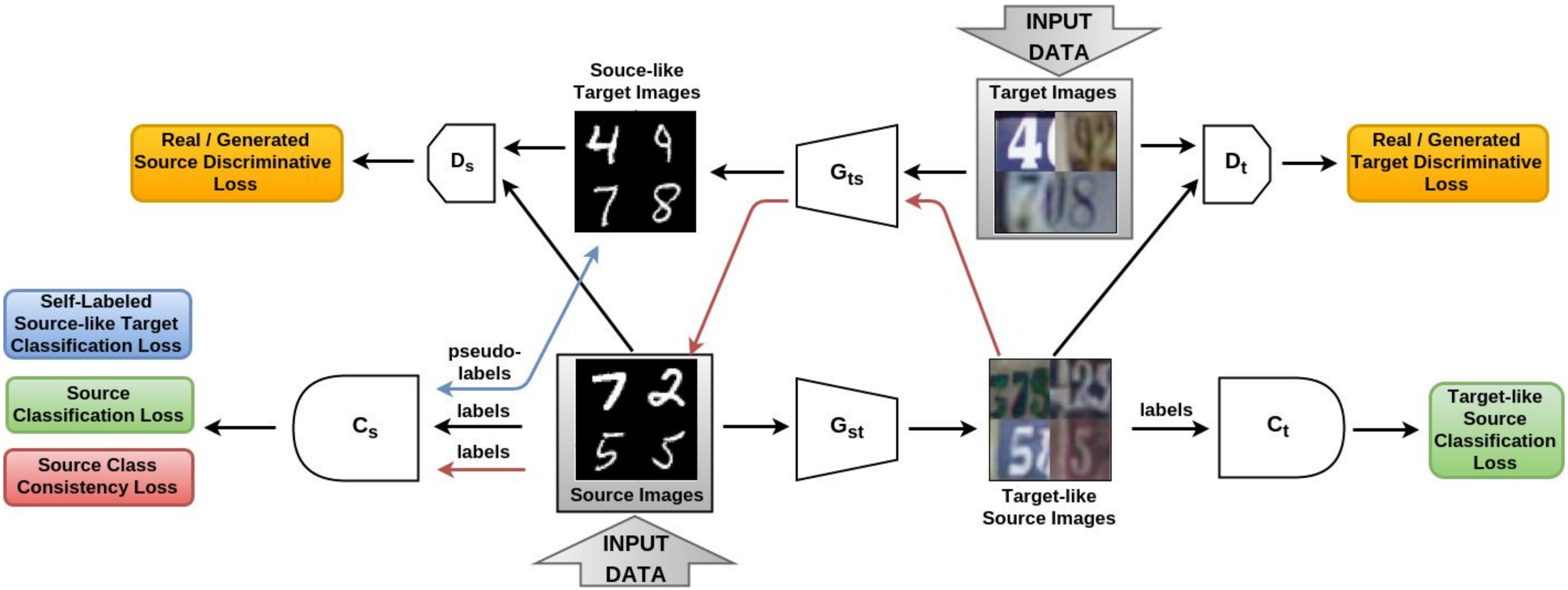
Zhu, J. Y., Park, T., Isola, P., & Efros, A.. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In ICCV 2017.

Unpaired Image-to-Image Translation



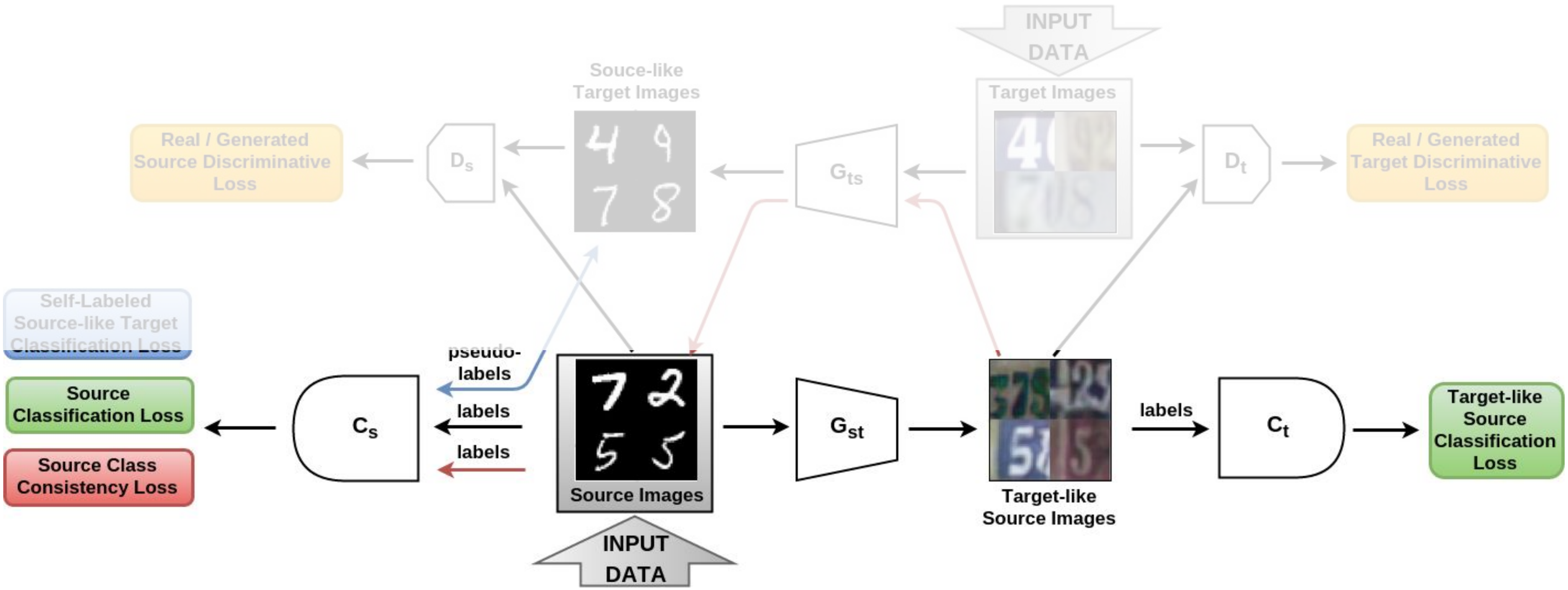
Zhu, J. Y., Park, T., Isola, P., & Efros, A.. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In ICCV 2017.

SBADA-GAN



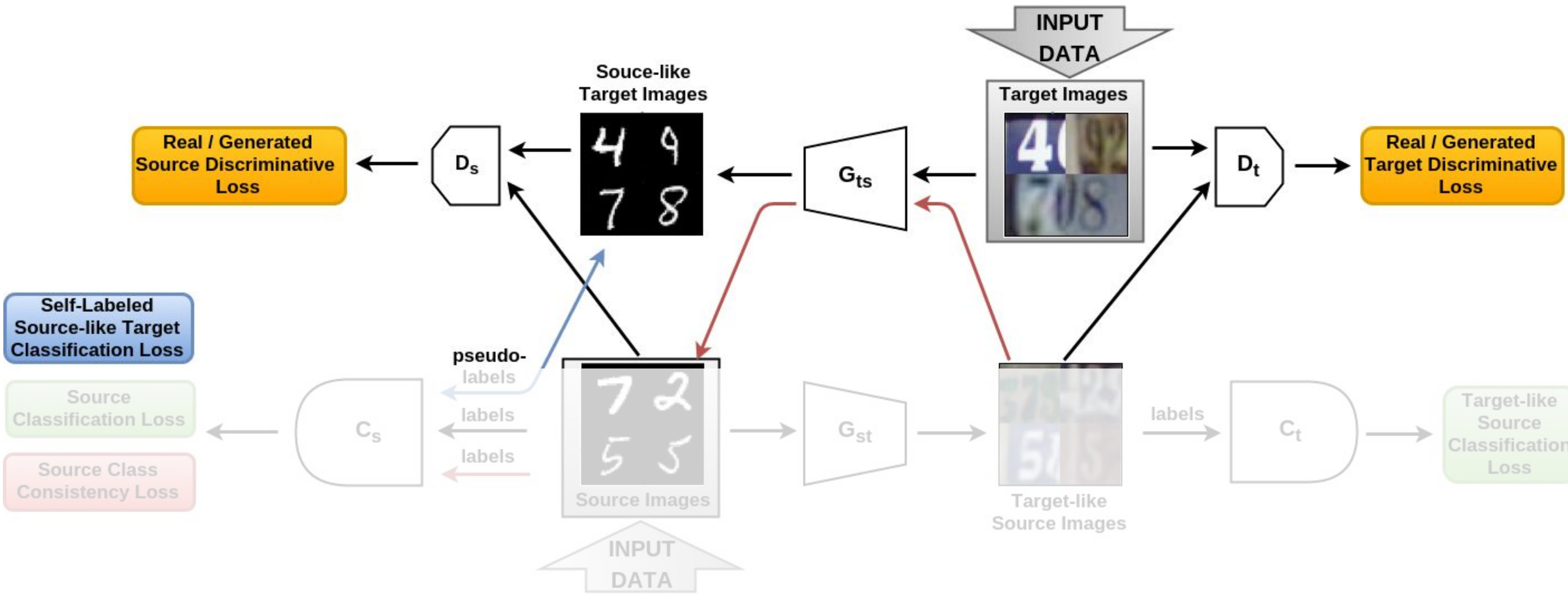
Russo, P., Carlucci, F. M., Tommasi, T., & Caputo, B. "From source to target and back: symmetric bi-directional adaptive gan". In CVPR 2018.

SBADA-GAN: source images path



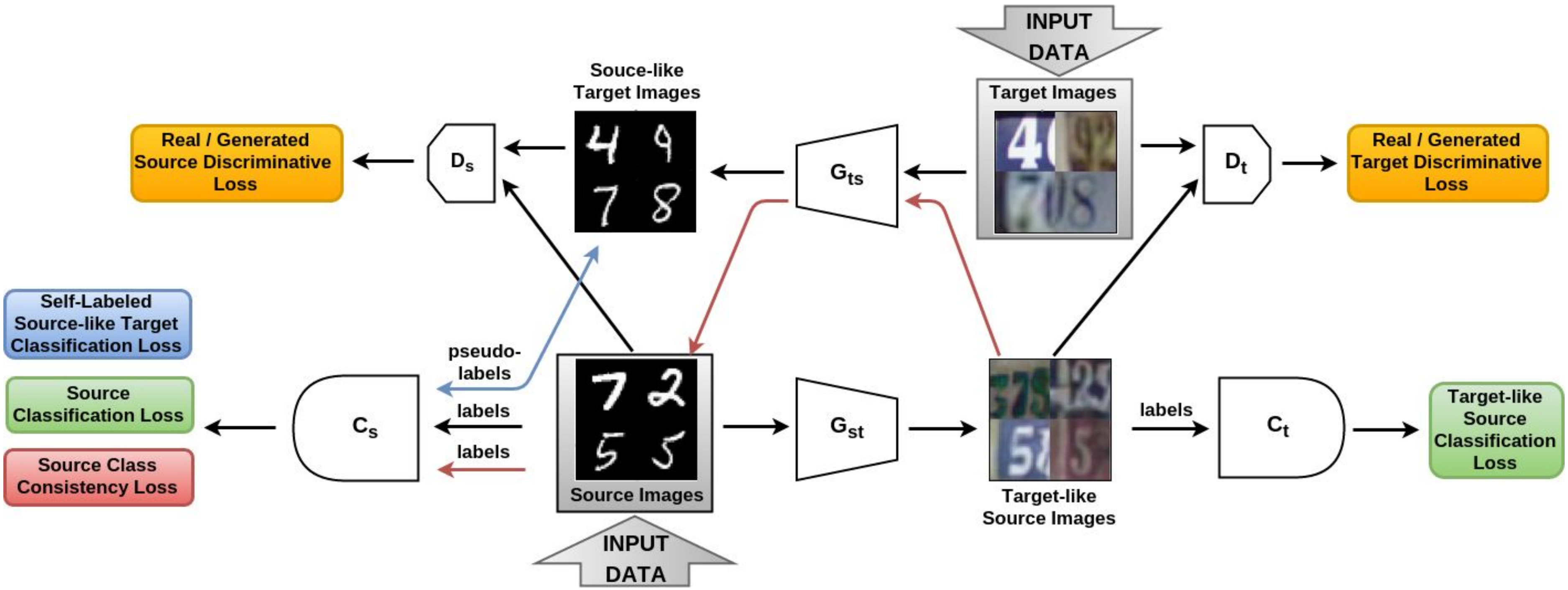
Russo, P., Carlucci, F. M., Tommasi, T., & Caputo, B. "From source to target and back: symmetric bi-directional adaptive gan". In CVPR 2018.

SBADA-GAN: target images path



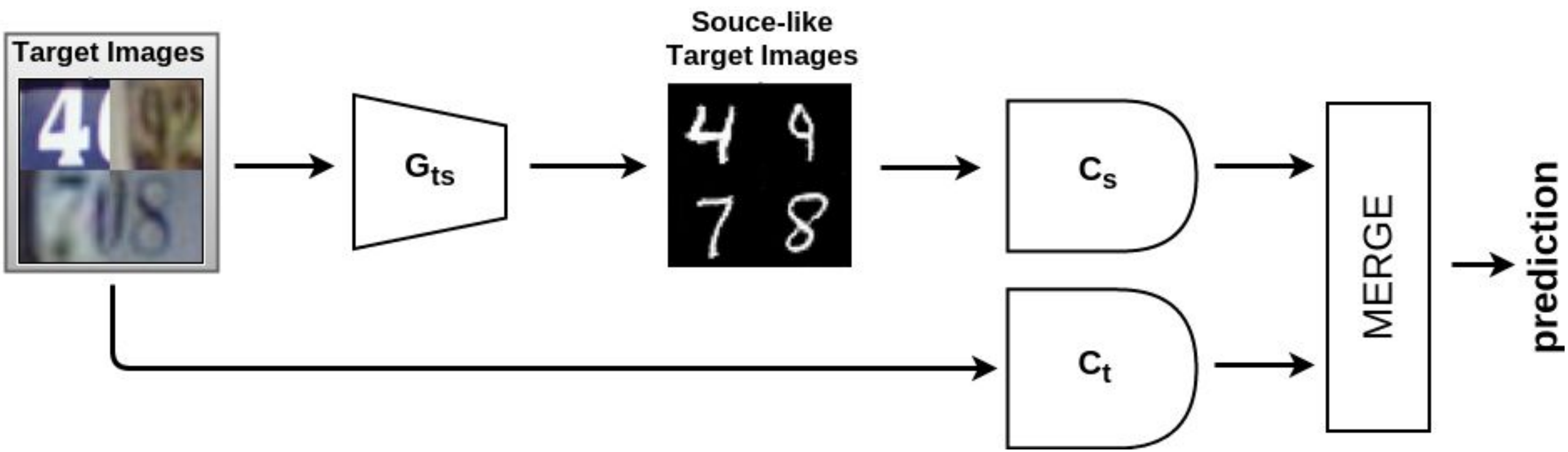
Russo, P., Carlucci, F. M., Tommasi, T., & Caputo, B. "From source to target and back: symmetric bi-directional adaptive gan". In CVPR 2018.

SBADA-GAN



Russo, P., Carlucci, F. M., Tommasi, T., & Caputo, B. "From source to target and back: symmetric bi-directional adaptive gan". In CVPR 2018.

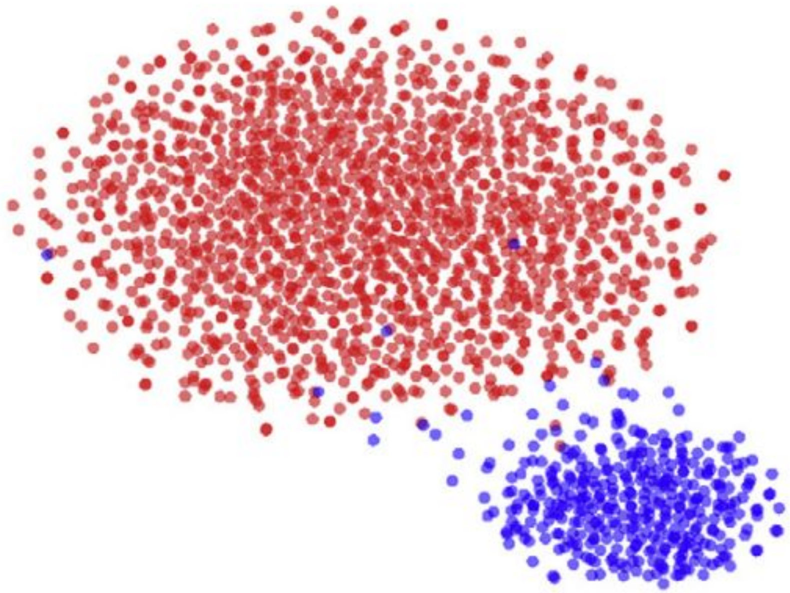
SBADA-GAN



Russo, P., Carlucci, F. M., Tommasi, T., & Caputo, B. "From source to target and back: symmetric bi-directional adaptive gan". In CVPR 2018.

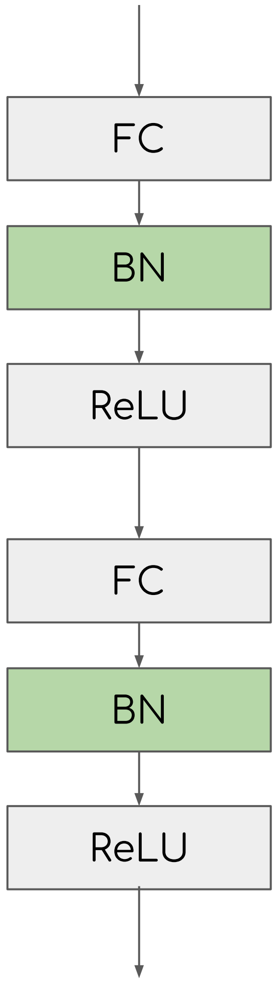
Domain Adaptation through Batch Normalization

How the problem looks like



Li et al. (2017) “Revisiting Batch Normalization For Practical Domain Adaptation”, ICLR-WS.

Some Background: Batch Normalization



Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1\ldots m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_\mathcal{B} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

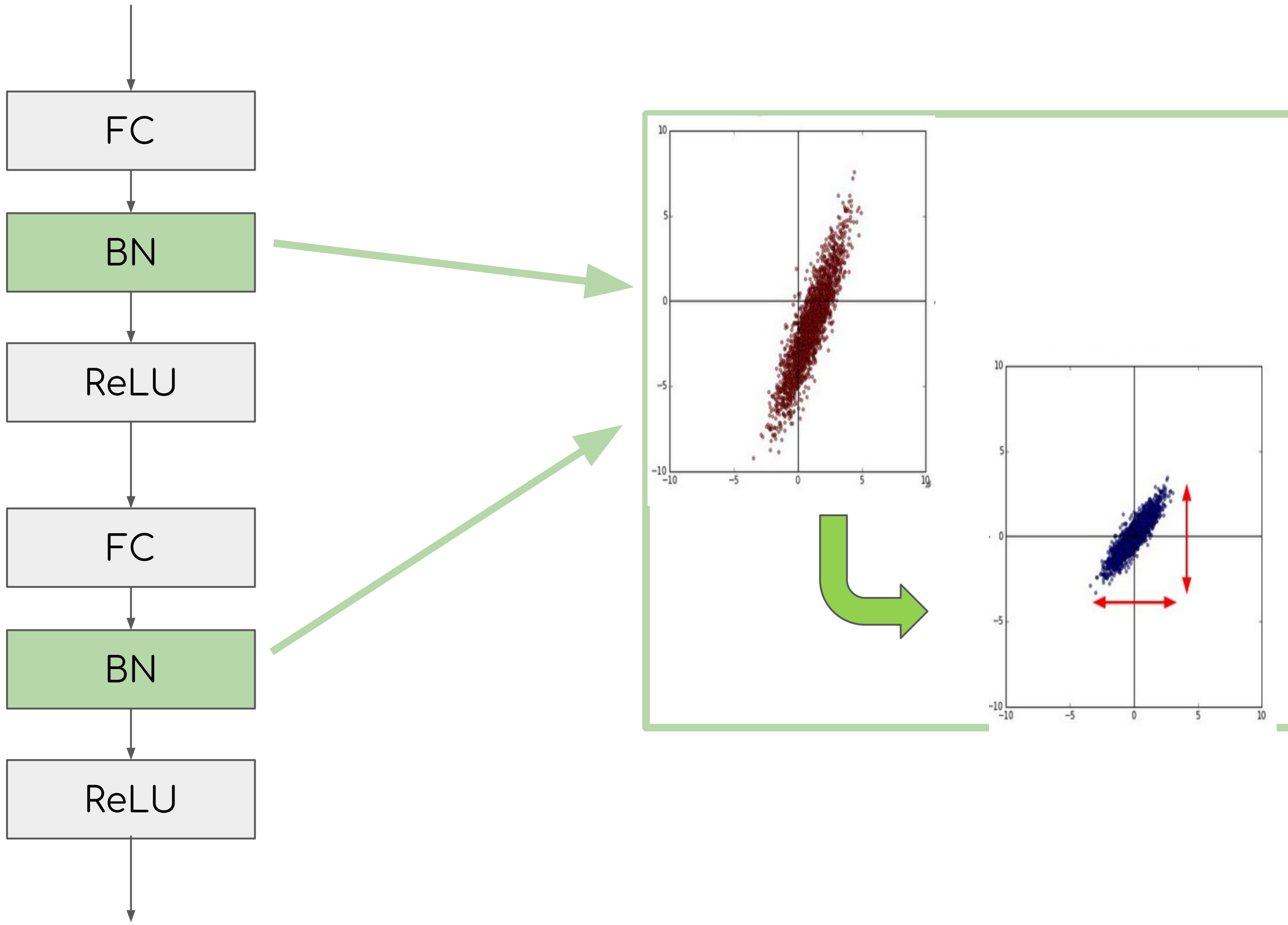
$$\sigma_\mathcal{B}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_\mathcal{B})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_\mathcal{B}}{\sqrt{\sigma_\mathcal{B}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

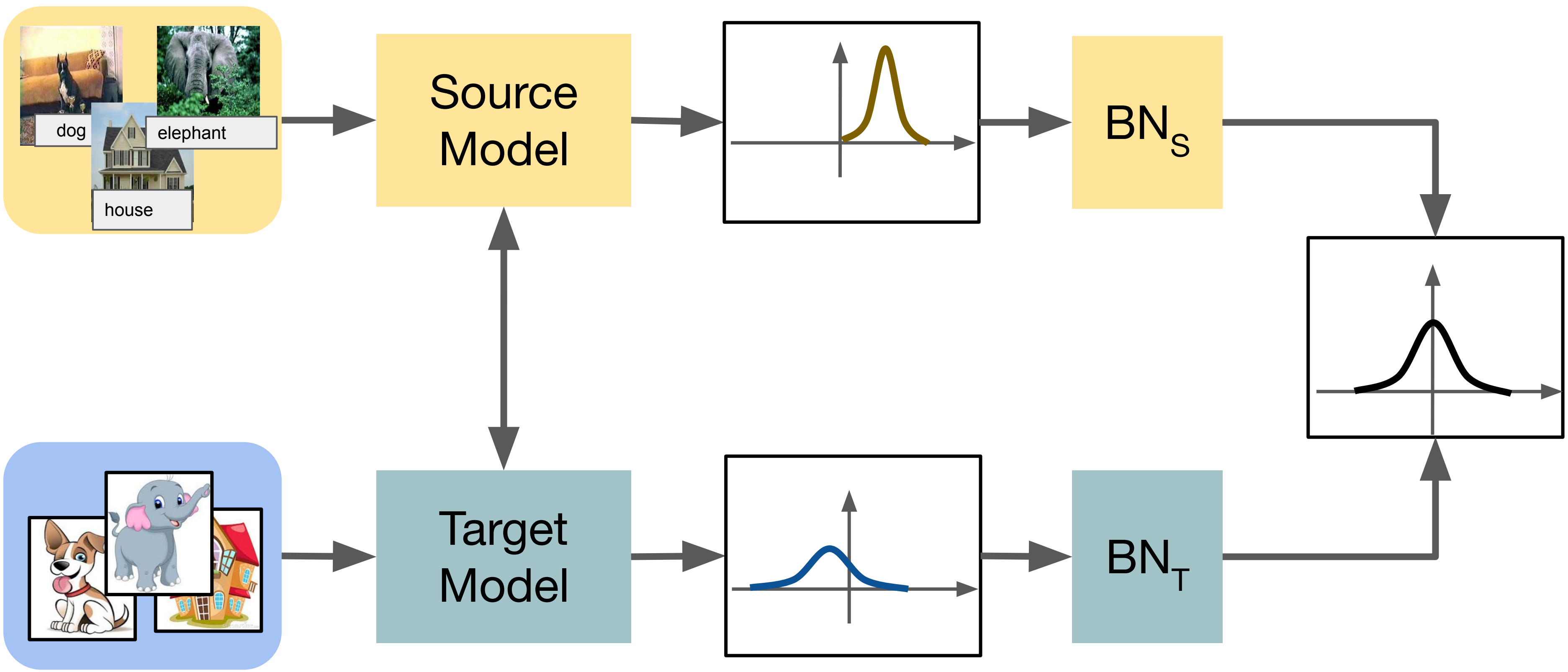
Ioffe, S., & Szegedy, C. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift: Automatic Domain Alignment Layers". *ICML 2015.*
http://cs231n.github.io/neural-networks-2/#batchnorm

Some Background: Batch Normalization



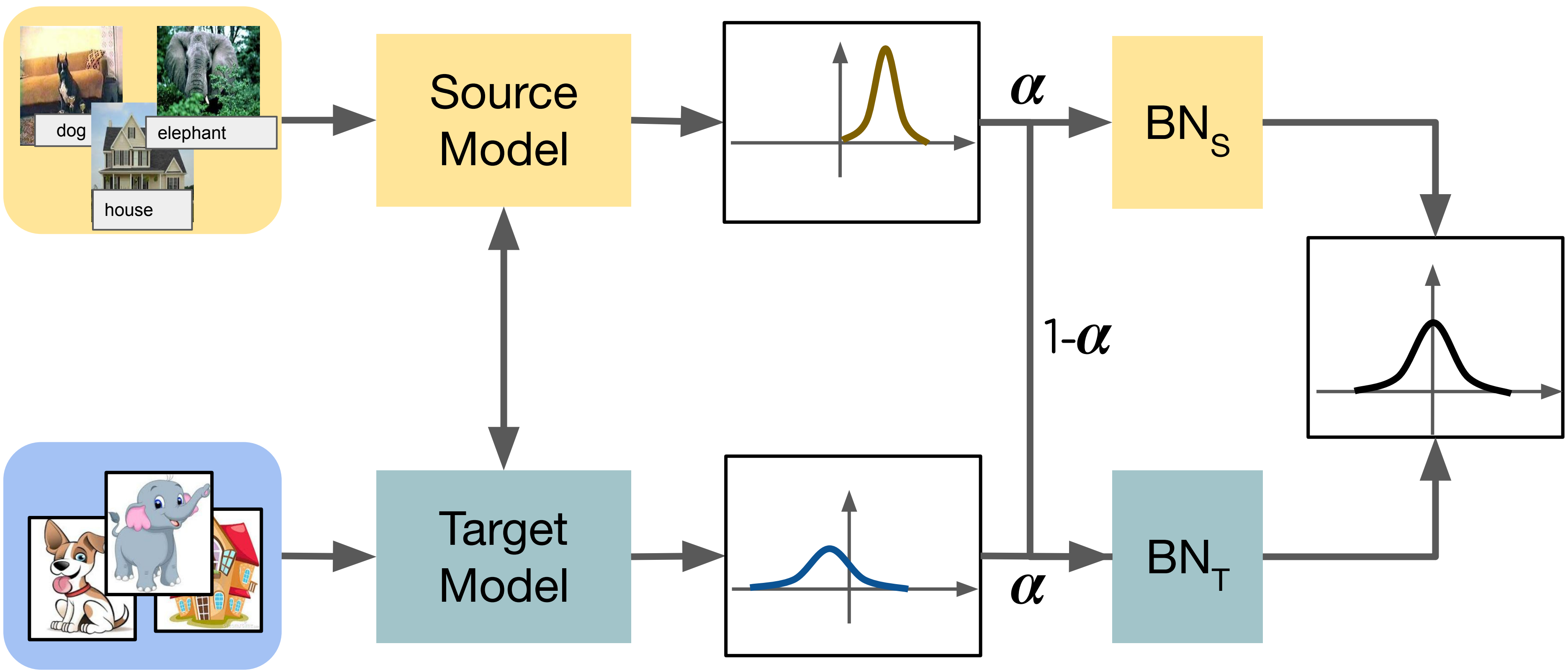
Ioffe, S., & Szegedy, C. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift: Automatic Domain Alignment Layers”. *ICML 2015.*
http://cs231n.github.io/neural-networks-2/#batchnorm

DA through Batch Normalization



Carlucci, F. M., Porzi, L., Caputo, B., Ricci, E., & Rota Bulò, S. "AutoDIAL: Automatic Domain Alignment Layers". *ICCV 2017.*
Carlucci, F. M., Porzi, L., Caputo, B., Ricci, E., & Rota Bulò, S. "Just dial: Domain alignment layers for unsupervised domain adaptation". ICIAP 2017.

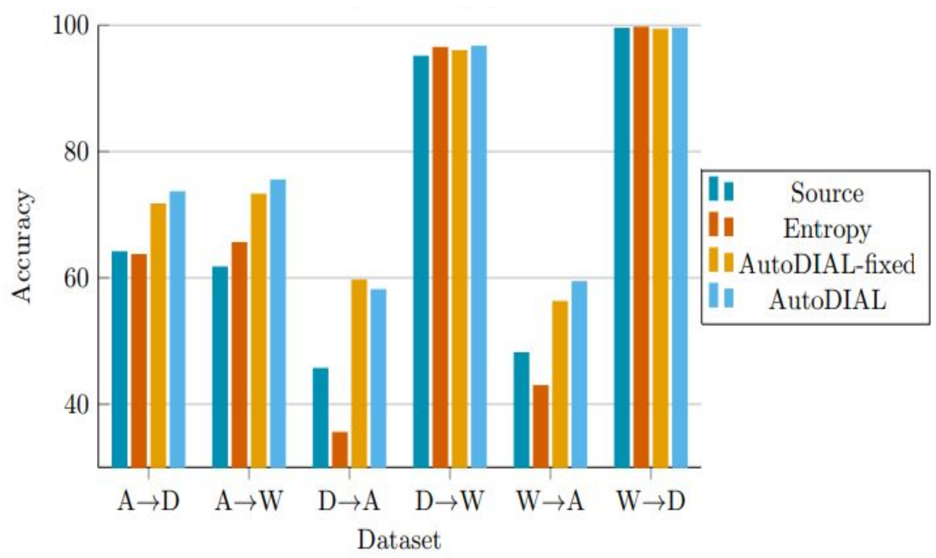
DA through Batch Normalization

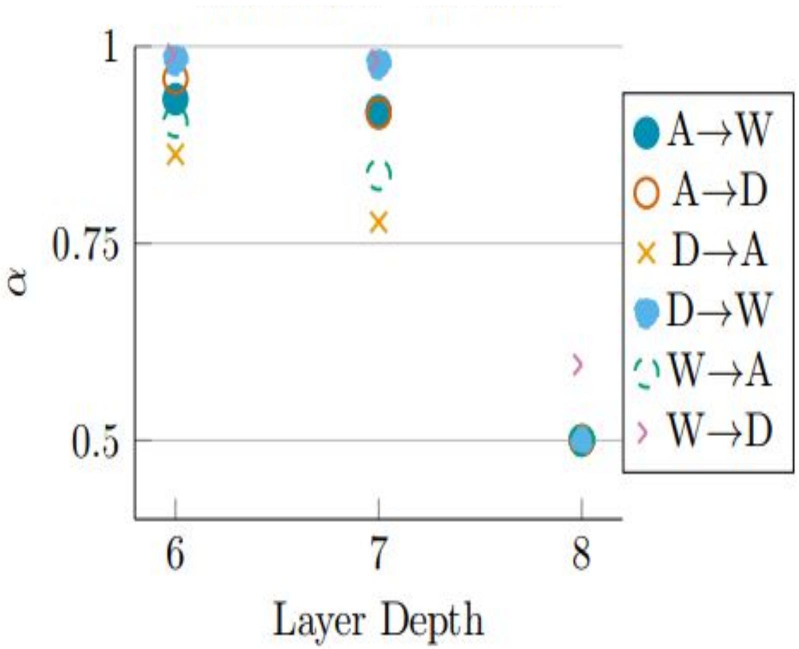


Carlucci, F. M., Porzi, L., Caputo, B., Ricci, E., & Rota Bulò, S. "AutoDIAL: Automatic Domain Alignment Layers". *ICCV 2017.*
Carlucci, F. M., Porzi, L., Caputo, B., Ricci, E., & Rota Bulò, S. "Just dial: Domain alignment layers for unsupervised domain adaptation". ICIAP 2017.

Results

Office 31 dataset - AlexNet

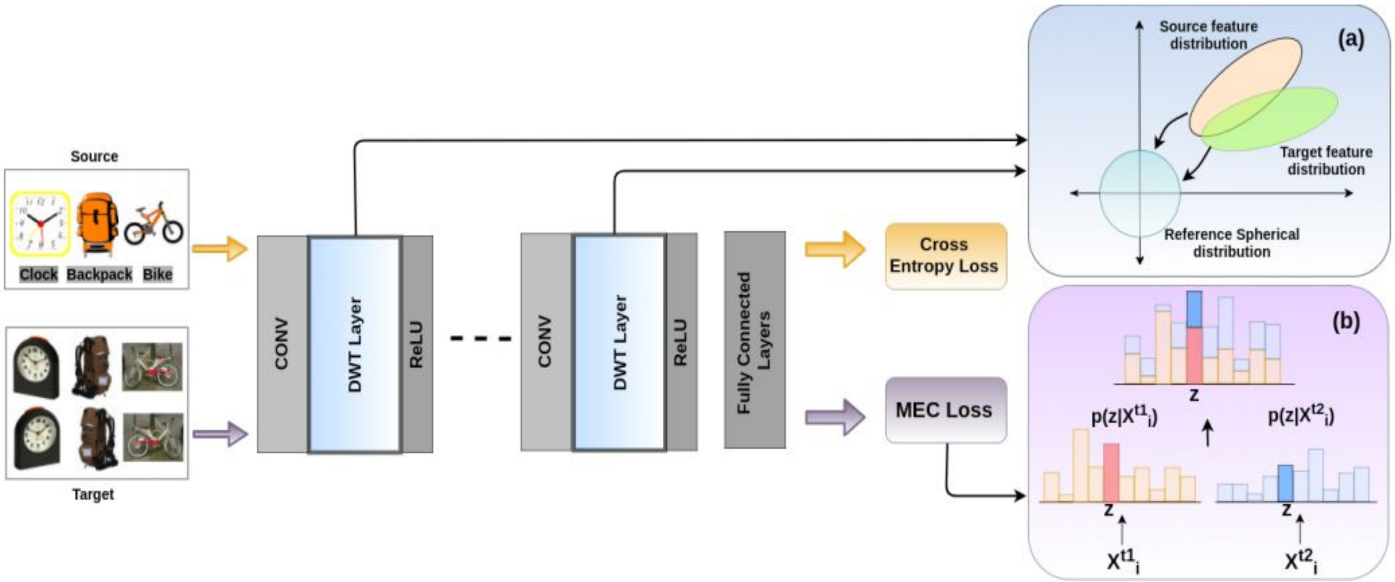




Carlucci, F. M., Porzi, L., Caputo, B., Ricci, E., & Rota Bulò, S.. "AutoDIAL: Automatic Domain Alignment Layers". *ICCV 2017.*

Whitening vs Batch Normalization

- **Key Idea:** improve over DIAL with Domain Whitening Layers (DWT)
  - Domain-alignment layers based on feature whitening
  - Exploit target data with a novel consensus loss (integrate entropy and consistency in a single loss)

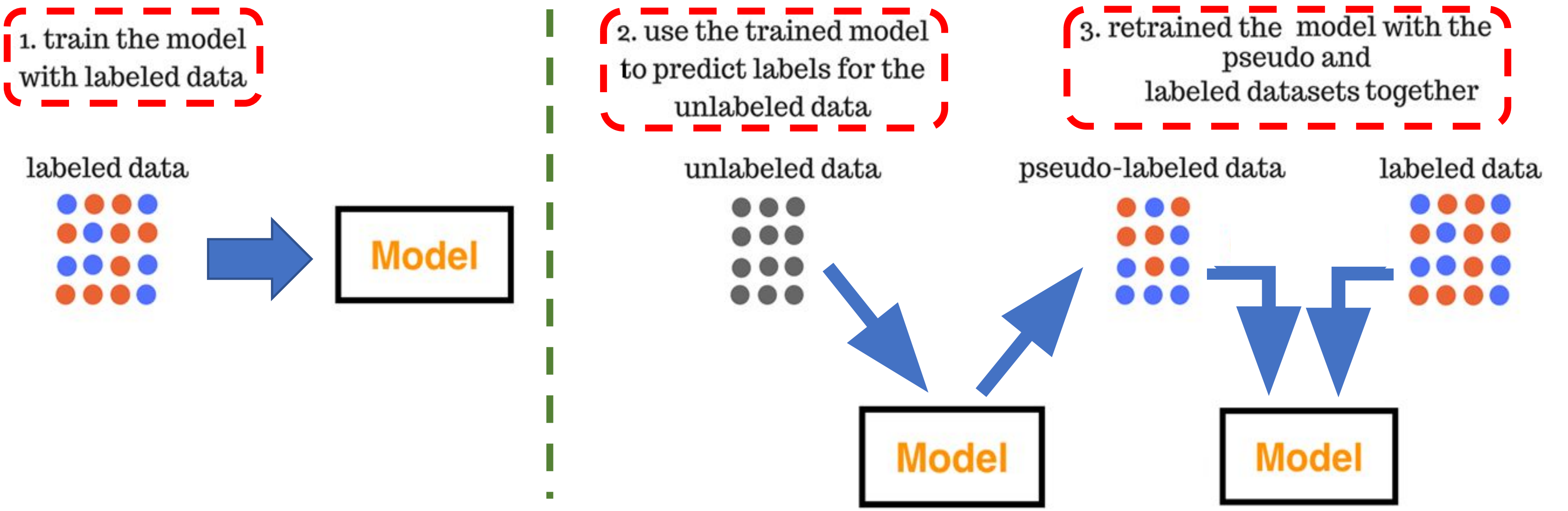


Roy, S., Siarohin, A., Sangineto, E., Bulo, S. R., Sebe, N., Ricci, E."Unsupervised Domain Adaptation using Feature-Whitening and Consensus Loss". CVPR 2019.

Domain Adaptation
Via Pseudo-Labelling

Pseudo-Labeling: A Naive Semi-Supervised Learning Method

First proposed by Lee in 2013, in which network is trained in a supervised fashion with labeled and unlabeled data simultaneously.



Dong-Hyun Lee. "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks." *Workshop on Challenges in Representation Learning, @ ICML* 2013.

Cont'd.

- For unlabeled data, just picking up the class which has the maximum predicted probability – *pseudo-labels* – and use that as if they were true labels.

$$y'_i = \begin{cases} 1 & \text{if } i = \text{argmax}_{i'} f_{i'}(x) \\ 0 & \text{otherwise} \end{cases}$$

- Because the total number of labeled data and unlabeled data is quite different and the training balance between them is quite important for the network performance, the overall loss function is

$$L = \frac{1}{n} \sum_{m=1}^{n} \sum_{i=1}^{C} L(y_i^m, f_i^m) + \alpha(t) \frac{1}{n'} \sum_{m=1}^{n'} \sum_{i=1}^{C} L(y_i'^m, f_i'^m),$$

$$\text{Loss per Batch} = \text{Labeled Loss} + Weight * \text{Unlabeled Loss}$$

Pseudo-labelled data is noisy!

UDA reduces to the problem of **Learning with Noisy Labels**

Issue: Deep nets can easily overfit noise

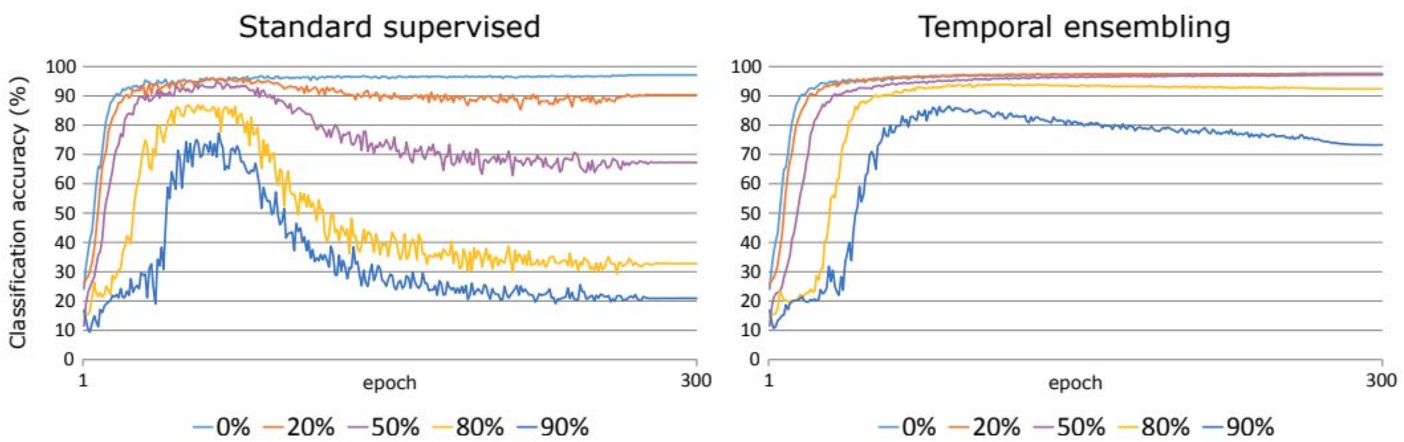
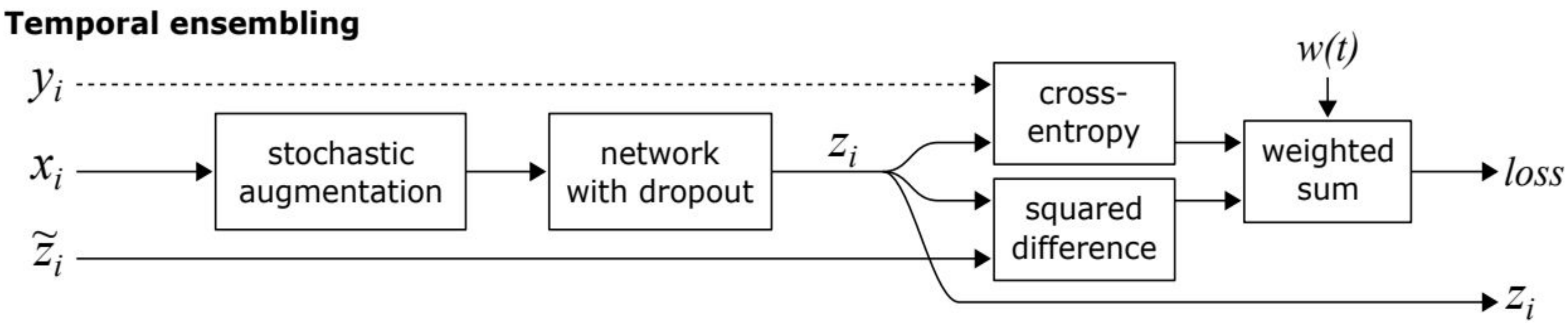


Figure 2: Percentage of correct SVHN classifications as a function of training epoch when a part of the labels is randomized. With standard supervised training (left) the classification accuracy suffers when even a small portion of the labels give disinformation, and the situation worsens quickly as the portion of randomized labels increases to 50% or more. On the other hand, temporal ensembling (right) shows almost perfect resistance to disinformation when half of the labels are random, and retains over ninety percent classification accuracy even when 80% of the labels are random.

Samuli Laine and Timo Aila. "Temporal ensembling for semi-supervised learning." *ICLR 2017*.

Temporal ensembling (Averaging label predictions)



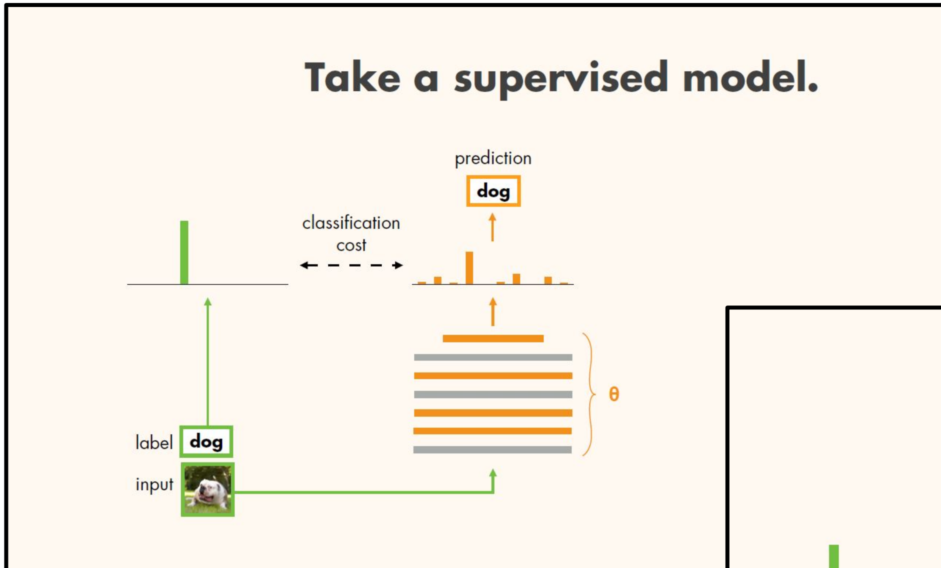
- After every training epoch, the network outputs z_i are accumulated into ensemble outputs Z_i by
  - updating $Z_i \leftarrow \alpha Z_i + (1 - \alpha)\, z_i$, where α is a momentum term that controls how far the ensemble reaches into training history.
- Because of dropout regularization and stochastic augmentation, Z thus contains a weighted average of the outputs of an ensemble of networks f from previous training epochs, with recent epochs having larger weight than distant epochs.
- For generating the training targets $\tilde{z}_i$, we need to correct for the startup bias in Z by dividing by factor $(1 - \alpha^t)$.

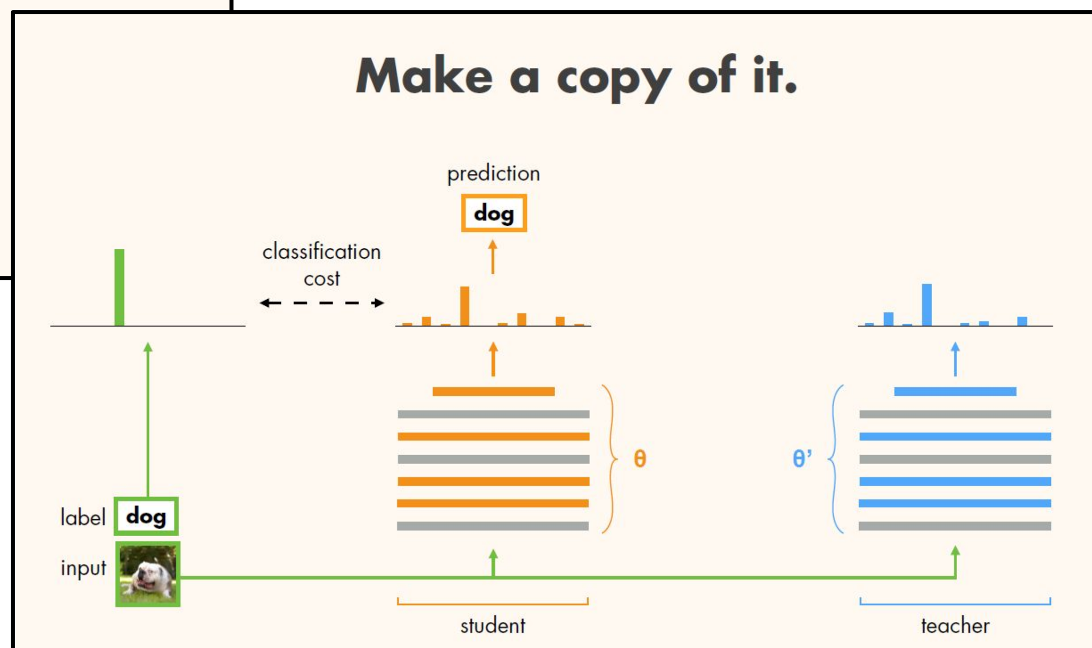
Mean Teachers
(Averaging model weights)

-
- The teacher model is an average of consecutive student models, so we call it **Mean Teacher** method.
- Averaging model weights over training steps tends to produce a more accurate model than using the final weights directly
- Instead of sharing the weights with the student model, the teacher model uses the exponential moving average (EMA) weights of the student model: it can aggregate information after every step instead of every epoch.
- In addition, since the weight averages improve all layer outputs, not just the top output, the target model has better intermediate representations.
- As a result, Mean Teacher improves test accuracy and enables training with fewer labels than Temporal Ensembling, without changing the network architecture.

Tarvainen, Antti, and Harri Valpola. "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.“ *NIPS* 2017.

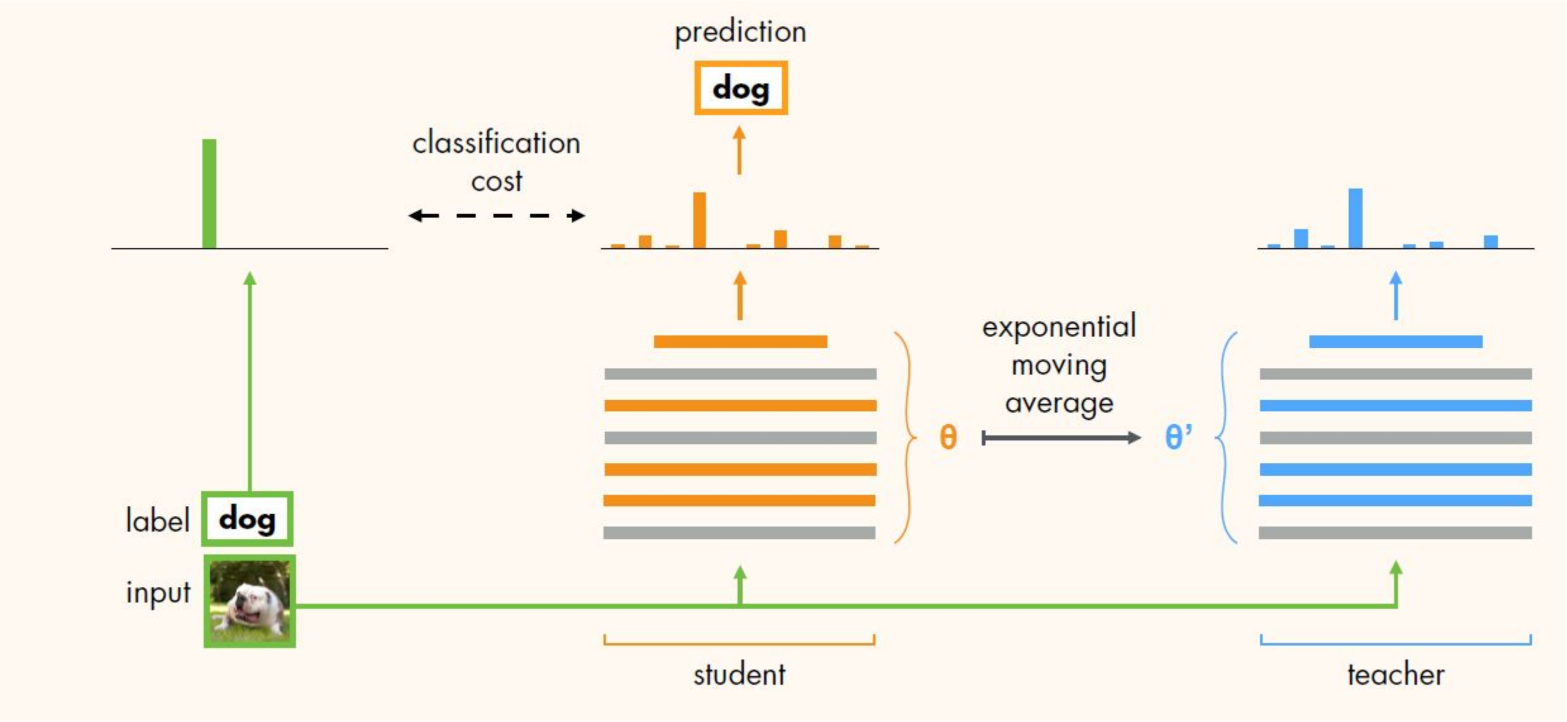
Mean Teachers
(Averaging model weights)





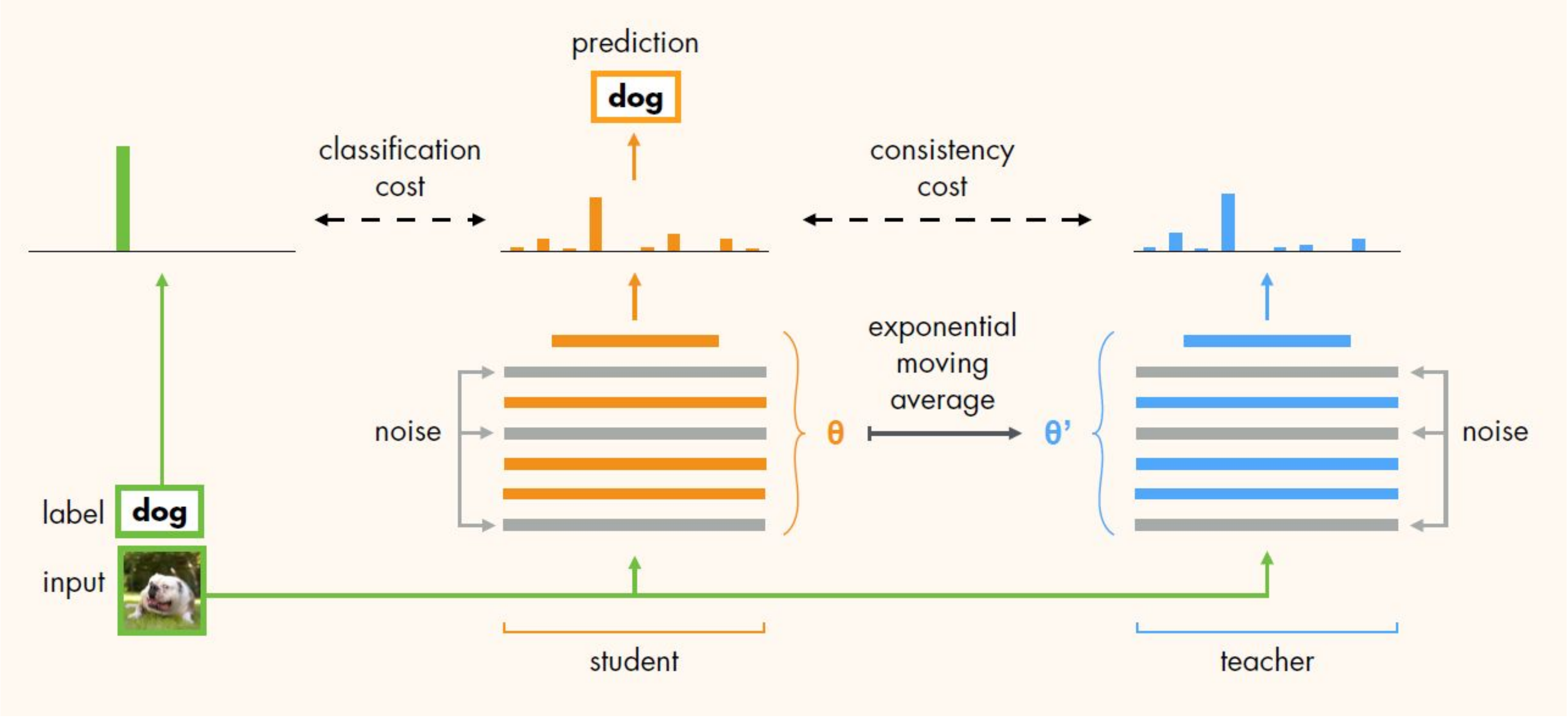
Mean Teachers
(Averaging model weights)

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t$$



Mean Teachers
(Averaging model weights)

$$J(\theta) = \mathbb{E}_{x,\eta',\eta}\left[\left\| f(x,\theta',\eta') - f(x,\theta,\eta) \right\|^2\right]$$

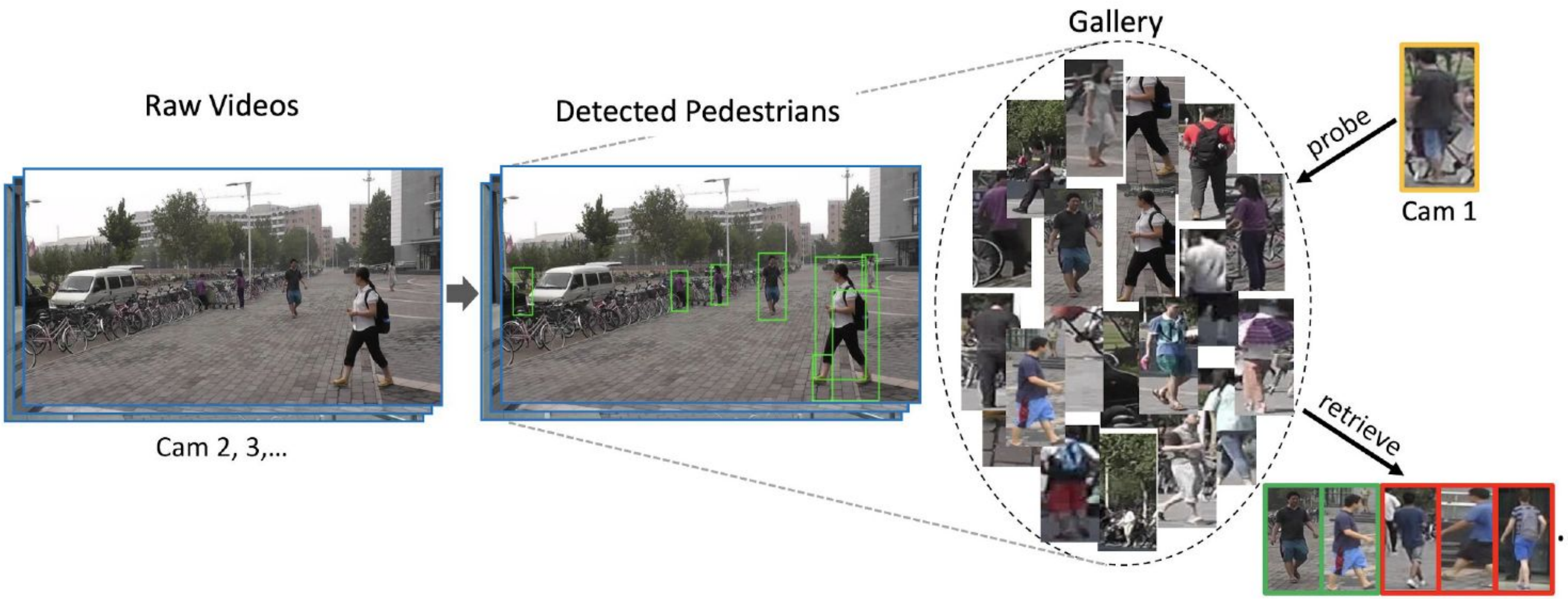


Mutual Mean Teaching
UDA for Person -Re-identification

- Person re-identification (re-ID) aims at identifying the same persons' images across different cameras.
- However, domain diversities between different datasets pose an evident challenge for adapting the re-ID model trained on one dataset to another one.

Y. Ge, Chen, and Li. "Mutual Mean-Teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification." *ICLR* 2020. https://www.youtube.com/watch?v=IQFL3nIYavk

Person Re-identification



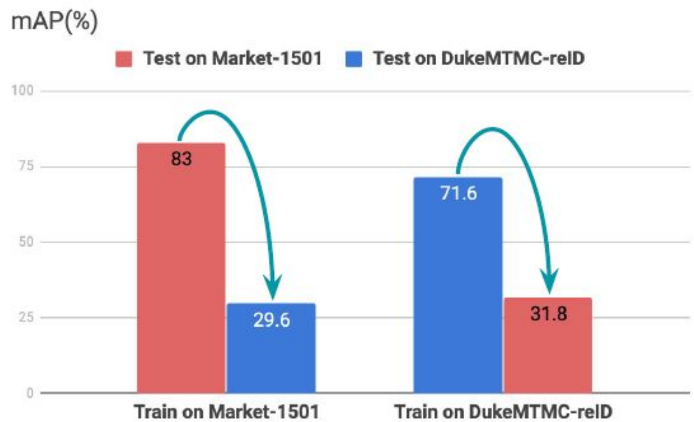
(a) Pedestrian Detection

(b) Person Re-identification

Person Re-identification & Domain Shift

Market-1501[2]

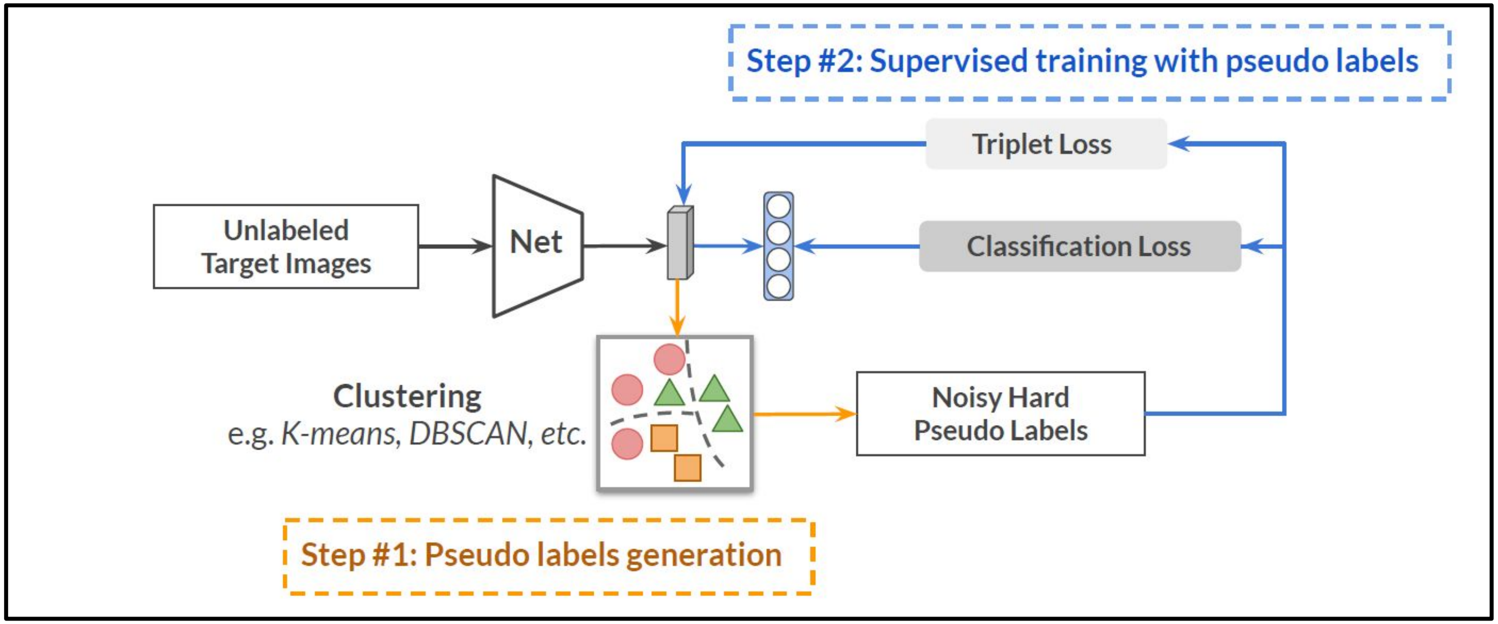
Captured in Tsinghua University

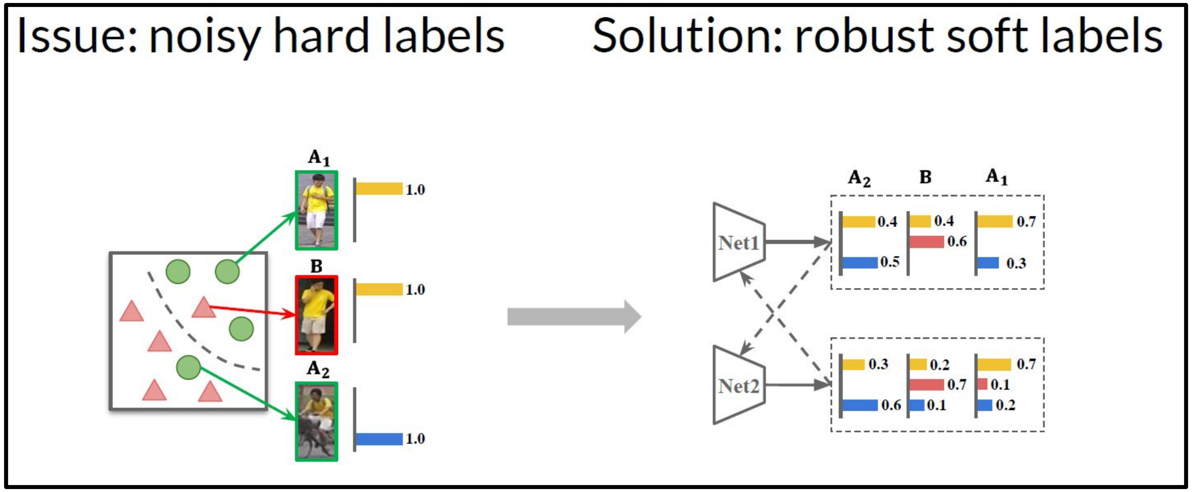


DukeMTMC-reID[3]

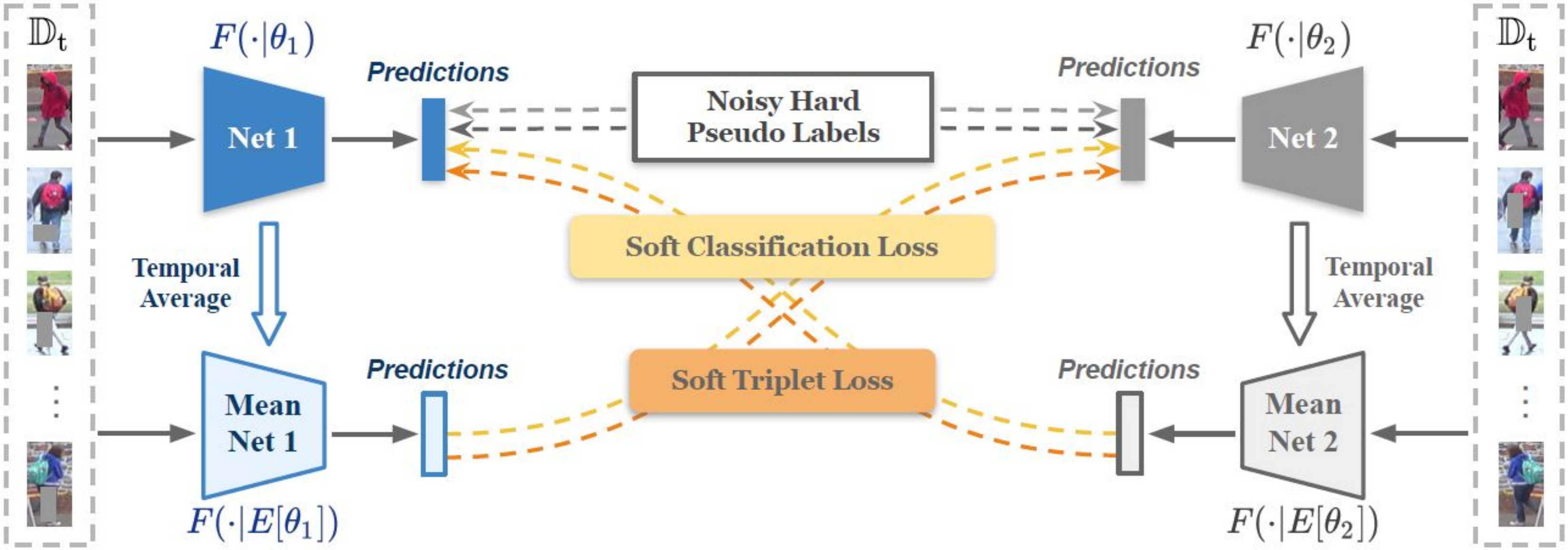
Captured in Duke University

General Pipeline for CLustering Based Pseudo Labelling for UDA





MMT



Y. Ge, Chen, and Li. "Mutual Mean-Teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification." *ICLR* 2020. https://www.youtube.com/watch?v=lQFL3nIYavk