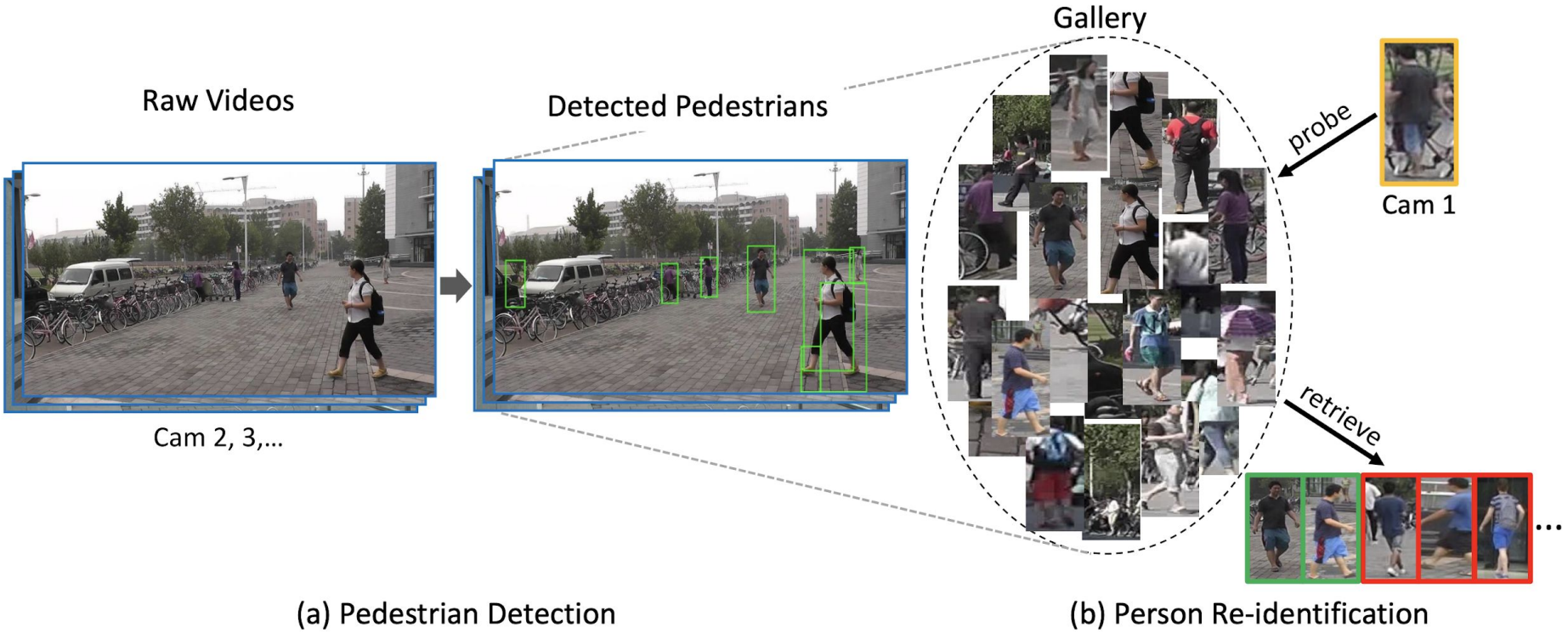
Improved Mutual Mean-Teaching for Unsupervised Domain Adaptive Re-ID

(2nd Place Solution to VisDA-2020 Challenge)



Yixiao Ge, Shijie Yu, Dapeng Chen
The Chinese University of Hong Kong
yxge@link.cuhk.edu.hk

Person re-identification (re-ID)



Zheng L, et al. Person re-identification: Past, present and future[J]. arXiv preprint arXiv:1610.02984, 2016.

Visual Domain Adaptation Challenge (VisDA-2020)

Synthetic (PersonX[A])

Source domain (labeled)

Adaptation

1) There are **larger domain gaps** between synthetic and real scenarios than those between real and real scenarios;

2) The **distribution** of unlabeled target-domain data is much **more realistic** than previous benchmarks.

Real-world images[B]

Target domain (unlabeled)

[A] Sun X., et al. Dissecting person re-identification from the viewpoint of viewpoint. CVPR, 2019.
[B] https://github.com/Simon4Yan/VisDA2020

Existing Methods for UDA re-ID

- **Domain-translation-based** methods (e.g. SPGAN, PTGAN, SDA, etc.)

Main steps: 1) translate the source-domain images to have the target-domain style while well preserving their original IDs or inter-sample relations; 2) adapt the network to the target domain by training with source-to-target translated images.

Features: fully explore the potential of source-domain images and their ground-truth identities.

- **Pseudo-label-based** methods (e.g. SSG, PAST, MMT, etc.)

Main steps: alternates between 1) generating pseudo labels by either clustering instance features or measuring similarities with exemplar features; 2) training the network with target-domain data and their pseudo labels.

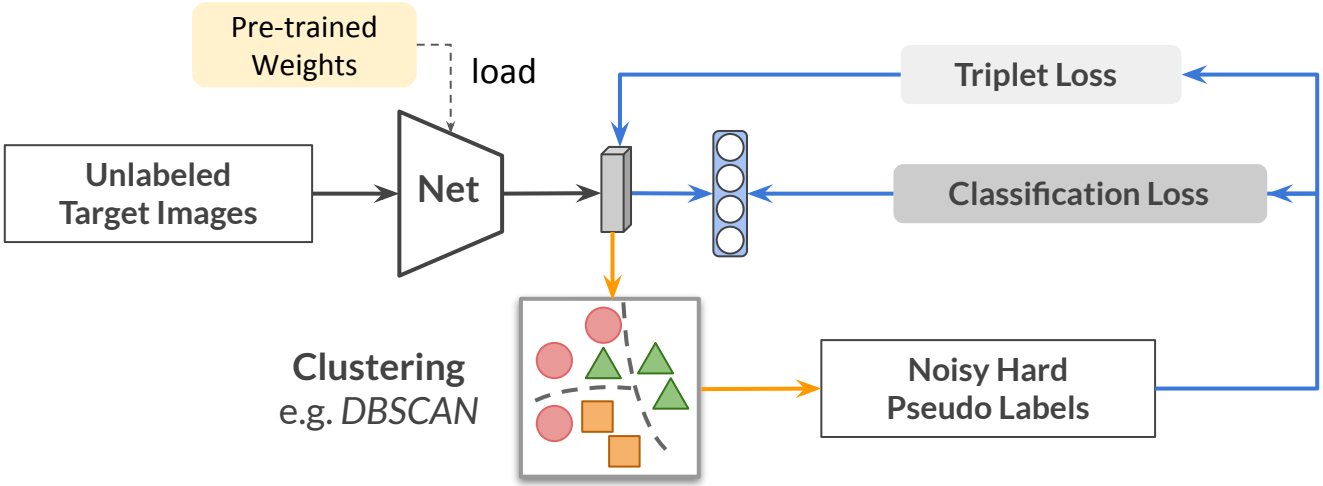
Features: model relations among target-domain instances and achieve SoTA performance.

A Clustering-based Baseline for UDA re-ID (i) Source-domain Pre-training

- **Training objectives**
  - Classification loss: $\mathcal{L}_{\text{cls}}^{s}(\mathcal{F}^{s}, \mathcal{C}^{s}) = \mathbb{E}_{x^{s} \sim \mathbb{X}^{s}} \left[\ell_{\text{ce}}(\mathcal{C}^{s}(\boldsymbol{f}^{s}), y^{s}) \right]$
  - Triplet loss: $\mathcal{L}_{\text{tri}}^{s}(\mathcal{F}^{s}) = \mathbb{E}_{x^{s} \sim \mathbb{X}^{s}} \left[\ell_{\text{bce}}(\mathcal{T}(\boldsymbol{f}^{s}), \mathbf{1}) \right]$

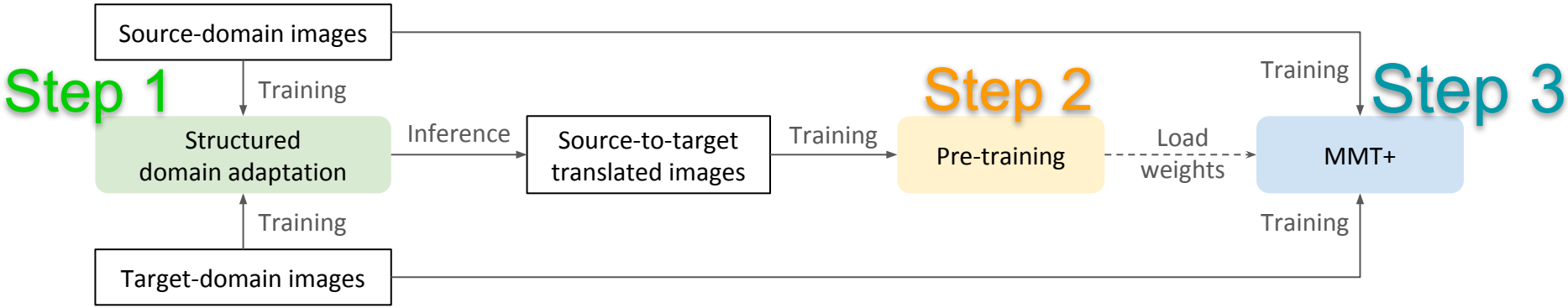
$$\mathcal{T}(\boldsymbol{f}^{s}) = \frac{\exp(\|\boldsymbol{f}^{s} - \boldsymbol{f}_{n}^{s}\|)}{\exp(\|\boldsymbol{f}^{s} - \boldsymbol{f}_{p}^{s}\|) + \exp(\|\boldsymbol{f}^{s} - \boldsymbol{f}_{n}^{s}\|)}$$

A Clustering-based Baseline for UDA re-ID (ii) Target-domain Training

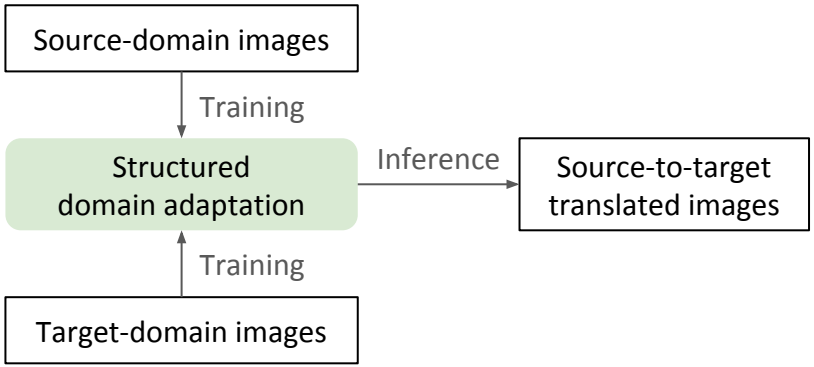


- **Training objectives**
  - Classification loss: $\mathcal{L}_{\text{cls}}^{t}(\mathcal{F}^{t}, \mathcal{C}^{t}) = \mathbb{E}_{x^{t} \sim \mathbb{X}^{t}} \left[\ell_{\text{ce}}(\mathcal{C}^{t}(\boldsymbol{f}^{t}), \hat{y}^{t}) \right]$
  - Triplet loss: $\mathcal{L}_{\text{tri}}^{t}(\mathcal{F}^{t}) = \mathbb{E}_{x^{t} \sim \mathbb{X}^{t}} \left[\ell_{\text{bce}}(\mathcal{T}(\boldsymbol{f}^{t}), \mathbf{1}) \right]$

Our Training Pipeline



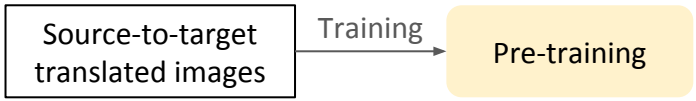
Our Training Pipeline



Step 1:

Structured Domain Adaptation (SDA) framework is trained to translate source-domain images to the target domain.

Our Training Pipeline



Step 2:

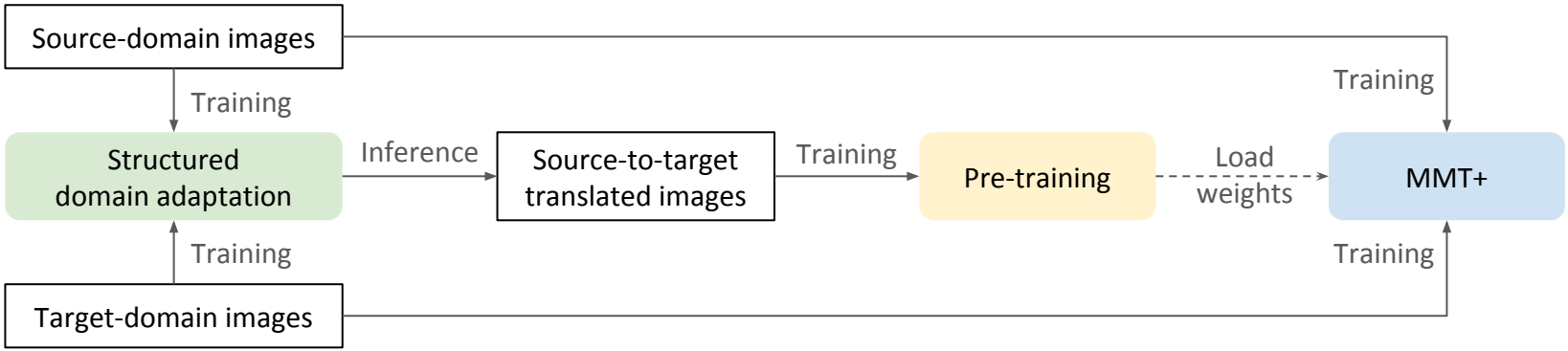
Source-to-target translated images serve as training samples to pre-train the network with ground-truth identities. The network can then be roughly adapted to the target domain.

Our Training Pipeline

Step 3:

The pre-trained network is further fine-tuned on the target domain with the improved Mutual Mean-Teaching (MMT+) framework. Both labeled source-domain raw images and unlabeled target-domain images are used for training.

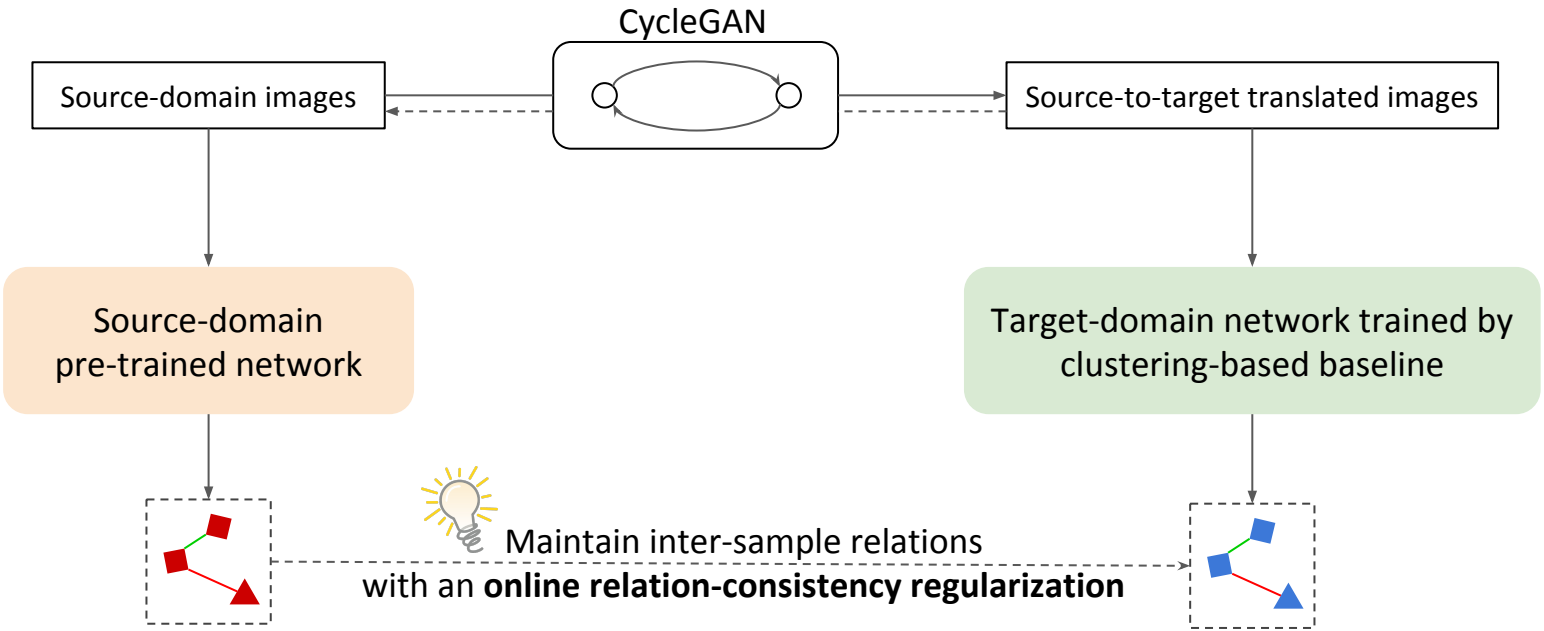
Our Training Pipeline



- Structured Domain Adaptation (**SDA**) framework
- Improved Mutual Mean-Teaching (**MMT+**) framework

A joint pipeline to properly make use of both the domain-translation-based and pseudo-label-based frameworks, which are complementary to each other.

Step 1: Structured Domain Adaptation



$$\mathcal{L}_{\mathrm{rc}}(\mathcal{G}^{s\to t}) = \mathbb{E}_{x^s\sim\mathbb{X}^s}\left[\ell_{\mathrm{bce}}(\mathcal{T}(\boldsymbol{f}^{s\to t}), \mathcal{T}(\boldsymbol{f}^{s}))\right]$$

Ge Y, et al. Structured Domain Adaptation with Online Relation Regularization for Unsupervised Person Re-ID. arXiv preprint arXiv:2003.06650, 2020.

Step 1: Structured Domain Adaptation -- Translated Images

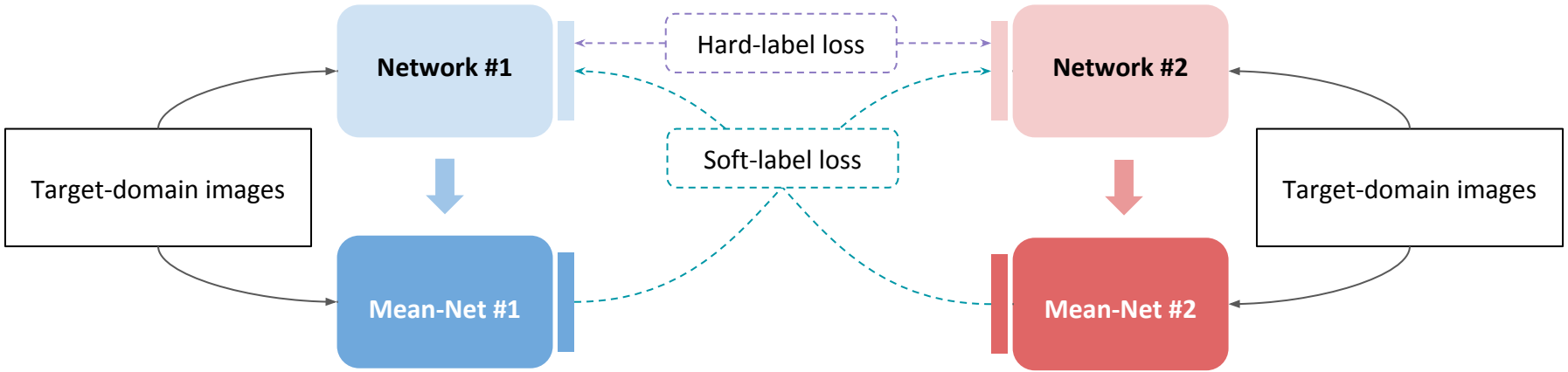
Step 2: Pre-training with Source-to-target Translated Images

- **Training objectives**
  - Classification loss: $\mathcal{L}_{\mathrm{cls}}^{s}(\mathcal{F}^{s}, \mathcal{C}^{s}) = \mathbb{E}_{x^{s\to t}\sim\mathbb{X}^{s\to t}}\left[\ell_{\mathrm{ce}}(\mathcal{C}^{s}(\hat{\boldsymbol{f}}^{s\to t}), y^{s})\right]$
  - Triplet loss: $\mathcal{L}_{\mathrm{tri}}^{s}(\mathcal{F}^{s}) = \mathbb{E}_{x^{s\to t}\sim\mathbb{X}^{s\to t}}\left[\ell_{\mathrm{bce}}(\mathcal{T}(\hat{\boldsymbol{f}}^{s\to t}), \mathbf{1})\right]$

- **Useful Training Tricks**
  - Auto-augmentation (avoid over-fitting on the source domain);
  - Adopt GeM pooling instead of average pooling;
  - Forward the target-domain images into networks without backpropagation (adapt BNs), e.g. pseudo code

```
# for each iteration
source_out = model(source_img)
target_out = model(target_img)
loss = criterion(source_out, source_ids)
loss.backward()
```

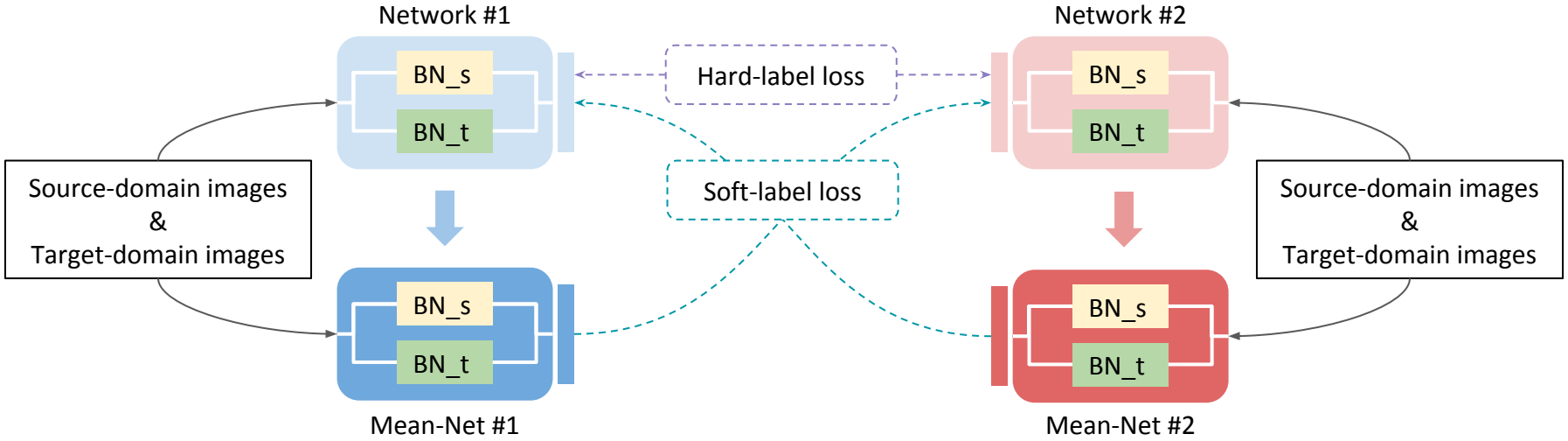
Step 3: Improved Mutual Mean-Teaching (MMT+)



Original Mutual Mean-Teaching (MMT) framework

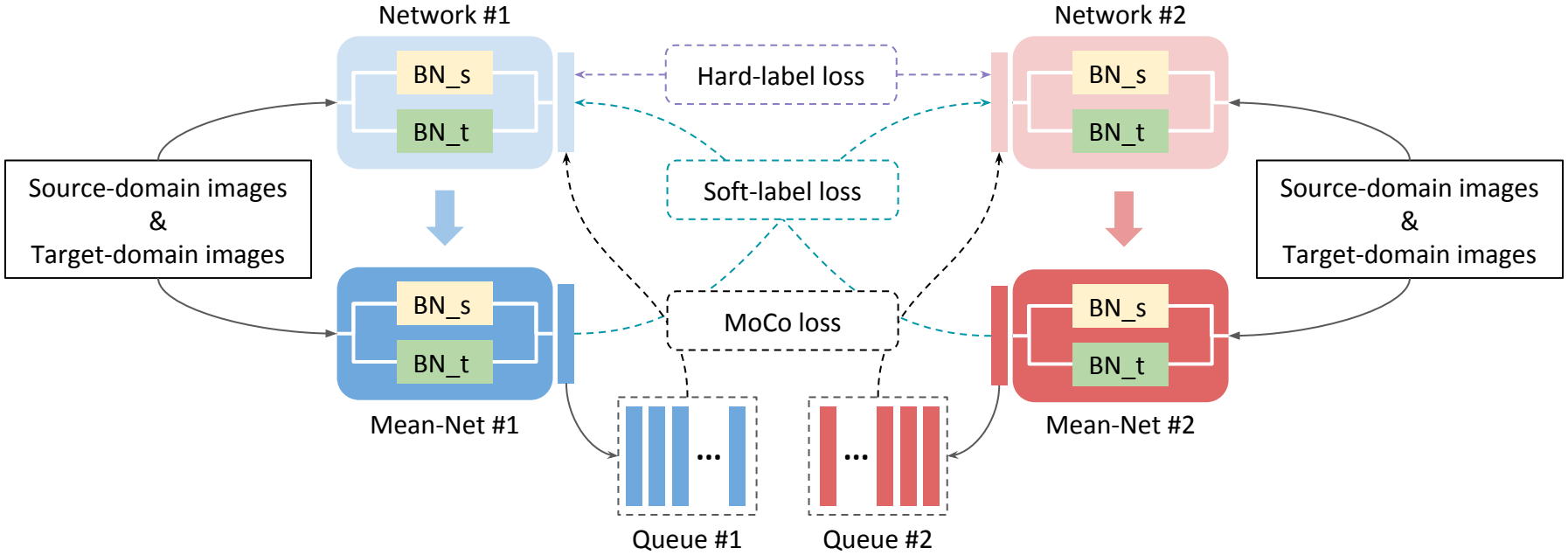
Ge Y, et al. Mutual Mean-Teaching: Pseudo Label Refinery for Unsupervised Domain Adaptation on Person Re-identification. ICLR, 2020.

Step 3: Improved Mutual Mean-Teaching (MMT+)



- Modeling inter-samples relations **across two domains** by jointly training with two domains' images (minimize the domain gaps in mini-batches with domain-specific BatchNorms)

Step 3: Improved Mutual Mean-Teaching (MMT+)



- To further mitigate the effects caused by pseudo label noise, we propose to adopt a MoCo[A] loss to **maintain the instance discrimination**. Note that the mean-net in MMT is almost the same as momentum encoder in MoCo, so MoCo loss can be easily added without extra costs.

[A] He K., et al. Momentum Contrast for Unsupervised Visual Representation Learning. CVPR, 2020.

Step 3: Improved Mutual Mean-Teaching (MMT+)

- **Training objectives**
  - Hard-label loss: $\mathcal{L}_{\text{hard}}^{t}(\mathcal{F}_1^t, \mathcal{F}_2^t, \mathcal{C}_1^t, \mathcal{C}_2^t) = \mathbb{E}_{x\sim\mathbb{X}}\left[\ell_{\text{ce}}(\mathcal{C}_1^t(\mathcal{F}_1^t(x)), y) + \ell_{\text{ce}}(\mathcal{C}_2^t(\mathcal{F}_2^t(x)), y)\right]$
    (Classification loss)
  - Soft-label loss: $\mathcal{L}_{\text{soft}}^{t}(\mathcal{F}_1^t, \mathcal{F}_2^t, \mathcal{C}_1^t, \mathcal{C}_2^t) = -\mathbb{E}_{x\sim\mathbb{X}}\Big[\mathbb{E}[\mathcal{C}_2^t](\mathbb{E}[\mathcal{F}_2^t](x)) \cdot \log \mathcal{C}_1^t(\mathcal{F}_1^t(x)) + \mathbb{E}[\mathcal{C}_1^t](\mathbb{E}[\mathcal{F}_1^t](x)) \cdot \log \mathcal{C}_2^t(\mathcal{F}_2^t(x))\Big]$
    (Soft cross-entropy loss)
  - MoCo loss: $\mathcal{L}_{\text{moco}}^{t}(\mathcal{F}_1^t, \mathcal{F}_2^t) = -\mathbb{E}_{x\sim\mathbb{X}}\left[\log \frac{\exp(\langle \mathcal{F}_1^t(x), \mathbb{E}[\mathcal{F}_1^t](x)\rangle/\tau)}{\exp(\langle \mathcal{F}_1^t(x), \mathbb{E}[\mathcal{F}_1^t](x)\rangle/\tau) + \sum_{\boldsymbol{k}_1^-}\exp(\langle \mathcal{F}_1^t(x), \boldsymbol{k}_1^-\rangle/\tau)} + \log \frac{\exp(\langle \mathcal{F}_2^t(x), \mathbb{E}[\mathcal{F}_2^t](x)\rangle/\tau)}{\exp(\langle \mathcal{F}_2^t(x), \mathbb{E}[\mathcal{F}_2^t](x)\rangle/\tau) + \sum_{\boldsymbol{k}_2^-}\exp(\langle \mathcal{F}_2^t(x), \boldsymbol{k}_2^-\rangle/\tau)}\right]$

Step 3: Improved Mutual Mean-Teaching (MMT+)

- **Useful Training Tricks**

  - Adopt arcface or cosface metrics to replace conventional hard-label classification loss;
  - Hard/soft triplet loss may be redundant as arcface/cosface itself is strong enough;
  - Adopt GeM pooling instead of average pooling;
  - Do not use auto-augmentation which benefits pre-training only (random erasing is ok);
  - Reset optimizers upon re-clustering.

Post-processing for Inference

- **Model ensemble**
  - ResNeSt50, ResNeSt101, DenseNet169-IBN, ResNeXt101-IBN;
  - Features encoded by the above backbones are concatenated and L2-normalized.

- **Negative camera similarities[A]**
  - Train a camera classification network with a backbone of ResNeSt50;
  - Camera similarities are subtracted by person similarities with a weight of 0.1:

$$s(q,k) = \|\boldsymbol{f}_q - \boldsymbol{f}_k\| - 0.1\|\mathcal{F}_c(q) - \mathcal{F}_c(k)\|$$

- **K-reciprocal re-ranking[B]**
  - k1=30, k2=6, lambda_value=0.3

[A] Zhu, X., et al.: Voc-reid: Vehicle re-identification based on vehicle- orientation-camera. CVPRW, 2020.
[B] Zhong, Z., et al.: Re-ranking person re-identification with k-reciprocal encoding. CVPR, 2017.

Ablation Studies -- SDA

Table 2. Ablation study on the effectiveness of source-to-target images translated by SDA. The results are evaluated on the target_val set. The pre-training for this ablation study adopts the backbone of ResNet50-IBN [12].

Images for pre-training	mAP(%)	top-1(%)
Raw source-domain images	61.0	71.6
Source-to-target images translated by SPGAN [5][a]	68.2	75.1
Source-to-target images translated by SDA [9]	**71.2**	**79.3**

[a]Downloaded from https://github.com/Simon4Yan/VisDA2020.

Comparison of the performance of source-domain pre-training

Ablation Studies -- MMT+

Table 3. Ablation study on the effectiveness of the improved MMT. The results are evaluated on the target_val set. The experiments in this ablation study adopts the backbone of ResNet50-IBN [12]. All the post-processing techniques except the ensembling as described in Sec. 4.2 are used.

Pseudo-label-based method	mAP(%)	top-1(%)
Original MMT [7]	78.4	86.5
Our MMT+	**81.2**	**87.3**

Comparison between original MMT and MMT+

Ablation Studies -- Backbones

Table 4. Performance of different backbones for MMT+. The results are evaluated on the target_val set. All the post-processing techniques as described in Sec. 4.2 are used.

Backbone for MMT+	mAP(%)	top-1(%)
ResNeSt50 [25]	83.6	89.4
ResNeSt101 [25]	82.7	89.1
DenseNet169-IBN [11, 12]	84.1	90.5
ResNeXt101-IBN [21, 12]	83.5	88.9
Ensemble	**86.3**	**91.2**

Comparison of different backbones for MMT+

Conclusion

- **A joint training pipeline**
  - take advantage of both domain-translation-based and pseudo-label-based methods;
  - two kinds of methods are complementary to each other.

- **Improved MMT**
  - model cross-domain relations by jointly training with two domains' images;
  - further mitigate the effects caused by noisy pseudo labels by maintaining instance discrimination with a MoCo loss.

- **Top-3 teams in VisDA-2020 challenge**

Team Name	mAP(%)	top-1(%)
Vimar Team	76.56	84.25
Ours	**74.78**	**82.86**
Xiangyu	72.39	83.85

Improved Mutual Mean-Teaching for Unsupervised Domain Adaptive Re-ID

Yixiao Ge, Shijie Yu, Dapeng Chen

The Chinese University of Hong Kong
yxge@link.cuhk.edu.hk

Code & Models available at

https://github.com/yxgeee/VisDA-ECCV20