Deep Metric Learning by Online Soft Mining and Class-Aware Attention

Xinshao Wang, Yang Hua, Elyor Kodirov, Guosheng Hu, and Neil M. Robertson
Queen's University Belfast, UK   AnyVision, UK

1 Introduction

- **Deep Metric Learning** (DML): DML aims to learn a deep embedding space such that **relative locations** of input samples are based on their **semantic similarities**, as in Figure 1.

Fig 1: An excerpt of t-SNE plot on CUB-200-2011 test set.

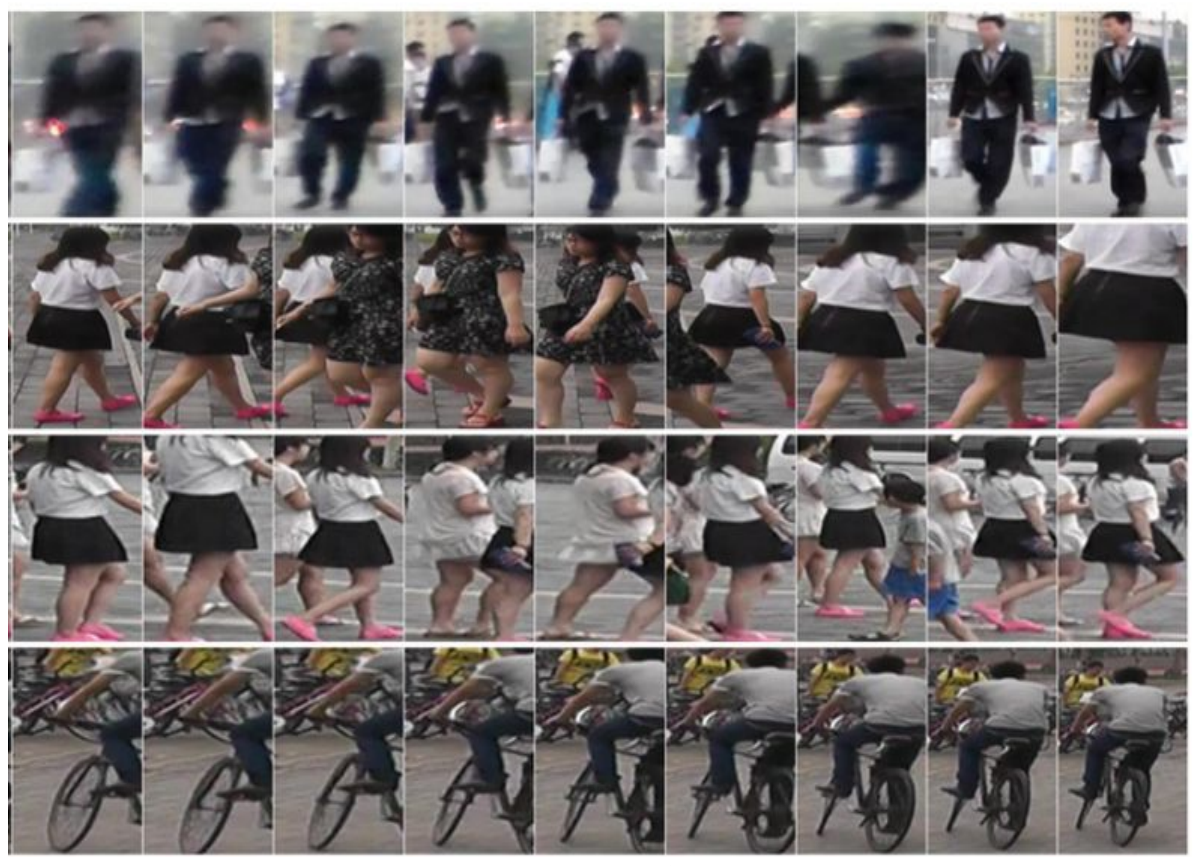


Fig 2: Illustration of trivial samples and outlying ones.

- DML learns representations, thus being fundamental and owning diverse applications.
- **Existing Problems of** DML (Figure 2)
  - Not making full use of all samples in the mini-batch
    a. Attention is necessary due to a large fraction of **trivial samples**
    b. Previous Solution: **binary attention**, i.e., hard sample mining using binary scores
  - Not taking care of **outlying samples** in the training sets
    a. Motion blur
    b. Occlusion
    c. Distractive objects
    d. Truncated objects

Trivial samples: image pairs that can be verified easily and have zero losses.
Outlying samples: Images that do not match their labels well.

2 Methodology

- Traditional contrastive loss for learning an embedding CNN *f*.

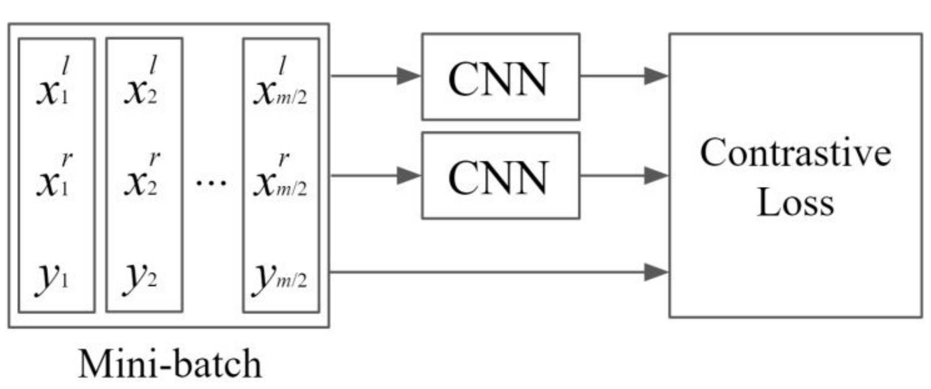


Fig 3: Traditional contrastive loss:
Each input is an images pair with a binay label.

$$\mathbf{f}_i^l = f(\mathbf{x}_i^l) \in \mathbb{R}^D, \mathbf{x}_i^l \in \mathbb{R}^{h\times w\times 3}, y_i \in \{0,1\}$$

$$d_i = \|\mathbf{f}_i^l - \mathbf{f}_i^r\|_2$$

$$L_{cont}^{\alpha}(\mathbf{x}_i^l, \mathbf{x}_i^r; f) = y_i d_i^2 + (1-y_i) max(0, \alpha - d_i)^2$$

- Weighted contrastive loss with OSM and CAA.

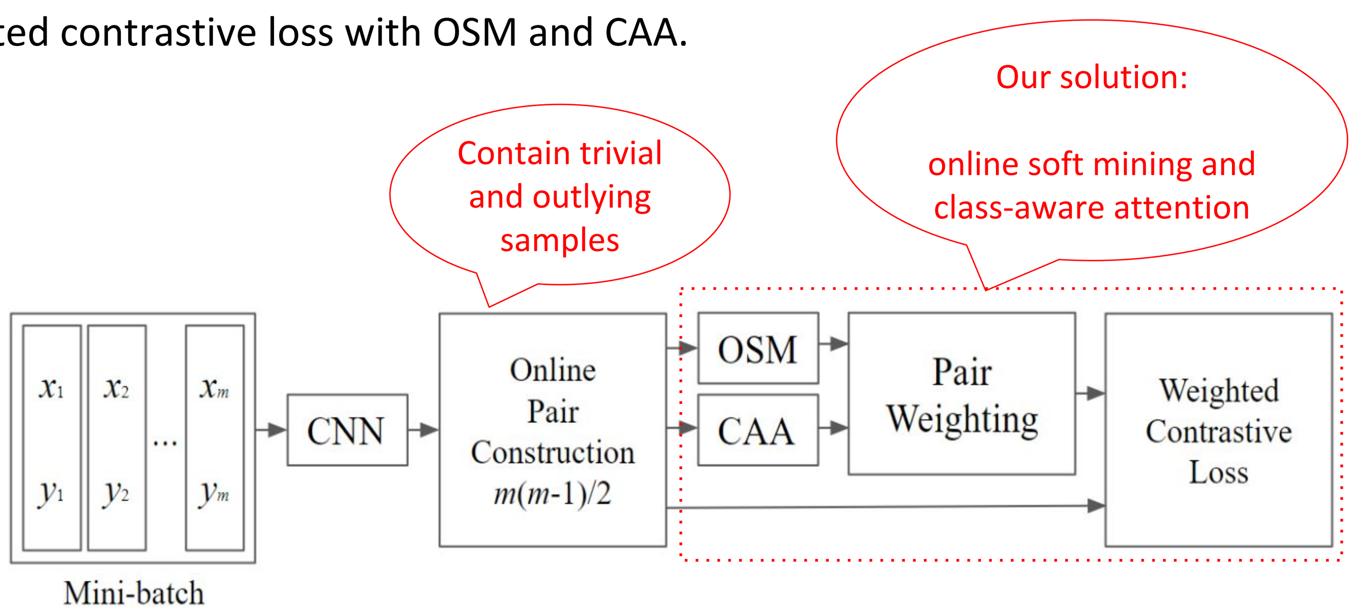
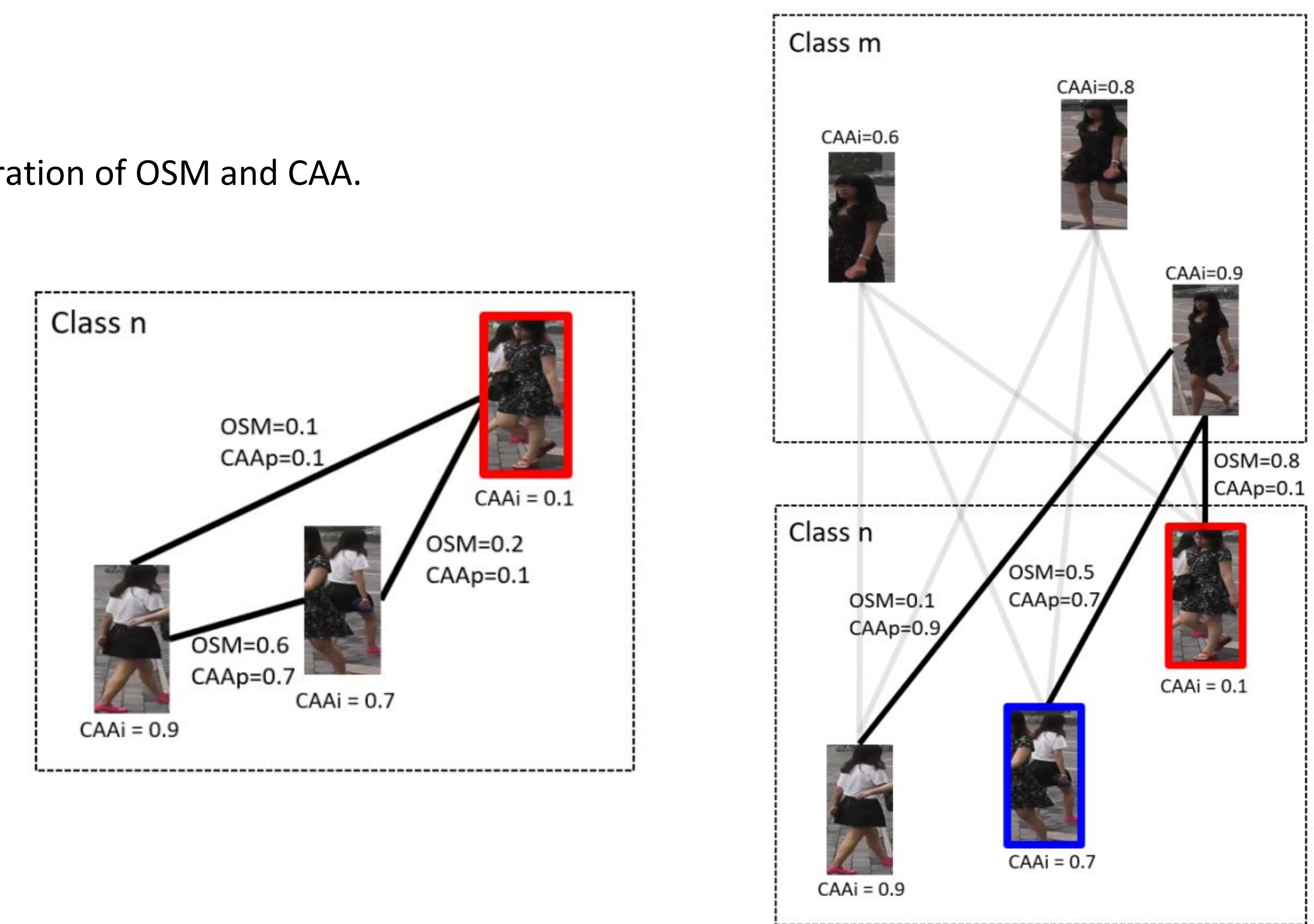


Fig 4: We propose OSM and CAA to take care of both trivial and outlying images.

- Illustration of OSM and CAA.



- Online Soft Mining (OSM) for Positives and Negatives
  - Higher scores to local/closer positives motivated by learning extended manifolds

$$d_{ij} = \|\mathbf{f}_i - \mathbf{f}_j\|_2 \quad s_{ij}^+ = \exp(-d_{ij}^2 \big/ \sigma_{\mathrm{OSM}}{}^2)$$

  - Higher scores to more difficult negatives

$$s_{ij}^- = \max(0, \alpha - d_{ij})$$

- Class-Aware Attention

$$a_i = \frac{\exp(\mathbf{f}_i^\top \mathbf{c}_{y_i})}{\sum_{k=1}^{C} \exp(\mathbf{f}_i^\top \mathbf{c}_k)}.$$

- Final Weight of Each Pair

$$w_{ij}^+ = s_{ij}^+ * a_{ij} \qquad w_{ij}^- = s_{ij}^- * a_{ij} \qquad a_{ij} = \min(a_i, a_j)$$

3 Experiments

- Video-based Person Re-ID
  - Intrinsically, Person ReID is an image retrieval problem with some constraints (pose/camera-invariant).
  - Each input is a video/tracklet instead of an image.
  - Training and testing classes are disjoint.

Table 1: Results on MARS in terms of CMC(%) and mAP(%).

Methods	Attention	1	5	20	mAP
IDE (ResNet50)	No	62.7	–	–	44.1
IDE (ResNet50)+XQDA	No	70.5	–	–	55.1
IDE (ResNet50)+XQDA+Re-ranking	No	73.9	–	–	68.5
CNN+RNN	No	43.0	61.0	73.0	–
CNN+RNN+XQDA	No	52.0	67.0	77.0	–
AMOC+EpicFlow	No	68.3	81.4	90.6	52.9
ASTPN	Yes	44.0	70.0	81.0	–
SRM+TAM	Yes	70.6	90.0	**97.6**	50.7
RQEN	Yes	73.7	84.9	91.6	51.7
RQEN+XQDA+Re-ranking	Yes	77.8	88.8	94.3	71.1
DRSA	Yes	82.3	–	–	65.8
CAE	Yes	82.4	92.9	–	67.5
Ours	Yes	**84.7**	**94.1**	97.0	**72.4**
Ours + Re-ranking	Yes	**86.0**	**94.4**	97.1	**81.0**

- Fine-grained Image Recognition
  - Two different evaluation settings: raw images and cropped images.
  - Image retrieval performance is evaluated, Recall@*K* (%) = CMC-*K* (%).
  - Training and testing classes are disjoint.

Table 2: Results on CARS196 and CUB-200-2011 in terms of Recall@*K* (%). 1st Group: Raw images are used for training and testing . 2nd Group: Cropped images are used for training and testing. * indicates cascaded models.

	CARS196						CUB-200-2011					
K	1	2	4	8	16	32	1	2	4	8	16	32
Contrastive	21.7	32.3	46.1	58.9	72.2	83.4	26.4	37.7	49.8	62.3	76.4	85.3
Triplet	39.1	50.4	63.3	74.5	84.1	89.8	36.1	48.6	59.3	70.0	80.2	88.4
LiftedStruct	49.0	60.3	72.1	81.5	89.2	92.8	47.1	58.9	70.2	80.2	89.3	93.2
Binomial Deviance	–	–	–	–	–	–	52.8	64.4	74. 7	83.9	90.4	94. 3
Histogram Loss	–	–	–	–	–	–	50.3	61.9	72.6	82.4	88.8	93.7
Smart Mining	64.7	76.2	84.2	90.2	–	–	49.8	62.3	74.1	83.3	–	–
HDC*	73.7	83.2	89.5	93.8	96.7	98.4	53.6	65.7	77.0	85.6	91.5	**95.5**
Ours	**74.0**	**83.8**	**90.2**	**94.8**	**97.3**	**98.6**	**55.3**	**67.3**	**77.5**	**85.8**	**91.8**	95.4
PDDM+Triplet	46.4	58.2	70.3	80.1	88.6	92.6	50.9	62.1	73.2	82.5	91.1	94.4
PDDM+Quadruplet	57.4	68.6	80.1	89.4	92.3	94.9	58.3	69.2	79.0	88.4	93.1	95.7
HDC*	83.8	89.8	93.6	96.2	97.8	98.9	60.7	72.4	81.9	89.2	93.7	96.8
Ours	**85.5**	**91.5**	**95.1**	**97.2**	**98.5**	**99.2**	**62.3**	**73.2**	**83.3**	**89.6**	**94.1**	**96.9**

4 Summary

- We address two problems in deep metric learning:
  - OSM for making full use of samples
  - CAA for alleviating the disturbance from outlying samples
- Our approach surpasses the state-of-the-arts by a large margin on two domain tasks:
  - Weighted contrastive loss for incorporating OSM and CAA
  - Intuitive, effective and easy to implement