Dual-Triplet Metric Learning for Unsupervised Domain Adaptation in Video Face Recognition

George Ekladious[§], Hugo Lemoine[§], Eric Granger
Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
Dept. of Systems Engineering, École de technologie supérieure
Montreal, Canada
{george.ekladious, eric.granger}@etsmtl.ca
hugo.lemoine-st-andre.1@ens.etsmtl.ca

Kaveh Kamali, Salim Moudache
Nuvoola Inc.
Montreal, Canada
{kaveh.kamali, salim.moudache}@Nuvoola.com

***Abstract*—The scalability and complexity of deep learning models remains a key issue in many of visual recognition applications. For instance, in video surveillance, fine tuning of a model with labeled image data from each new camera is required to reduce the domain shift between videos captured from the source domain (laboratory setting) and the target domain (operational environment). In many video surveillance applications, like face recognition and person re-identification, a pair-wise matcher is typically employed to assign a query image captured using a video camera to the corresponding reference images in a gallery. The different configuration, viewpoint, and operational conditions of each camera can introduce significant shifts in pair-wise distance distributions, resulting in a decline in recognition performance for new cameras. In this paper, a new deep domain adaptation (DA) method is proposed to adapt the CNN embedding of a Siamese network using unlabeled tracklets captured with a new video camera. To this end, a dual-triplet loss is introduced for metric learning, where two triplets are constructed using video data from a source camera, and a new target camera. In order to constitute the dual triplets, a mutual-supervised learning approach is introduced where the source camera acts as a teacher, providing the target camera with an initial embedding. Then, the student relies on the teacher to iteratively label the positive and negative pairs collected during, e.g., initial camera calibration. Both source and target embeddings continue to simultaneously learn such that their pair-wise distance distributions become aligned. For validation, the proposed metric learning technique is used to train deep Siamese networks under different training scenarios, and is compared to state-of-the-art techniques for still-to-video FR on the COX-S2V and a private video-based FR dataset. Results indicate that the proposed method can provide a level of accuracy that is comparable to the upper bound performance, in training scenario where labeled target data is employed to fine-tune the Siamese network.**

***Index Terms*—Video Surveillance, Face Recognition, Unsupervised Domain Adaptation, Triplet Loss, Deep Learning.**

I. Introduction

Learning discriminant representations from facial images, and efficiently calibrating the resulting embeddings to new capturing devices, conditions, and operational environments is required in a wide range of video analytics and surveillance applications, such as watch-list screening, biometric authentication, video captioning, and web-base search and retrieval. Representation learning methods from still facial images have been studied extensively, where availability of extremely large public datasets facilitates the design of deep learning models that can achieve human-level performance [15].

Video-based face representations, on the other hand, are harder to learn for two main reasons. First, facial images are extracted from videos under unconstrained capture conditions, which can introduce significant variability in facial appearances according to pose, illumination, scale, resolution, expression, etc. Second, there are fewer publicly-available datasets for video FR, and these are smaller in size compared datasets for still FR. This can limit the ability to train reliable deep representations. For instance, large-scale labeled video database publicly available to date, such as YouTube face dataset [21], contains 3.4K videos in total from 1.5K different subjects, as apposed to the still VGG face2 dataset with 3.3M faces from 9K subjects [14].

To address the above challenges, one can reduce the variability of facial images captured so with video cameras prior to matching, and hence exploit powerful still-based face representations. For instance, auto-encoder networks may be employed to learn discriminant face embedding, and to reconstruct high-quality canonical face images (frontal, well-illuminated, less blurred faces with neutral expression) from faces captured under various video conditions [13]. This approach can be challenging for real-time applications, and may require data from the target domain, and can be impractical to calibrate new video sources using low-shot calibration data. Another approach to address the challenges of video-based face representation learning is to generate video-based data from a large still data (i.e., integrate the effects of video capturing conditions to still images), and then use this generated data to design new representations for the video-based face applications. For instance, one can artificially blur training data to account for capture conditions in real-world video surveillance applications. By using training data composed of both still images and artificially blurred data, a deep CNN is encouraged to learn blur insensitive features [3]. Domain-Specific Face Synthesis methods have been proposed where the domain specific variations, e.g., pose, illumination, etc., are projected onto the reference still faces of each individual of interest

[§]Equal contribution

in the gallery, so that they resemble these individuals under the capture conditions relevant to the operational domain [10] [6]. These methods can be complex to calibrate for new video cameras, and may fail to cover the complete range of capturing conditions that can occur during operations.

Recently, the representation learning of video-based faces has been approached from the domain adaptation (DA) perspective. High quality still images that are considered as the source data, and images captured in video under different environment and capturing conditions are considered as the target data. Some methods used labeled target datasets to fine tune a model initially trained on source data [20]. More practical methods employed unsupervised domain adaptation (UDA) to align the discriminant source model to the target data using unlabeled target data [4], [9], [20]. There models are either based on adversarial, discrepancy, or reconstruction approaches [19]. They mostly apply to classical "single stream" face classification models, and they are not designed specifically for the multiple-stream networks trained through metric learning. These models include deep Siamese networks employed for pair-wise face matching still-to-video FR [2].

Some UDA models for deep distance metric learning were recently proposed [8] [18]. These method are either not applicable to face representation learning since they are designed for closed-set and small-set problems, such as handwritten digit recognition [8]. Moreover, they would require a mixture of techniques for automatic calibration of surveillance cameras, at the expense of higher computational complexity that might hinder practical real-time video surveillance applications [17] [6]. More recently, self-supervised learning approaches have been proposed to automatically label target data by leveraging temporal and contextual information in videos tracklets [16] [22] [1]. These methods require abundance of unlabeled target data,and availability of co-occurring tracklets (i.e., tracklets from different subjects in the same scene that are produced by accurate face detectors and trackers).

This paper addresses the aforementioned limitations of adapting video-based face representation of data from new video capture conditions and environments. UDA is possible given unlabeled target data or where the detection and tracking information are not accurate or informative to produce labeled tracklets. The contributions of this paper are as follows:

- A new domain adaptation framework – called Dual-Triplet Metric Learning (DTML) – is introduced that applies a novel dual-triplet loss function and a mutual supervised learning approach. The proposed framework allows for adapting deep pair-wise matchers to different domains by aligning their distance distributions.
- A mutual-supervised learning approach is proposed, where the source (teacher) iteratively labels the unlabeled target (student) data.
- The proposed dual-triplet loss and mutual-supervised learning approach is applied to the still-to-video FR, and provided a level of accuracy that is comparable to the state-of-the-art methods for video face representation learning, but with a capability for UDA.

It is important to mention that the proposed framework can be applied to different modalities, while in this paper we assess the method using the video-based FR as a specific use case.

II. Related Work

This section provides some background on UDA methods for deep metric learning, video-based representation and self-supervised learning. The relation of our proposed approach with existing methods are also discussed.

A. UDA with Deep Metric Learning:

The distance metric learning approach has been extensively applied to the computer vision area, so that examples belonging to the same label (within-class samples) are close as possible in some embedding space, and samples from different labels (between-class samples) are as far from one another as possible. Recently, triplet and Siamese networks were employed for metric learning which have been successfully applied for few-shot learning [5] [8]. Designing a discriminant and robust distance metric requires abundant of labeled data, and accordingly, UDA methods are required to adapt an existing metric to a new domain where data are unlabeled or scarce (or both cases, as for the applications discussed in this current work).

There are a few works on UDA for deep distance metric learning are recently proposed [8] [18]. In [8], the adversarial learning approach [4] is applied to decrease the domain discrepancy between the datasets and simultaneously a magnet loss is applied to align the class centers for the source and target embedding. This method works only for closed-set problems (where the source and target share the label space). In [18], on the other hand, open-set problems can be tackled by introducing a separation loss so that different source and target sets are separated in the embedding space.

We argue that, for such specific case of UDA (i.e., where the adapted model is a distance metric rather than feature-based model), the embedding should be optimized in the distance space rather than in the feature space. In other words, the straight forward objective function (when designing an UDA algorithm for adapting a distance metric) should help to distinguish between the different "distance" types (i.e., within-class and between class distances), and simultaneously it should be hard to identify the source of a distance sample (i.e., being constituted from samples coming from the source or the target domain). With such objective, the produced distance metrics can be employed for both close- and open-set problems, since it is concerned with the ultimate pairwise distances (rather than with absolute feature representations like that with the existing methods [8] [18]). This new concept is followed to design our proposed method.

B. Video-based Face Representation Learning:

Face representations, based on still images, are usually designed by training deep CNNs, in general, and Siamese CNNs, in particular. Such techniques, however, can provide unreliable performance when applied to design video-based

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

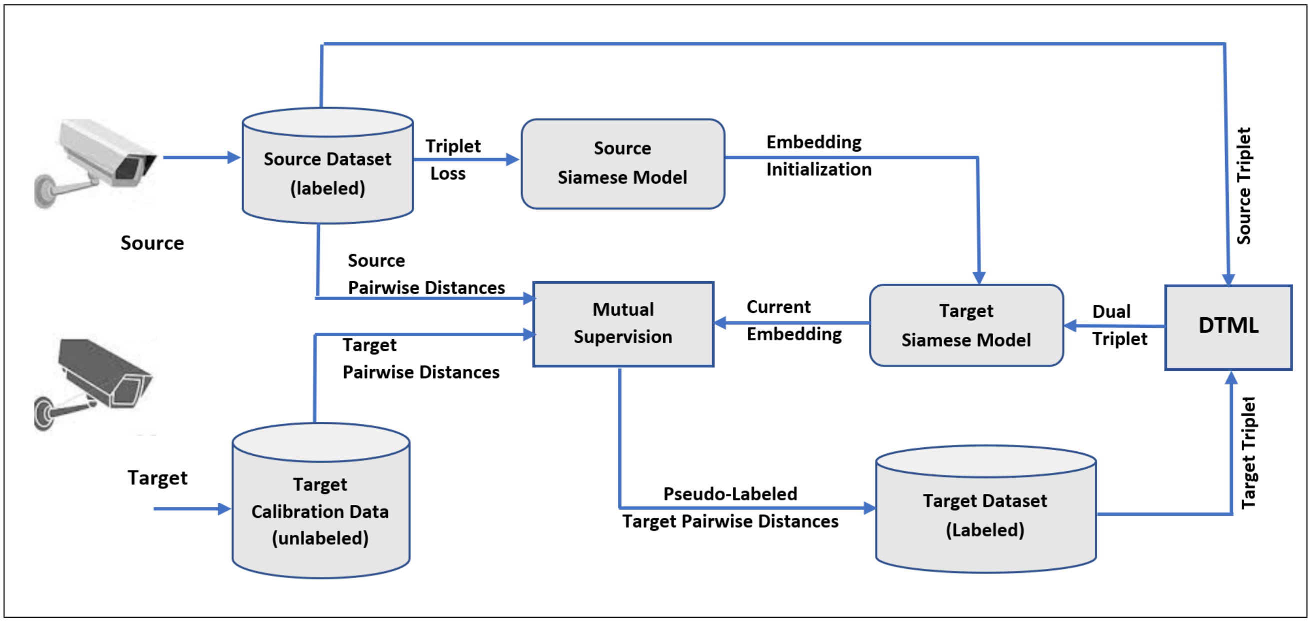


Fig. 1: Dual-Triplet Metric Learning for Unsupervised Domain Adaptation.

face representations. Accordingly, more complex models were proposed to provide improved performance, with the expense of decreased efficiency, and that can hinder the real-time applicability of the designed representations. More importantly, state-of-the-art methods require information that can be unavailable during operation (or they can be expensive to obtain) and that can make these methods impractical, especially to efficiently calibrate existing models for new video sources (i.e., cameras) using a few unlabeled data from new target subjects.

For instance, in [3], reliable detection of facial landmarks is required and that may fail due to occlusion. Also, besides the complex ensemble structure that can hinder the real-time processing, the method involves a fine tuning step that requires large amount of data from the operational target domain. In [12], Haar-like features are extracted so that facial landmark extraction is no longer required, and in [11] a lighter network structure is proposed for improved efficiency. These methods, however, still require synthetic generation of video-like face images and fine tuning using considerable amount of data from the target domain. Some methods like in [20], require labeled data from the target domain for tuning.

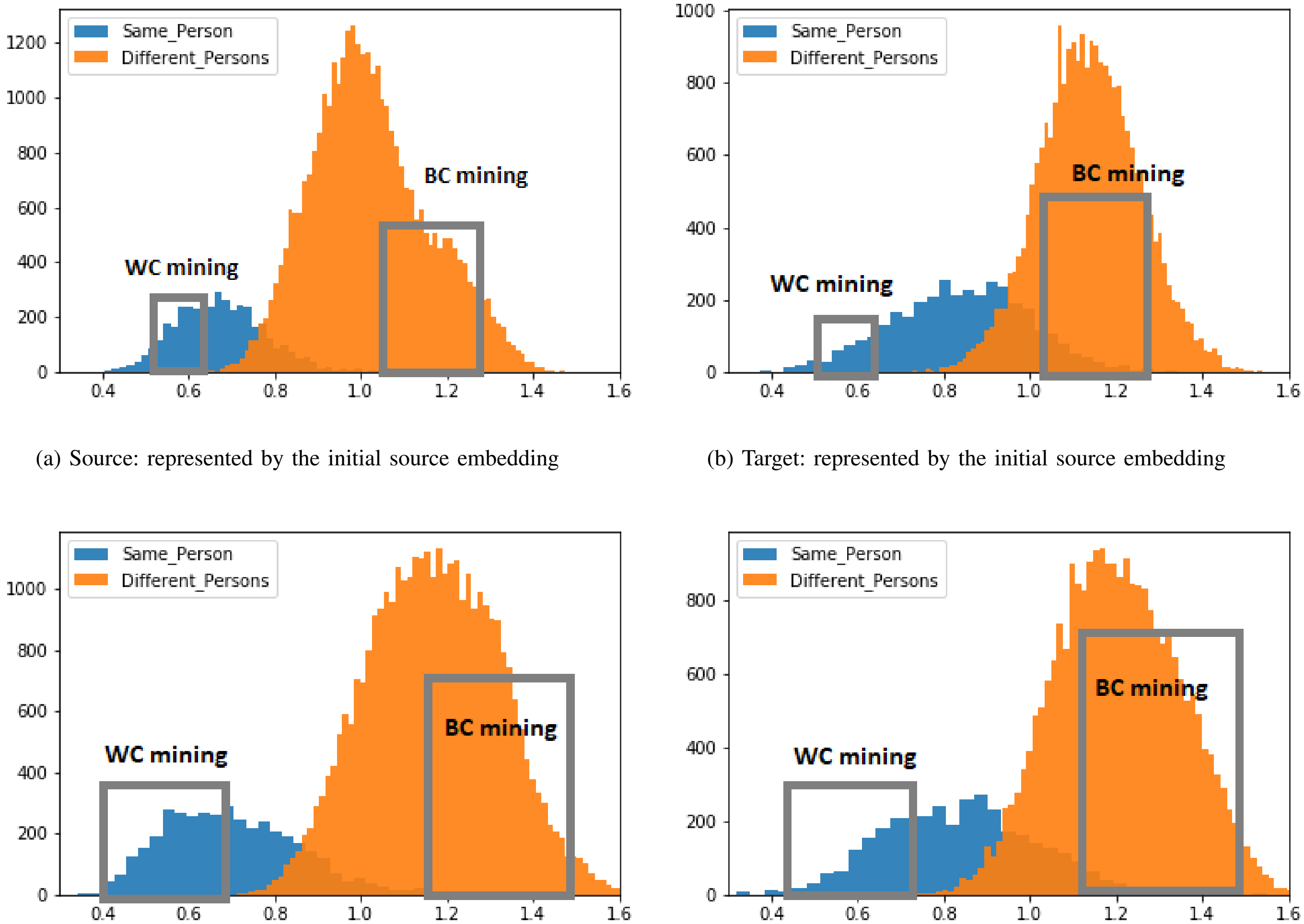
A more recent trend is to employ self-supervised learning to automatically label the target data by levering temporal and contextual information exist in videos, e.g., tracklets [16] [22] [1]. Although provide reasonable performance, these methods mostly require large unlabeled data to train representations from scratch. Importantly, these methods require availability of co-occurring tracklets (tracklets from different subjects in the same scene) so that negative samples can be obtained to constitute triplets. The proposed method works even with singleton tracklets (only tracklets from a single subject appear in a scene), or where the face detector and tracker are not reliable to produce labeled tracklets.

III. Dual-Triplet Metric Learning

Figure 1 illustrates the proposed Dual-Triplet Metric Learning (DTML) framework for Unsupervised Domain Adaptation. This framework facilitates the calibration of a new video source (camera) when added to an existing operational video-based face system (e.g., a video surveillance Network). Moreover, the proposed framework can be employed to adapt an existing model that works for a specific surveillance network to be operational in a different environment or even within a completely new or a different network.

The calibration data consist of some Regions of Interest (ROIs) with faces of unknown people extracted from videos captured by the new camera. The capturing conditions of the environment and the capturing device are represented in the extracted ROIs and used to calibrate a video-based face representation of an existing video source. The Source can be considered as a teacher as it provides the target (student) with the initial knowledge (embedding) it acquires through supervised learning, and also they (the teacher and the student) continue to learn a shared knowledge (embedding) using their different data (labeled data of the source and unlabeled calibration data of the target).

To this end, the source labeled data are used to learn an initial representation with employing the ordinary triplet-loss approach [15]. For each new (target) camera or environment, the initial source embedding is loaded to the target and gets improved with minimizing a dual-triplet loss. In order to constitute the target part of the dual-triplet loss, a "mutual-supervised" process is employed, where pairwise distances between target calibration samples are computed and statistics of the pairwise source distances are used to label the target



(a) Source: represented by the initial source embedding

(b) Target: represented by the initial source embedding

(c) Source: represented by a shared embedding through DTML

(d) Target: represented by a shared embedding through DTML

Fig. 2: Illustration of the Mutual Supervision learning using the distributions of distances between samples from same-person (within-class (WC) samples) and distances between samples from different-persons (between-class (BC) samples). Left column (Fig.a,c) show the distance distributions for the source data (teacher) and right column (Fig.b,d) show the distributions for the target data (student). Upper row (Fig.a,b) show the distance distributions for the initial source representation, while the bottom row (Fig.c,d) show the distributions where a shared target representation is learned using the proposed dual-triplet and mutual-supervision learning method.

distances (as being within-class or between-class distances). During training, the source and target pairwise distance distributions become more similar over time and that implies the following: 1) a distance metric that works for the source also works for the target, and 2) the resulting metric can label pairwise distances from the target as accurate as for the source distances.

A. Dual-Triplet Loss:

The dual-triplet loss L consists of two terms: a source term L_s and a target term L_t:

$$L = L_s + \lambda \cdot L_t. \tag{1}$$

where λ is the parameter that balancing the two objectives.

A source triplet L_s is constituted from labeled source data, using an anchor a, a positive sample p, a negative sample n, and a margin α:

$$L_s = \max(||f(a) - f(p)|| - ||f(a) - f(n)|| + \alpha, 0) \tag{2}$$

To constitute a target triplet, it can be impossible to use the absolute representation of anchor, positive and negative samples, since the target data can be unlabeled. To resolve this limitation, we are only interested in labeling the pairwise distances as being either within-class (wc) or between-class

(bc). Once the distances are labeled, the target triplet is constituted as follows:

$$L_t = \max(||wc|| - ||bc|| + \alpha, 0) \tag{3}$$

The source triplet aims at designing a discriminant distance metric since it is based on labeled data from the source domain. The target triplet aims at separating the within-class and between-class target distances with the same margin as that for the source distances, so that the pairwise distance distributions of both domains are similar (and hence, the resulting distance metric is valid for both domains).

It is important to mention that the proposed DTML method can be applied when a few labeled samples are available from the target domain (e.g., where co-occurring tracklets are available or when some calibration data are manually annotated). In that case, the target triplets can be constituted directly from the labeled samples as that with the source triplets. In case no labeled data are available from the target domain or when only singleton tarcklets data are available (i.e., frames belong to a single person appear in the scene, so no dissimilar pairwise labels are provided by tracklets), the following mutual-supervised method is required.

B. Mutual Supervision:

The objective of the mutual-supervised learning method is to leverage the labeled samples of the source to extract positive (within-class) and negative (between-class) pairwise distance samples from the unlabeled target samples.

Figure 2 illustrates the proposes mutual-supervised learning method. Firstly, both source and target training samples are represented using the initial embedding trained using the labeled source data in supervised mode. Then, pairwise distances from both datasets are generated by computing the Euclidean distance between the feature vectors of each pair. Since the source dataset is labeled, it is straightforward to label the source pairwise distances as within-class (WC) or between-class (BC) if they belong to same-person or different persons, respectively. Distributions of the WC and BC pairwise distance samples are generated (see Figure 2.a). Statistics of these distributions are used to identify two mining windows: 1) within-class mining window (WC_{mw}) and 2) between-class mining window (BC_{mw}):

$$WC_{mw} = [\mu_{wc} - \sigma_{wc}, \mu_{wc}]. \tag{4}$$

$$BC_{mw} = [\mu_{bc}, \mu_{bc} + \sigma_{bc}]. \tag{5}$$

where μ_{wc}, σ_{wc} and μ_{bc}, σ_{bc} are the mean and standard deviation of the WC and BC pairwise distances, respectively.

These mining windows are computed to achieve a trade-off between confidence of labeling (picking distance samples that are close enough to the center of the distributions and far from the confusion areas where WC and BC distances can overlap) and also to avoid picking too much easy samples (samples that exist towards the tail of the distributions as these samples most likely lie beyond the margin so they do not contribute to the loss function).

Once the mining windows are computed based on the source pairwise distance distributions, they are used to locate (label) the target distance samples (see Figure 2.b). Initially, the source and target distributions are not aligned (as a result of the domain shift), so using the source mining windows maybe not accurate enough and also may locate a small number of samples. These samples are used to constitute the target triplet loss term, then dual triplet is used to train a new target representation.

The above process is repeated for each training batch and eventually the source and target pairwise distance distributions become more aligned as WC and BC samples from both domains are enforced to be separated by the same margin, and a shared representation is used to represent samples from both domains.

Figure 2 (c and d) illustrate how the source and target distributions are getting aligned through the DTML with mutual-supervised training. When better aligned, the source mining Windows produce more accurate pseudo-labels of the target pairwise distance and also locate larger number of samples from each bucket (the WC and BC buckets). Also, it is important to note that the inaccuracies of the target loss term (as a result of the imperfect pseudo-labeling) can be compensated by the existence of the perfect source loss term.

IV. Experimental Methodology

Although the proposed DTML framework and mutual-supervised learning approach may be applied for different modalities, we assess the methods here for the specific still-to-video (S2V) face recognition (FR) application. To this end, face ROIs are captured by video cameras and matched against high quality frontal still face images of some users enrolled to the system.

Two video-based face FR datasets are used for the experimentation: 1) the public COX face dataset and 2) a private video-based dataset that we created internally for performance assessment.

The COX dataset is used to assess the proposed approach ability to adapt a model that works for an existing camera in a surveillance network to be operational for a new camera added to the network. On the other hand, the private dataset is utilized to simulate the case where a model designed for am exiting surveillance network is leveraged and adapted for a different network or operational environment.

For the COX dataset, the standard experimental protocol described in [7] is followed in this experimental study, so results can be compared to the state-of-the-art methods. The dataset consists of 1000 subjects with still images are captured for each subject and then each person is captured by 3 video cameras with different views. As described in [7], samples from 300 subjects are considered a training set, and the remaining 700 subjects are used for the testing. To simulate the camera calibration scenario, we split the training set to 200 subjects for training the initial source models, and 100 subjects for calibrate the new camera. This split is important to simulate the case where subjects used during developing the initial solution (i.e., in the lab) are different than the subjects who appear in the operational field during a calibration session. This setup also simulates the "open-set" scenario, where subjects appear during operation are not seen during the design and calibration phases.

The private video-based FR dataset consists of 100 subjects where only video-based face images are captured by a commercial IP video surveillance camera during real operational setup. Since there is no still templates are collected during operation, we manually selected best face image from each subject (i.e., nearly frontal, best size and quality, etc) and used them as a still images gallery. The training set consists of 30 subjects and the testing set consists of the remaining 70 subjects. Since we simulate the case where a model from an existing network is leveraged and adapted for a new network or environment, so here we use a model tuned for Cam 1 from the COX dataset as a source, and employ the proposed approach to adapt this model to the camera that we used to create our private dataset. Accordingly, the whole training set (30 subjects) are used for calibration.

For all experiments, samples are firstly represented using the VGG Face representation [14]. Then, a source embedding is trained using the ordinary triplet loss. This embedding is used to test the case where only source models are leveraged without any domain adaptation step. To this end, the S2V performance for the subjects of the testing set is computed and considered as a lower-bound performance.

To simulate the case where labeled data are available from the calibrated camera (e.g., through expensive manual annotation), the DTML is employed but the target labels are taken directly from the dataset (instead of the labels produced by the mutual-supervised method) and consisedred as an upper-bound performance.

To simulate the camera calibration process using our proposed UDA approach, the calibration and training data (from source and target cameras, respectively) are represented by the learned source representation and used to constitute the dual-triplet terms. Then, the DTML algorithm runs to learn an embedding for the calibrated camera. For such case, three variants of the DTML method are employed to test the impact of the dual-triplet loss terms, three scenarios are implemented:

1) L_s: where only data from the source camera are used and only the the source triplet is used to tune the network. This scenario simulates the case where we don't use calibration data and only keep improving the source representation.
2) L_t: where only data from the target camera are used and only the the target triplet is used to tune the network. This scenario simulates the case where we only rely on calibration data to adapt the source model for the new camera.
3) $L_s + L_t$: where data from both the source and calibrated cameras are used for adaptation, which is the exact DTML proposed method.

	Cam1 $\rightarrow$ Cam3		Cam3 $\rightarrow$ Cam2	
Methods	AUC	Acc	AUC	Acc
VGG-Face [14]	-	68.1	-	76.0
Source model: without DA	0.90	81.7	0.94	87.2
Proposed UDA: DTML-A	0.95	88.7	0.97	91.0
Upper-bound: supervised DA	0.98	93.3	0.99	95.3
TBE-CNN [3]	-	88.2	-	95.7
CCM-CNN [11]	-	88.6	-	92.1
HaarNet [12]	-	89.3	-	97.0

TABLE I: AUC and accuracy of the proposed DTML with mutual supervision and baseline methods on the COX dataset.

For all experiments, the batch size is set to 100 (5 persons per batch, 20 images per person). The source labels are used to load training images from five different persons per batch, while the proposed mutual-supervision method is employed to pseudo-label pair-wise image distances so that equivalent number of similar and dissimilar images are loaded from the target dataset. Then, DTML runs for 40 epochs, and performance is tested using rank 1 accuracy and the Area Under ROC Curves (AUC). Parameter λ (see Eq. 1) is used to balance the contributions of the source and target triplet terms to the overall loss function. Extensive experimentation show that equal contributions lead to best results, so we set $\lambda = 1$ in all reported results.

V. Experimental Results and Discussion

Table I shows the S2V FR results using the COX dataset. It is obvious that tuning the basic feature representation (VGG-Face) using the source data only provides improved accuracy, even without applying any domain adaptation step. This observation is expected, as cameras within same network have some similarities, and accordingly training for one camera can help for the other cameras.

Applying the proposed DTML algorithm with mutual-supervision achieved a significant improvement for both domain shift problems (around 5% increase in AUC for both cases). While the source model performance is very low compared to the upper-bound performance (in case that labels of the calibration data are available, e.g., through expensive manually annotation, perfect co-occurring tracklets, etc.), the proposed method could perform a reliable domain adaptation in an unsupervised fashion (i.e., UDA) and achieved a performance close to the upper-bound.

Comparing with the performance of state-of-the-art methods, the proposed method achieves comparable results (for the Cam3 $\rightarrow$ Cam2 case), and somewhat less performing (for the Cam1 $\rightarrow$ Cam3 case). Noting that these methods are either less efficient (e.g., because they involve complex ensembles of CNNs, expensive generation of synthetic video-like ROIs, etc.), or they rely on information that can be unavailable during camera calibration (e.g., facial marks that can be unavailable due to occlusion, abundant of calibration data, etc.). On the other hand, the proposed DTML with mutual-supervision approach employs simple CNN structures,

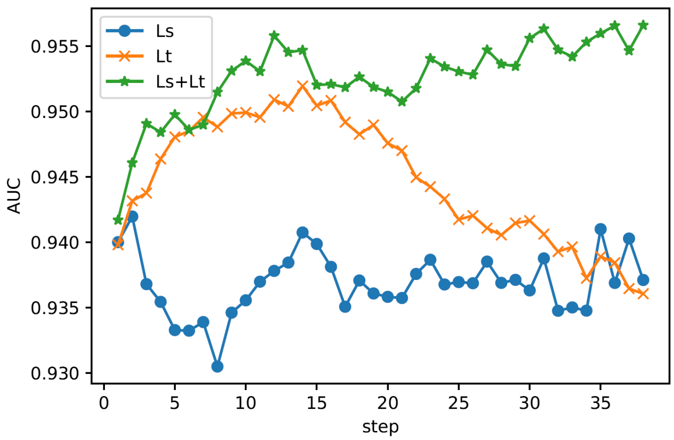


Fig. 3: AUC performance on COX dataset: Cam3 $\rightarrow$ Cam2.

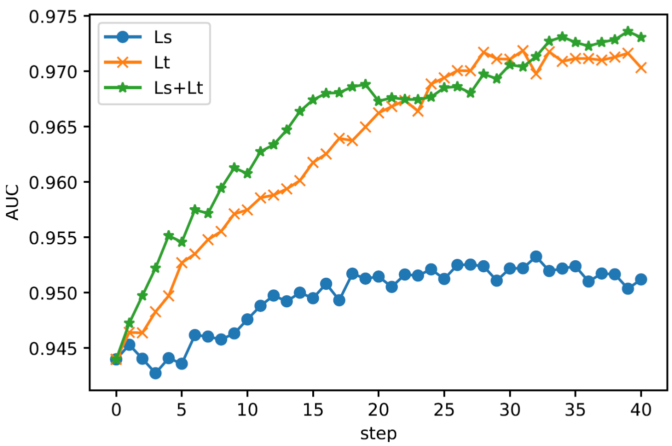


Fig. 4: AUC performance on COX dataset: Cam1 $\rightarrow$ Cam3.

and only uses a small amount of unlabeled calibration data without the need to generate synthetic captures or extract facial landmarks.

Figures 3 and 4 illustrates the impact of the dual triplet terms. In Figure 3, it is clear that using the source data only (i.e., using loss triplet term L_s only) is not helpful to classify samples from the target camera. Also, using the target loss (L_t) alone, although gained performance increase during the initial training iteration, the performance is dramatically decreased afterwards due to model overfitting. Having both the source and target triplets ($L_s + L_t$) guarantee continuous performance increase over the training period.

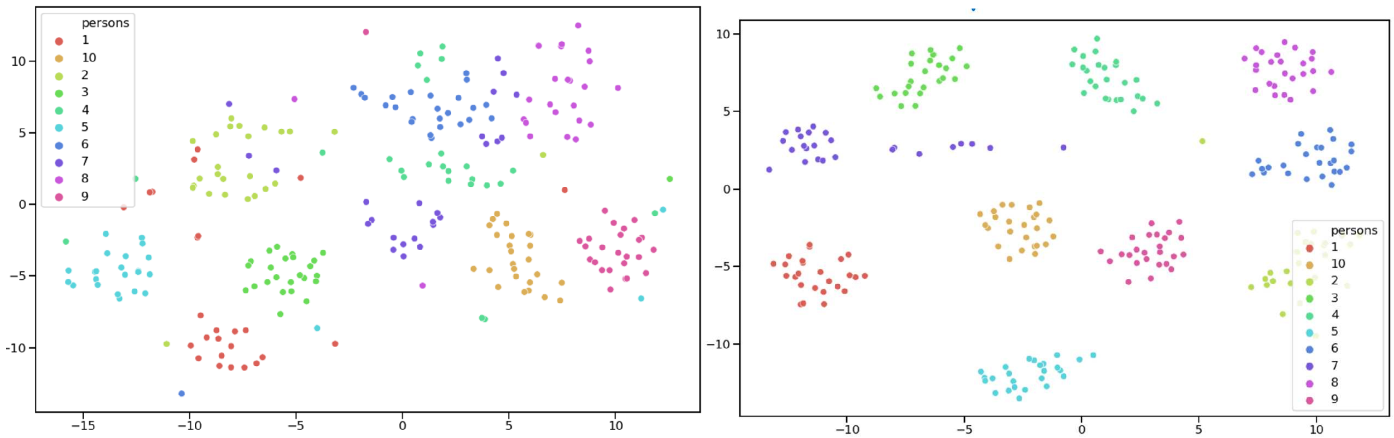
In Fig 4, it is clear that source data alone provides a limited performance gain. Similar to the above observation, the best performance is achieved when triplets from both domains are optimized. Although using triplets from only the target domain provided adequate performance for this domain shift problem, this result wouldn't be achieved without having the help of the source as the target data are initially unlabeled and they are labeled using the source embedding (in order to constitute the target triplets).

Fig 5 illustrates the impact of applying the proposed UDA method to adapting a model designed for an existing camera to a new camera so that surveillance networks are extensible. Fig 5.a shows the TSNE representation of the testing samples from the first ten subjects captured by Cam 2 (target camera, newly added to the network) of the COX dataset, where the embedding is generated using a model tuned for Cam 3 (source camera, already exist in the network). It is clear that the source embedding is not suitable enough for the target camera. Fig 5.b shows the the TSNE representation of the target camera after adapting the model using the proposed UDA approach. It is clear that the adapted model provides discriminant representation.

Fig 6 illustrates the impact of applying the proposed method to leveraging a model of an existing network to be functional for a new network or environment (e.g., where collecting and labeling enough data from the new network to train a model from scratch is infeasible). Fig 6.a shows the TSNE representation of the testing samples from the first ten subjects captured by the target camera (a commercial IP camera used to capture video-based face image in a realistic operation conditions in an uncontrolled environment). The embedding is generated using a model that is fine tuned for Cam 1 from the COX dataset. Although we have chosen the source camera from the COX dataset (that is closest to the target camera as it has the least domain shift between source and target), it is obvious most clusters (subjects representation) are split into two separate sets and there is a significant overlap between subjects (see the top part of Fig 6.a). When we applied the proposed UDA approach, using unlabeled samples from the target camera, the clusters are well separated (see 6.b) and accordingly the FR accuracy has significantly improved (True positive rate at False Alarm rate (FAR=1%) increased from 52% from 73%).

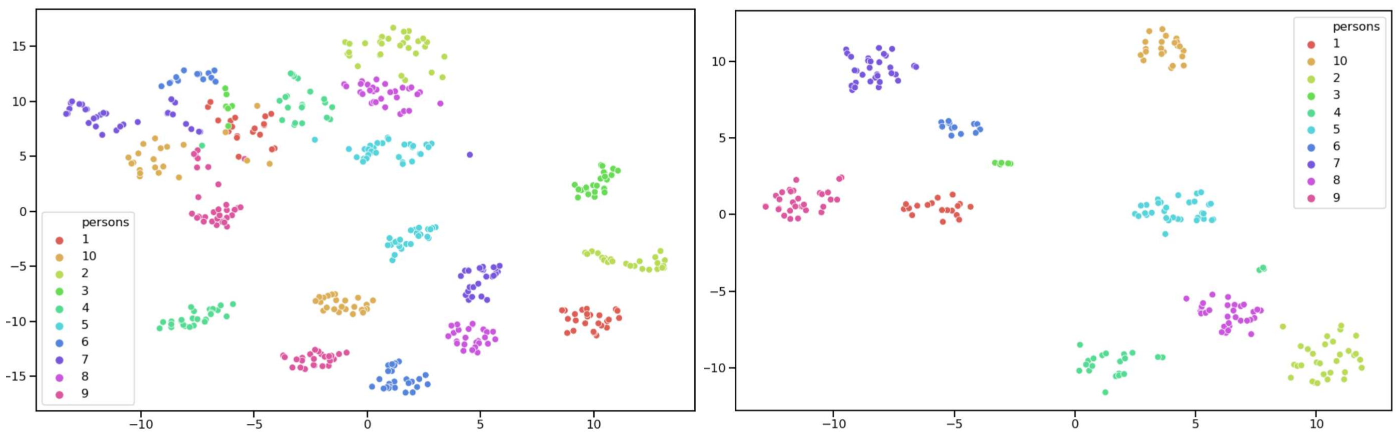
Figure 2 (that is used to illustrate the proposed mutual-supervised learning method in in Section III.B; where distance distribution plots are generated using Cam 1 from the COX dataset as a source and our private video-face dataset as a target) shows the separability of WC and BC distance distributions of the target data where the source initial embedding is adapted using the proposed UDA approach. With UDA, the source and target distributions are well aligned. More specifically, after adaptation, a simple threshold, e.g., 0.9 provides a good trade-off between false positives and false negatives for both of the source and target domains, as apposed to to initial state (Figures 2 (a and b) where the source and target distance distributions are not aligned and a threshold that works for the source (e.g., 0.8) results in high FAR rates when used by the target.

It is important to mention that, although the adapted embedding provides more aligned and separable pairwise distance distributions (i.e., target within-class (WC) and between-class (BC) distributions are better separated and aligned with the source distributions), it is not expected to rely only on the quality of resulting embedding to provide accurate classification results when the simple Euclidean distance is used as a distance metric, and also when a simple threshold is used



(a) Before domain adaptation (b) After domain adaptation.

Fig. 5: The t-SNE representation for first ten persons in the testing set of our COX dataset,



(a) Before domain adaptation (b) After domain adaptation

Fig. 6: The t-SNE representation for first ten persons in the testing set of our Private Video Face dataset.

for classification in the embedding space. Accordingly, we further feed the embedding of the still (template) and video (query) samples to a two-layer fully connected network and we trained this outer network separately using the pseudo-labeling approach described in Section III.B. This step has improved the recognition accuracy from 73% to 84% with $FAR = 1\%$ (for the private FR video-based dataset). More importantly, instead of feeding the two streams of the absolute embedding (from both the still and video samples), we further generate the following dissimilarity feature representation:

$$\delta(X) = |X^Q - X^T|. \tag{6}$$

where X^Q and X^T are the feature representation generated by embedding adapted with the proposed DTML approach for the query (video) and template (still) samples, respectively. The resulting dissimilarity representation is accordingly of the same dimensionality as that for the original representation. We noticed a significant improvement in accuracy due this transformation step (recognition accuracy has increased from 84% to 90% with $FAR = 1\%$). Future work will explore employing the DTML approach with having the dissimalrity representation and the outer layers trained in an end-to-end fashion.

Also, it is important to note that the proposed method is only applicable to transfer problems where the domain shift between the source and target domain is small enough so that mutual-supervised training is possible. For instance, if the initial source and target distance distribution are not aligned enough, the proposed mining approach can locate completely wrong samples, or may fail to locate any sample from either the WC or the BC buckets, so in this case the mutual-supervised learning mechanism will not work correctly (see Figure 2). Future work will explore different methods to enforce the distance distributions of the different domains to be aligned without the need for the pseudo-labeling step, by employing the adversarial learning concept.

VI. Conclusion

A general method for UDA of deep distance metrics is proposed in this paper. The proposed method is applicable to pair-wise matching problems with domain shift, where the target domain can provide a small amount of data, and also the cases where only unlabeled target data are available. Two main aspects of the proposed method are discussed: the new dual-triplet optimization and mutual-supervision process. Using a dual-triplet form the source and target domains allows to mitigate the issue model corruption because, even with limited target data, it can also allow for pairwise distance distributions of both domains more similar. The mutual-supervision feature provides a novel tool to constitute triplets from the unlabelled target samples. The method is applied to provide unsupervised domain adaptation for the still-to-video FR systems and achieved a level of accuracy comparable to the state-of-the-art more complex methods, that also can be impractical given the limitations of the camera calibration use-case. Future directions for our research include validating this approach on a wider range of applications and datasets, and exploring the relationship with methods in knowledge distillation.

Acknowledgement

This research was partially supported by the Mathematics of Information Technology and Complex Systems (MITACS) and the Natural Sciences and Engineering Research Council of Canada (NSERC) organizations.

References

[1] R. G. Cinbis, J. Verbeek, and C. Schmid. Unsupervised metric learning for face identification in tv video. In *ICCV 2011*.

[2] M. A. A. Dewan, E. Granger, G.-L. Marcialis, R. Sabourin, and F. Roli. Adaptive appearance model tracking for still-to-video face recognition. *Pattern Recognition*, 49:129–151, 2016.

[3] C. Ding and D. Tao. Trunk-branch ensemble convolutional neural networks for video-based face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):1002–1014, 2018.

[4] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv*, stat.ML(1409.7495v2), Sept. 2014.

[5] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In A. Feragen, M. Pelillo, and M. Loog, editors, *Similarity-Based Pattern Recognition*, pages 84–92, Cham, 2015. Springer International Publishing.

[6] S. Hong, W. Im, J. Ryu, and H. S. Yang. SSPP-DAN- deep domain adaptation network for face recognition with single sample per person. *arXiv*, cs.CV(1702.04069v4), Feb. 2017.

[7] Z. Huang, S. Shan, R. Wang, S. Lao, A. Kuerban, and X. Chen. A benchmark and comparative study of video-based face recognition on cox face database. *IEEE Transactions on Image Processing*, 14(12):5967–5981, 2015.

[8] I. H. Laradji and R. Babanezhad. M-adda: Unsupervised domain adaptation with deep metric learning. *CoRR*, abs/1807.02552, 2018.

[9] Z. Luo, J. Hu, W. Deng, and H. Shen. Deep unsupervised domain adaptation for face recognition. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 453–457, May 2018.

[10] F. Mokhayeri, E. Granger, and G. Bilodeaun. Domain-specific face synthesis for video face recognition from a single sample per person. *IEEE Transactions on Information Forensics and Security*, 14(3), 2019.

[11] M. Parchami, S. Bashbaghi, and E. Granger. Cnns with cross-correlation matching for face recognition in video surveillance using a single training sample per person. In *Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, 2017.

[12] M. Parchami, S. Bashbaghi, and E. Granger. Video-based face recognition using ensemble of haar-like deep convolutional neural networks. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, pages 4625–4632, 2017.

[13] M. Parchami, S. Bashbaghi, E. Granger, and S. Sayed. Using deep autoencoders to learn robust domain-invariant representations for still-to-video face recognition. In *Proceedings of the 14th IEEE International Conference on dvanced Video and Signal Based Surveillance (AVSS)*, 2017.

[14] Q. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising face across pose and age. In *2018 International Conference on Automatic Face and Gesture Recognition*, 2018.

[15] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 815–823, 2015.

[16] V. Sharma, M. Tapaswi, M. S. Sarfraz, and R. Stiefelhagen. Self-supervised learning of face representations for video face clustering. *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*, pages 1–8, 2019.

[17] K. Sohn, S. Liu, G. Zhong, X. Yu, M.-H. Yang, and M. Chandraker. Unsupervised domain adaptation for face recognition in unlabeled videos. *arXiv*, cs.CV, cs.AI(1708.02191v1), Aug. 2017.

[18] K. Sohn, W. Shang, X. Yu, and M. Chandraker. Unsupervised domain adaptation for distance metric learning. In *International Conference on Learning Representations*, 2019.

[19] M. Wang and W. Deng. Deep visual domain adaptation: A survey. *CoRR*, abs/1802.03601, 2018.

[20] G. Wen, H. Chen, D. Cai, and X. He. Improving face recognition with domain adaptation. *Neurocomputing*, 287:45 – 51, 2018.

[21] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534, June 2011.

[22] B. Wu, S. Lyu, B. Hu, and Q. Ji. Simultaneous clustering and tracklet linking for multi-face tracking in videos. In *2013 IEEE International Conference on Computer Vision*, pages 2856–2863, Dec 2013.