Domain Adaptation in Video Recognition

Eric Granger
Dept. of Systems Engineering
ETS Montreal

January 2024

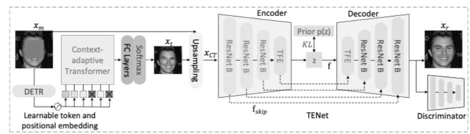
LIVIA LABORATOIRE D'IMAGERIE, DE VISION ET D'INTELLIGENCE ARTIFICIELLE · ILLS International Laboratory on Learning Systems · ÉTS Le génie pour l'industrie

1

1

Research Interests

- machine learning – domain adaptation, incremental, and weakly-supervised, and multimodal learning
- computer vision
- pattern recognition in static and dynamically changing environments
- information fusion



2

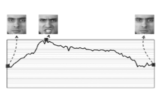
2

2

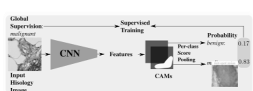
Application Areas

Video Analytics and Surveillance:
- real-time object detection, tracking, re-identification and fusion

Affective Computing in Healthcare:
- spatiotemporal expression recognition
- multimodal fusion



Analysis of Medical Images
- breast cancer grading and localization in histology



3

3

3

Overview

1) Person Re-Identification
- domain adaptation in dissimilarity space
- visual-infrared ReID with intermediate domains

2) Facial Expression Recognition
- subject-based MSDA
- source-free adaptation

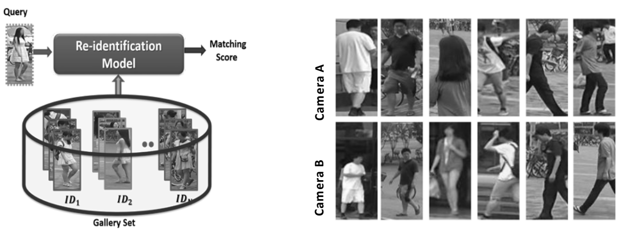
4

4

4

Video Analytics & Surveillance – Re-Identification

Task: Match individuals or objects captured over a distributed set of non-overlapping camera viewpoints



Challenges: low resolution, motion blur, occlusions, variation in pose and illumination, misalignment over different camera views

5

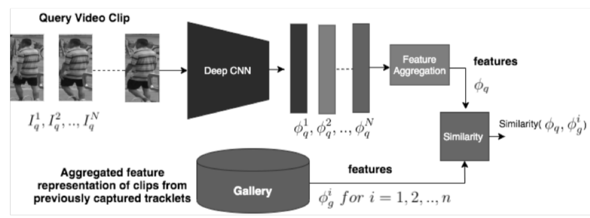
Source: T. Wang et al., Person Re-Identification by Video Ranking, ECCV2014. 5

5

Video Analytics & Surveillance – Re-Identification

DL models for video-based similarity matching:

- **Train:** metric learning of the embedding network for pairwise similarity
- **Test:** given a clip of probe and gallery images, predict their similarity



6

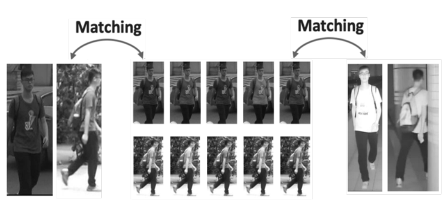
Source: M Kiran et al., 2021. Holistic guidance for occluded person re-identification. BMVC 2021. 6

6

Video Analytics & Surveillance – Cross-Modal ReID

Visible-Infrared ReID

- match persons/objects across RGB and IR cameras
- **challenge**: the large shift between RGB and IR data distributions



7
7

7

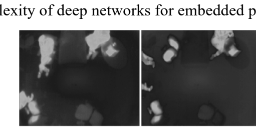
Chaire de recherche industrielle Distech Controls sur les réseaux de neurones embarqués pour le contrôle de bâtiments connectés

Objectives:

- Control of **intelligent building occupancy analysis** using low-cost distributed sensors and AI
- Reducing energy footprint and increasing comfort in buildings
- **Applications for using low - resolution sensors**

Challenges:

- Integration of information from various low-resolution sensors: IR, RGB, etc.
- Adapting systems to changing environmental conditions
- Reducing the complexity of deep networks for embedded platforms



8

8

Common Challenges

Improving performance:

- domain shifts and fusion across different cameras and modalities
- variations for different people, objects, and capture conditions (pose, occlusion, illumination, scale, motion blur, etc.)
- robustness of models trained on image data using limited and ambiguous annotations

Reducing complexity:

- state-of-art deep learning (DL) models are complex and can grow with the number of cameras and modalities
- cost of collecting and annotating large-scale datasets

9

9

9

Domain Adaptation Methods

Objective: learn robust domain-invariant representations from source domain (SD) and target domain (TD) samples

Common approaches:

1. **discrepancy-based:** fine-tune model with source and target data to diminish shift between domain distributions
   - e.g., use a *statistical criterion* (MMD, CORAL, KL divergence, etc.) to align the SD and TD distributions
2. **adversarial-based:** rely on domain discriminator to predict if samples are drawn from SD or TD, and encourage domain confusion
   - e.g., *non-generative models* map SD to TD representation space using a discriminator and domain confusion loss

10

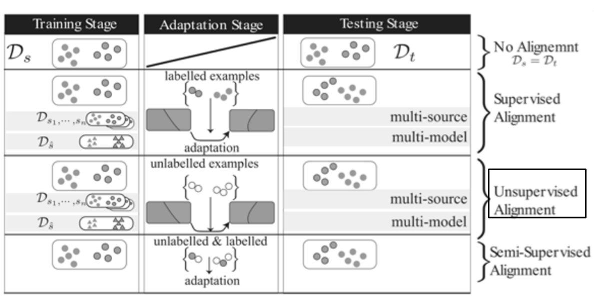
Source: A. Khamis, *et al.*, "Earth Movers in The Big Data Era: A Review of Optimal Transport in Machine Learning," *ArXiv:2305.05080*, May 2023.

10

10

Domain Adaptation Methods

Common Settings:



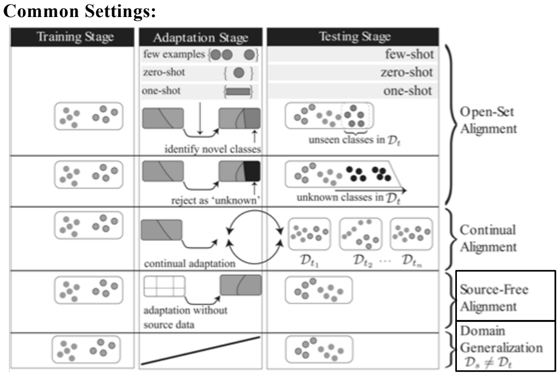
Source: A. Khamis, *et al.*, "Earth Movers in The Big Data Era: A Review of Optimal Transport in Machine Learning," *ArXiv:2305.05080*, May 2023.

11

11

Domain Adaptation Methods

Common Settings:



Source: A. Khamis, *et al.*, "Earth Movers in The Big Data Era: A Review of Optimal Transport in Machine Learning," *ArXiv:2305.05080*, May 2023.

12

12

Unsupervised Domain Adaptation

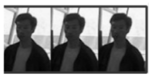
Unsupervised DA Setting: adapt a ML model using labeled source and unlabeled target samples to improve performance in the target domain
– SD and TD learning tasks are the same, but data distributions differ

Examples: video-based face recognition

1) **SD:** still ROI camera 0

TD: video ROIs camera 3

2) **SD:** video ROIs camera 1



TD: video ROIs camera 3

SD: source domain
TD: target domain

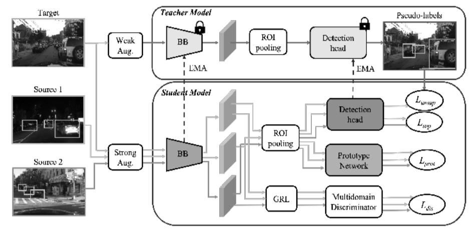
13
13

13

Multi-Source Domain Adaptation

Objective: Adapt ML model using 2+ source datasets to improve target domain accuracy and robustness

Example: Prototype-based Mean Teacher (PMT) object detection model for MSDA



Source: Belal, et al., Multi-Source Domain Adaptation for Object Detection with Prototype-based Mean-teacher, WACV 2024

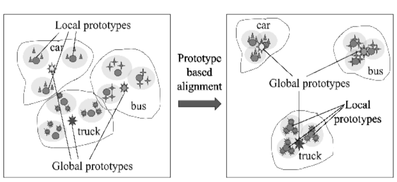
14
14

14

Multi-source Domain Adaptation

Objective: Adapt ML model using 2+ source datasets for improved accuracy and robustness

Example: Prototype-based Mean Teacher (PMT) object detection model for MSDA



- Prototype-based feature alignment with 3 source domains (and 3 classes)
- After alignment, class confusion and intra-class distance to global prototypes are reduced.

Source: Belal, et al., Multi-Source Domain Adaptation for Object Detection with Prototype-based Mean-teacher, WACV 2024

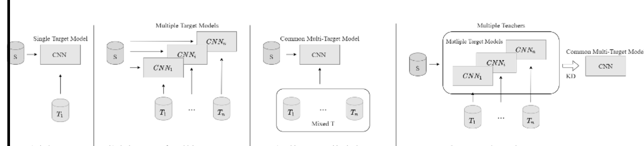
15
15

15

Multi-Target Domain Adaptation

Single-Target DA (STDA): adapt a model to a single target domain

Multi-Target DA (MTDA): adapt a common (compact) model to perform well in 2+ target domains



Source:

- Nguyen-Meidine, et al., Incremental multi-target domain adaptation for object detection with efficient domain transfer, *Pattern Recognition*, 2022.
- Remigereau, et al., Knowledge Distillation for Multi-Target Domain Adaptation in Real-Time Person Re-Identification, *ICIP* 2022.

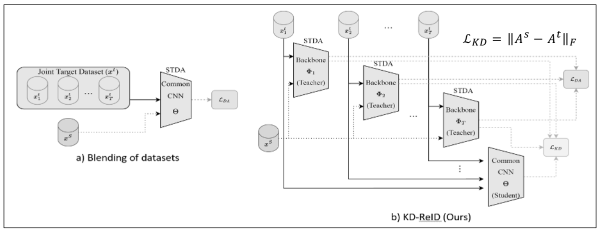
16
16

16

Multi-Target Domain Adaptation

Objective: Adapt a common (compact) ML model that performs well in 2+ target domains

Example: KD-ReID distils knowledge from specialized teachers, one per target domain, into a single smaller student backbone



Remigereau, et al., Knowledge Distillation for MTDA in Real-Time Person ReID, ICIP 2022.
Nguyen-Meidine, et al., Unsupervised MTDA Through Knowledge Distillation, WACV 2021.

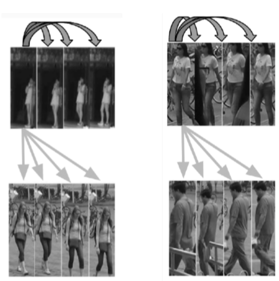
Unsupervised Domain Adaptation in the Dissimilarity Space for Person Re-Identification

Djebril Mekhazni, Amran Bhuiyan, George Ekladious & Eric Granger

ECCV 2020: European Conf. on Computer Vision

UDA in the Dissimilarity Space

- **Assumptions:** target data is unlabeled, but we can leverage knowledge of tracklets from cameras



within class (wc): with the same person

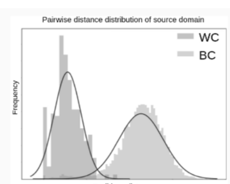
between class (bc): with different persons

Source tracklets

Target tracklets

UDA in the Dissimilarity Space

- **We can therefore extract dissimilarity distributions:**



Distributions in the dissimilarity space:

$d_i^{wc}(\mathbf{x}_i^u, \mathbf{x}_i^v) = ||\phi(\mathbf{x}_i^u) - \phi(\mathbf{x}_i^v)||_2, \; u \neq v$

$d_{i,j}^{bc}(\mathbf{x}_i^u, \mathbf{x}_j^z) = ||\phi(\mathbf{x}_i^u) - \phi(\mathbf{x}_j^z)||_2, \; i \neq j \; \& \; u \neq z$

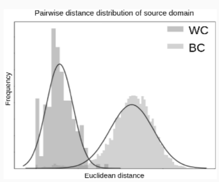
$\mathbf{x}_i^u$ — u^{th} sample $\mathbf{x}$ of identity i

$\phi(\mathbf{x})$ — Features of the sample $\mathbf{x}$

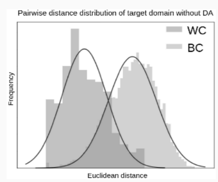
D Mekhazni, A Bhuiyan, G Ekladious & E Granger, Unsupervised Domain Adaptation in the Dissimilarity Space for Person Re-Identification, ECCV 2020.

UDA in the Dissimilarity Space

• **Dissimilarity distributions:** a typical case



Well separated **Source** Distributions



Big overlap in **Target** Distributions

Because of the overlap, class behavior is hard to estimate on **Target** data.

21

21

UDA in the Dissimilarity Space

Proposed discrepancy-based approach:

- Maximum Mean Discrepancy (MMD) in the dissimilarity space to align pairwise distances between source and target domain

MMD: distance between two distributions A and B.

$$\mathrm{MMD}(A,B) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(a_i,a_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(b_i,b_j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(a_i,b_j)$$

a_i: i^{th} sample of A
n: # of samples in A

b_i: i^{th} sample of B
m: # of samples in B

k: gaussian kernel

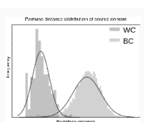
Objective : Minimize MMD.

22

UDA in the Dissimilarity Space

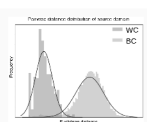
Apply MMD loss in the dissimilarity space, not feature space

- Align pairwise distances between SD and TD
- **d** : distances distribution

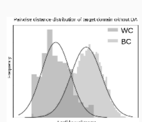


Source Distributions

$$\mathcal{L}^{\mathrm{wc}}_{MMD} = MMD(\mathbf{d}^{\mathrm{wc}}_s, \mathbf{d}^{\mathrm{wc}}_t)$$

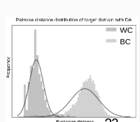


Source Distributions



Target Distributions

$$\mathcal{L}^{\mathrm{bc}}_{MMD} = MMD(\mathbf{d}^{\mathrm{bc}}_s, \mathbf{d}^{\mathrm{bc}}_t)$$



23

Target Distributions

UDA in the Dissimilarity Space

- **Overall Dissimilarity-MMD Loss:**

$$\mathcal{L} = \underbrace{\mathcal{L}_{\mathrm{ces}} + \lambda \cdot \mathcal{L}_{\mathrm{tri}}}_{\mathcal{L}_{\mathrm{Supervised}}} + \underbrace{\mathcal{L}_{MMD} + \mathcal{L}^{\mathrm{wc}}_{MMD} + \mathcal{L}^{\mathrm{bc}}_{MMD}}_{\mathcal{L}_{D-MMD}}$$

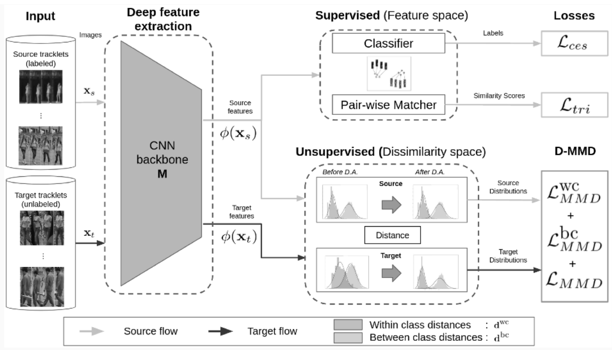
Supervised Ensure stability

Adaptation Based on strong source reference

24

UDA in the Dissimilarity Space

- D-MMD loss for adaptation of deep learning model:



25

UDA in the Dissimilarity Space

Example of results – comparison with state-of-art:

- Person ReID accuracy on Duke and MSMT target datasets, with Market1501 as source dataset

Methods	Source: Market1501							
	DukeMTMC				MSMT17			
	r-1	r-5	r-10	mAP	r-1	r-5	r-10	mAP
Lower Bound	23.7	38.8	44.7	12.3	6.1	12.0	15.6	2.0
BUC [Lin et al., 2019]	47.4	62.6	68.4	27.5	-	-	-	-
ECN [Zhong et al., 2019]	63.3	75.8	80.4	40.4	25.3	36.3	42.1	8.5
D-MMD (Ours)	**63.5**	**78.8**	**83.9**	**46.0**	**29.1**	**46.3**	**54.1**	**13.5**

Conclusion: Dissimilarity space was a viable alternative for image retrieval (metric learning) problems

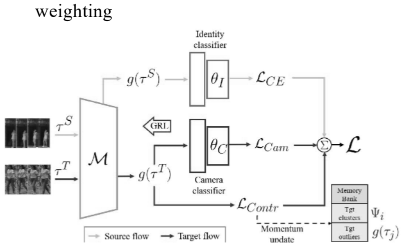
26

26

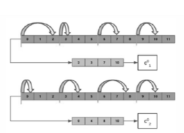
UDA in the Dissimilarity Space

Camera Alignment and Weighted Contrastive Learning for UDA in Video Person ReID

- A camera classifier performs adversarial alignment between target camera distributions
- Estimates the reliability of contrastive loss for image pairs using *k*NN weighting



Clip sampling: provide strong positive pairs that provide a helpful guidance



D. Mekhazni, et al., Camera Alignment and Weighted Contrastive Learning for Domain Adaptation in Video Person ReID, WACV 2023.

27

UDA in the Dissimilarity Space

Camera Alignment and Weighted Contrastive Learning for Domain Adaptation in Video Person ReID

Method	Setting	iLIDS → PRID (2 cameras)		PRID → iLIDS (2 cameras)		iLIDS → MARS (6 cameras)	
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
Lower Bound (sup. S only)	–	49.0	60.0	12.7	20.9	19.8	10.1
DGM+IDE [40], ICCV'19	OneEx	–	–	–	–	36.8	16.8
Stepwise [22], ICCV'17	OneEx	–	–	–	–	41.2	19.6
EUG [36], CVPR'18	OneEx	–	–	–	–	**62.2**	42.5
TAUDL [3], BMVC'18	Unsup	85.3	–	56.9	–	46.8	21.4
UTAL [17], TPAMI'19	Unsup	54.7	–	35.1	–	49.9	35.2
UGA [35], ICCV'19	Unsup	80.9	–	57.3	–	58.1	39.3
BUC [19], AAAI'19	Unsup	–	–	–	–	61.1	38.0
Soft Sim. [20], CVPR'20	Unsup	–	–	–	–	61.9	43.6
SPCL* [8], NeurIPS'20	UDA	77.6	82.1	41.9	47.6	37.6	20.4
Ours ($\mathcal{L}_{\text{cam}}^{\text{CLC}}$)	UDA	70.8	77.3	32.0	42.6	31.5	16.3
Ours ($\mathcal{L}_{\text{cam}}^{\text{CLC}} + \mathcal{L}_{\text{contr}}^{\text{kNN}}$)	UDA	**86.5**	**89.9**	**58.3**	**66.7**	**62.2**	**44.8**
Upper Bound (sup. $S \cup T$)	Tuning	92.1	94.5	76.0	84.0	86.9	81.8

Source: D. Mekhazni, et al., Camera Alignment and Weighted Contrastive Learning for Domain Adaptation in Video Person ReID, WACV 2023.

28

28

Bidirectional Multi-Step Domain Generalization for Visible-Infrared Person Re-Identification

M. Alehdaghi, P. Shamsolmoali, R.M.O. Cruz & E. Granger

submitted to CVPR 2024

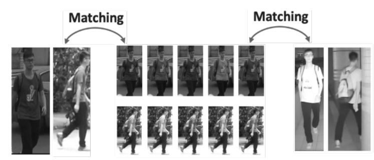
Visible-Infrared ReID Using Privileged Information

Cross-Modal ReID – match persons/objects across RGB and IR cameras

- challenging because of the large shift between RGB and IR data distributions

Our approach: reduce the domain gap – leverage related privileged information (PI) as intermediate domains to train the CNN backbone:

- learning under privileged information (LUPI) paradigm
- generate privileged intermediate representations that connect the RGB and IR modalities during training epochs



30

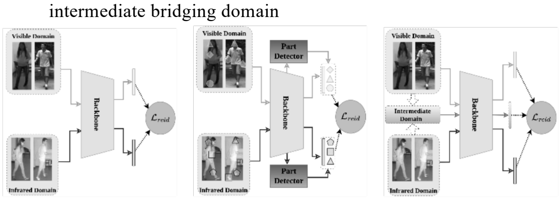
Visible-Infrared ReID Using Privileged Information

Overview of training architectures for V-I ReID:

(a) a global features representation

(b) a local part-based representation to preserve locality alongside the global features

(c) approaches based on an intermediate modality that generates an intermediate bridging domain



(a) Global representation. (b) Part-based representation. (c) Intermediate modality.

domain gap is too large

- One-step approach
- More accurate, but the intermediate domain may not capture enough common discriminant information

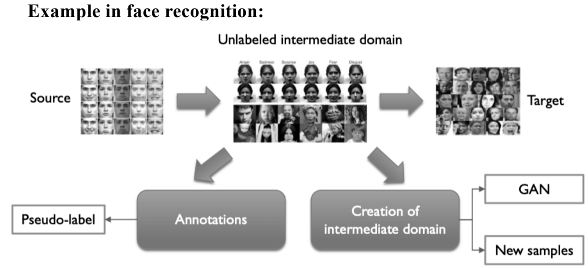
31

Gradual Domain Adaptation

Motivation:

- select intermediate domains with smaller domain shift
- gradual and multi-step UDA can improve accuracy when there is a large domain shift

Example in face recognition:



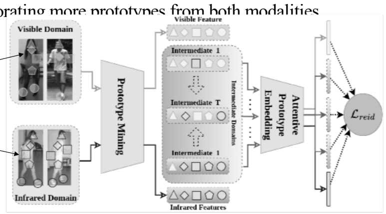
Bidirectional Multi-Step Domain Generalization

Main idea – minimize the cross-modal gap by identifying shared prototypes that capture key discriminative features across modalities

Extract prototypes from both modalities in each step to create multiple intermediate bridging domains

1) create multiple virtual intermediate domains using a combination of body part prototypes extracted from both I and V modalities
2) bidirectional multi-step learning progressively improves feature representations in each step by incorporating more prototypes from both modalities

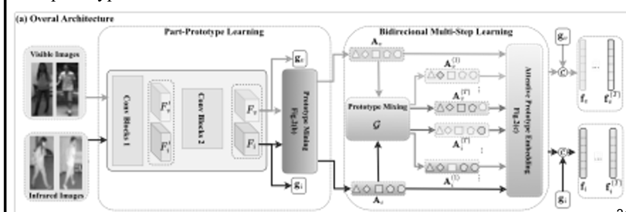
Part- Prototypes: feature representations are linked to specific body parts, allowing for effective integration of features from both modalities



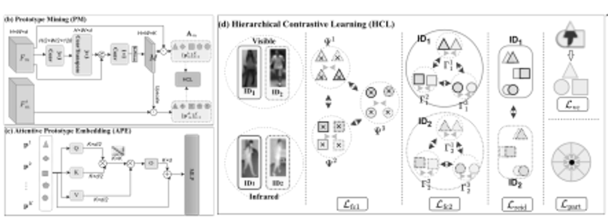
Bidirectional Multi-Step Domain Generalization

Overall Training Architecture:

- **Prototype learning module (left):** extracts body part prototype representations from V and I images
  - uses a shallow U-Net to create a region mask for each prototype
- **Bidirectional multi-step learning module (right):** learns discriminant features using multiple intermediate domains created by mixing prototype information



Bidirectional Multi-Step Domain Generalization



(b) Prototype mining (PM): mines prototypes from spatial features

(c) Attentive prototype embedding (APE) attention mechanism to aggregate mixed prototypes to produce output features

- emphasize important channels in their feature map and weights prototype features based on similarities between them

(d) Hierarchical contrastive learning (HCL) allows the prototypes to focus on similar semantics for all individuals without losing ID-discriminative information

Bidirectional Multi-Step Domain Generalization

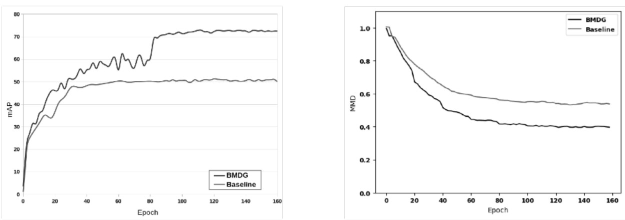
Results: Accuracy of the proposed BMDG and state-of-the-art methods on the SYSU-MM01 (single-shot setting) and RegDB datasets. All numbers are percent.

Family			SYSU-MM01						RegDB					
			All Search			Indoor Search			Visible → Infrared			Infrared → Visible		
	Method	Venue	R1	R10	mAP	R1	R10	mAP	R1	R10	mAP	R1	R10	mAP
Intermediate	SMCL[36]	ICCV'21	67.39	92.84	61.78	68.84	96.55	75.56	83.93	-	79.83	83.05	-	78.57
	MMN [53]	ICM'21	70.60	96.20	66.90	76.20	**99.30**	79.60	91.60	97.70	84.10	87.50	96.00	80.50
	RPIG [2]	ECCVw'22	71.08	96.42	67.56	82.35	98.30	82.73	87.95	98.3	82.73	86.80	96.02	81.26
	FTMI [31]	MVA'23	60.5	90.5	57.3	-	-	-	79.00	91.10	73.60	78.8	91.3	73.7
	G2DA [33]	PR'23	63.94	93.34	60.73	71.06	97.31	76.01	-	-	-	-	-	-
	SEFL [11]	CVPR'23	75.18	96.87	70.12	78.40	97.46	81.20	91.07	-	85.23	92.18	-	86.59
	BMDG (ours)[a]	-	75.43	97.42	72.86	82.35	98.02	82.16	92.59	98.11	89.18	94.08	97.0	88.67
	BMDG (ours)[b]	-	76.39	**97.90**	**78.22**	**83.59**	98.96	**83.87**	**94.76**	98.91	**92.21**	**94.56**	98.31	**93.07**

Bidirectional Multi-Step Domain Generalization

Results: The mAP and domain shift (MMD distance) between I and V modalities over training epochs.

(a) Training curve of BMDG vs baseline on-step model
(b) MMD over the center of I and V features from IDs of the SYSU-MM01 dataset



(a) Learning Curve. (b) BMDG.

37

37

Bidirectional Multi-Step Domain Generalization

Ablations: Accuracy (R1% and mAP%) of BMDG for different numbers of part prototypes (K) and intermediate steps (T)

T	Number of part prototypes (K)					
	3	4	5	6	7	10
0	68.25	69.24	69.40	70.27	70.03	68.12
1	69.97	70.81	72.07	71.97	71.31	69.28
2	71.20	72.35	73.98	73.61	72.45	71.25
3	73.32	73.94	74.11	74.98	73.22	71.67
4	-	74.08	74.15	**75.43**	73.51	71.99
6	-	-	-	75.37	73.52	72.15
10	-	-	-	-	-	72.07

T	Number of part prototypes (K)					
	3	4	5	6	7	10
0	65.98	67.03	67.23	68.11	67.55	65.66
1	67.42	68.72	69.28	69.46	68.32	69.28
2	69.51	70.08	70.72	71.14	69.97	71.25
3	71.44	71.69	71.82	72.02	71.06	71.67
4	-	71.98	72.15	**72.86**	71.19	69.00
6	-	-	-	72.40	71.17	69.54
10	-	-	-	-	-	69.46

Accuracy of part-based ReID methods with BMDG on the SYSU-MM01, under single-shot setting with training.

Method	R1 (%)	mAP (%)
DDAG [46]	53.62	52.71
DDAG with BMDG	55.36	54.05
MPANet [41]	66.24	62.89
MPANet with BMDG	68.74	64.25
SAAI [10]	71.87	68.16
SAAI with BMDG	73.69	70.08

38

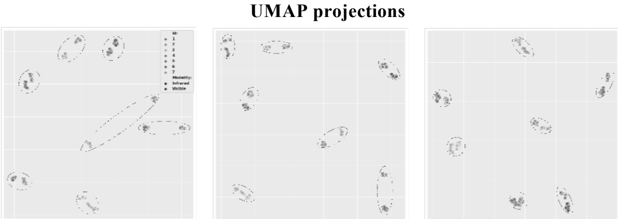
38

Bidirectional Multi-Step Domain Generalization

Results: Distributions of V and I features learned for 7 identities from the SYSU-MM01 dataset

- the baseline, one-step using part prototypes, BMDG methods
- each colour shows the identity

UMAP projections



(a) Baseline. (b) One Step. (c) BMDG.

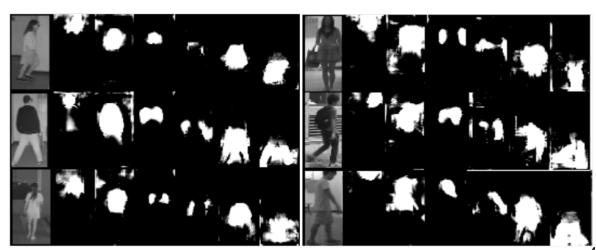
39

39

Bidirectional Multi-Step Domain Generalization

Results: Prototype regions extracted by the PM module for (a) infrared and (b) visible images. The region mask of prototypes focuses on similar body parts without accounting for identity

- mask size is 18×9, which is then resized to fit the original input image



(a) Infrared. (b) Visible.

40

40

Overview

1) Person Re-Identification
- domain adaptation in dissimilarity space
- visual-infrared reID with intermediate domains

2) Facial Expression Recognition
- subject-based multi-source UDA
- source-free adaptation with missing classes

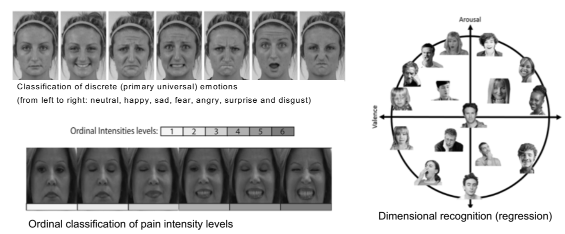
41

41

41

Affective Computing – Expression Recognition

Expression recognition is the fundamental problem of affective computing and can take several forms



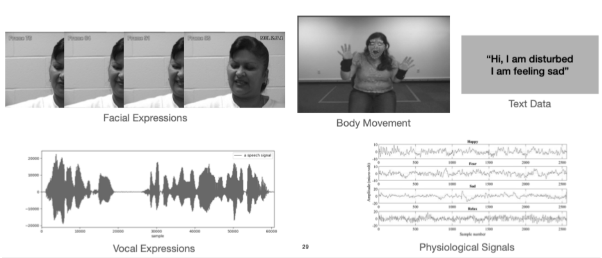
Source: Du, et al., "Compound facial expressions of emotion," Proc. National Academy of Sciences, 2014.

42

42

Affective Computing – Expression Recognition

Expression recognition: several potential modalities

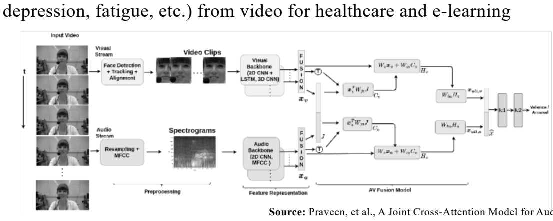


43

43

Affective Computing – Expression Recognition

Tasks: spatiotemporal recognition of expressions (linked to pain, stress, depression, fatigue, etc.) from video for healthcare and e-learning



Source: Praveen, et al., A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition, *IEEE Trans on Biometrics, Behavior, and Identity Science*, 2023.

Challenges:
- weakly-supervised learning of videos with limited and ambiguous annotations
- rapid adaptation to different persons and capture conditions
- fusion of facial, vocal and other modalities
- spatiotemporal localization and attention mechanisms

44

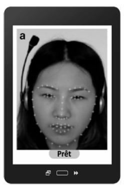
44

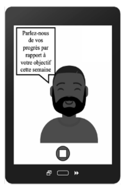
44

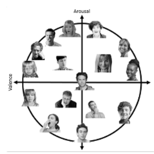
Chaire de recherche FRSQ double Concordia-ÉTS-CIUSSS-NIM en IA et santé numérique pour le changement des comportements de santé

Objectives:

- predict a subject's affective state in health diagnosis and monitoring
- estimating non-verbal cues to personalize eHealth interventions in behavior change programs
- spontaneous recognition of facial, textual, and vocal expressions related to engagement, ambivalence, hesitation, motivation, etc.







45

45

45

Affective Computing – Expression Recognition

Challenges in Real-world Applications

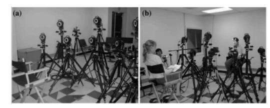
Variability of expressions across different individuals, cultures, and capture conditions

- shift in distributions between the design (source) and operational (target) datasets
- real-world capture: occlusion, pose, illumination, complex background, noise

pain (left) vs. no pain (right) frames

State-of-art DL models are complex and may require large datasets for training

- few relevant public datasets are available in health
- high cost of collecting and annotating large-scale datasets that contain controlled expressions
- label ambiguity from annotators



Adapt DL model to a specific person and capture condition using unlabeled video data from the operational environment

46

46

46

Subject-Based Domain Adaptation for Facial Expression Recognition

O. Zeeshan, M. H. Aslam, S. Belharbi, A. L Koerich, M. Pedersoli, S. Bacon & E. Granger

submitted to Face and Gesture 2024

47

47

Subject-based UDA for Facial Expression Recognition

UDA methods: several have been proposed to adapt deep FER models across source and target data sets

Challenge:

- the high intra- and inter-person variability in FER, so it would help to account for different subjects
- state-of-the-art methods do not scale well to a larger number of source domains

Our general approach:

- consider that each subject corresponds to a domain, not entire datasets, but ensure that it can scale well to many sources
- employ an MSDA method – leverage multiple subject-specific source domains allows for an accurate representation of the intra- and inter-person variability

48

48

48

Subject-based UDA for Facial Expression Recognition

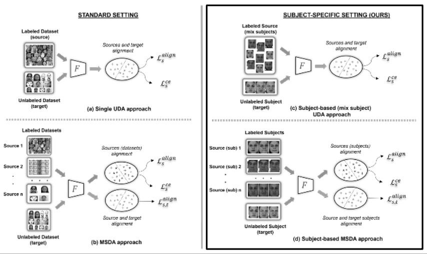
Settings for adaptation of a deep FER model:

a) **Single UDA**, where one labeled dataset is adapted to a single unlabeled dataset

b) **MSDA** aligns multiple source datasets and then adapts to the single target domain

c) **Subject-based UDA** considers that a single labeled source dataset as a mix with different subject IDs is aligned with unlabeled target subjects

d) **Subject-based UDA** considers each subject as a separate domain, mitigating the domain shift among the sources and then aligning the source with the target subject



blue: labeled source data
green: unlabeled target data
grey: data from both domains

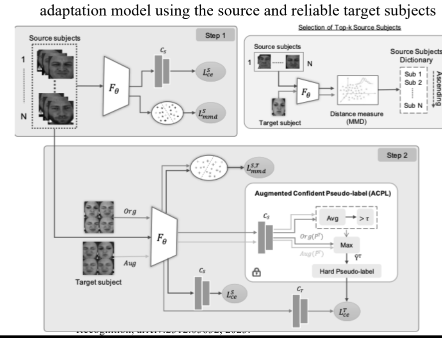
49

49

Subject-based UDA for Facial Expression Recognition

Proposed subject-based MSDA method

1) align labeled source subjects using discrepancy and supervision loss
2) apply the ACPL strategy to generate reliable target pseudo-labels and train the adaptation model using the source and reliable target subjects



Can also select the top k closest sources to the target using MMD for the adaptation process

50

50

Subject-based UDA for Facial Expression Recognition

Step 1: align labeled source subjects using discrepancy and supervision losses

$$\mathcal{L}_{ce}^{s} = \frac{1}{N^s}\sum_{i=1}^{N^s} [C_s(F_\theta(x_i^s)), y_i^s] \qquad \mathcal{L}_{mmd}^{s} = MMD(S1, S2), \qquad \mathcal{L}_{align}^{s} = \mathcal{L}_{ce}^{s} + \mathcal{L}_{mmd}^{s}$$

Step 2: generate reliable target pseudo-labels using Augmented Confident Pseudo-Label (ACPL),

- generate softmax probability from X^T and $\hat{X}^T$:

$$P_n^T = p_1, \ldots, p_n \quad , \qquad \hat{P}_n^T = \hat{p}_1, \ldots, \hat{p}_n$$

- average these probabilities:

$$a_n = \frac{p_1 + \hat{p}_1}{2}, \ldots, \frac{p_n + \hat{p}_n}{2} \qquad p_n^* = \max(p_n, \hat{p}_n)$$

- Applying confident threshold and assigning label:

$$\hat{Y}^T = \underset{i \in n}{\operatorname{argmax}}\{(p_i^*).1(a_i > \tau)\} \qquad N_{cl}^T = (X^T, \hat{Y}^T)$$

and adapt the model using the target and reliable sources

$$\mathcal{L}_{ce}^{t} = \frac{1}{N_{cl}^T}\sum_{i=1}^{N_{cl}^T} [C_T(F_\theta(x_i^t)), \hat{y}_i^t] \qquad \mathcal{L}_{mmd}^{s,t} = MMD(S,T), \qquad \mathcal{L}_{adapt}^{t} = \mathcal{L}_{ce}^{s} + \mathcal{L}_{ce}^{t} + \mathcal{L}_{mmd}^{s,t}$$

51

51

Subject-based UDA for Facial Expression Recognition

Results: BioVid heat and pain (PartA) with 87 subjects

- 10 subjects treated as a target domain, the remaining 77 subjects as source domains
- ResNet18 is used in all of the experiments

Setting	Methods	Sub-1	Sub-2	Sub-3	Sub-4	Sub-5	Sub-6	Sub-7	Sub-8	Sub-9	Sub-10	Avg
Source combined	Source-only	0.62	0.61	0.65	0.55	0.51	0.71	0.7	0.52	0.54	0.55	0.59
	Subject-based (UDA)	0.73	0.64	0.73	0.59	0.54	0.75	0.76	0.53	0.51	0.58	0.63
Multi-Source	M3SDA	0.67	0.66	0.61	0.58	0.55	0.50	0.67	0.56	0.54	0.67	0.60
	CMSDA	**0.93**	0.47	0.81	**0.87**	0.53	0.84	0.57	0.54	0.74	**0.70**	0.70
	SImpAI	0.80	0.69	0.55	0.75	0.52	0.81	0.71	0.61	0.59	0.56	0.65
	Subject-based (MSDA)	**0.93**	0.69	0.84	0.64	0.57	0.85	0.81	0.58	0.60	0.60	0.71
	Subject-based with top-k	**0.93**	**0.71**	**0.86**	**0.87**	**0.88**	**0.92**	**0.86**	**0.77**	**0.84**	0.68	**0.83**
Oracle	Fully-supervised	0.99	0.91	0.98	0.97	0.98	0.97	0.96	0.95	0.99	0.98	0.96

Selected a max of top k=30 closest source subjects from each target subject

Source: Zeeshan et al., Subject-Based Domain Adaptation for Facial Expression Recognition, arXiv:2312.05632, 2023.

52

52

Subject-based UDA for Facial Expression Recognition

Results: UNBC-McMaster Shoulder Pain with 25 subjects

- 5 subjects treated as a target domain, the remaining 20 subjects as source domains
- ResNet18 is used in all of the experiments

Setting4	Sub-1	Sub-1	Sub-2	Sub-3	Sub-4	Sub-5	Avg
Source combined	source-only	0.74	0.84	0.81	0.68	0.83	0.78
	Subject-based (UDA)	0.76	0.87	0.84	0.70	0.85	0.80
Multi-Source DA	M³SDA	0.78	0.87	0.92	0.66	0.81	0.80
	CMSDA	0.80	0.86	0.83	0.71	0.85	0.81
	SImpAI	0.80	0.88	0.81	0.70	0.87	0.81
	Subject-based (MSDA)	**0.81**	**0.91**	**0.94**	**0.72**	**0.92**	**0.86**
Oracle	Fully-supervised	0.99	0.91	0.98	0.97	0.98	0.96

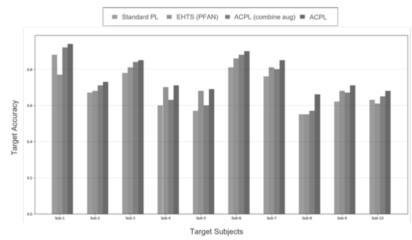
Selected top *k=10* closest source subjects from each target subject

53
53

Future Steps

Better analysis of techniques like ACPL for generating target pseudo-labels:

- **Cyan:** standard way of generating pseudo-label
- **Orange**: EHTS (PFAN) approach.
- **Green:** ACPL strategy by combining different augmentation, i.e., horizontal-flip, vertical-flip, increase sharpness, and rotation-90◦
- **Blue:** ACPL technique with only horizontal-flip augmentation.



54
54

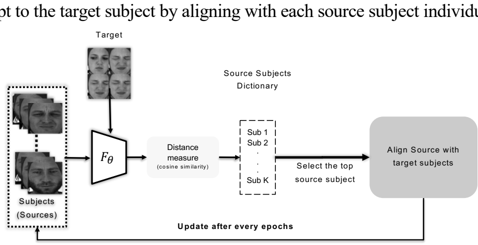
Future Steps

Subject-based DA using self-paced learning

Most to least similar subjects to the target domains used to adapt the models

Subject-based MSDA (curriculum learning)

- select top-k closest source subjects, using some distance between source and target distributions
- adapt to the target subject by aligning with each source subject individually



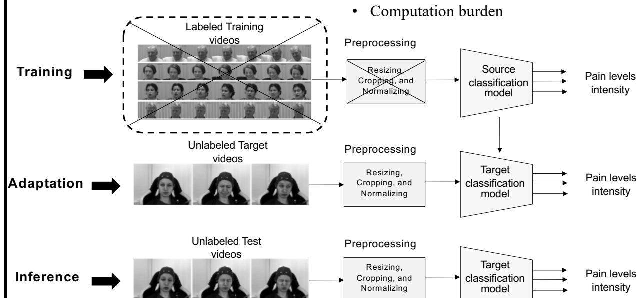
55

Future Steps

Source-Free Adaptation

Why source-free UDA?

- Data privacy protection
- Data storage and transmission cost
- Computation burden



Walter, S., The Biovid Heat Pain Database data. *CYBCO 2013.*

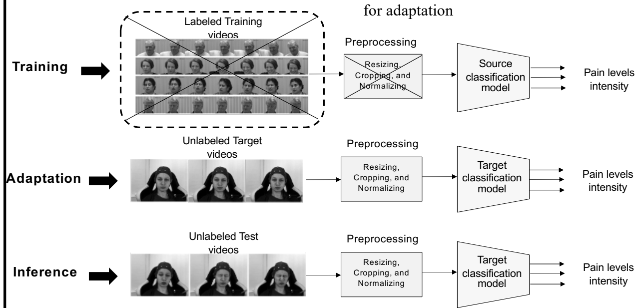
56

Future Steps

Source-Free Adaptation with Neutral Videos

Our Case

- Subject-based adaptation of models
- Only short neutral video is available for adaptation



Walter, S., The Biovid Heat Pain Database data. *CYBCO 2013.*

57

57

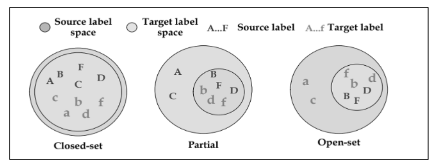
Future Steps

Source-Free Adaptation with Neutral Videos

Challenges:

- subject-based adaptation of large DL models based on a short calibration video
- adaptation on a few neutral videos from one subject
- setting: UDA where target data has missing classes

Existing Settings



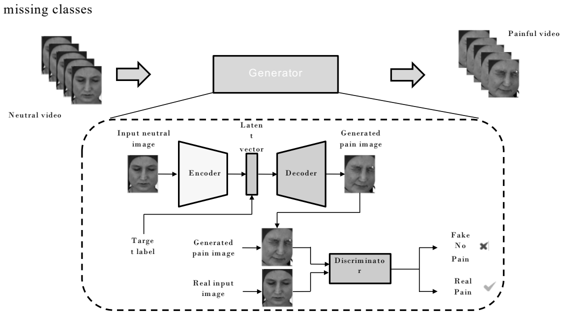
58

58

Future Steps

Source-Free Adaptation

Fine-tuning approach: SF-UDA methods that rely on discriminative image generation for missing classes



Generate videos with pain intensity levels in the target domain and perform standard SFUDA (with all target classes)

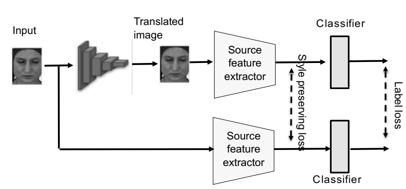
59

59

Future Steps

Source-Free Adaptation

- **Target-to-source image translation approach to preserve the source model**
- Transform the target data to look like the source data by changing the contrast and background



Steps:

- Train the transformation to transfer the input target image to the unseen source domain
- The transferred image is fed into the frozen source model
- **Style preserving loss:** Aligning the input image features with translated image features
- The same classifier is used to predict labels
- **Label loss:** Minimizing the label difference between input image and translated image

60

60

Conclusion

Many challenges remain in video-based recognition

- model: limited robustness to variations
- domain and modality shifts: divergence between domain data

But many opportunities to improve performance with the abundance of target videos?

- rely on tracklet, clip, and cluster information
- spatiotemporal dependency in videos, optical flow, etc.
- deep DA using unlabeled or weakly-labeled videos
- cross-domain (e.g., camera) and multi-modal adaptation and generalization

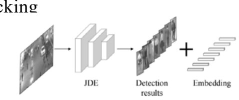
61

61

Potential Areas for Collaboration

Developing DL models for visual recognition based on image data with limited annotations:

- rapid adaptation/calibration of DL models for deployment
- video-base emotion recognition
- methods weaky-supervised learning
- weakly-supervised spatial and temporal localization for visual interpretation
- joint detection & embedding (JDE) for cost-effective ReID and multi-object tracking



Align distributions to handle multiple cameras scenarios

Le génie pour l'industrie

62

62