Adaptation of Deep Neural Networks for Video Recognition with Weakly-Labeled Data

Eric Granger
Dept. of Systems Engineering
ETS Montréal

- FRQS Co-chair in AI and Health, and ÉTS
- Industrial Co-chair on Embedded NNs for Intelligent Connected Buildings

ILLS/DATAIA Workshop
May 25, 2023

LIVIA LABORATOIRE D'IMAGERIE, DE VISION ET D'INTELLIGENCE ARTIFICIELLE

ILLS International Laboratory on Learning Systems

ÉTS

Overview

1) Personal Presentation

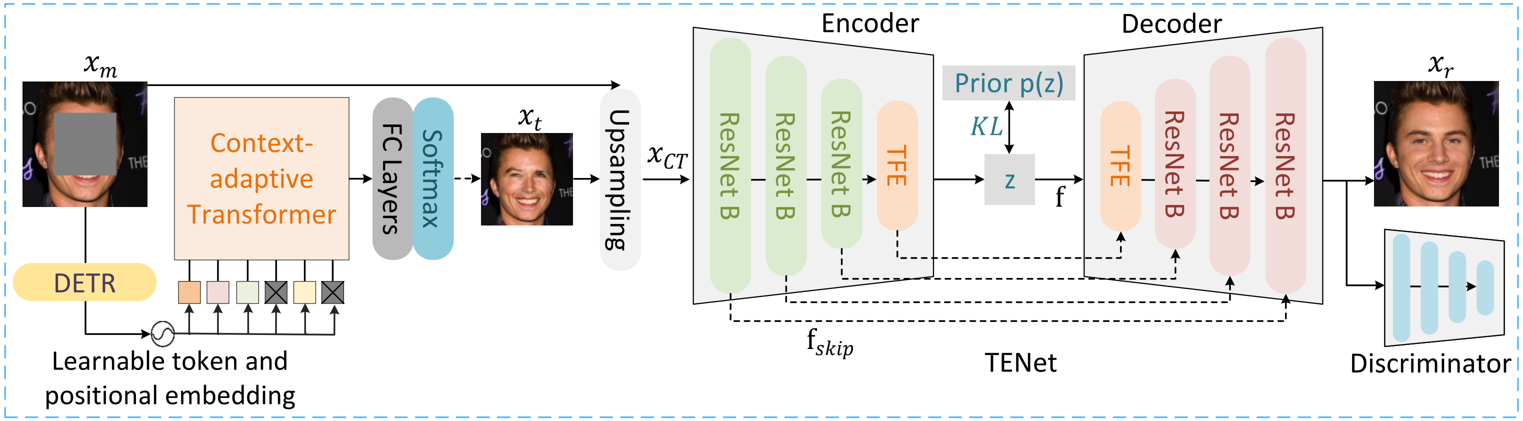
2) Recent Research:
   - unsupervised domain adaptation
   - cross-modal recognition
   - weakly-supervised object localization

3) Potential Areas of Collaboration

ÉTS Le génie pour l'industrie

Research Interests

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



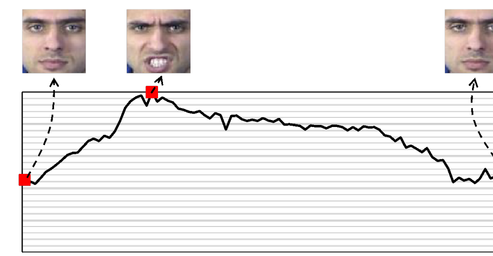
Application Areas

Video Analytics and Surveillance:

- real-time object detection, tracking, re-identification and fusion
- face analysis and recognition

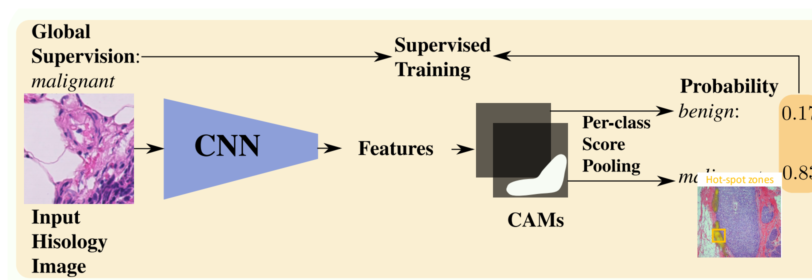
Affective Computing in Healthcare:

- spatio-temporal expression recognition
- A-V fusion of facial and vocal modalities



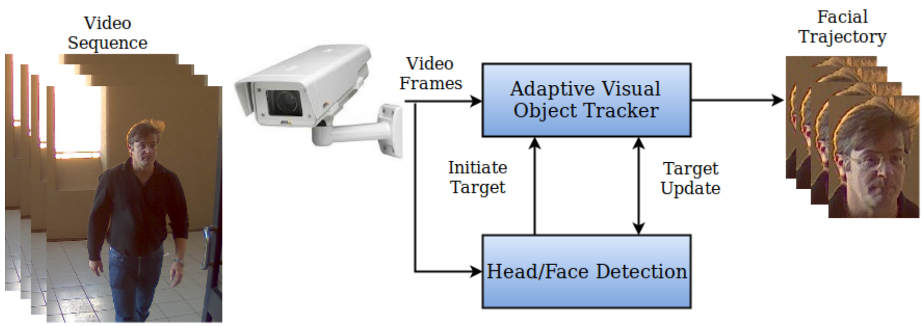
Analysis of Medical Images

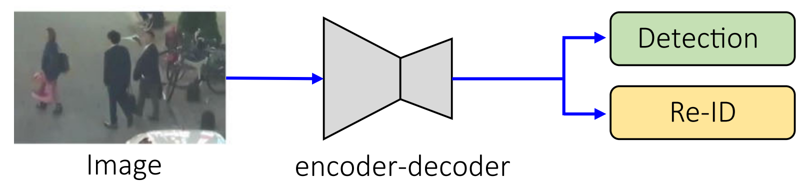
- breast cancer grading and localization in histology



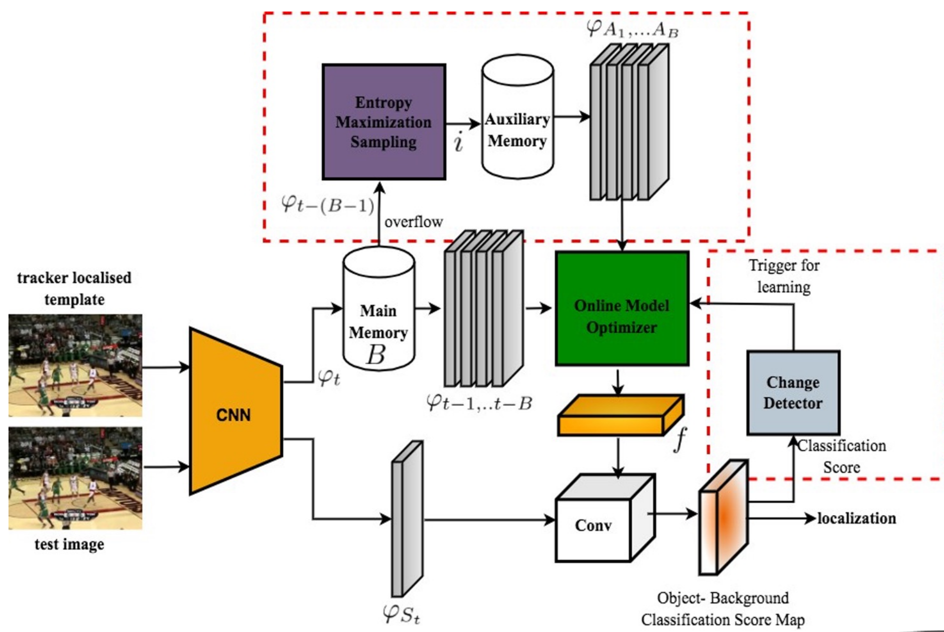
Video Analytics & Surveillance – Detection & Tracking

Front-end processing: joint detection and tracking of multiple objects appearing in a video camera, and output tracklets





Adaptive Siamese FC networks for tracking with change detection

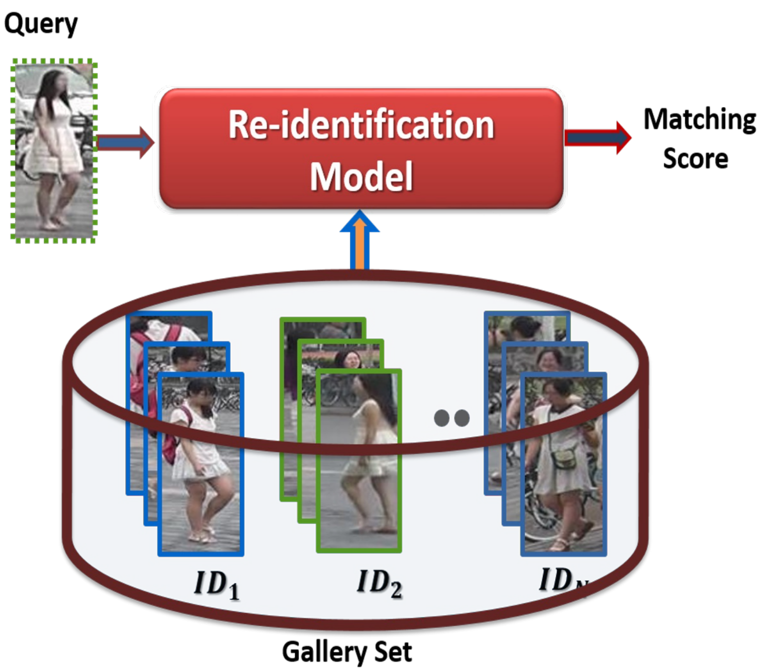


T. Wang et al., Dynamic Template Selection Through Change Detection for Adaptive Siamese Tracking, IJCNN 2022.
Zhang, Y., et al., FairMOT: On the fairness of detection and re-identification in multiple object tracking. IJCV 2021.

ÉTS Le génie pour l'industrie

Video Analytics & Surveillance – Re-Identification

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

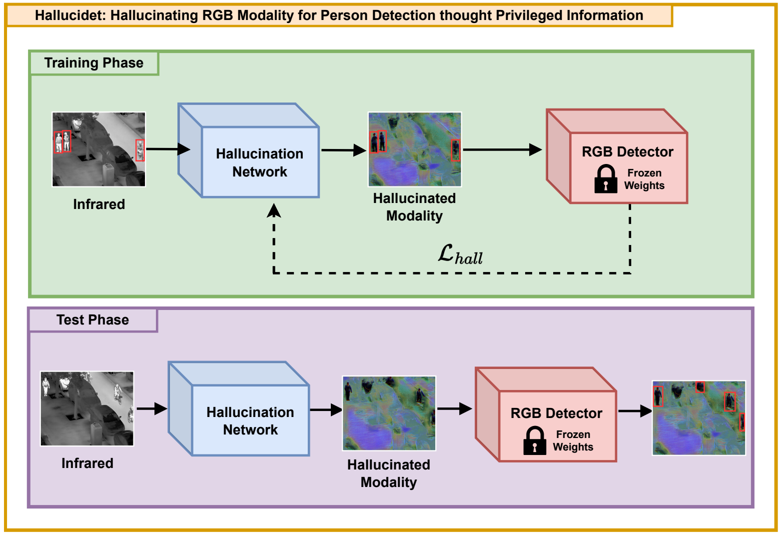


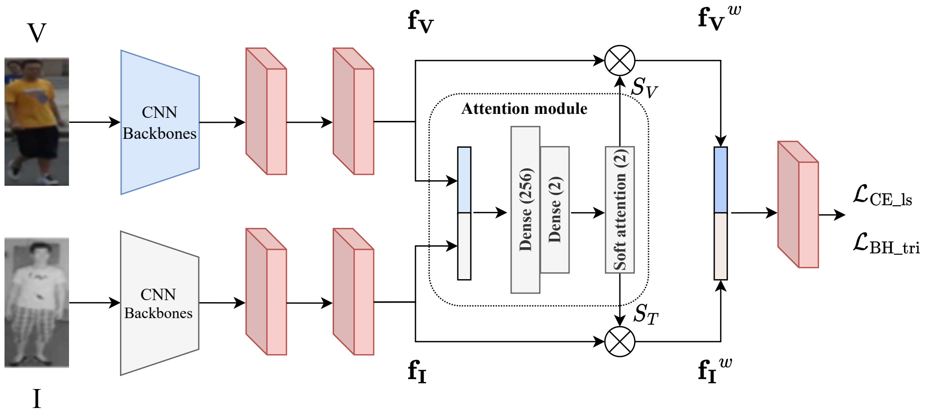
Challenges: low resolution, motion blur, occlusions, variation in pose and illumination, misalignment over different camera views

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

Video Analytics and Surveillance – Multimodal Recognition

- Leverage RGB data to improve generalization for object detection in IR
- Fusion for RGB-IR ReID from corrupted multimodal data





LLVIP Dataset: A high-resolution RGB/IR dataset for object detection.

Le génie pour l'industrie

Jia, Xinyu, et al. LLVIP: A Visible-Infrared Paired Dataset for Low-Light Vision. ICCV 2021.
Josi, Arthur, et al., Fusion for V-I Person ReID in Real-World Surveillance Using Corrupted Multimodal Data." *arXiv* 2023.

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

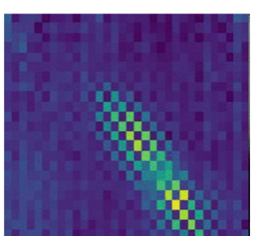
Objectives:

- Contrôle des bâtiments connectés à l'aide de capteurs distribués à coût modique et de l'IA
- Réduction de l'empreinte énergétique et augmentation du confort dans les bâtiments

Challenges:

- Intégration de l'information de divers capteurs (IR, RGB, D) à basse résolution
- Adaptation et calibration automatique des systèmes aux changements des conditions environnementales
- Réduction de la complexité des réseaux profonds pour des plateforme embarqués

caméra IR de Distech à basse résolution



personnes détectés

contrôleur Distech

Chaire de recherche industrielle Distech Controls sur les réseaux de neurones embarqués dans un contrôleur pour bâtiments connectés

Applications for intelligent building occupancy analysis using low resolution RGB and IR thermal sensors

- Adaptation and calibration of models to real-world data
- Multi-person tracking for people counting and XY localization
- Recognizing persons over multiple non-adjacent cameras
- Action/event recognition
- Model compression and acceleration
- Privacy preservation

ÉTS
Le génie pour l'industrie

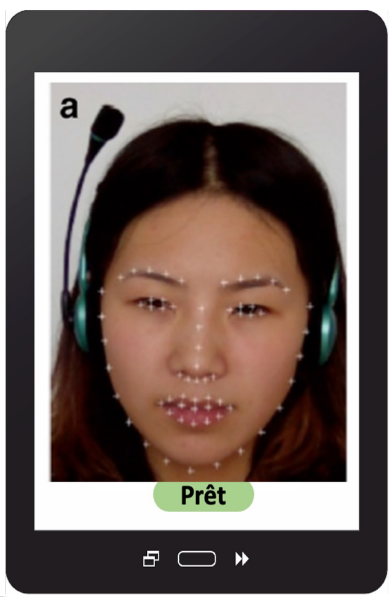
Chaire de recherche FRSQ double Concordia-ÉTS-CIUSSS-NIM en IA et santé numérique pour le changement des comportements de santé

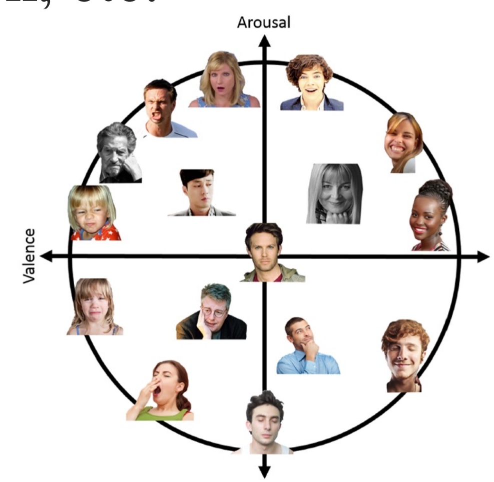
CHAIRE DOUBLE FRSQ EN IA ET EN SANTÉ NUMÉRIQUE

FRSQ DOUBLE CHAIR IN AI AND DIGITAL HEALTH

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 and vocal expressions related to engagement, ambivalence, hesitation, motivation, etc.

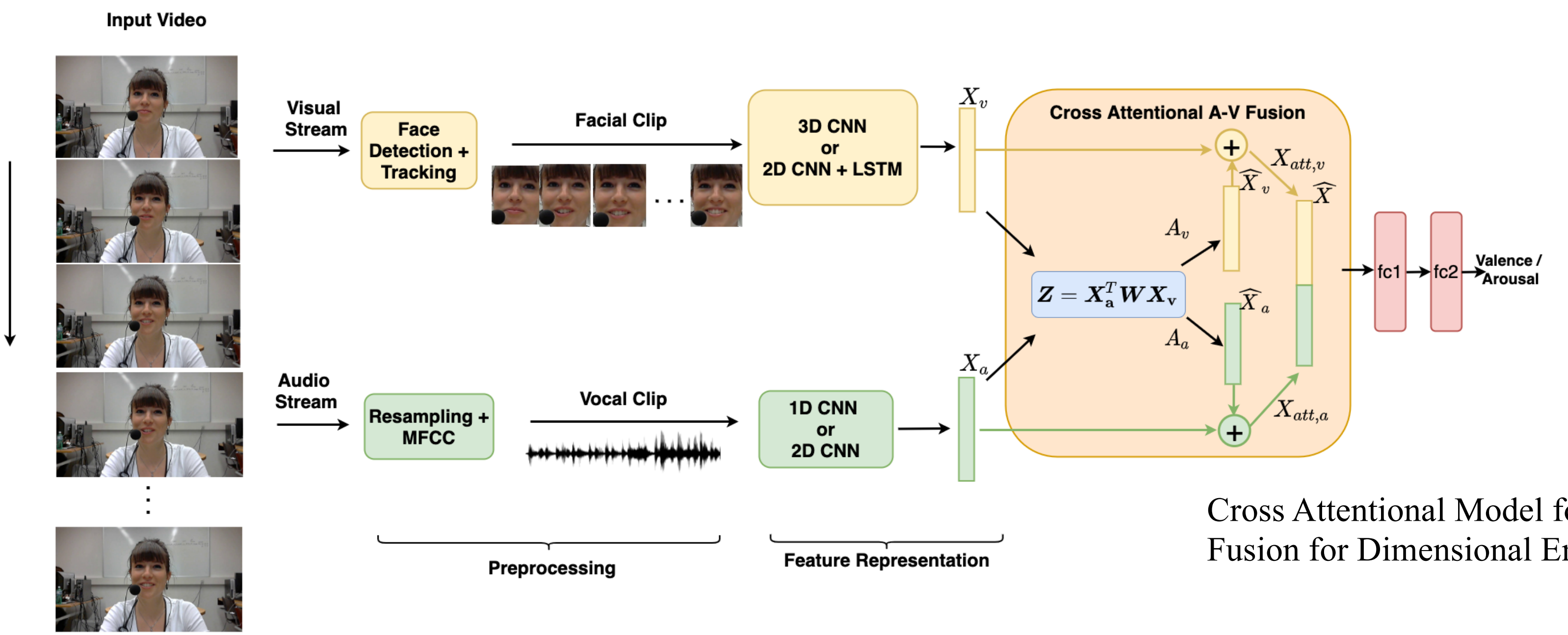




ÉTS
Le génie pour l'industrie

Affective Computing – Emotion Recognition

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



Cross Attentional Model for Audio-Visual Fusion for Dimensional Emotion Recognition

- weakly-supervised learning of videos with limited and ambiguous annotations
- rapid adaptation to different person and capture conditions
- A-V fusion of facial and vocal (and other) modalities
- spatial and temporal localisation and attention mechanisms

Rajasekhar, et al. "Cross Attentional Audio-Visual Fusion for Dimensional Emotion Recognition," FG 2021

ÉTS Le génie pour l'industrie

Challenges in Real-World Environments

Accuracy:

- domain shifts across different cameras and modalities
- variations for different people, objects, and capture conditions (pose, occlusion, illumination, scale, motion blur, etc.)
- robustness of models trained using a limited amount of annotated of image data
- SOA DL models require some labeled data for supervised training

Complexity:

- SOA DL models are complex, growing with the number of cameras and modalities

ÉTS
Le génie pour l'industrie

Focus of Talk

Applications: develop accurate ML/DL models for video-based recognition using data with limited annotations

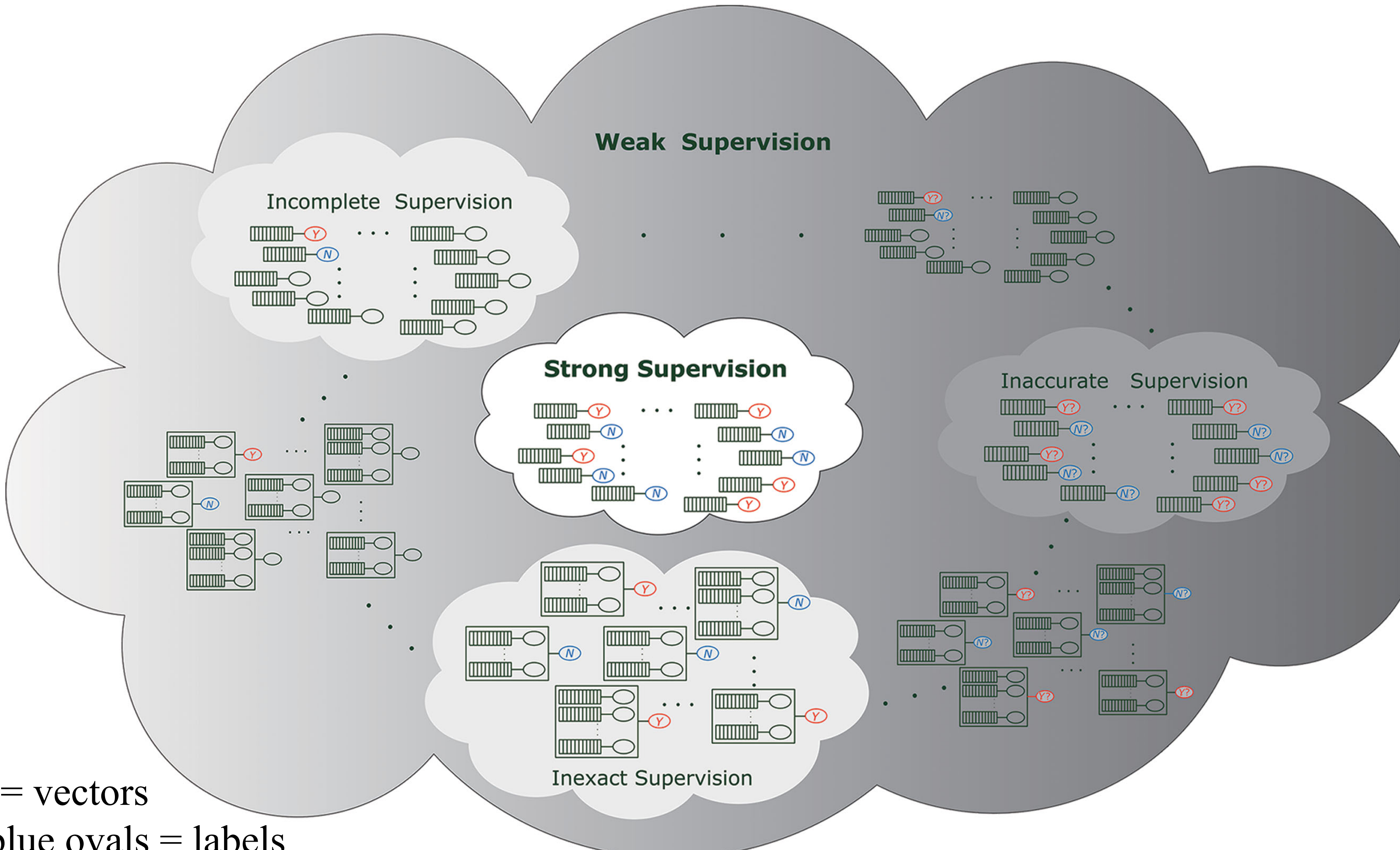
- **in monitoring/surveillance**: recognition of persons and action over different cameras and modalities
- **in healthcare:** recognition of expressions in e-health

Leveraging large amounts of videos data with limited annotation, using:

- tracklet, clip and cluster information
- domain-specific generation
- domain adaptation and generalization
- weakly-supervised learning

ÉTS Le génie pour l'industrie

Weak Supervised Learning Scenarios



- bars = vectors
- red/blue ovals = labels
- "?" = inaccurate labels

Source: Z. Zhou. 'A brief introduction to weakly supervised learning.' *National Science Review*, 5(1):44–53, 2018.

Le génie pour l'industrie

Weak Supervised Learning Scenarios

1) *Incomplete supervision*: *when only a small subset of training data has labels, although unlabelled data is abundant*
   - **active learning (AL):** query an expert to label most relevant samples
   - **semi-supervised learning (SSL):** train a model using both fully labeled and unlabeled examples

2) *Inexact supervision*: *when training on labelled data with coarse labels*
   - **multiple instance learning (MIL):** uses training examples grouped into sets (bags). Supervision is only provided for an entire set

3) *Inaccurate supervision*: *when labels may suffer from errors or noise*
   - **data-editing methods:** determine outlier annotations
   - **crowdsourcing with majority vote:** synthesis of responses from a large population of annotators

ÉTS Le génie pour l'industrie

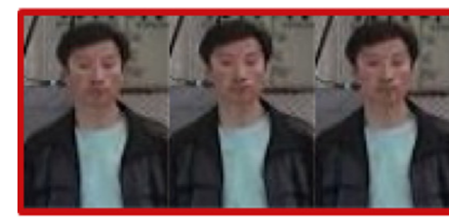
Domain Adaptation

Unsupervised Setting: Given a set of labeled SD samples, adapt a model using unlabeled TD samples to improve recognition in the TD.

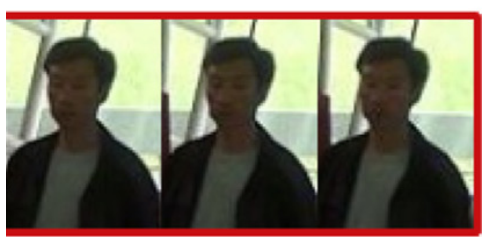
– adapt a model with both labeled (SD) and unlabeled (TD) data
– SD and TD learning tasks are the same, but data distributions differ
– example: video-based face recognition

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

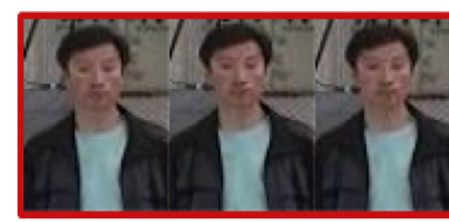
→ **TD:** video ROIs camera 3



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

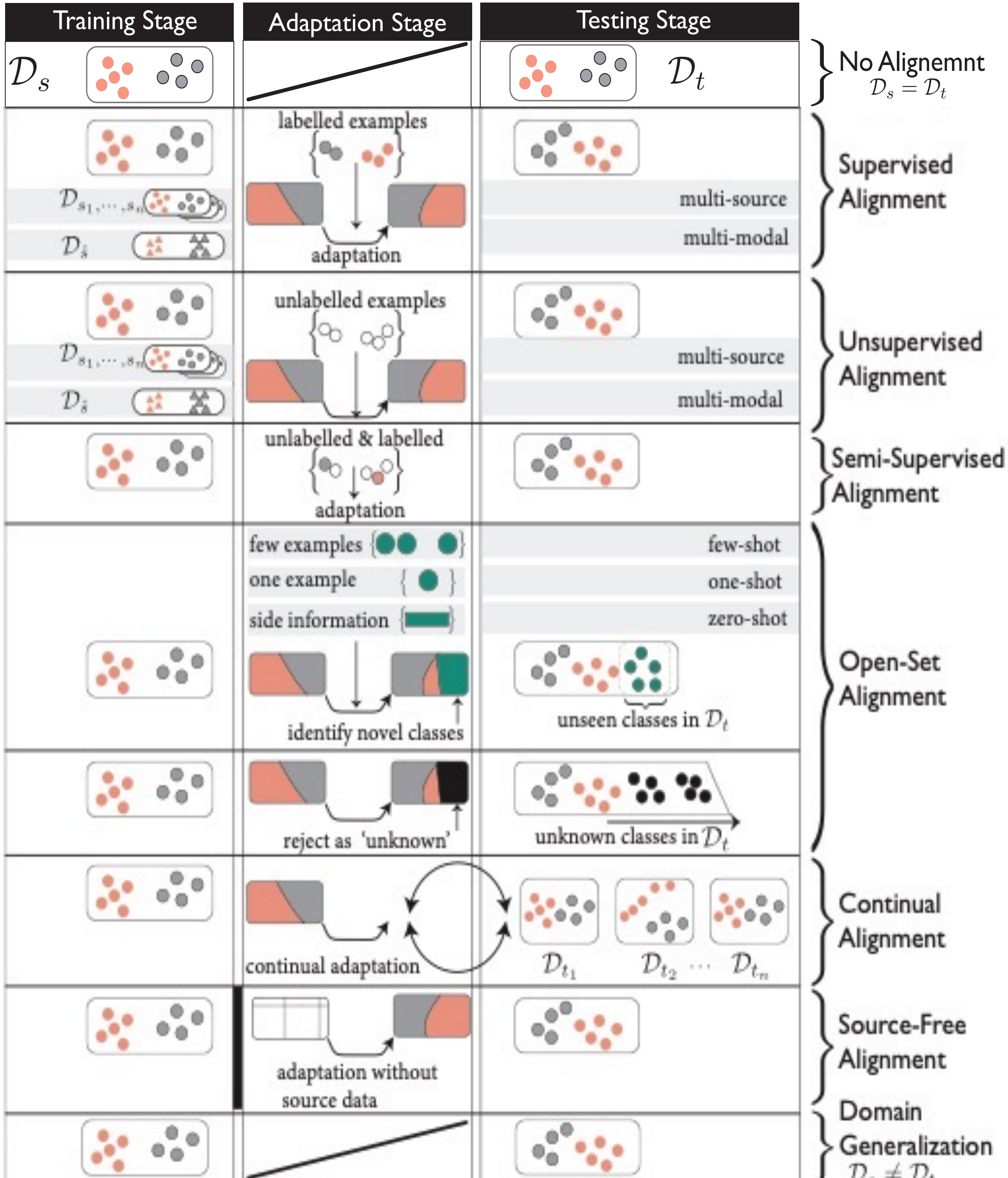


→ **TD:** video ROIs camera 3



SD: source domain
TD: target domain

Le génie pour l'industrie



Domain Adaptation Methods: learn robust domain-invariant representations from source and target samples

Approaches:

- discrepancy-based
- adversarial-based
- reconstruction-based

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

Le génie pour l'industrie

Overview

1) **Personal Presentation**

2) **Recent Research:**
   - unsupervised domain adaptation
   - cross-modal recognition
   - weakly-supervised object localization

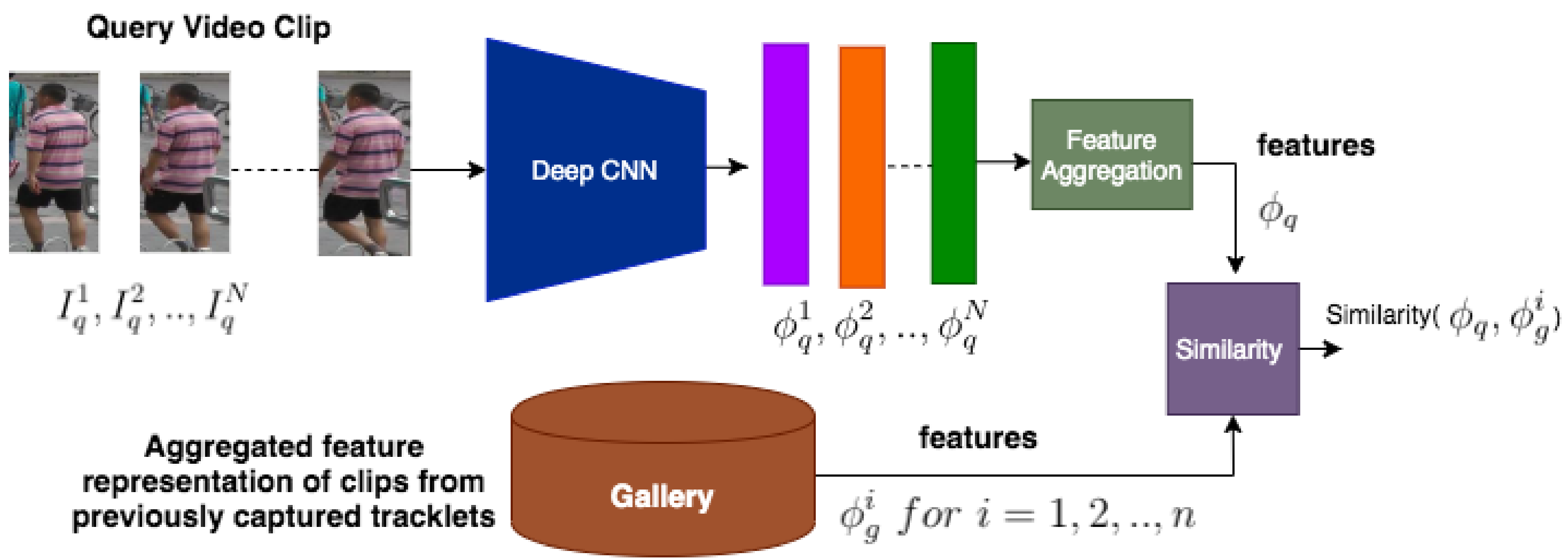
3) **Potential Areas of Collaboration**

ÉTS
Le génie pour l'industrie

UDA in the Dissimilarity Space

DL models for video-based similarity matching:

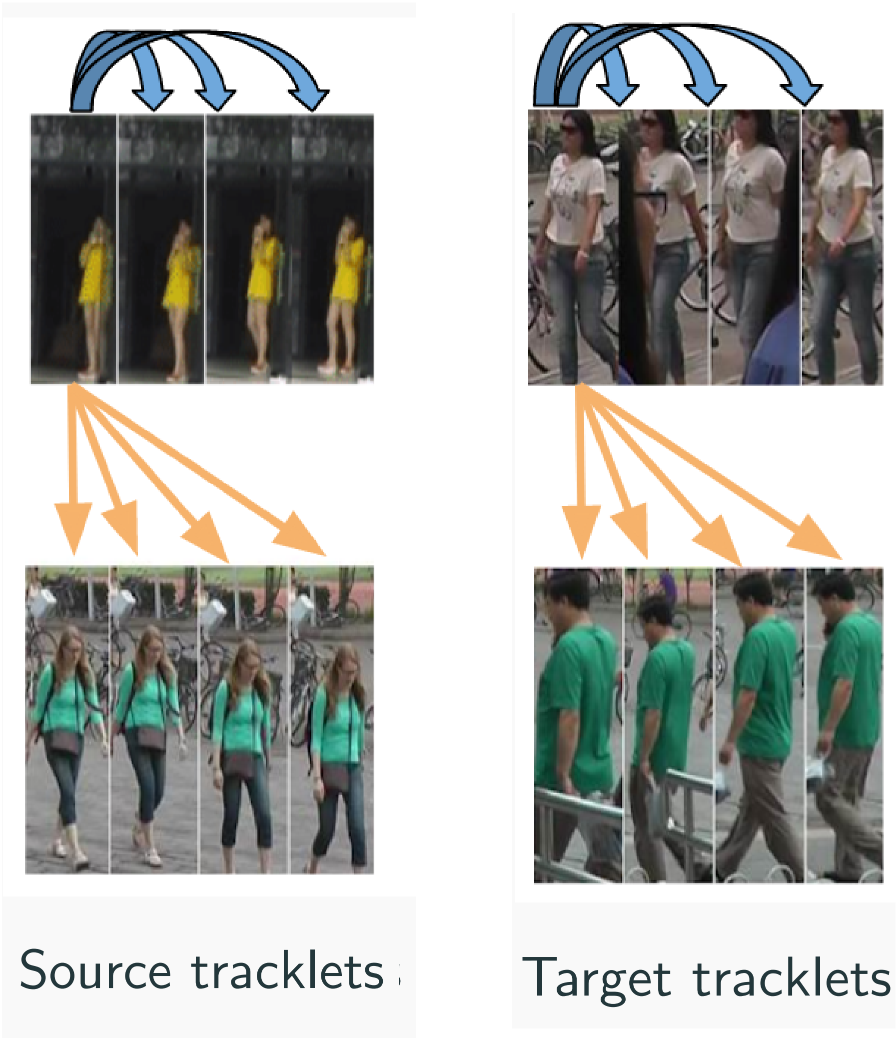
- metric learning of the embedding network for pairwise similarity
- given a clip of probe and gallery images, predict their their similarity



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

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



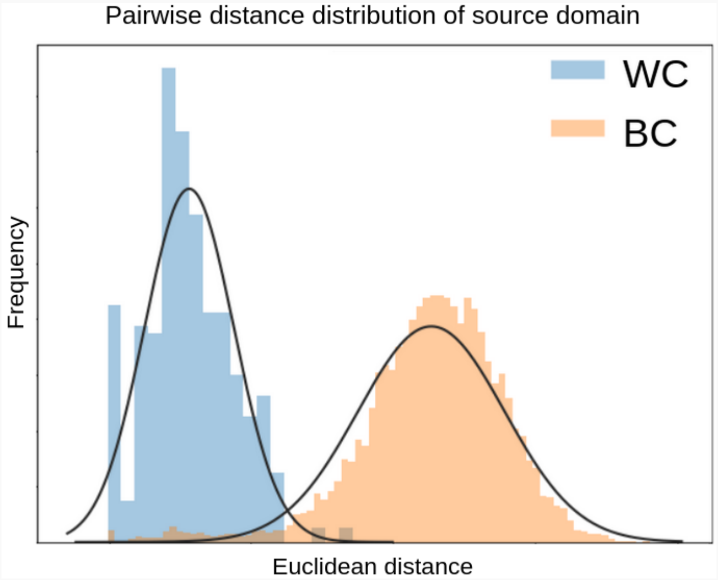
- **within class (wc):** with the same person
- **between class (bc):** with different persons

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

Le génie pour l'industrie

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 $\mathbf{x}$

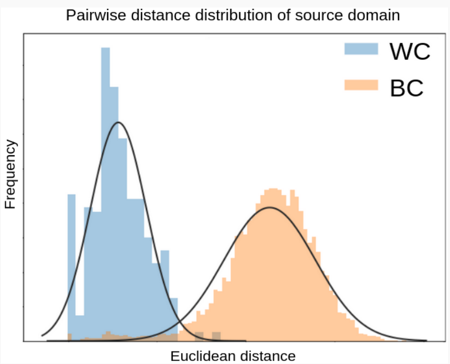
D Mekhazni, A Bhuiyan, G Ekladious & E Granger, Unsupervised Domain Adaptation in the Dissimilarity Space for Person Re-Identification, ECCV 2020.

Le génie pour l'industrie

UDA in the Dissimilarity Space

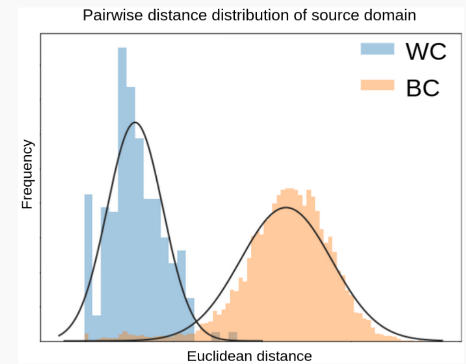
- Apply Maximum Mean Discrepancy (MMD) loss in the dissimilarity space (not the feature space)
- Align pairwise distances between source and target domain

(**d** : distances distribution)

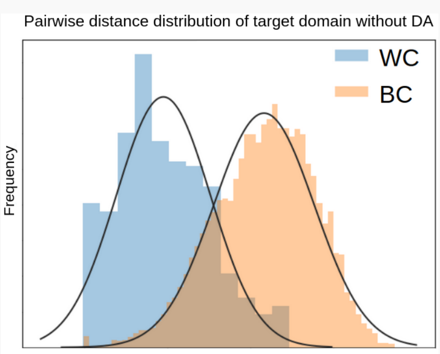


Source Distributions

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

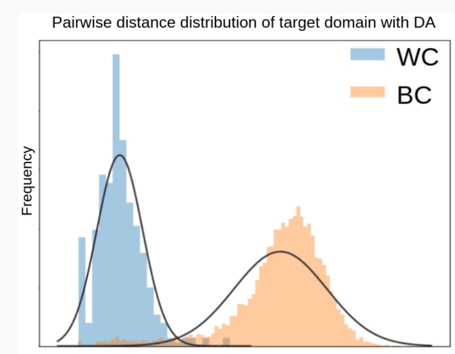


Source Distributions



Target Distributions

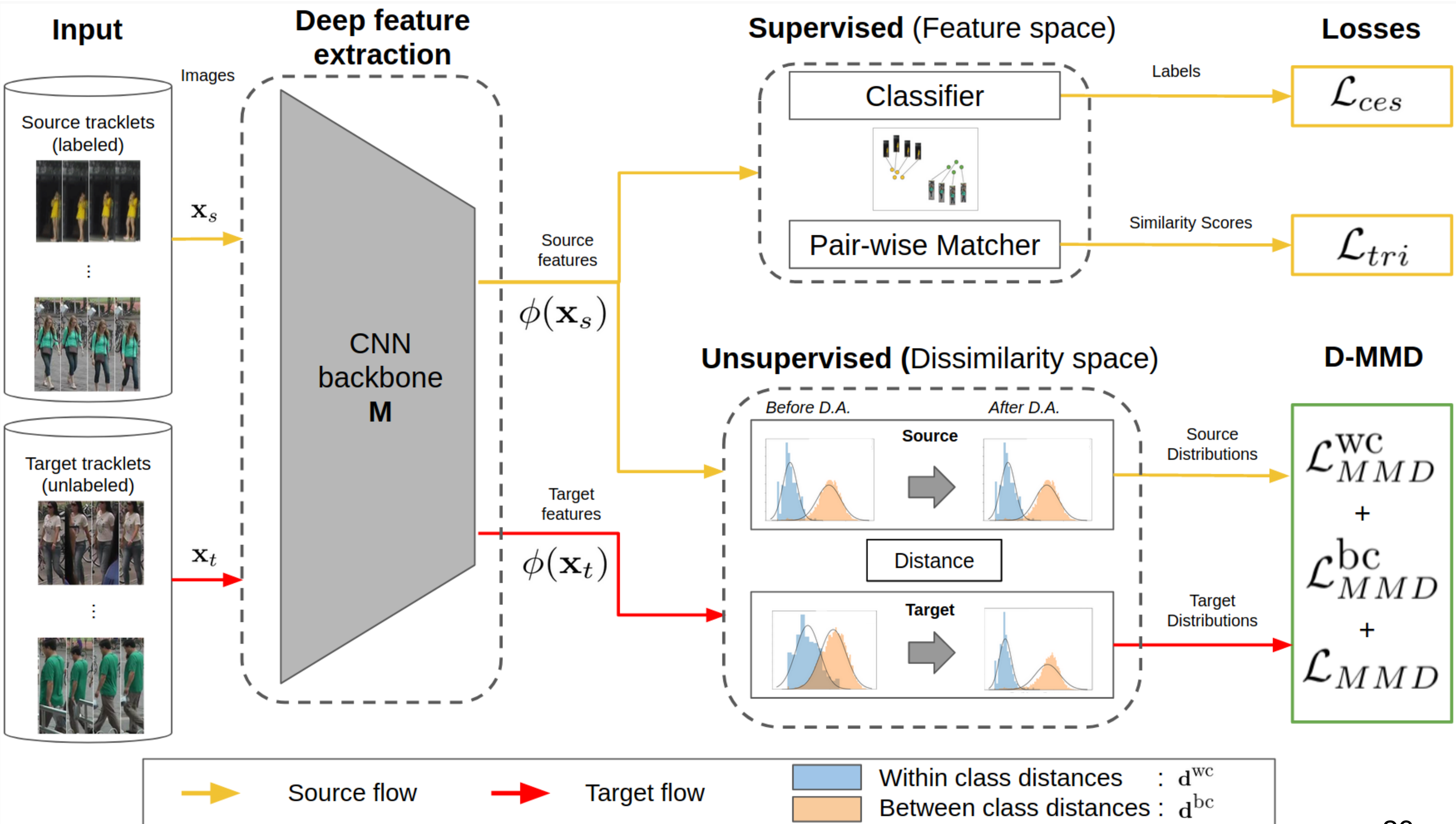
$$\mathcal{L}_{MMD}^{\text{bc}} = MMD(\mathbf{d}_s^{\text{bc}}, \mathbf{d}_t^{\text{bc}})$$



Target Distributions

UDA in the Dissimilarity Space

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



UDA in the Dissimilarity Space

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

- Video-based 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:

- Results suggest that the dissimilarity space may be a viable alternative for metric learning problems

UDA in the Dissimilarity Space

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

- addresses shift across cameras in target domain through adversarial alignment
- Estimates the reliability of contrastive loss for image pairs via *k*NN weighting

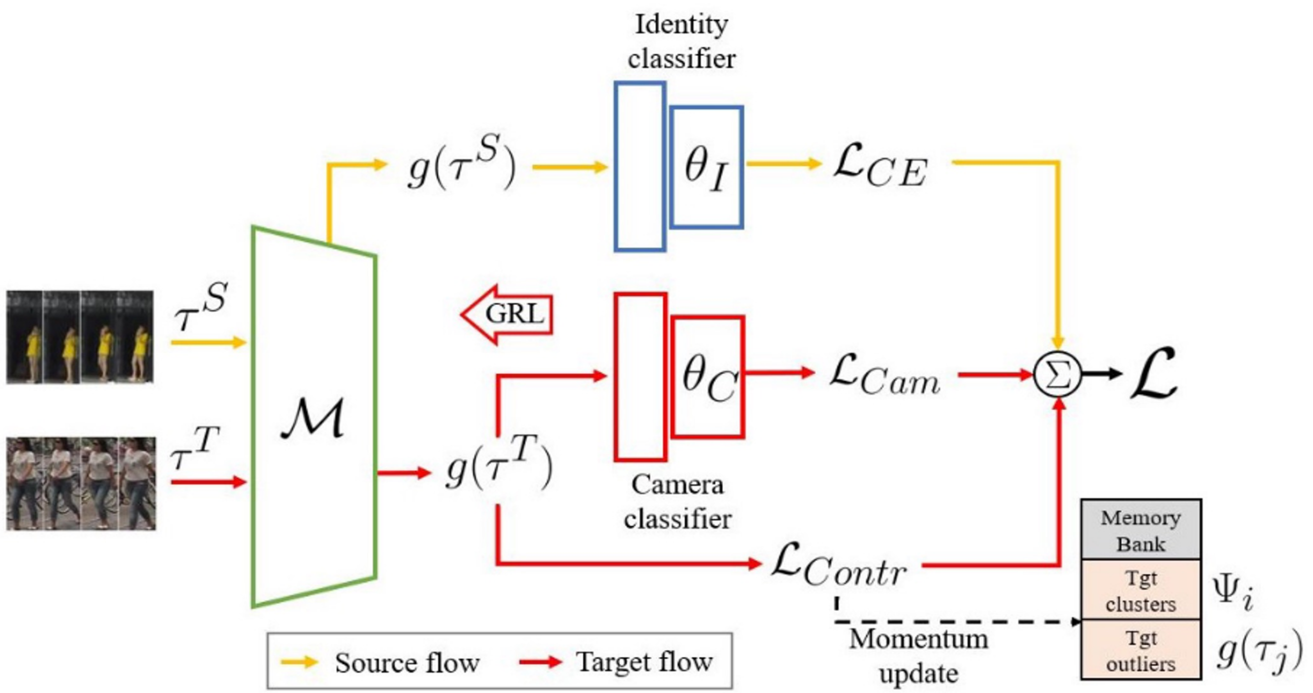
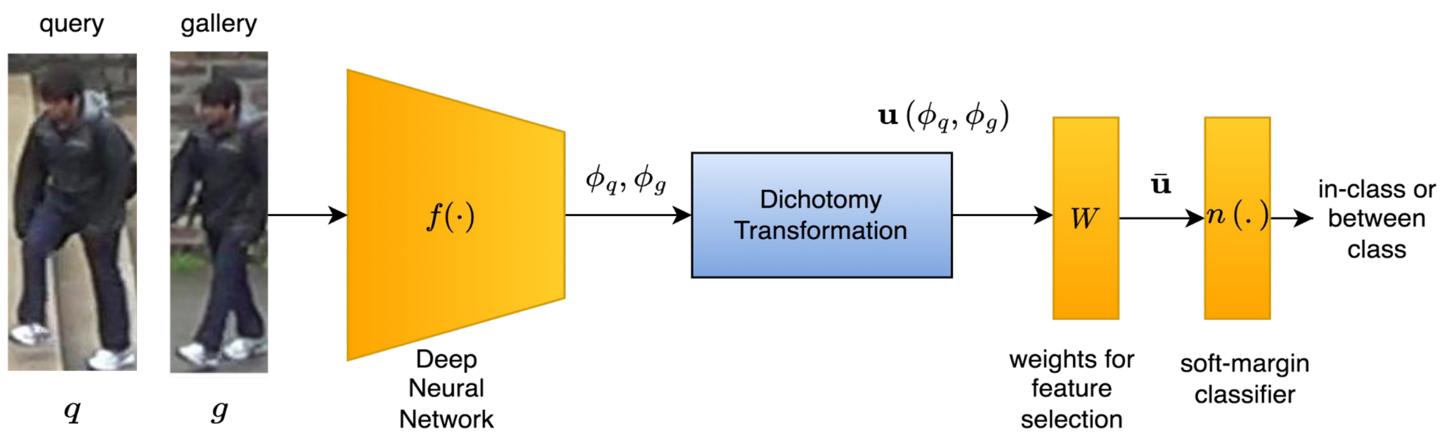


Figure 4: Overall training framework.

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

UDA in the Dissimilarity Space

DisReID : end-to-end training of the embedding network and a linear soft-margin classifier (matcher) in the the dissimilarity space

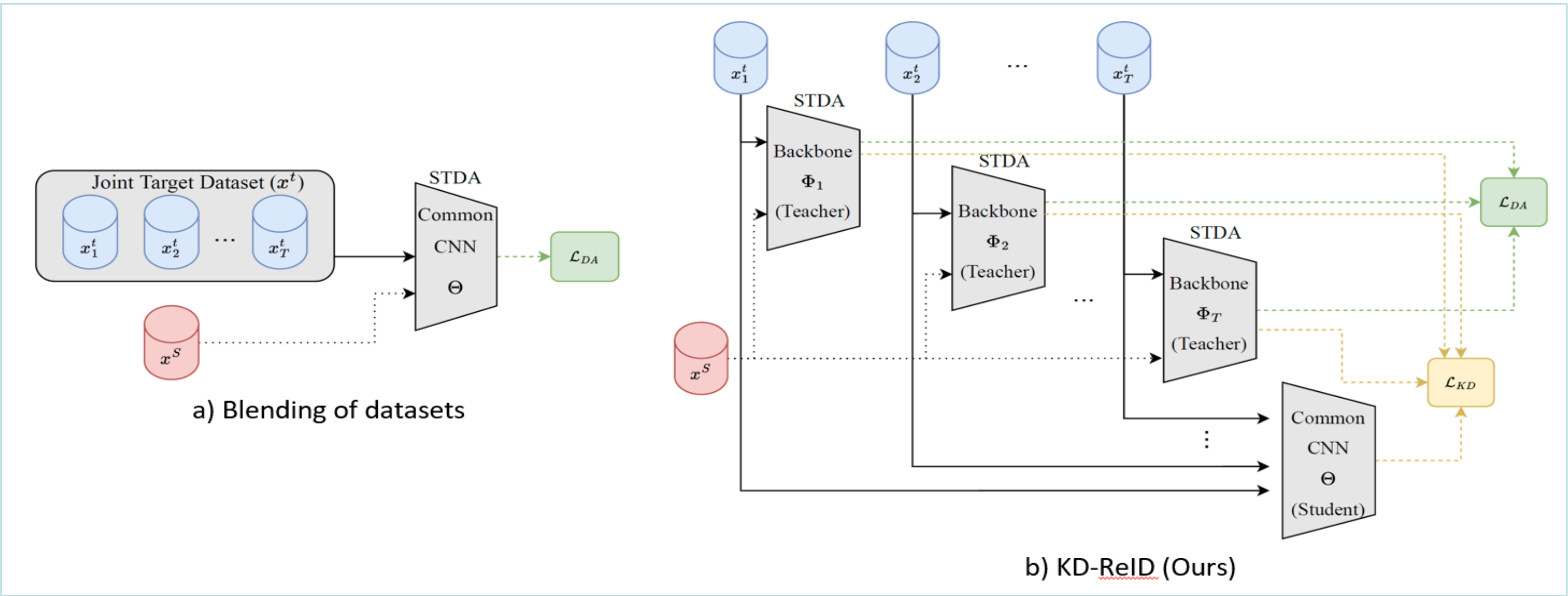


- Losses are jointly optimized along with L_2 norm on the weights of the linear classifier to train a linear soft-margin classifier.

- DisReID can improve ReID performance with compact DL backbones

Multi-Target Domain Adaptation

- **Objective:** MTDA method to train compact classification and ReID models through knowledge distillation

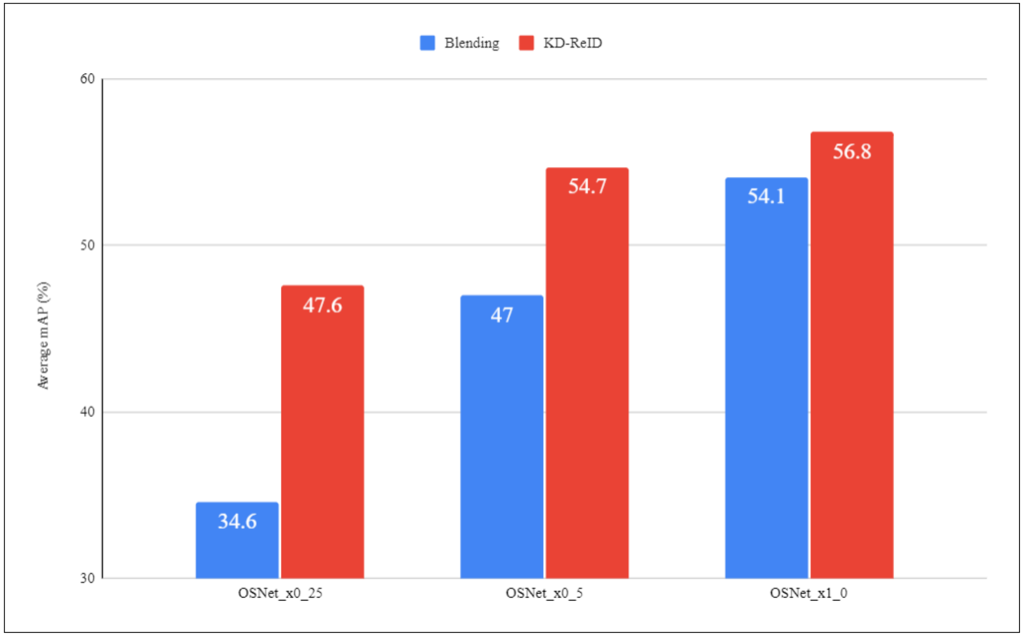


KD-ReID combines the knowledge from large specialized backbones (teachers), one per target domain, into a single small CNN (Student) using Knowledge Distillation

Multi-Target Domain Adaptation

Examples of results: performance of MTDA methods when MSMT17 is used as the source dataset

MTDA Method – Base STDA Method	Accuracy on Target Data (%)								Complexity	
	Market1501		DukeMTMC		CUHK03		Average			
	mAP	R1	mAP	R1	mAP	R1	mAP	R1	# Parameters	FLOPs
Lower Bound: Superv. on Source Only	27.7	54.6	30.1	49.5	27.8	32.0	28.5	45.3	12.2 M	1.19 G
One Model per Target – D-MMD (Teachers)	51.4	74.9	51.4	69.3	61.8	65.9	54.9	70.0	*T* x 27.7 M	2.70 G
Blending Targets – D-MMD	40.3	64.5	42.2	61.8	54.2	58.0	45.6	61.4	12.2 M	1.19 G
KD-ReID – D-MMD (Ours)	48.9	71.9	48.9	66.9	**58.0**	**61.7**	51.9	66.5	12.2 M	1.19 G
KD-ReID – Mixed D-MMD & SPCL (Ours)	**55.2**	**76.3**	50.5	68.8	53.5	57.8	**53.1**	**67.6**	12.2 M	1.19 G
Upper Bound: Superv. Fine-Tuning on Targets	65.7	86.1	60.5	77.2	65.9	68.5	64.0	77.3	12.2 M	1.19 G



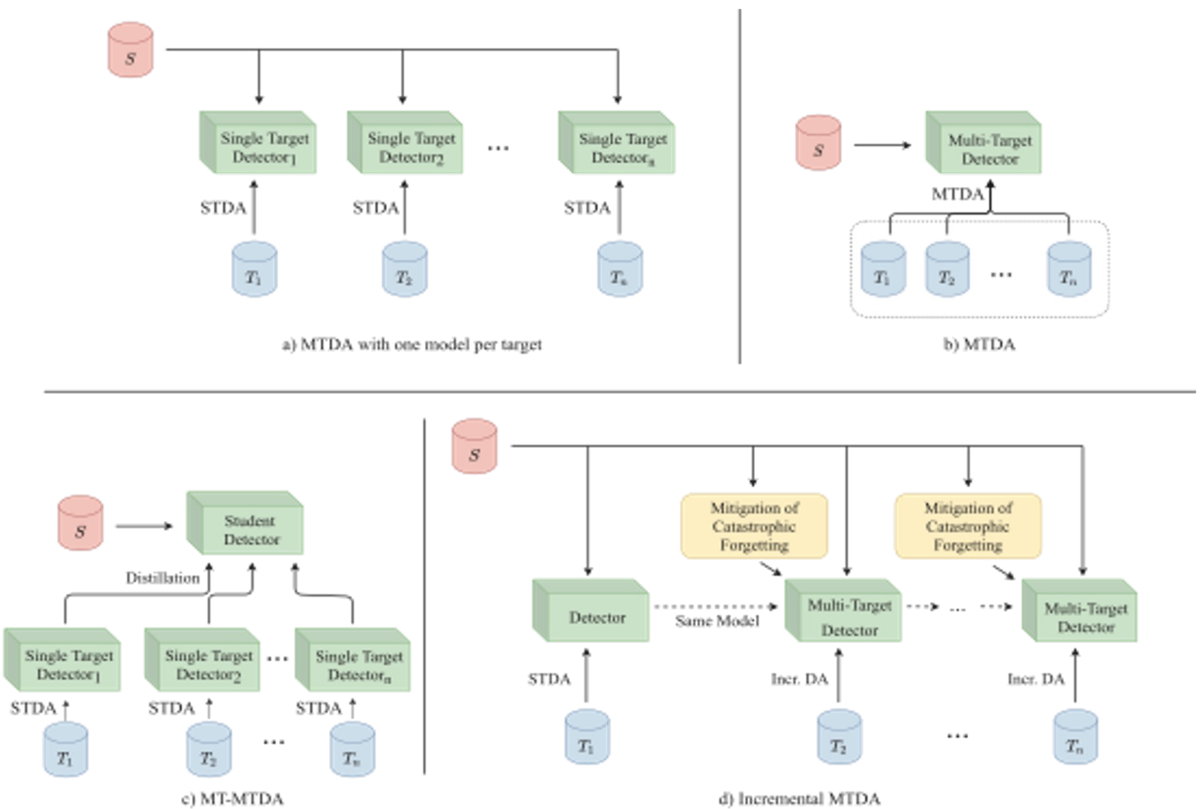
F Remigereau, et al., Knowledge Distillation for MTDA in Real-Time Person ReID, ICIP 2022.

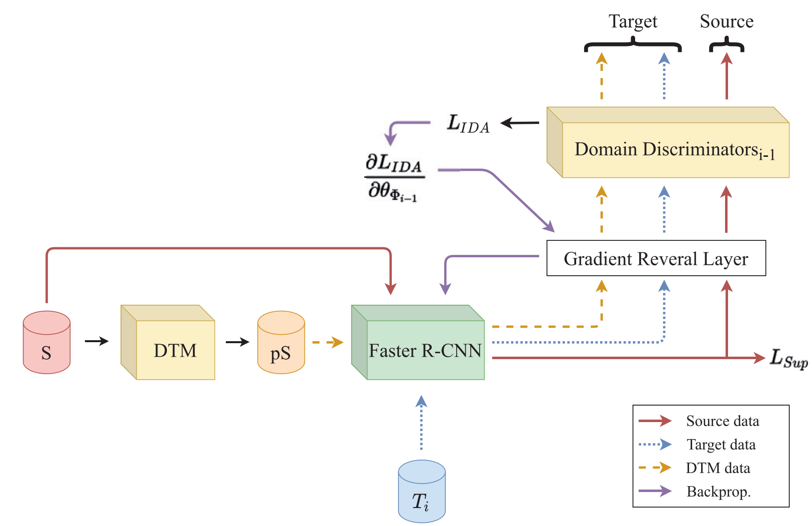
LT Nguyen-Meidine, et al., Unsupervised MTDA Through Knowledge Distillation, WACV 2021.

Le génie pour l'industrie

Multi-Target Domain Adaptation

- **An incremental MTDA method that allows to progressively train a compact object detection model**





A common detector is adapted incrementally one target at time, using a duplicated OD model for distillation to limit catastrophic forgetting.

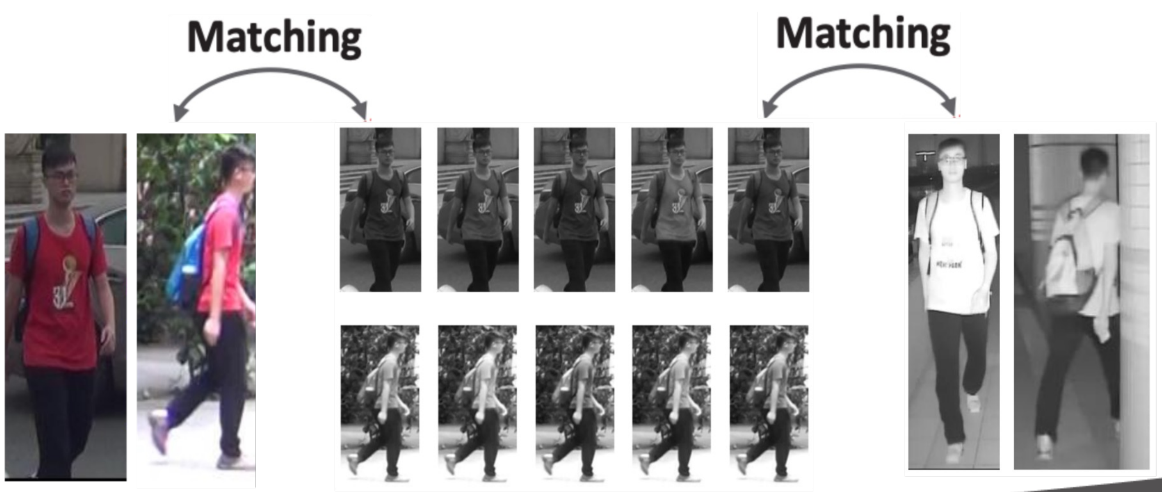
Visible-Infrared ReID Using Privileged Information

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

Challenge of V-I ReID: large shift between RGB and IR data distributions

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

- learning under privileged information (LUPI) paradigm
- generate privileged intermediate images, which connects the RGB and IR modalities during training



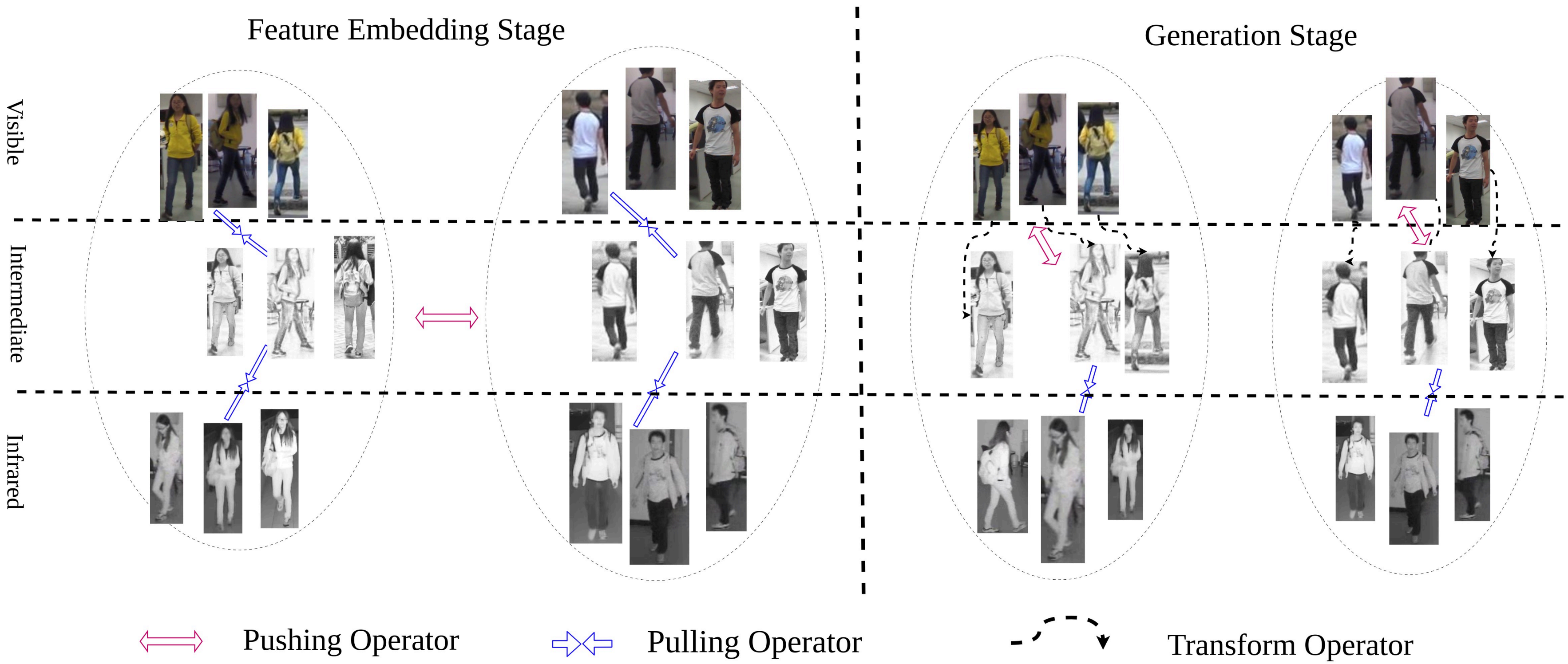
M Alehdaghi et al., Adaptive Generation of Privileged Intermediate Information for Visible-Infrared ReID, ECCVw 2022.

ÉTS Le génie pour l'industrie

Visible-Infrared ReID Using Privileged Information

Training strategy:

- (left) to generate the privileged images, the feature embedding stage pushes the extracted features towards the intermediate domain
- (right) meanwhile the generation stage transforms V images to an intermediate domain that approaches I images.

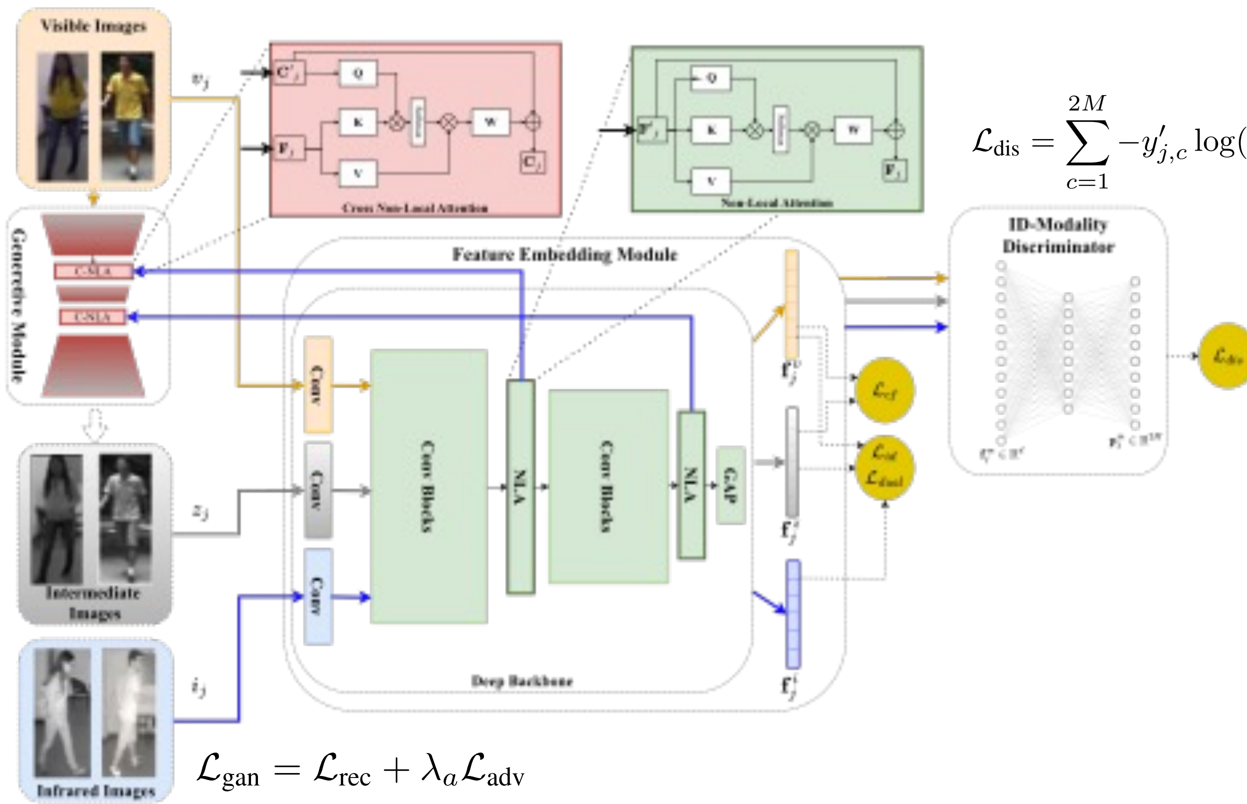


M Alehdaghi et al., Adaptive Generation of Privileged Intermediate Information for Visible-Infrared ReID, ECCVw 2022.

ÉTS Le génie pour l'industrie

Visible-Infrared ReID Using Privileged Information

Joint learning of generator, feature embedding, and ID-modality discrimination



$$\mathcal{L}_{\text{dis}} = \sum_{c=1}^{2M} -y'_{j,c} \log(p'_{j,c})$$

$$\mathcal{L}_{\text{gan}} = \mathcal{L}_{\text{rec}} + \lambda_a \mathcal{L}_{\text{adv}}$$

Feature Embedding Module

- color-free loss
- intermediate dual triplet loss
- cross-entropy

$$\mathcal{L}_{cf} = \|\mathbf{f}_j^v - \mathbf{f}_j^z\|_1$$

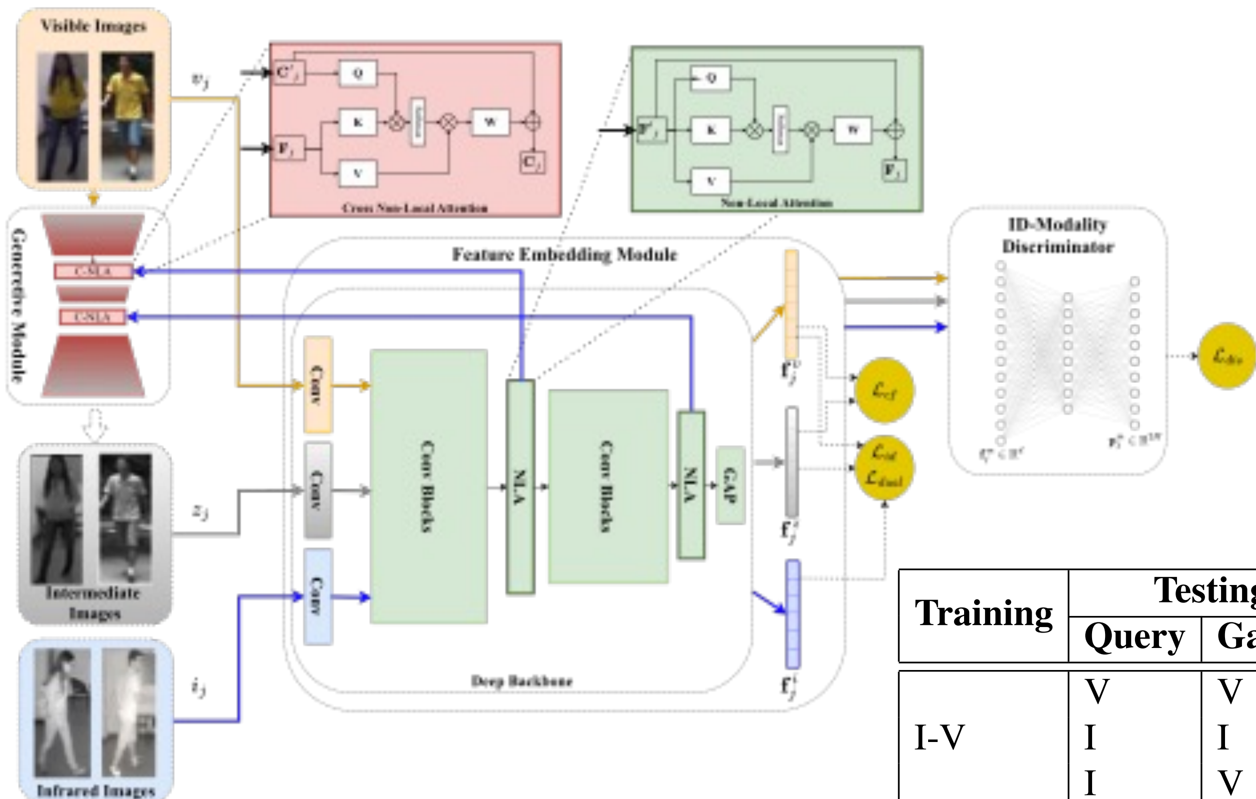
$$\mathcal{L}_{dual} = \mathcal{L}_{tri}(\mathbf{f}_{a\in\mathcal{V}}, \mathbf{f}_{p\in\mathcal{I}}, \mathbf{f}_{n\in\mathcal{Z}}) + \mathcal{L}_{tri}(\mathbf{f}_{a\in\mathcal{I}}, \mathbf{f}_{p\in\mathcal{Z}}, \mathbf{f}_{n\in\mathcal{V}})$$

M Alehdaghi et al., Adaptive Generation of Privileged Intermediate Information for Visible-Infrared ReID, ECCVw 2022.

ÉTS Le génie pour l'industrie

Visible-Infrared ReID Using Privileged Information

Joint learning of generator, feature embedding, and ID-modality discrimination



Training	Testing		R1 (%)	mAP(%)
	Query	Gallery		
I-V	V	V	97.40	91.82
	I	I	95.96	80.49
	I	V	58.69	41.57
I-V-Z	V	V	97.68	90.19
	I	I	95.34	81.62
	I	V	73.17	56.35

The impact on accuracy of the proposed intermediate module on the SYSU-MM01 dataset

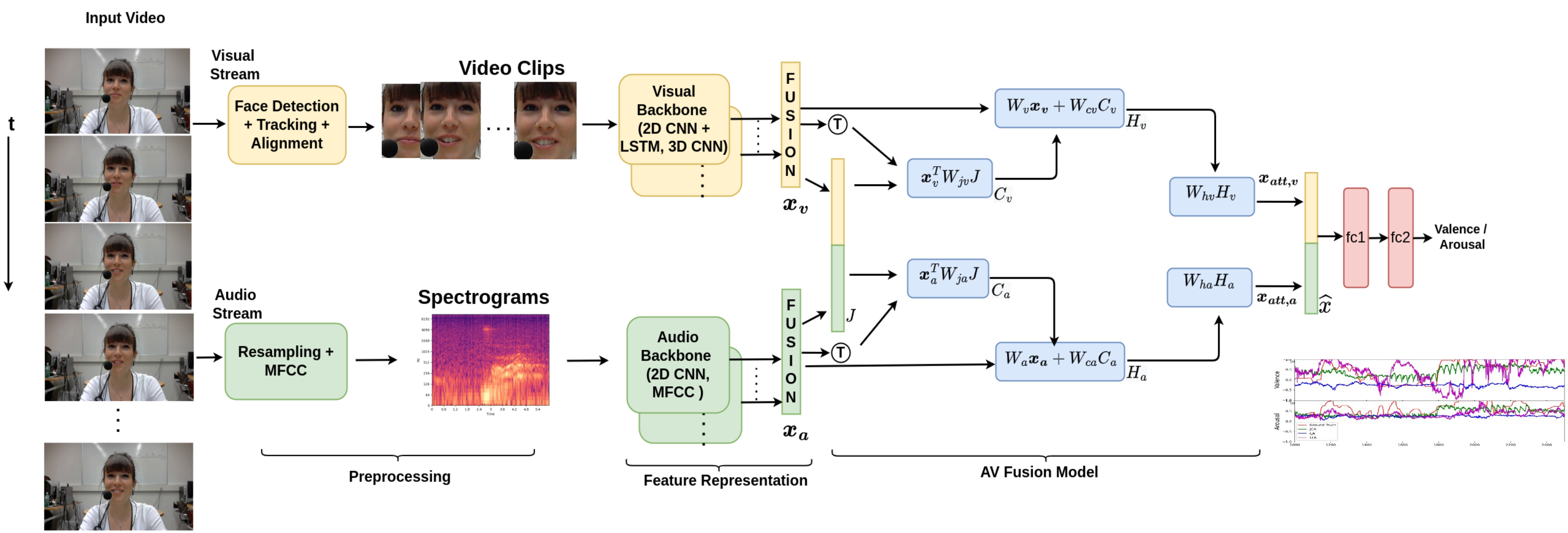
M Alehdaghi et al., Adaptive Generation of Privileged Intermediate Information for Visible-Infrared ReID, ECCVw 2022.

ÉTS Le génie pour l'industrie

Multimodal A-V Fusion of Faces and Voices

A Joint Cross-Attention Model for A-V Fusion in Dimensional Emotion Recognition

- Joint modeling of inter- and intra-modal relationships to capture the semantic relevance among A-V features



Joint cross attention maps

$$H_v = ReLU(W_v X_v + W_{cv} C_v^\top)$$

$$H_a = ReLU(W_a X_a + W_{ca} C_a^\top)$$

Attended features

$$X_{att,a} = W_{ha} H_a + X_a$$

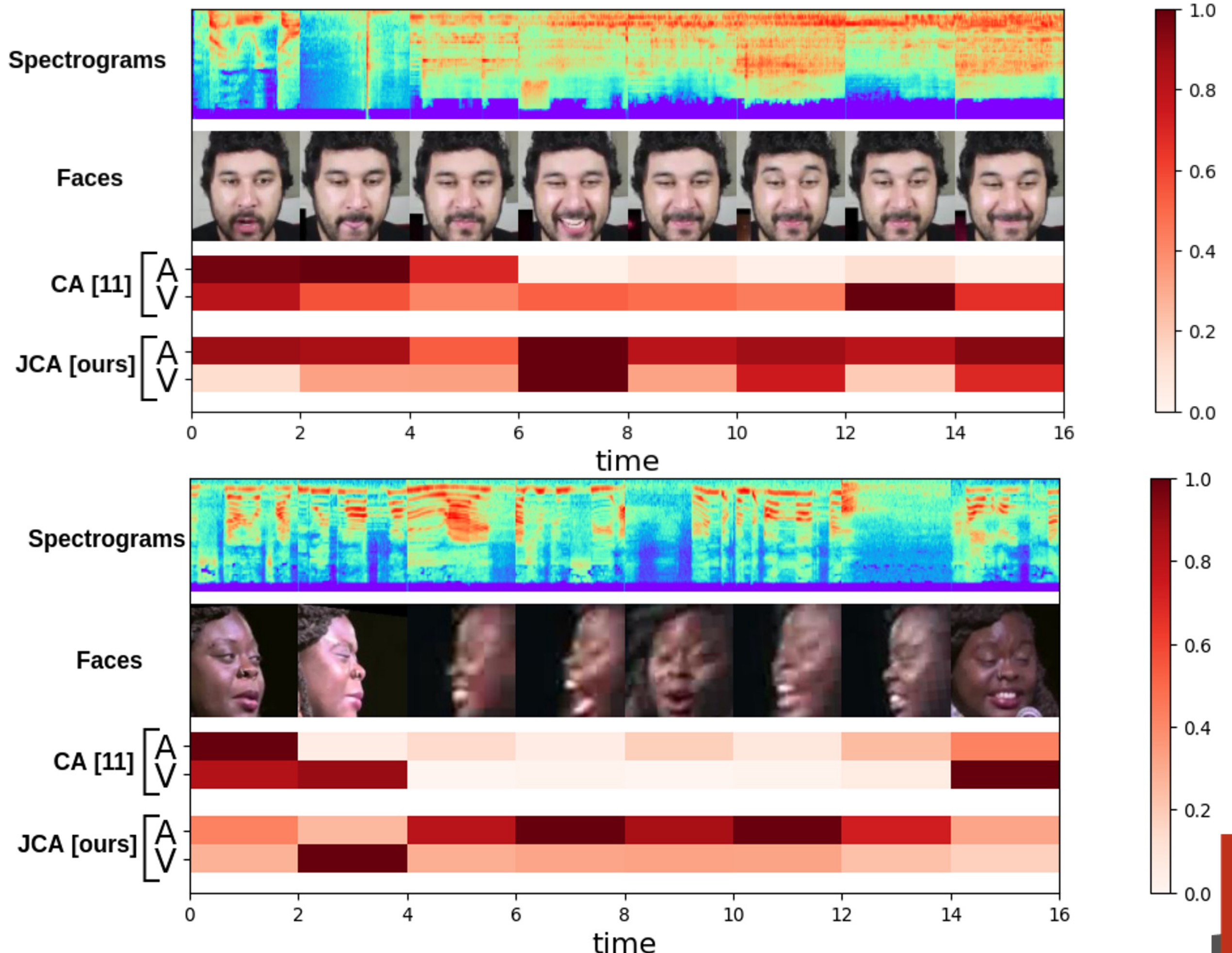
$$X_{att,v} = W_{hv} H_v + X_v$$

R G Praveen, et al., "Audio-Visual Fusion for Emotion Recognition in the Valence-Arousal Space Using Joint Cross-Attention." *IEEE Trans. on Biometrics, Behavior, and Identity Science* (2023).

Multimodal A-V Fusion of Faces and Voices

A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition:

- Visualization of attention scores of proposed A-V fusion (JCA) and CA models on video of Affwild2 dataset.



R G Praveen, et al., "Audio-Visual Fusion for Emotion Recognition in the Valence-Arousal Space Using Joint Cross-Attention." *IEEE Trans. on Biometrics, Behavior, and Identity Science* (2023).

ÉTS Le génie pour l'industrie

Multimodal A-V Fusion of Faces and Voices

A Joint Cross-Attention Model for Audio-Visual Fusion in Dimensional Emotion Recognition:

- Results: comparison to state-of-the-art on RECOLA and AffWild 2 data

Method – A/V backbone	Valence			Arousal		
	Audio	Visual	Fusion	Audio	Visual	Fusion
[He et al., 2015] – A: LLDs; V: LLDs	0.400	0.441	0.609	0.800	0.587	0.747
[Han et al., 2017] – A: LLDs + SM; V: geometric features + S.M.	0.480	0.592	0.554	0.760	0.350	0.685
[Tzirakis et al., 2017] – A: 1D-CNN; V: Resnet50	0.428	0.637	0.502	0.786	0.371	0.731
[Ortega et al., 2019] – A:LLDs; V: 2D-CNN	-	-	0.565	-	-	0.749
[Schoneval et al., 2021] – A: Finetuned VGGish; V: Distilled CNN	0.460	0.550	0.630	0.800	0.570	0.810
Cross Attention (Ours) – A: 2D-CNN; V: I3D	0.463	0.642	0.687	0.822	0.582	0.831
Joint Cross-Attention (Ours) – A: 2D-CNN; V: I3D	0.463	0.642	**0.728**	0.822	0.582	**0.842**

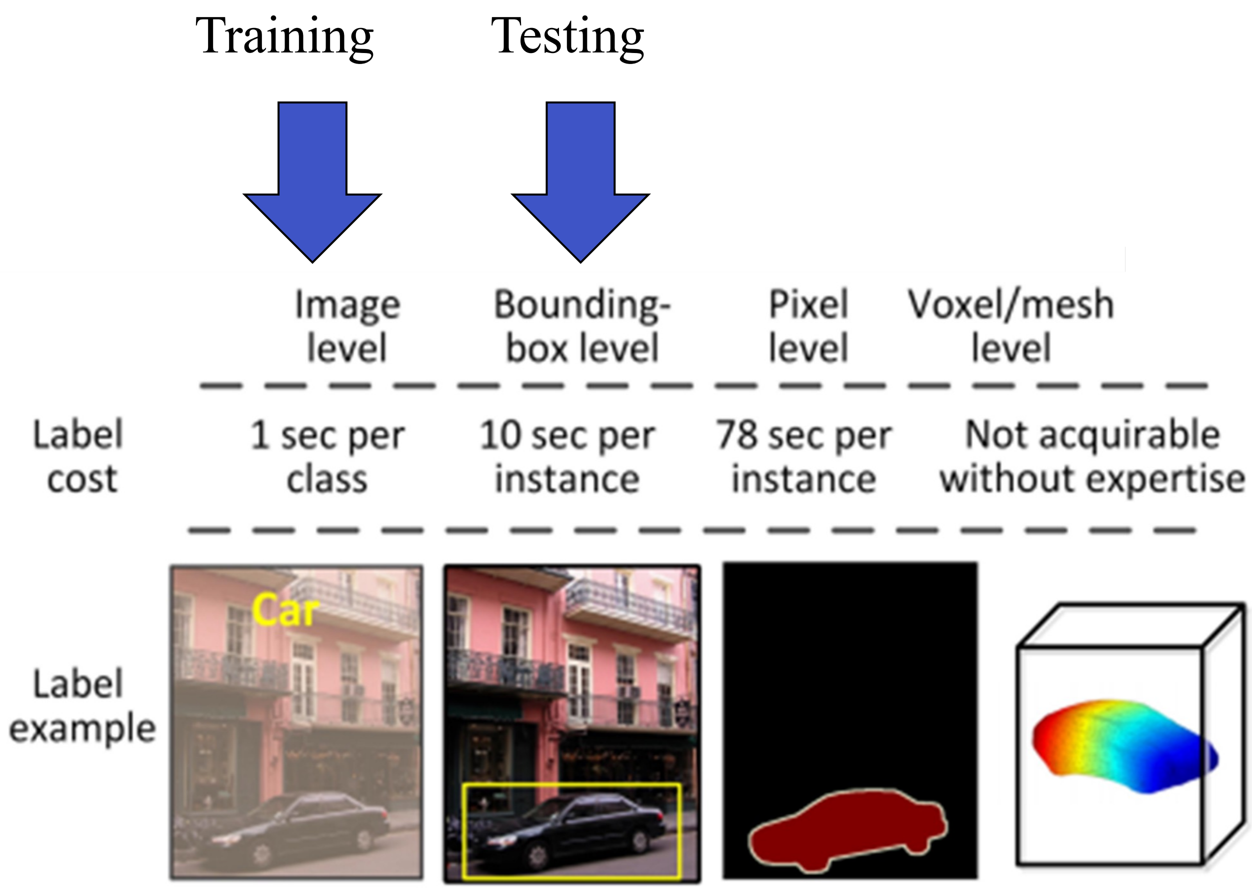
Results on RECOLA

Method – A/V backbone	Valence			Arousal		
	Audio	Visual	Fusion	Audio	Visual	Fusion
[Kuhnke et al., 2020] – A: Resnet18; V: R(2plus1)D	0.355	0.463	0.493	0.359	0.570	0.613
[Zhang et al., 2021] – A: VGGish; V: Resnet50 + TCN	-	0.425	0.469	-	0.647	**0.649**
Cross-Attention (Ours) – A: Resnet18; V: I3D	0.355	0.412	0.541	0.359	0.534	0.517
Joint Cross-Attention (Ours) – A: Resnet18; V: I3D	0.355	0.412	**0.657**	0.359	0.534	**0.580**

Results on Affwild2

R G Praveen, et al., "Audio-Visual Fusion for Emotion Recognition in the Valence-Arousal Space Using Joint Cross-Attention." *IEEE Trans. on Biometrics, Behavior, and Identity Science* (2023).

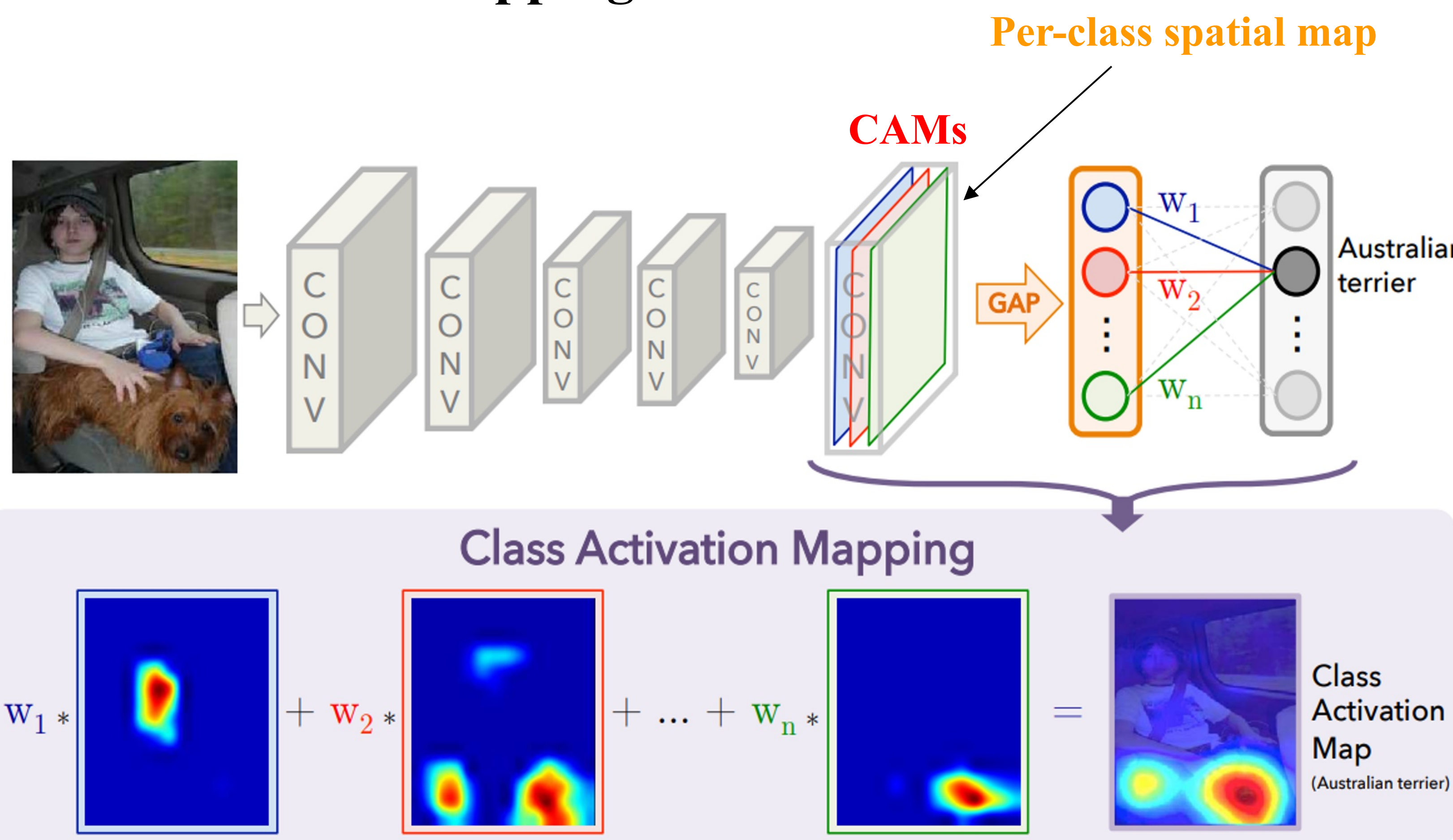
Weakly-Supervised Object Localization (WSOL)



Choe J. et al.. Evaluating Weakly Supervised Object Localization Methods Done Right, CVPR 2020.

Weakly-Supervised Localization

Class Activation Mapping Methods



Source: Zhou et al. Learning Deep Features for Discriminative Localization. CVPR 2016

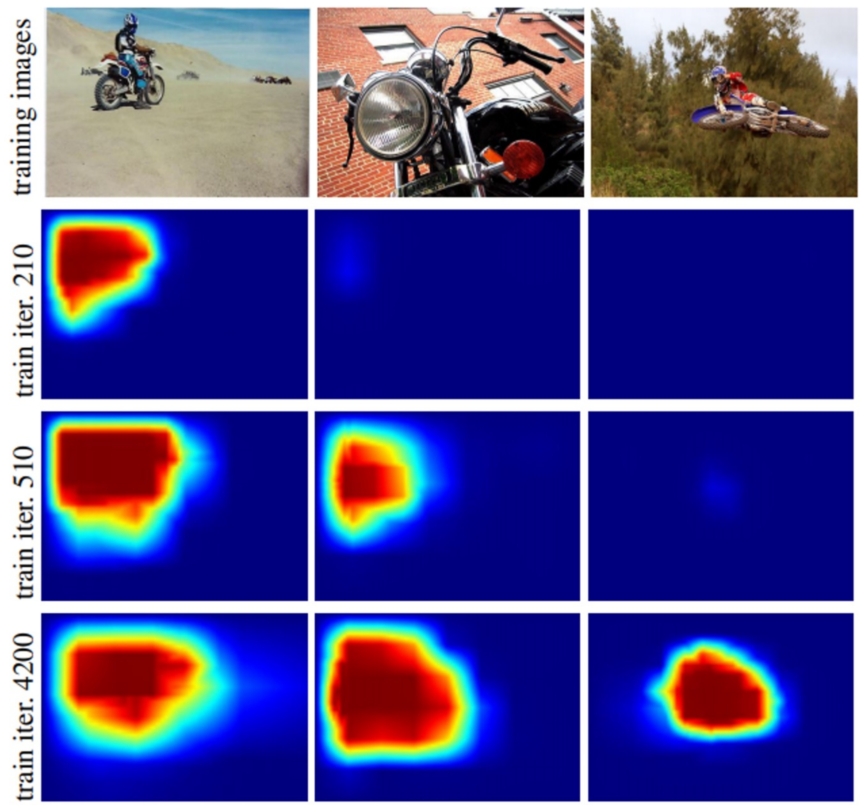
F-CAM for Improved Interpolation

- **A Challenge with CAMs:** low resolution (due to convolution and pooling) has negative impact on localization performance

Standard interpolated from CAM of 8x8 resolution (downscale factor of 32)

(a) Input (b) ResNet-18

Standard interpolated CAMs

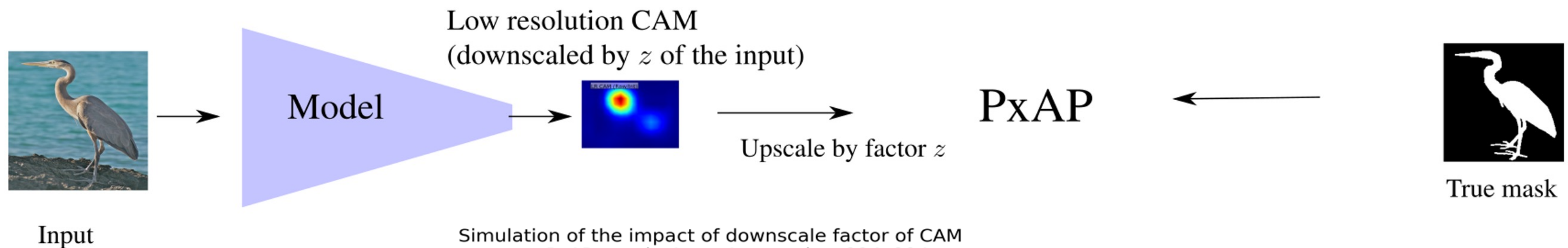


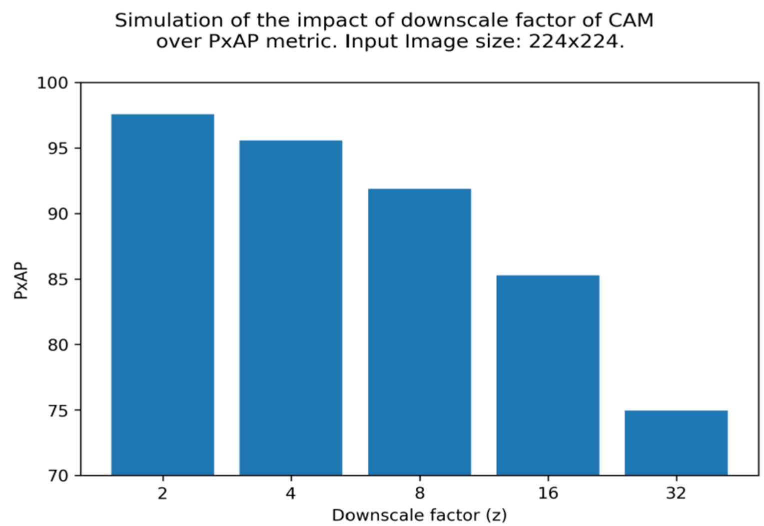
Source: F. Yu, V. Koltun, and T. Funkhouser, Dilated residual networks, CVPR 2017
Source: Oquab, M., et al., Is object localization for free?-weakly-supervised learning with CNNs. CVPR 2015

ÉTS Le génie pour l'industrie

F-CAM for Improved Interpolation

- **Challenges:** Impact of CAMs size on localization performance on CUB dataset



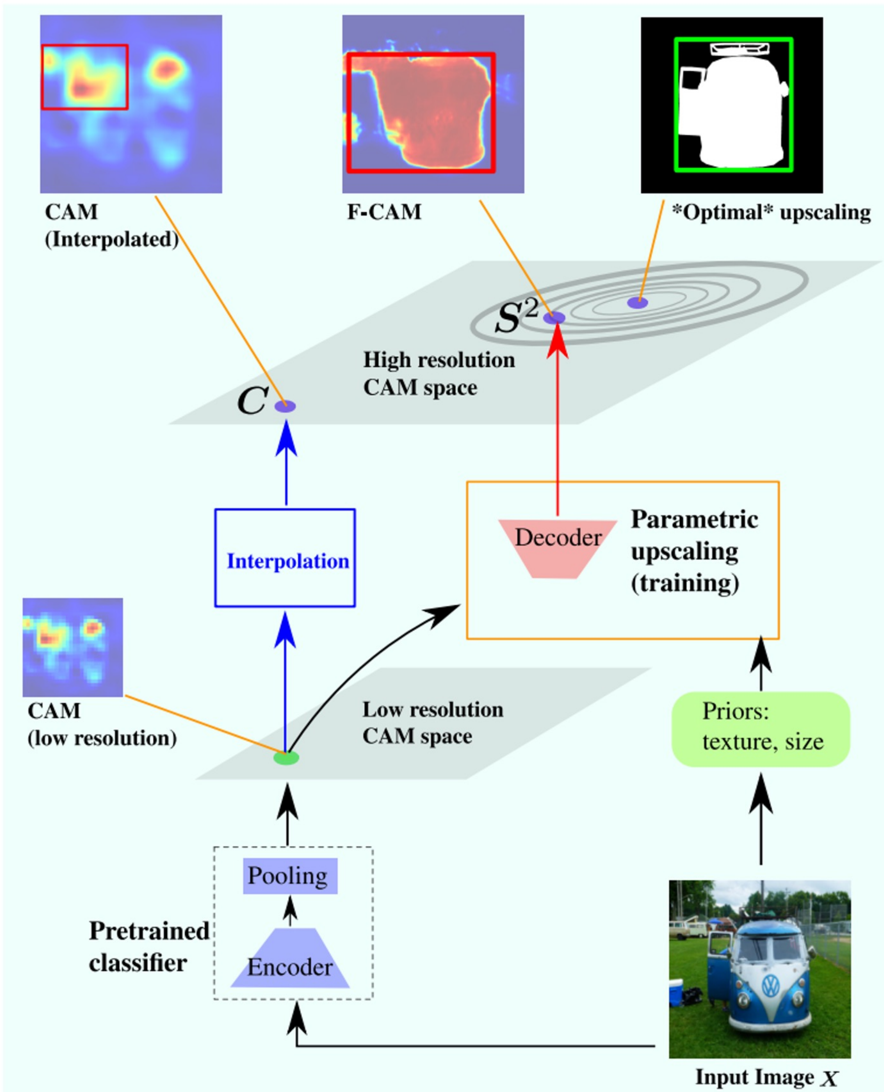


Results: increasing the downscaling factor (z) leads to a considerable decline in localization accuracy

Belharbi, S, et al., "F-CAM: Full resolution class activation maps via guided parametric upscaling." WAVC 2022.

F-CAM for Improved Interpolation

- **Proposed F-CAM with Guided Parametric Upscaling**



Encoder: any pre-trained CNN classifier,

L_c = classification loss (supervised)

Decoder: trained to perform parametric upscaling

L_D = pixel alignment loss (unsupervised)

= SR (CAM) + CRF (image) + ASC (size)

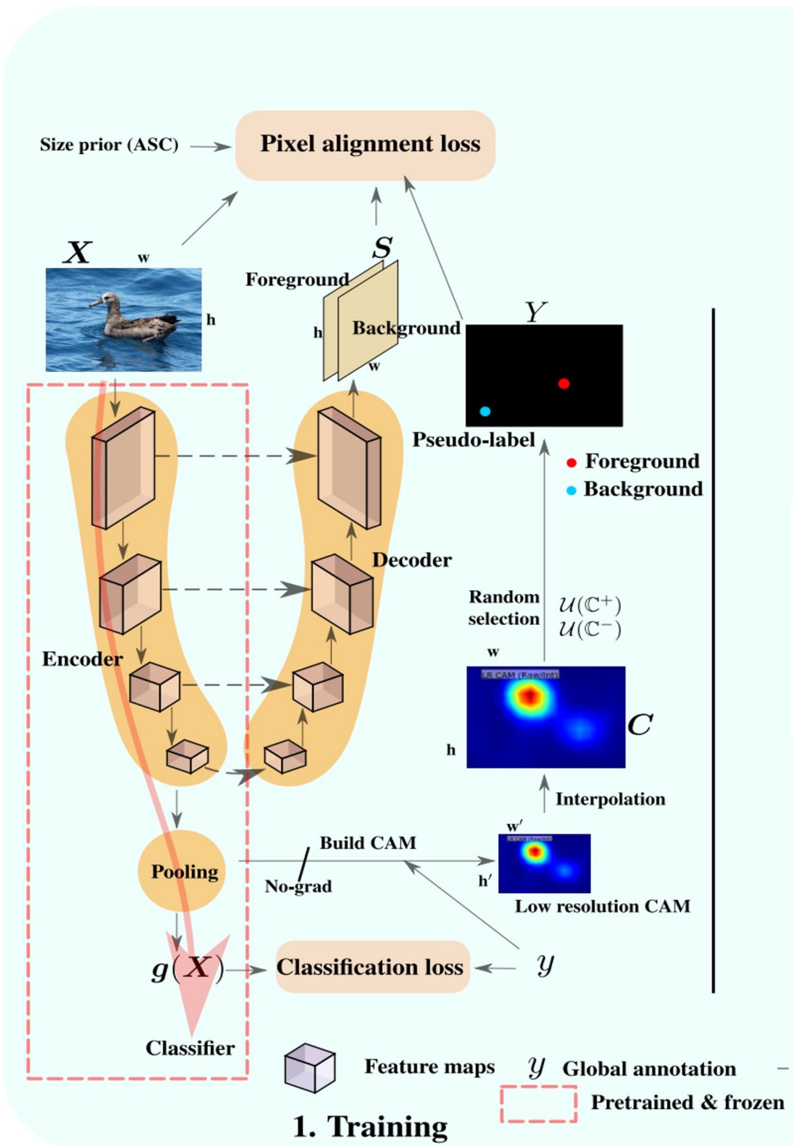
where

- SR: pseudo-labels (positive/negative evidence at pixel level)
- CRF: image properties
- ASC: unsupervised size constraint

Belharbi, S, et al., "F-CAM: Full resolution class activation maps via guided parametric upscaling." WAVC 2022.

F-CAM for Improved Interpolation

- **Proposed F-CAM:** training models the foreground and background



Overall loss for end-to-end training

$L_c = L_{CE}$

$L_D = L_{SR} + L_{CRF}$

$$\min_{\boldsymbol{\theta}} \quad -\log(\boldsymbol{g}(\boldsymbol{X})[y]) + \alpha \sum_{p \in \Omega'} \boldsymbol{H}(Y_p, \boldsymbol{S}_p) + \lambda\, \mathcal{R}(\boldsymbol{S}, \boldsymbol{X}) \,,$$

$$\text{s.t.} \quad \sum \boldsymbol{S}^r \geq 0 \,, \quad r \in \{1, 2\} \,,$$

ASC: area size constraint

Belharbi, S, et al., "F-CAM: Full resolution class activation maps via guided parametric upscaling." WAVC 2022.

F-CAM for Improved Interpolation

Experiments: Visual results on images from the CUB dataset

VGG16 | Inception v3 | ResNet50

CAM

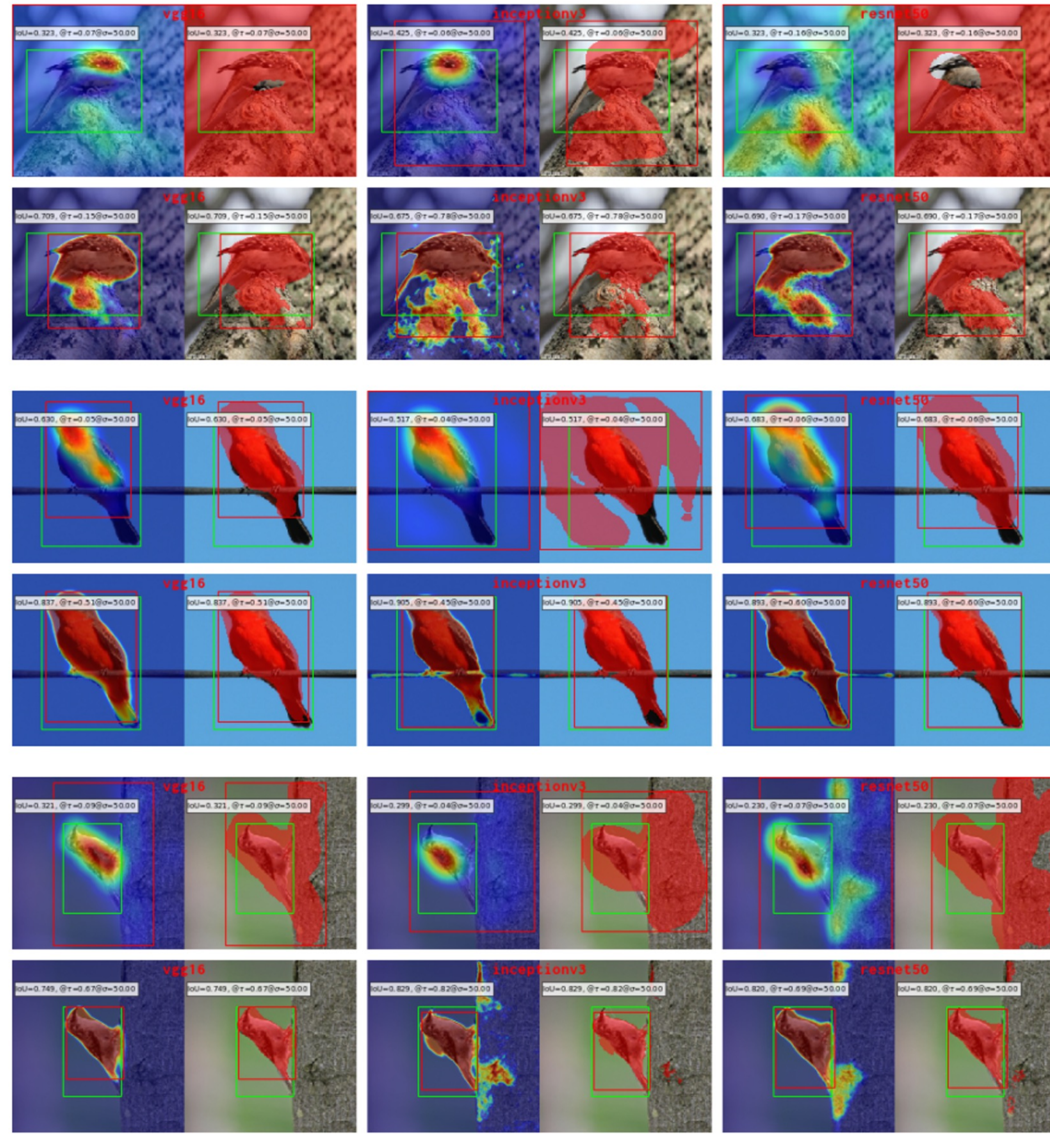
+ F-CAM

Grad CAM

+ F-CAM

Grad CAM++

+ F-CAM



Belharbi, S, et al., "F-CAM: Full resolution class activation maps via guided parametric upscaling." WAVC 2022.

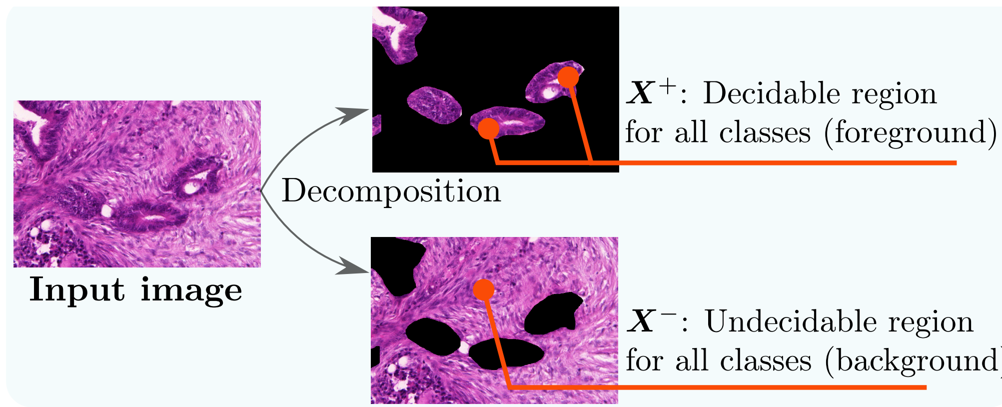
F-CAM:

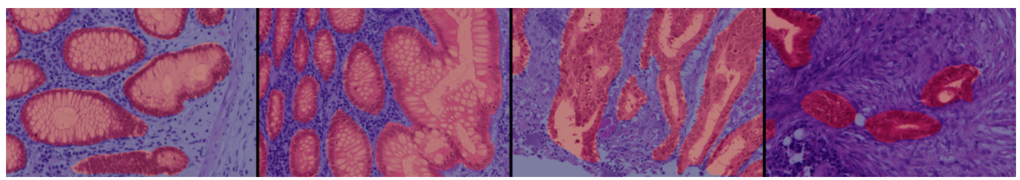
Some Results

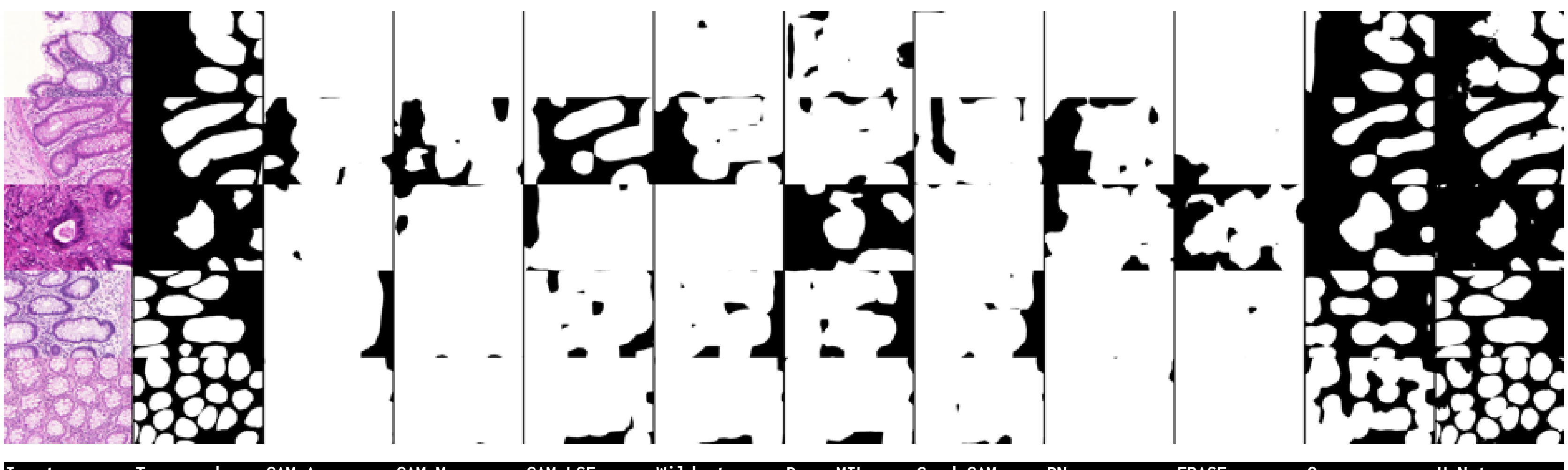
Methods	CUB (MaxBoxAcc) VGG	Inception	ResNet	Mean	OpenImages (PxAP) VGG	Inception	ResNet	Mean
CAM [57] *(cvpr,2016)*	71.1	62.1	73.2	68.8	58.1	61.4	58.0	59.1
HaS [34] *(iccv,2017)*	76.3	57.7	78.1	70.7	56.9	59.5	58.2	57.8
ACoL [53] *(cvpr,2018)*	72.3	59.6	72.7	68.2	54.7	63.0	57.8	58.4
SPG [54] *(eccv,2018)*	63.7	62.8	71.4	66.0	55.9	62.4	57.7	58.6
ADL [9] *(cvpr,2019)*	75.7	63.4	73.5	70.8	58.3	62.1	54.3	58.2
CutMix [51] *(eccv,2019)*	71.9	65.5	67.8	68.4	58.2	61.7	58.7	59.5
Best WSOL	76.3	65.5	78.1	70.8	58.3	63.0	58.7	59.5
FSL baseline	86.3	94.0	95.8	92.0	61.5	70.3	74.4	68.7
Center baseline	59.7	59.7	59.7	59.7	45.8	45.8	45.8	45.8
CSTN [22] *(icpr,2020)*	Resnet101 [14]: 76.0				–	–	–	–
TS-CAM [13] *(corr,2021)*	Deit-S [39]: 83.8				–	–	–	–
MEIL [21] *(cvpr,2020)*	73.8	–	–	–	–	–	–	–
DANet [47] *(iccv,2019)*	67.7	67.03	–	–	–	–	–	–
SPOL [44] *(cvpr,2021)*	–	–	96.4	–	–	–	–	–
CAM* [57] *(cvpr,2016)*	61.6	58.8	71.5	63.9	53.0	62.7	56.8	57.5
GradCAM [32] *(iccv,2017)*	69.3	62.3	73.1	68.2	59.6	63.9	60.1	61.2
GradCAM++ [7] *(wacv,2018)*	84.1	63.3	81.9	76.4	60.5	64.0	60.2	61.5
Smooth-GradCAM++ [25] *(corr,2019)*	69.7	66.9	76.3	70.9	52.2	61.7	54.3	56.0
XGradCAM [12] *(bmvc,2020)*	69.3	60.9	72.7	67.6	59.0	63.9	60.2	61.0
LayerCAM [15] *(ieee,2021)*	84.3	66.5	85.2	78.6	59.5	63.5	61.1	61.3
CAM* [57] + ours	87.3	82.0	90.3	86.5	67.8	71.9	72.1	70.6
GradCAM [32] + ours	87.5	84.4	90.5	87.4	68.6	70.0	70.9	69.8
GradCAM++ [57] + ours	91.5	84.6	91.0	89.0	64.8	67.1	66.3	66.0
Smooth-GradCAM++ [57] + ours	89.1	86.8	90.7	88.8	60.3	65.4	64.4	63.3
XGradCAM [57] + ours	86.8	84.4	90.4	88.8	68.7	71.3	70.4	70.1
LayerCAM [57] + ours	91.0	85.3	92.4	89.7	64.3	64.9	65.3	64.8
Best WSOL + ours	91.5	86.8	92.4	89.7	68.7	71.9	72.1	70.6

Table 1: Performance on MaxBoxAcc and PxAP metrics.

NEGEV: Extension of F-CAM to histology image analysis







S Belharbi et al., , "Deep interpretable classification and weakly-supervised segmentation of histology images via max-min uncertainty." *IEEE Trans on Medical Imaging* (2022)

Le génie pour l'industrie

Weakly-Supervised Video Object Localization

Video object localization allows to:

- locate object of interest in video
- understand video content
- improve subsequent tasks: video summarization, event detection, object detection, tracking, etc.

Localization in Unconstrained videos is challenging:

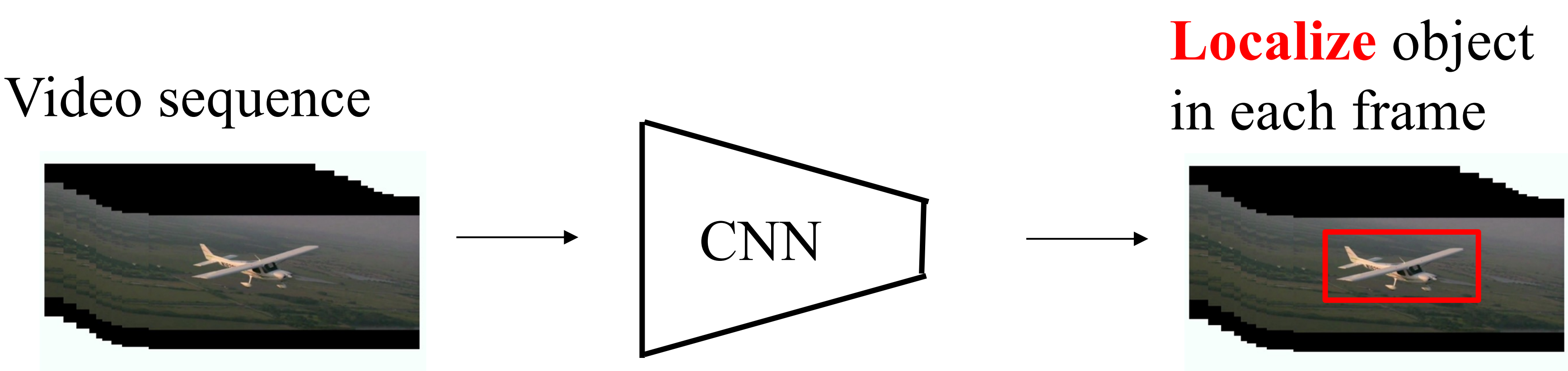
- moving and occluded objects
- camera motion and viewpoint changes
- decoding artifacts and editing effects

Belharbi, S, et al., TCAM: Temporal Class Activation Maps for Object Localization in Weakly-Labeled Unconstrained Videos. WACV 2023.

Weakly-Supervised Video Object Localization

Levels of supervision:

- annotating all the frames using bounding boxes (bbox) is an expensive process
- training a model with weak video labels, like video tags are less expensive
- *global video tag* = main object class in the video, not necessarily present in all the frames



Belharbi, S, et al., TCAM: Temporal Class Activation Maps for Object Localization in Weakly-Labeled Unconstrained Videos. WACV 2023.

Weakly-Supervised Video Object Localization

Challenges for State-of-Art Methods:

- Multiple sequential and independent stages
- Video tags (labels) are only used to cluster video
- ROI are not necessarily discriminative
- Motion cues (optical flow) are noisy, not always discriminative, and need post-processing
- Requires solving an optimization problem at inference time: slow inference: build a model per class/video

Weakly-Supervised Video Object Localization

Adapt CAMs to exploit the spatio-temporal dependency in videos

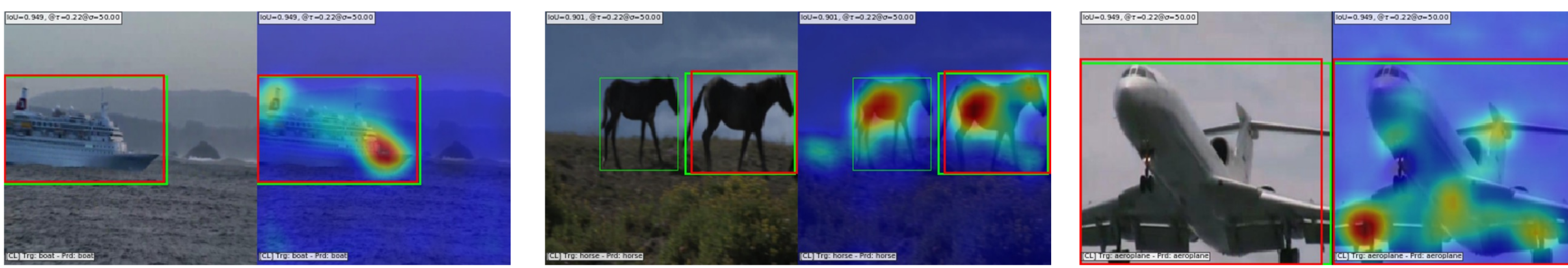
Advantages compared to SOTA of WSVOL (videos):

- single, discriminative model for all classes
- fast inference (single forward pass)

Advantage compared to CAMs for WSOL (still images):

- allows to leverage temporal information in videos

CAM results on still images



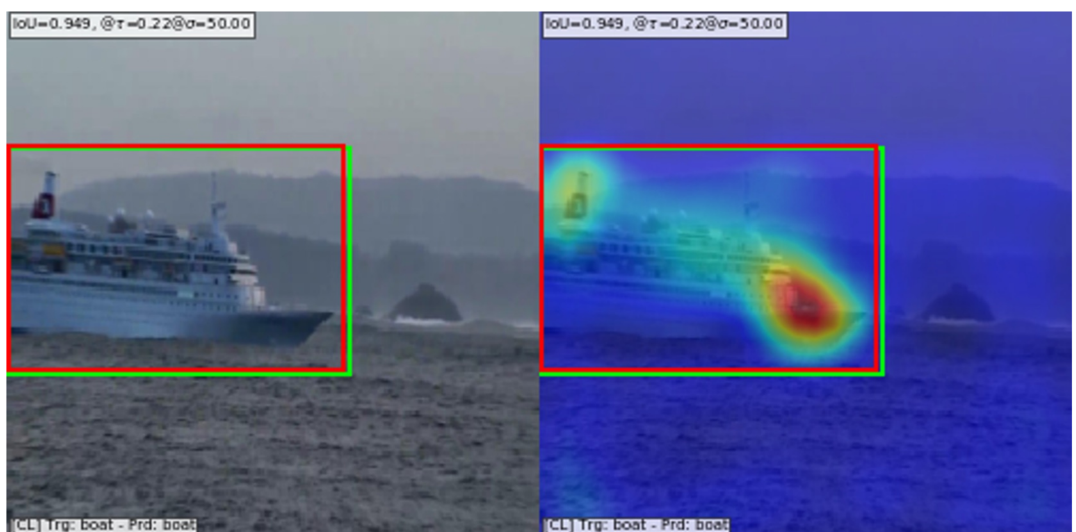
Belharbi, S, et al., TCAM: Temporal Class Activation Maps for Object Localization in Weakly-Labeled Unconstrained Videos. WACV 2023.

Le génie pour l'industrie

Temporal CAM (TCAM) Method

Adapt CAM methods to exploit the spatiotemporal dependency in videos

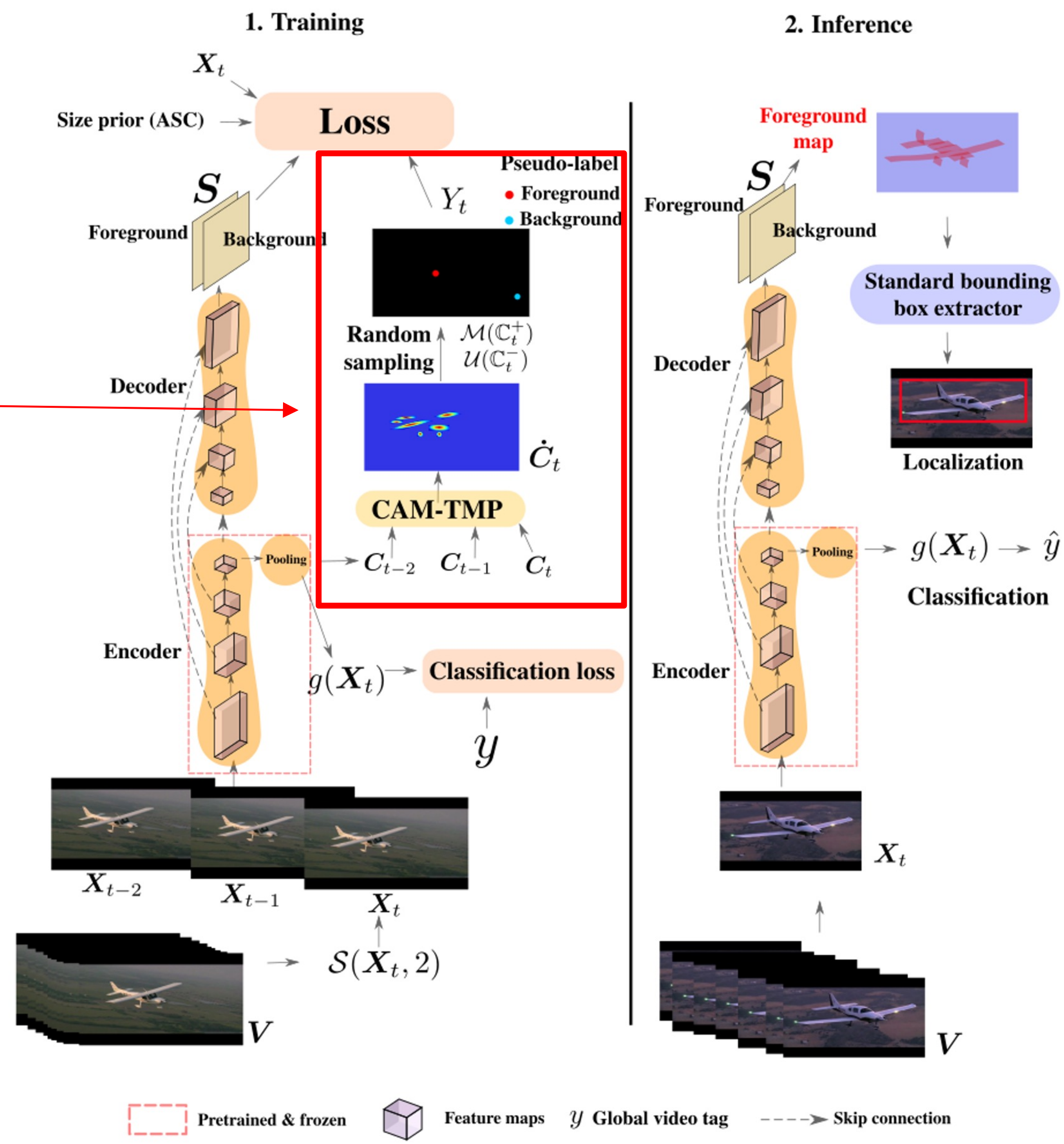
- leverage the slight variations in sets of consecutive frames
- aggregate diversified CAMs from *n* frames
- include a learnable decoder to produce accurate F-CAMs
- use aggregated CAMs to sample pixel pseudo-labels for training the decoder



Belharbi, S, et al., TCAM: Temporal Class Activation Maps for Object Localization in Weakly-Labeled Unconstrained Videos. WACV 2023.

Temporal CAM (TCAM) Method

Training: accounts for spatio-temporal dependency at a CAM level – it leverages sequences of n frames



$\boldsymbol{V}$: video

y : video tag (class)

$\boldsymbol{X}_t$: frame at time t

$\boldsymbol{C}_t$: CAM of fran$\boldsymbol{X}_t$

$\mathcal{S}(\boldsymbol{X}_t, 2)$: sampling function

$\dot{\boldsymbol{C}}_t$: aggregated CAM

Y_t : pixel pseudo-label mask

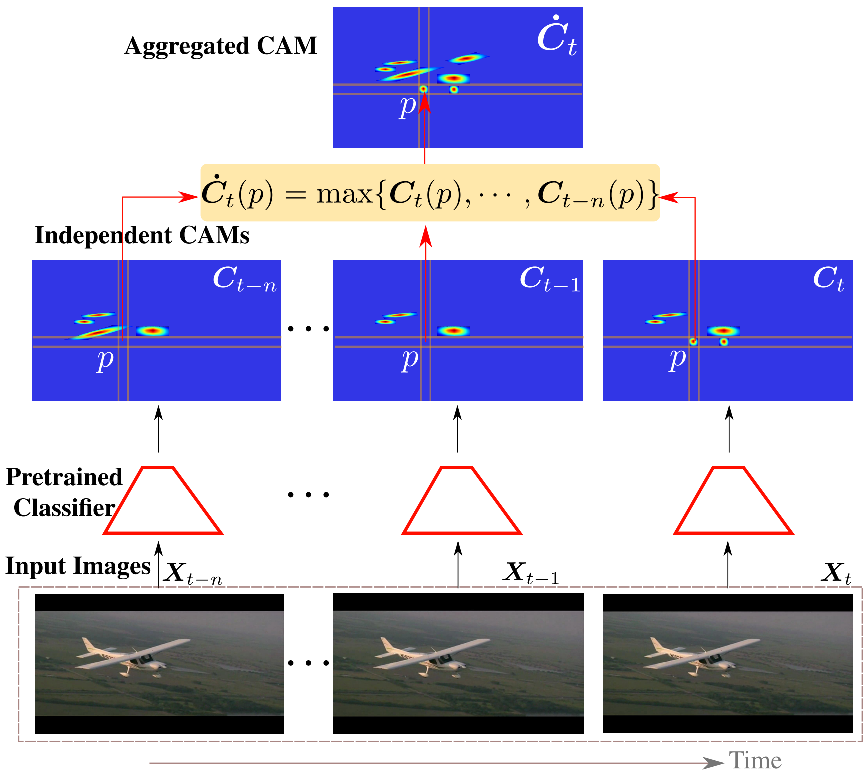
$\boldsymbol{S}$: output CAM

Belharbi, S, et al., TCAM: Temporal Class Activation Maps for Object Localization in Weakly-Labeled Unconstrained Videos. WACV 2023.

TCAM: Temporal Class Activation Maps

- **Object Localization in Weakly-Labeled Unconstrained Videos**

CAM-TMP
CAM Temporal Max-Pooling



Adapt CAM methods to exploit the spatiotemporal dependency in videos

- Leverage the slight variations in sets of consecutive frames
- The CAM-TMP module aggregates diversified CAMs from *n* frames
- It relies on the maximum activation at location *p* across the independent CAMs

Belharbi, S, et al., TCAM: Temporal Class Activation Maps for Object Localization in Weakly-Labeled Unconstrained Videos. WACV 2023.

Temporal CAM (TCAM) Method

Training: accounts for spatio-temporal dependency at a CAM level

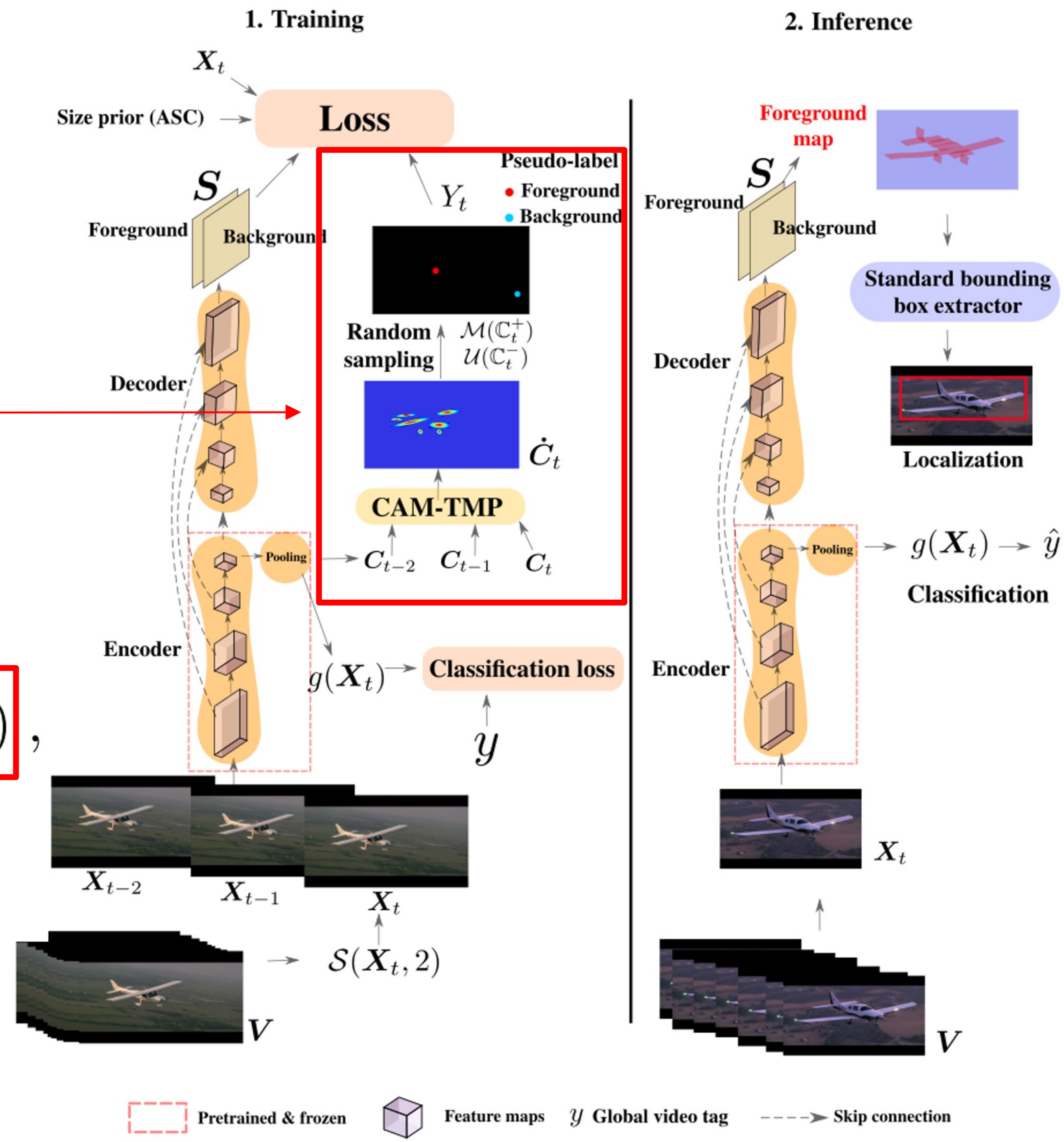
pixel pseudo-labels

CRF

$$\min_{\boldsymbol{\theta}} \quad \boxed{\sum_{p \in \Omega'_t} \boldsymbol{H}_p(Y_t, \boldsymbol{S}_t)} + \lambda \boxed{\mathcal{R}(\boldsymbol{S}_t, \boldsymbol{X}_t)},$$

$$\text{s.t.} \quad \boxed{\sum \boldsymbol{S}^r_t \geq 0}, \quad r \in \{0, 1\},$$

large size (FG/BG)



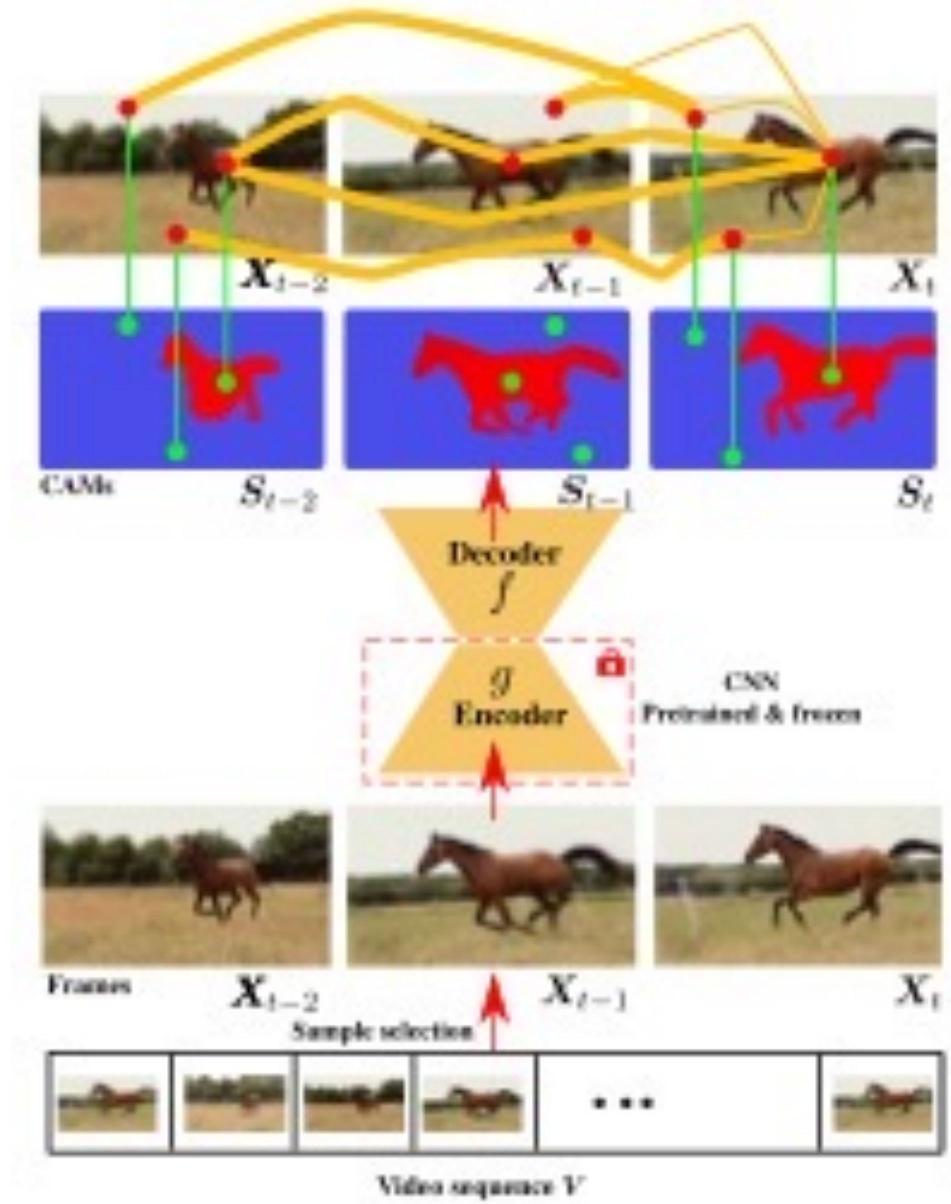
Belharbi, S, et al., TCAM: Temporal Class Activation Maps for Object Localization in Weakly-Labeled Unconstrained Videos. WACV 2023.

CoLo-CAM Method for Object Co-Localization

Multi-frame training using CoLo-CAM method with *n* = 3 frames.

- Each pixel (dot), is connected (orange line) to every pixel across frames to measure color similarity (connection thickness indicates similarity strength).

- CAM locations at pixels with similar colors are constrained to have similar activations (green lines are for alignment).



S Belharbi, et al., CoLo-CAM: Class Activation Mapping for Object Co-Localization in Weakly-Labeled Unconstrained Videos. *arXiv:2303.09044*.

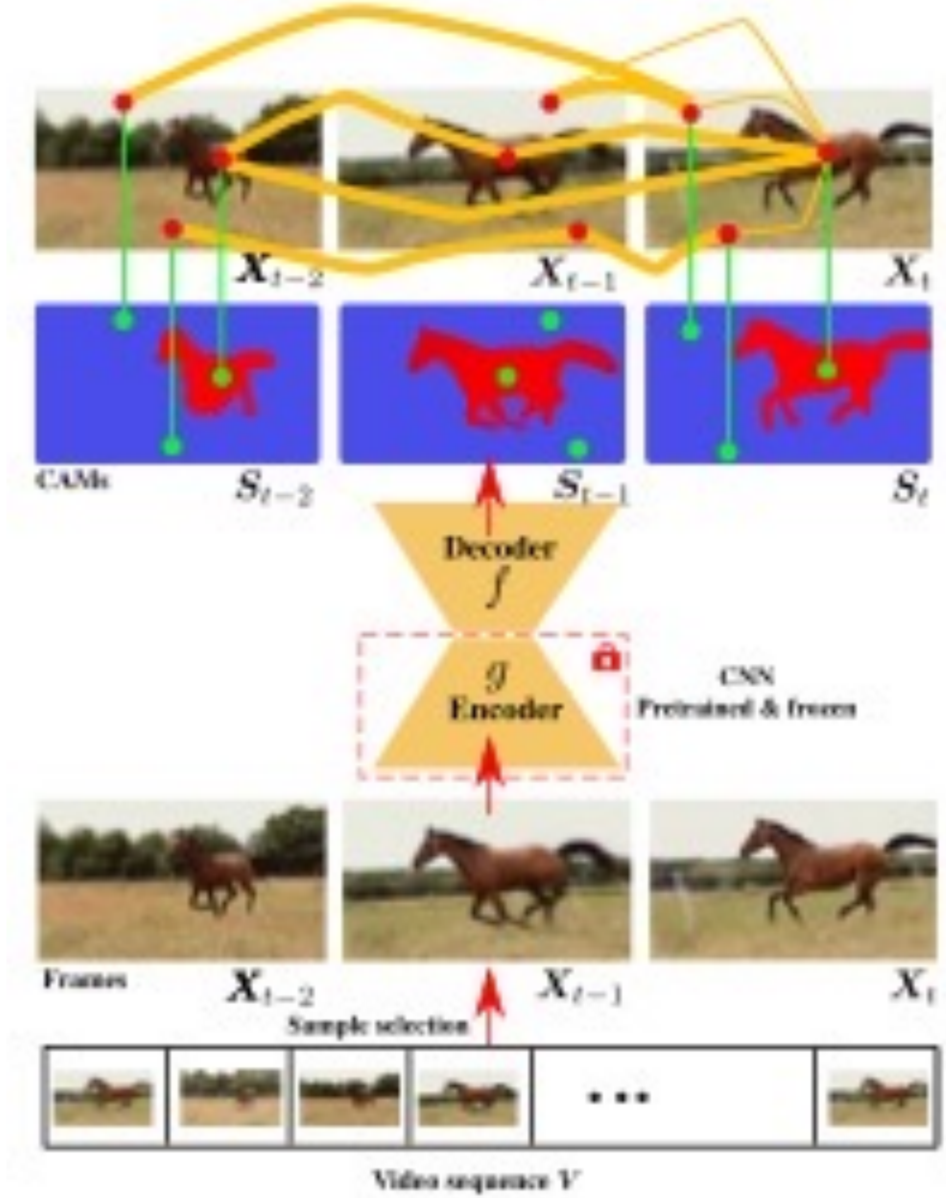
CAM Method for Object Co-Localization

Multi-frame training using CoLo-CAM method with *n* = 3 frames.

- CoLo-CAM can leverage spatiotemporal information in activation maps without any assumptions about object movement.
- Given a sequence of frames, explicit joint learning of localization is produced across maps based on color cues, by assuming an object has similar color across frames.

$$\min_{\boldsymbol{\theta}} \quad \sum_{p \in \Omega'_t} \boldsymbol{H}_p(\boldsymbol{Y}_t, \boldsymbol{S}_t) + \lambda\, \mathcal{R}(\boldsymbol{S}_t, \boldsymbol{X}_t) + \mathcal{R}_s(\boldsymbol{S}_t) + \frac{\lambda_c}{|[\mathcal{R}_c]|}\, \mathcal{R}_c(\{\boldsymbol{S}\}_t^n, \{\boldsymbol{X}\}_t^n) \,,$$

multi-frame loss



S Belharbi, et al., CoLo-CAM: Class Activation Mapping for Object Co-Localization in Weakly-Labeled Unconstrained Videos. *arXiv:2303.09044*.

CAM Method for Object Co-Localization

Experimental results: CorLoc on the YouTube-Object v1.0 dataset

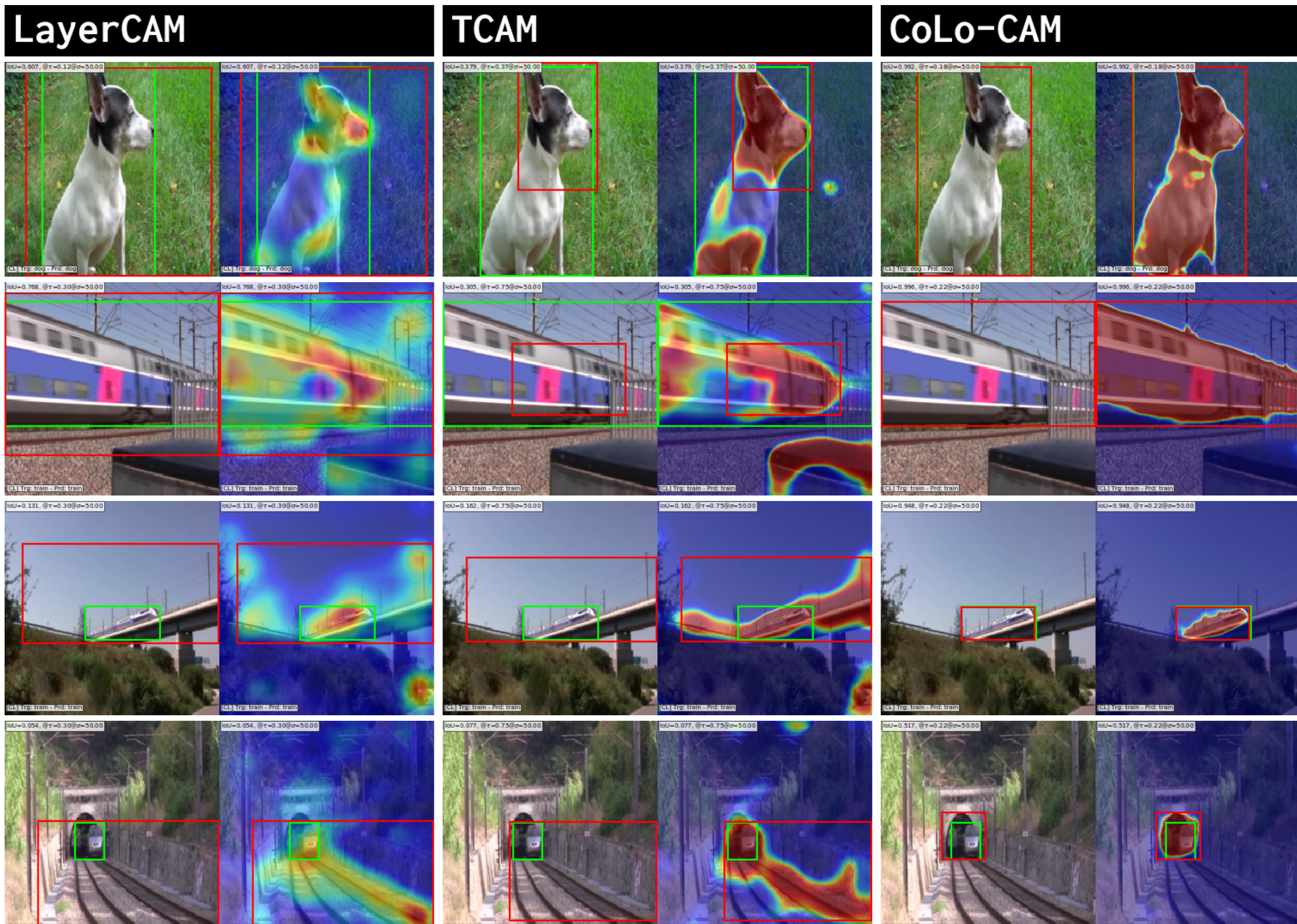
Method *(venue)*	Aero	Bird	Boat	Car	Cat	Cow	Dog	Horse	Mbike	Train	Avg	Time/Frame
[65] *(cvpr)*	51.7	17.5	34.4	34.7	22.3	17.9	13.5	26.7	41.2	25.0	28.5	N/A
[62] *(iccv)*	65.4	67.3	38.9	65.2	46.3	40.2	65.3	48.4	39.0	25.0	50.1	4s
[39] *(eccv)*	25.1	31.2	27.8	38.5	41.2	28.4	33.9	35.6	23.1	25.0	31.0	N/A
[46] *(iccv)*	56.5	66.4	58.0	76.8	39.9	69.3	50.4	56.3	53.0	31.0	55.7	N/A
[66] *(ivc)*	60.8	54.6	34.7	57.4	19.2	42.1	35.8	30.4	11.7	11.4	35.8	N/A
[79] *(eccv)*	71.5	74.0	44.8	72.3	52.0	46.4	71.9	54.6	45.9	32.1	56.6	N/A
POD [43] *(cvpr)*	64.3	63.2	73.3	68.9	44.4	62.5	71.4	52.3	78.6	23.1	60.2	N/A
[80] *(eccv)*	66.1	59.8	63.1	72.5	54.0	64.9	66.2	50.6	39.3	42.5	57.9	N/A
[32] *(iccv)*	76.3	71.4	65.0	58.9	68.0	55.9	70.6	33.3	69.7	42.4	61.1	0.35s
[19] (LowRes-Net$_{iter1}$) *(ijcv)*	77.0	67.5	77.2	68.4	54.5	68.3	72.0	56.7	44.1	34.9	62.1	0.02s
[19] (LowRes-Net$_{iter2}$) *(ijcv)*	79.7	67.5	68.3	69.6	59.4	75.0	78.7	48.3	48.5	39.5	63.5	0.02s
[19] (DilateU-Net$_{iter2}$) *(ijcv)*	85.1	72.7	76.2	68.4	59.4	76.7	77.3	46.7	48.5	46.5	65.8	0.02s
[19] (MultiSelect-Net$_{iter2}$) *(ijcv)*	84.7	72.7	78.2	69.6	60.4	80.0	78.7	51.7	50.0	46.5	67.3	0.15s
SPFTN (M) [93] *(tpami)*	66.4	73.8	63.3	83.4	54.5	58.9	61.3	45.4	55.5	30.1	59.3	N/A
SPFTN (P) [93] *(tpami)*	**97.3**	27.8	81.1	65.1	56.6	72.5	59.5	**81.8**	79.4	22.1	64.3	N/A
FPPVOS [81] *(optik)*	77.0	72.3	64.7	67.4	79.2	58.3	74.7	45.2	80.4	42.6	65.8	0.29s
CAM [101] *(cvpr)*	75.0	55.5	43.2	69.7	33.3	52.4	32.4	74.2	14.8	50.0	50.1	0.2ms
GradCAM [70] *(iccv)*	86.9	63.0	51.3	81.8	45.4	62.0	37.8	67.7	18.5	50.0	56.4	27.8ms
GradCAM++ [14] *(wacv)*	79.8	85.1	37.8	81.8	75.7	52.4	64.9	64.5	33.3	56.2	63.2	28.0ms
Smooth-GradCAM++ [60] *(corr)*	78.6	59.2	56.7	60.6	42.4	61.9	56.7	64.5	40.7	50.0	57.1	136.2ms
XGradCAM [28] *(bmvc)*	79.8	70.4	54.0	**87.8**	33.3	52.4	37.8	64.5	37.0	50.0	56.7	14.2ms
LayerCAM [38] *(ieee)*	85.7	**88.9**	45.9	78.8	75.5	61.9	64.9	64.5	33.3	56.2	65.6	17.9ms
TCAM [5] *(wacv)*	90.5	70.4	62.2	75.7	**84.8**	**81.0**	81.0	64.5	70.4	50.0	73.0	18.5ms
CoLo-CAM (ours)	90.4	74.0	**91.8**	**87.8**	78.7	80.9	**89.1**	74.1	**85.1**	**68.7**	**82.1**	18.5ms

CAM methods

- **Standard CAMs**: can yield discriminative CNNs with accurate localization
- Leveraging temporal information during training yielded new SOA results

CAM Method for Object Co-Localization

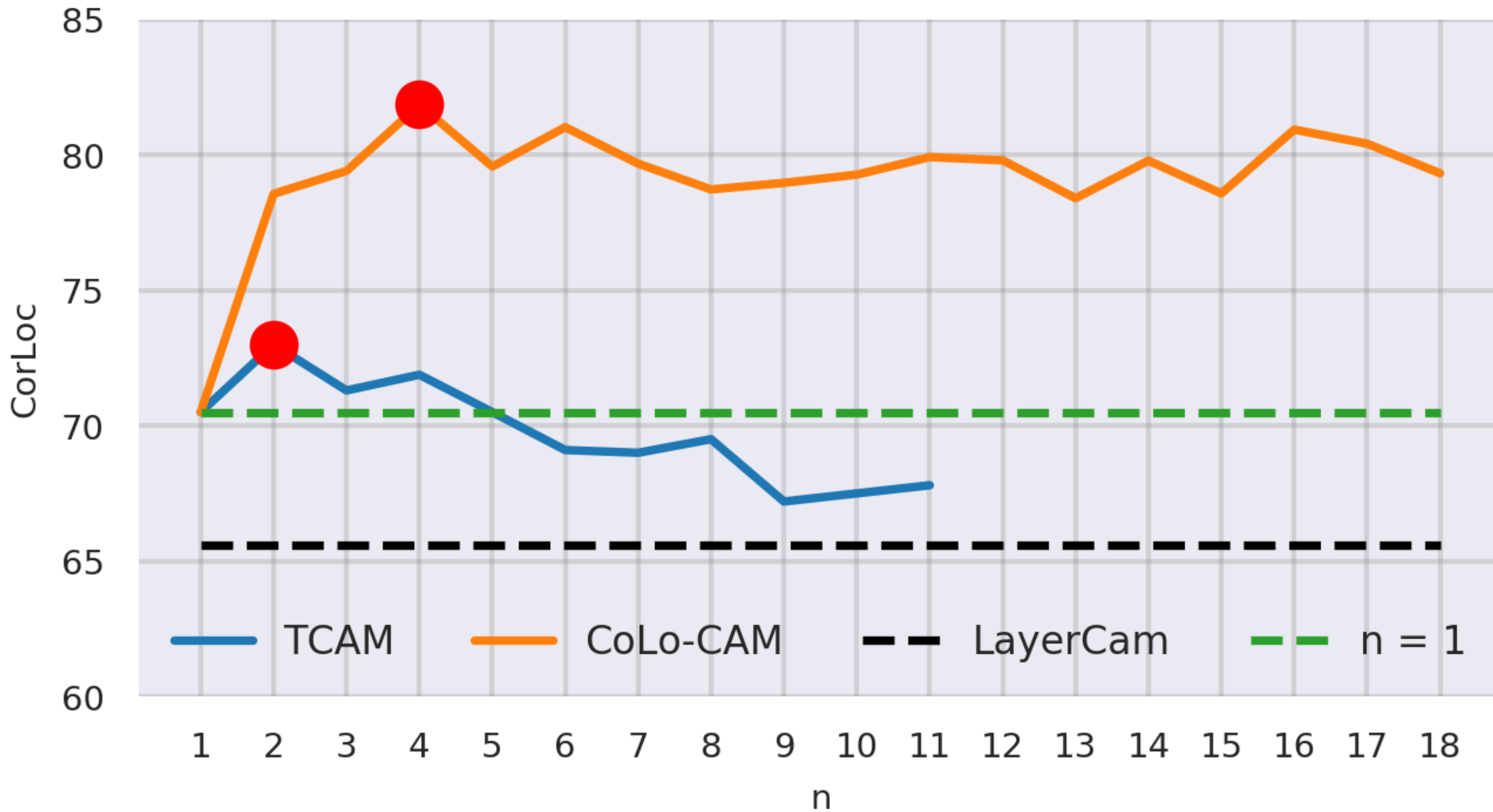
Experimental results: Localization examples of test sets frames YouTube-Object v1.0 and v2.2 datasets. Bounding box: GT (green), prediction (red).



S Belharbi, et al., CoLo-CAM: Class Activation Mapping for Object Co-Localization in Weakly-Labeled Unconstrained Videos. *arXiv:2303.09044*.

CAM Method for Object Co-Localization

Experimental results: Impact on CorLoc accuracy of the number of frames *n* on YTOv1 test set.



S Belharbi, et al., CoLo-CAM: Class Activation Mapping for Object Co-Localization in Weakly-Labeled Unconstrained Videos. *arXiv:2303.09044*.

Le génie pour l'industrie

Overview

1) **Personal Presentation**

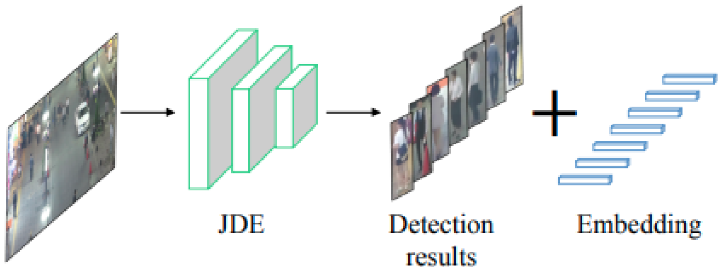
2) **Recent Research:**
   – unsupervised domain adaptation
   – cross-modal recognition
   – weakly-supervised object localization

3) **Potential Areas of Collaboration**

Potential Areas for Collaboration

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

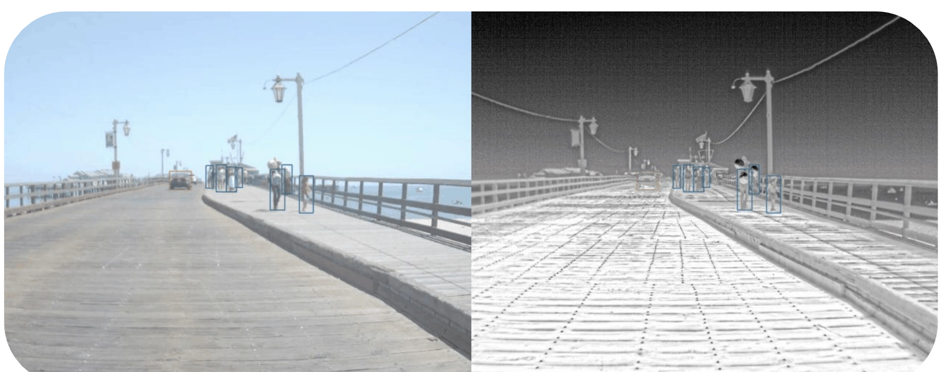
- rapid adaptation/calibration of DL models for deployment
- video-base emotion recognition
- methods weaky-supervised learning
- weakly-supervised spatial and temporal localization for visual interpretation
- joint detection & embedding (JDE) for cost-effective ReID and multi-object tracking

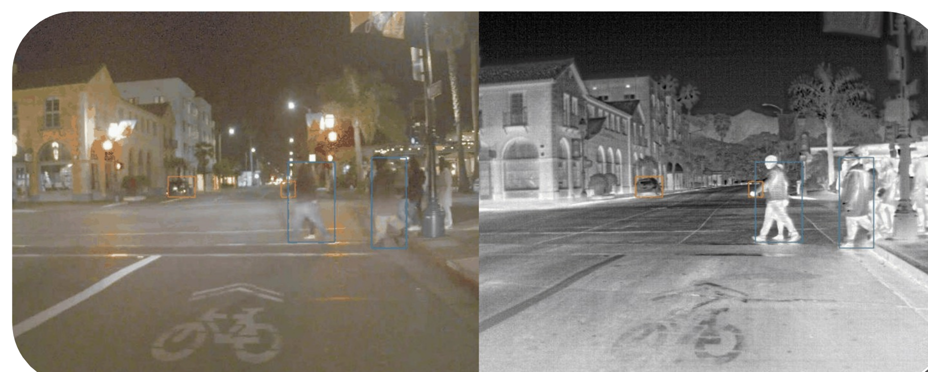


Align distributions to handle multiple cameras scenarios

Rapid Adaptation of DL Models for Deployment

- **Weakly-supervised DA** based e.g., on video tags
- **Domain generalization** to improve robustness
- **Multi-source DA** using several source datasets for robust adaptation
- **Multi-target DA:** adapt of one compact model for multiple targets
- **Cross-modality adaptation** across sensors, e.g., RGB-IR

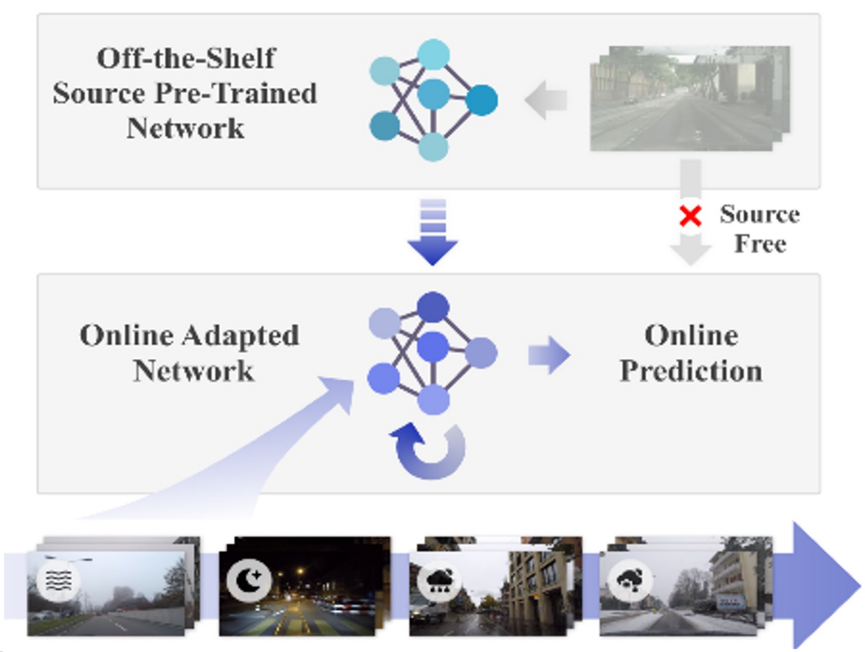


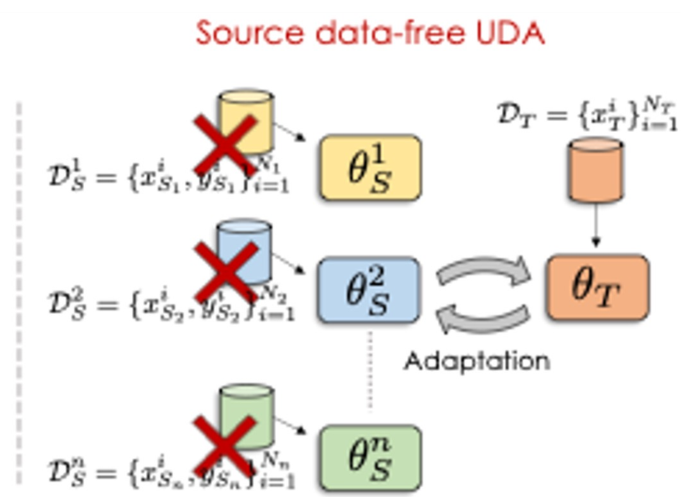


Jia, Xinyu, et al. "LLVIP: A Visible-Infrared Paired Dataset for Low-light Vision." ICCV 2021.

Rapid Adaptation of DL Models for Deployment

- **Source-free and test-time DA:** adapt rapidly without source data for efficiency and privacy
- **Continuous (incremental) DA:** adapt to new data
- **Gradual DA:** find on multiple intermediate domains, and multiple steps to manage larger domain shifts

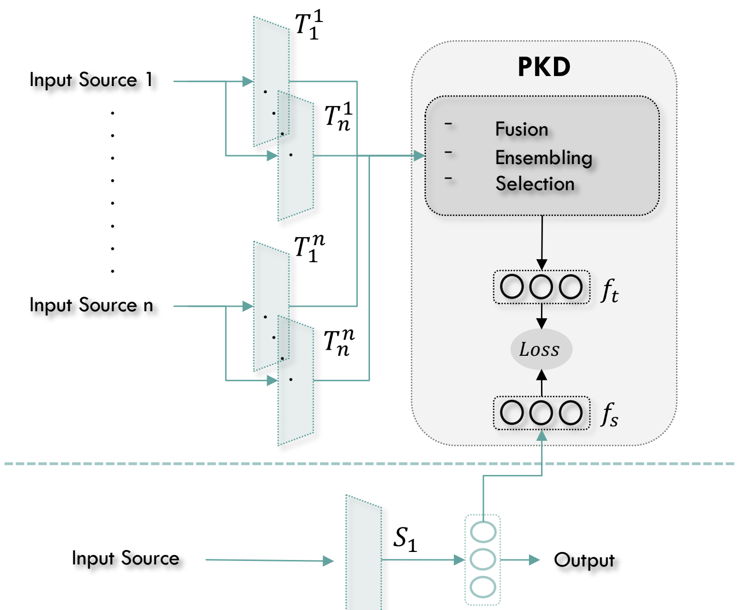




Tahmed, S. M., et al., Unsupervised multi-source domain adaptation without access to source data. CVPR 2021.

Video-Based Emotion Recognition

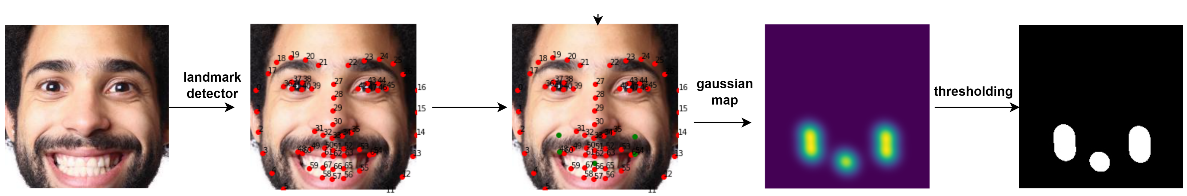
Privileged knowledge distillation: distill knowledge from teacher (w privileged information) to a student (w/o privileged information)



Modality	Train	Test	Type
Visual	✓	✓	Discriminant
Audio	✓	✓	Discriminant
Text	✓	O	Discriminant
Physiological	O	X	Discriminant
Age, Gender	O	O	Side Info
Pose, Eye Gaze	X	✓	Contextual
Gestures/Body Language	X	✓	Contextual

Test-time adaptation to persons/contexts with short neutral control video

Spatio-temporal localization based with constraints from action units.



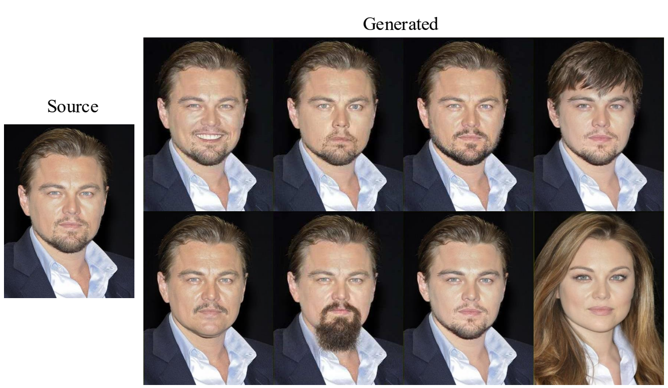
Aslam, H. et al., Privileged Knowledge Distillation for Dimensional Emotion Recognition in the Wild. CVPR 2023 workshops.

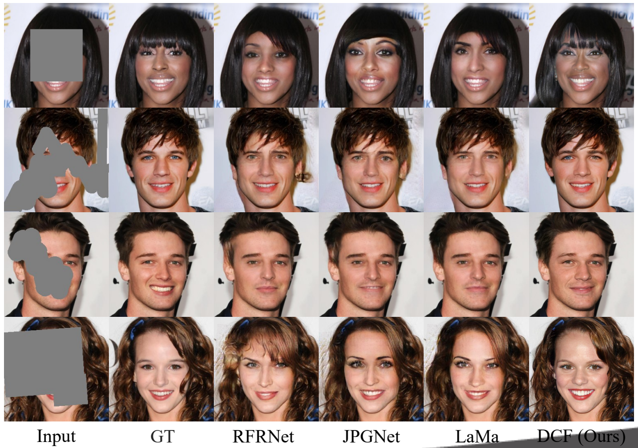
Image Processing

- **Adversarial Nets:**
  P Shamsolmoali, M Zareapoor, E Granger, H Zhou, R Wang, ME Celebi, J Yang. "Image synthesis with adversarial networks: A comprehensive survey and case studies." Information Fusion, 2021.

- **Image completion/inpainting:**
  P. Shamsolmoali, M Zareapoor, E Granger, Image Completion via Dual-Path Cooperative Filtering, ICASSP 2023.

- **Image-to-image translation, face style transfer:**
  Shamsolmoali P, Zareapoor M, Das S, Garcia S, Granger E, Yang J. GEN: Generative equivariant networks for diverse image-to-image translation. IEEE Trans. on Cybernetics, 2022.





Some Partners

- **Video analytics and surveillance:** Genetec, Nuvoola, Sportlogic, Computer Research Institute of Montreal (CRIM)
- **Health:** CIUSSE-Nord-de-l'Île-de-Montréal, Montreal Behavioural Medicine Centre, Centre de rehabilitation Lucie-Bruneau, Jackson Laboratory
- **Gaming:** Ubisoft LaForge
- **Building Automation and IoT:** Distech Controls
- **Communications:** Ericsson