TCAM: Temporal Class Activation Maps for Object Localization in Weakly-Labeled Unconstrained Videos

#1703

Soufiane Belharbi[1], Ismail Ben Ayed[1], Luke McCaffrey[2], **Eric Granger**[1]

1. LIVIA, Dept. of Systems Engineering, ETS Montreal, Canada
2. Goodman Cancer Research Centre, Dept. of Oncology, McGill University, Montreal, Canada

WACV WAIKOLOA HAWAII JAN 3-7 • 2023

ÉTS Le génie pour l'industrie

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

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

McGill

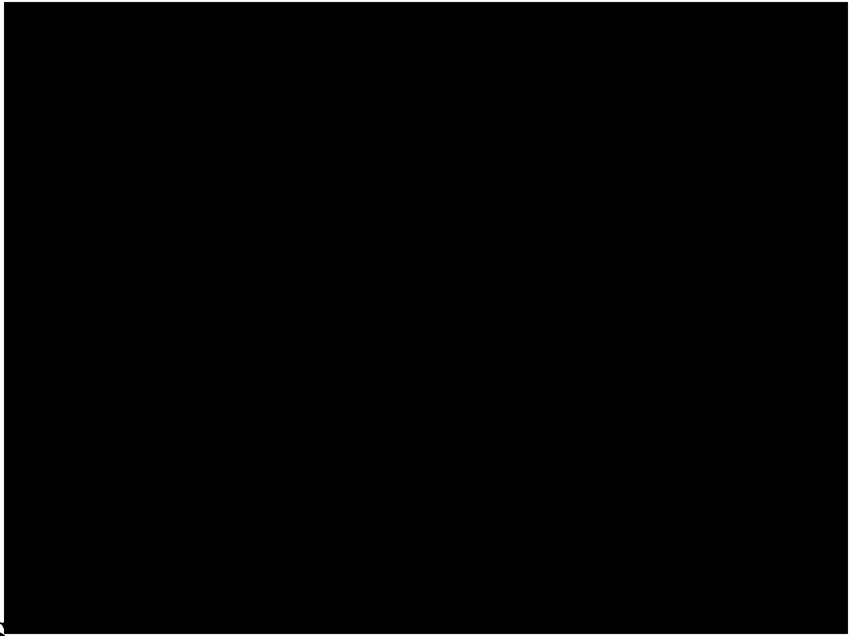
1. **Background:** Weakly-Supervised Video Object Localization

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.

Unconstrained videos are challenging:

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



Source: Unconstrained videos from the YouTube-Objects v1.0 dataset.

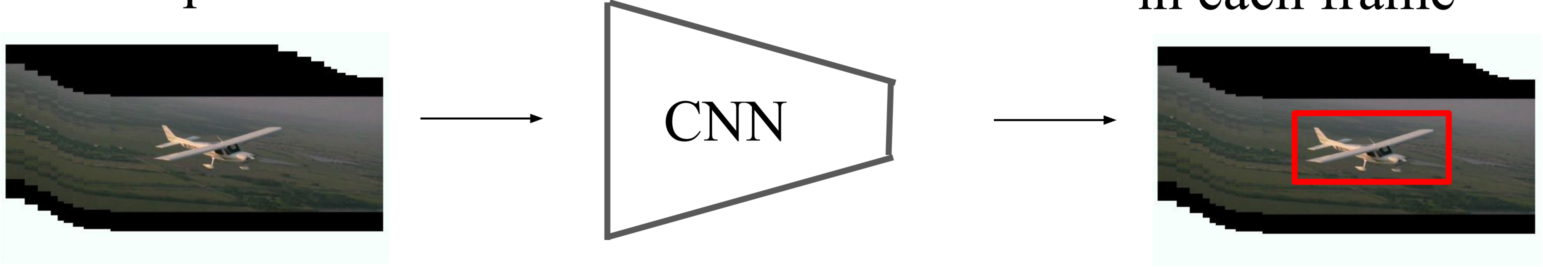
1. **Background:** Weakly-Supervised Video Object Localization

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

Video sequence

Localize object in each frame



1. **Background:** Weakly-Supervised Video Object Localization

Challenges for State-of-Art Methods:

- Multiple sequential, independent stages
- Video tags (labels) are only used to cluster video
- ROI are not necessarily discriminative
- Motion cues (optical flow) are not necessarily discriminative
- Localization involves solving an optimization problem over one or more videos →

  slow inference time, build model per class/video

1. **Background:** Weakly-Supervised Video Object Localization

CAM methods:

- successful in developing discriminative WSOL models for CNNs
- we adapt CAMs to exploit the spatio-temporal dependency in videos

1. **Background:** Weakly-Supervised Video Object Localization

CAM methods:

- successful in developing discriminative WSOL models for CNNs
- we adapt CAMs to exploit the spatio-temporal dependency in videos

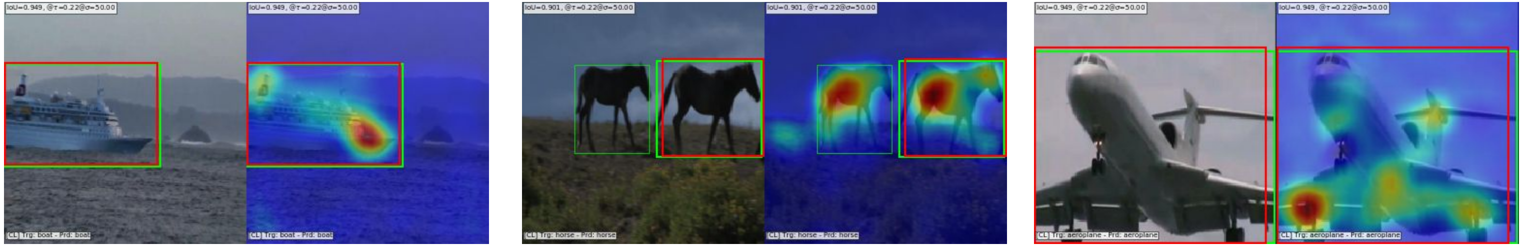
Advantages compared to SOTA of WSVOL (videos):

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

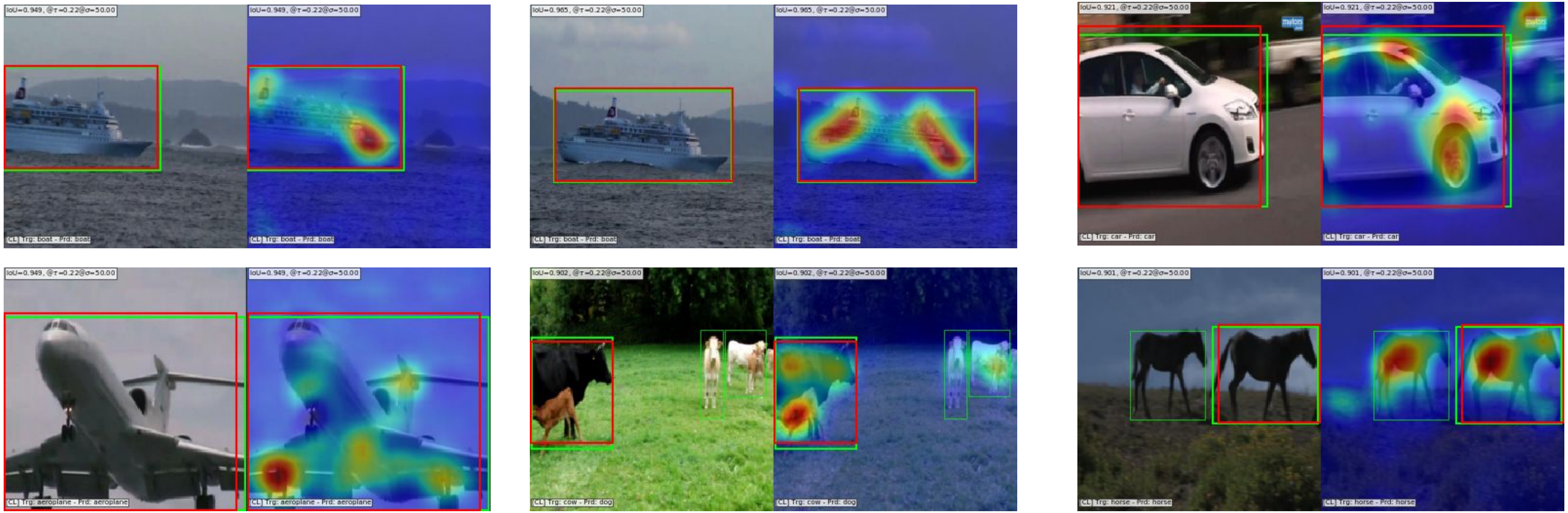
Advantage compared to CAMs for WSOL (still images):

- allows to leverage temporal information in videos

CAM results on still images

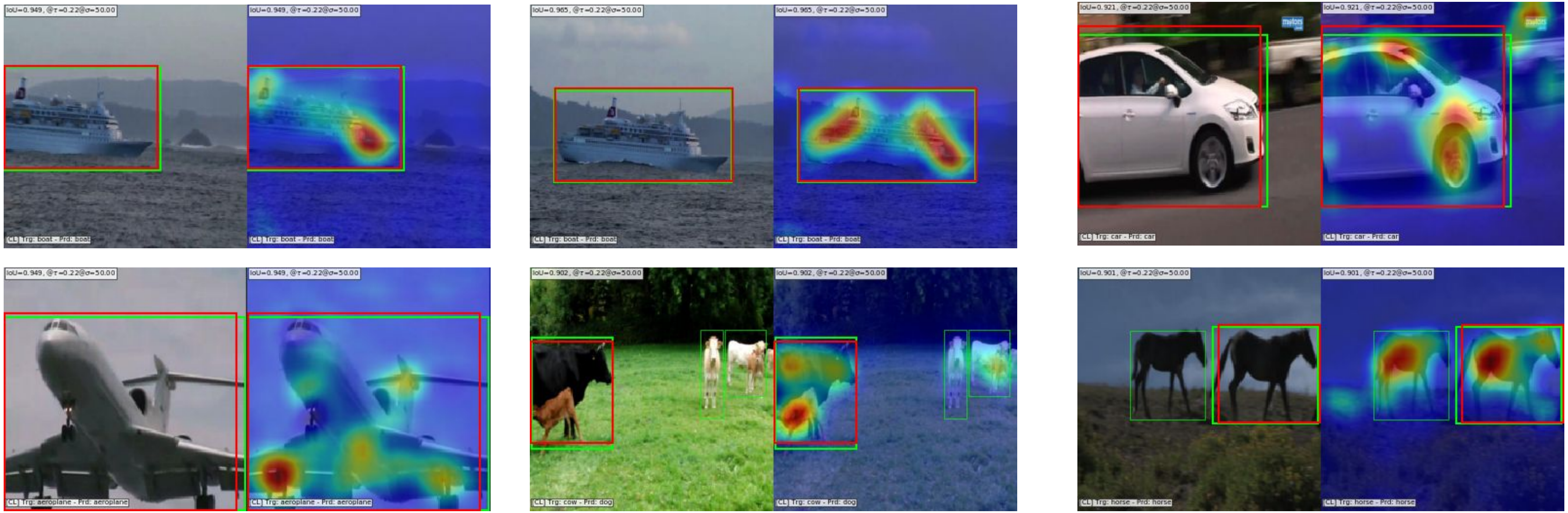


1. **Background:** CAM methods trained on still images yield decent localization performance



Source: Zhou, B. et al., Learning deep features for discriminative localization. In CVPR 2016.

1. **Background:** CAM methods trained on still images yield decent localization performance



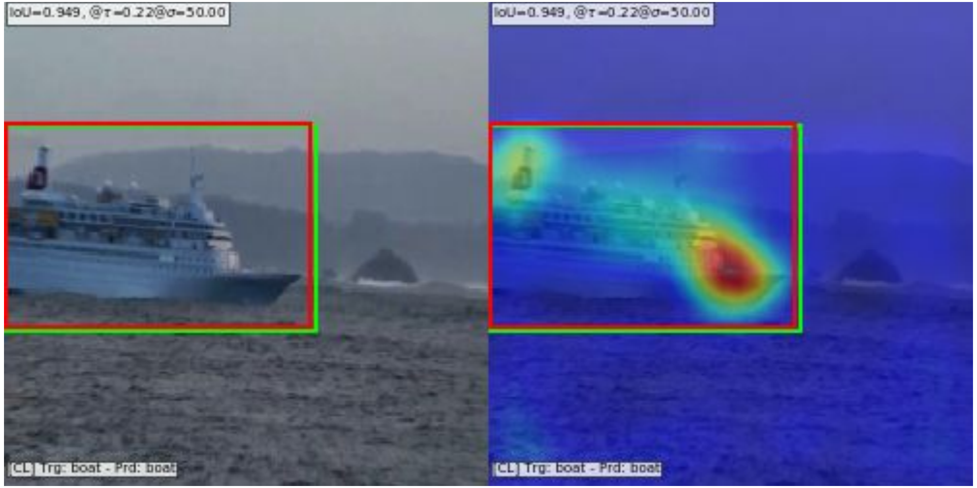
Do not account for spatio-temporal dependency

Source: Zhou, B. et al., Learning deep features for discriminative localization. In CVPR 2016.

2. Proposed TCAM method

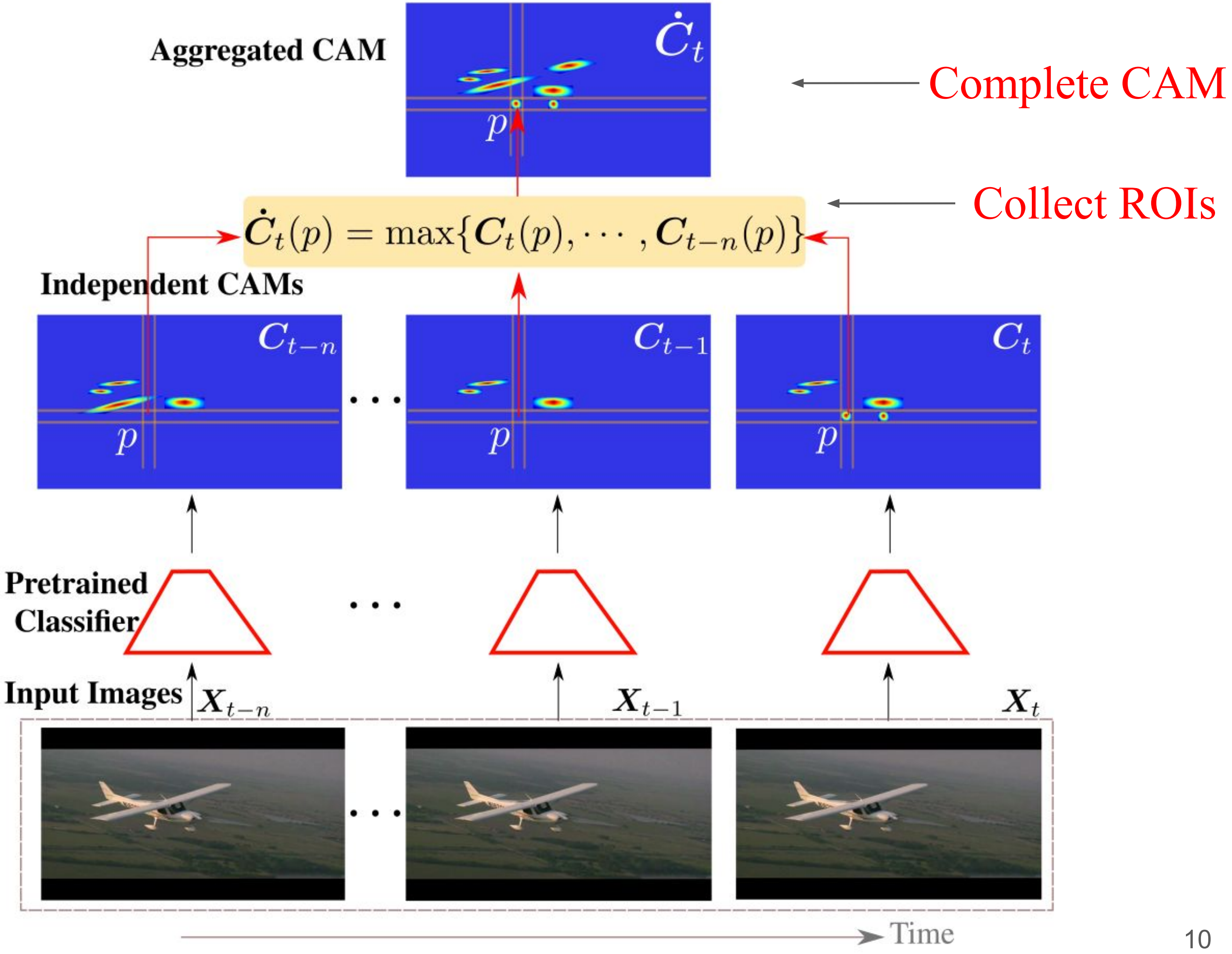
Main idea:

- leverage the slight variations in sets of consecutive frames
- aggregate diversified CAMs from *n* frames
- use aggregated CAMs for sample pixel pseudo-labels for training



2. Proposed TCAM method

CAM-Temporal **M**ax **P**ooling **(CAM-TMP)** for aggregation of n consecutive CAMs



2. Proposed TCAM

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

V : video

y : video tag (class)

X_t : frame at time t

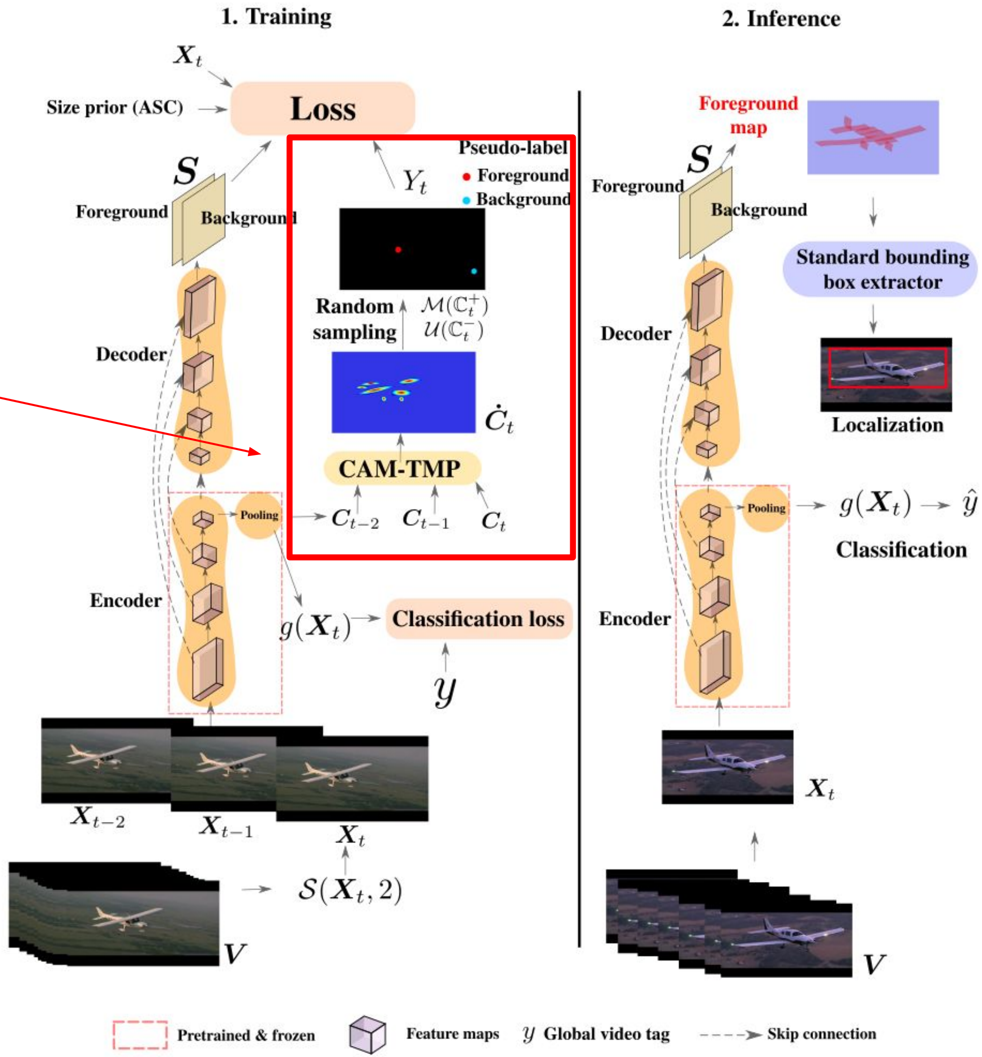
C_t : CAM of frame X_t

$\mathcal{S}(X_t, 2)$: sampling function

$\dot{C}_t$: aggregated CAM

Y_t : pixel pseudo-label mask

S : output CAM



2. Proposed TCAM

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

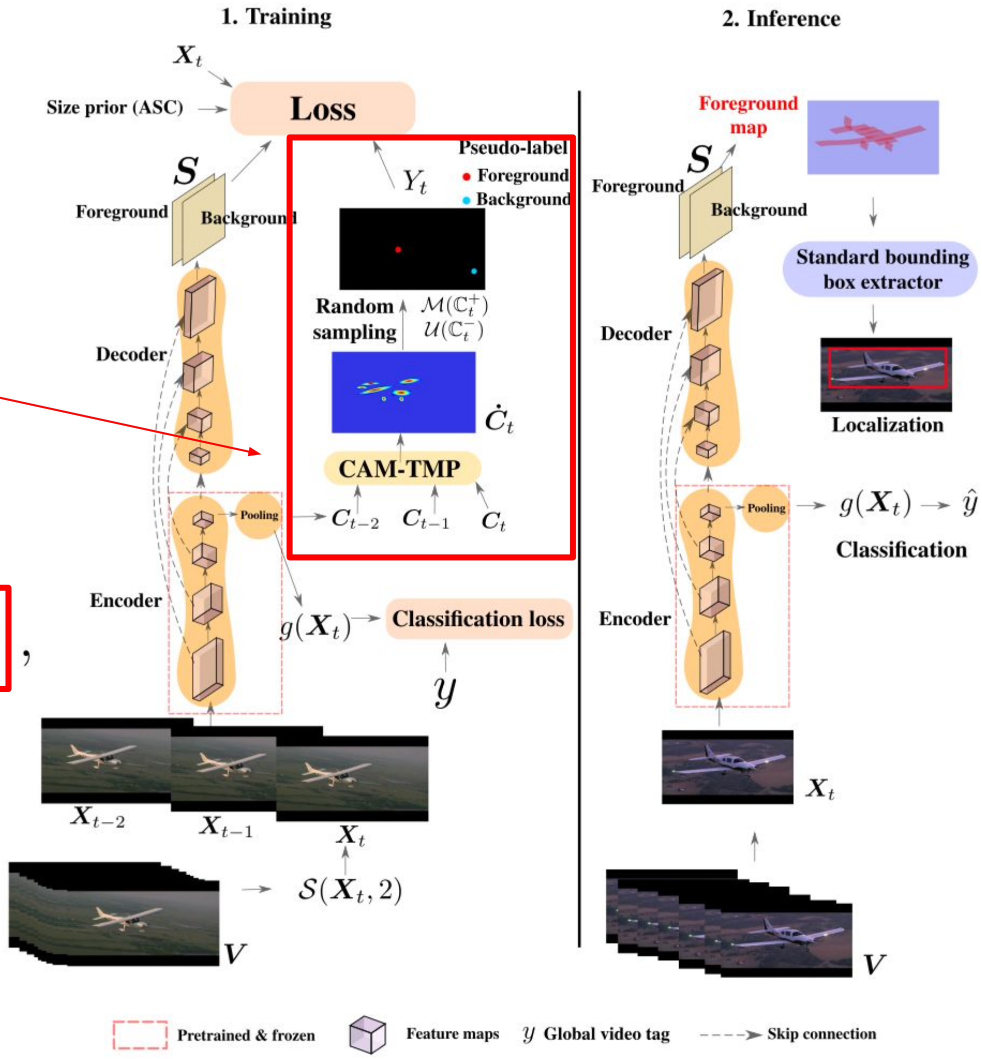
pixel pseudo-labels

CRF

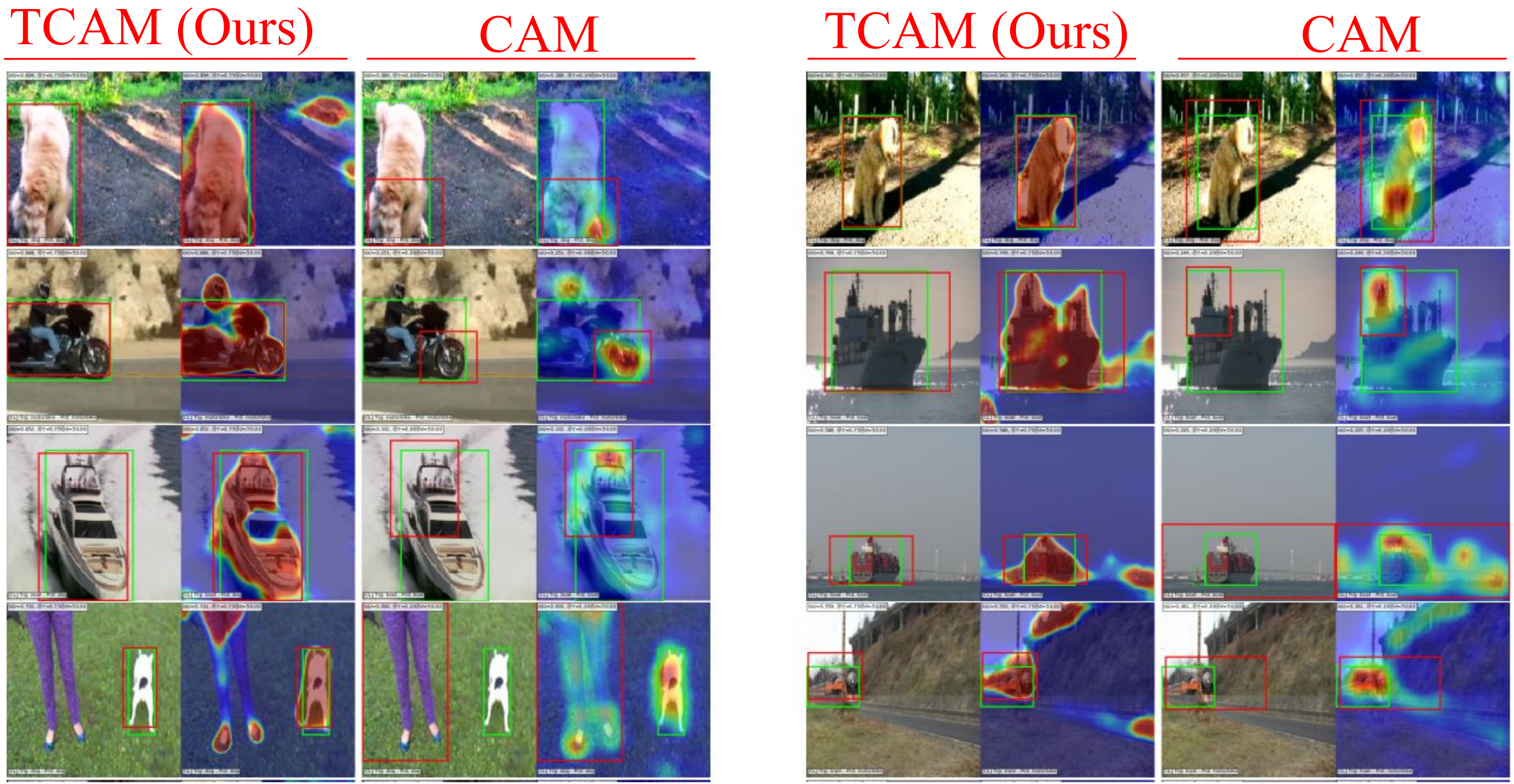
$$\min_{\boldsymbol{\theta}} \boxed{\sum_{p \in \Omega'_t} \boldsymbol{H}_p(Y_t, \boldsymbol{S}_t)} + \lambda \boxed{\mathcal{R}(\boldsymbol{S}_t, \boldsymbol{X}_t)},$$

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

large size (FG/BG)



3. Experimental results: YouTube-Object v1.0 and v2.2 datasets



3. **Experimental results:** YTOv1 dataset

- Standard CAM methods yield discriminative models for localization
- Leveraging temporal information with TCAM yields new state-of-the-art results

Method *(venue)*	Aero	Bird	Boat	Car	Cat	Cow	Dog	Horse	Mbike	Train	Avg	Time/Frame
(Prest et al., 2012) *(cvpr,2012)*	51.7	17.5	34.4	34.7	22.3	17.9	13.5	26.7	41.2	25.0	28.5	N/A
(Papazoglou and Ferrari, 2013) *(iccv,2013)*	65.4	67.3	38.9	65.2	46.3	40.2	65.3	48.4	39.0	25.0	50.1	4s
(Joulin et al., 2014) *(eccv,2014)*	25.1	31.2	27.8	38.5	41.2	28.4	33.9	35.6	23.1	25.0	31.0	N/A
(Kwak et al., 2015) *(iccv,2015)*	56.5	66.4	58.0	76.8	39.9	69.3	50.4	56.3	53.0	31.0	55.7	N/A
(Rochan et al., 2016) *(ivc,2016)*	60.8	54.6	34.7	57.4	19.2	42.1	35.8	30.4	11.7	11.4	35.8	N/A
(Tokmakov et al., 2016) *(eccv,2016)*	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 (Koh et al., 2016) *(cvpr,2016)*	64.3	63.2	73.3	68.9	44.4	62.5	71.4	52.3	78.6	23.1	60.2	N/A
(Tsai et al., 2016) *(eccv,2016)*	66.1	59.8	63.1	72.5	54.0	64.9	66.2	50.6	39.3	42.5	57.9	N/A
(Haller and Leordeanu, 2017) *(iccv,2017)*	76.3	71.4	65.0	58.9	68.0	55.9	70.6	33.3	69.7	42.4	61.1	0.35s
(Croitoru et al., 2019) (LowRes-Net$_{iter1}$) *(ijcv,2019)*	77.0	67.5	77.2	68.4	54.5	68.3	72.0	56.7	44.1	34.9	62.1	0.02s
(Croitoru et al., 2019) (LowRes-Net$_{iter2}$) *(ijcv,2019)*	79.7	67.5	68.3	69.6	59.4	75.0	78.7	48.3	48.5	39.5	63.5	0.02s
(Croitoru et al., 2019) (DilateU-Net$_{iter2}$) *(ijcv,2019)*	85.1	72.7	76.2	68.4	59.4	76.7	77.3	46.7	48.5	46.5	65.8	0.02s
(Croitoru et al., 2019) (MultiSelect-Net$_{iter2}$) *(ijcv,2019)*	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) (Zhang et al., 2020b) *(tpami,2020)*	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) (Zhang et al., 2020b) *(tpami,2020)*	**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 (Umer et al., 2021) *(optik,2021)*	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 (Zhou et al., 2016) *(cvpr,2016)*	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 (Selvaraju et al., 2017) *(iccv,2017)*	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++ (Chattopadhyay et al., 2018) *(wacv,2018)*	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++ (Omeiza et al., 2019) *(corr,2019)*	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 (Fu et al., 2020) *(bmvc,2020)*	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 (Jiang et al., 2021) *(ieee,2021)*	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 (ours)	90.5	70.4	62.2	75.7	**84.8**	**81.0**	**81.0**	64.5	70.4	50.0	**73.0**	18.5ms

CAM methods

3. **Experimental results:** YTOv2.2 dataset

Method ***(venue)***	**Aero**	**Bird**	**Boat**	**Car**	**Cat**	**Cow**	**Dog**	**Horse**	**Mbike**	**Train**	**Avg**	**Time/Frame**
(Haller and Leordeanu, 2017) *(iccv,2017)*	76.3	68.5	54.5	50.4	59.8	42.4	53.5	30.0	53.5	**60.7**	54.9	0.35s
(Croitoru et al., 2019) (LowRes-Net$_{iter1}$) *(ijcv,2019)*	75.7	56.0	52.7	57.3	46.9	57.0	48.9	44.0	27.2	56.2	52.2	0.02s
(Croitoru et al., 2019) (LowRes-Net$_{iter2}$) *(ijcv,2019)*	78.1	51.8	49.0	60.5	44.8	62.3	52.9	48.9	30.6	54.6	53.4	0.02s
(Croitoru et al., 2019) (DilateU-Net$_{iter2}$)*(ijcv,2019)*	74.9	50.7	50.7	60.9	45.7	60.1	54.4	42.9	30.6	57.8	52.9	0.02s
(Croitoru et al., 2019) (BasicU-Net$_{iter2}$)*(ijcv,2019)*	**82.2**	51.8	51.5	62.0	50.9	64.8	55.5	45.7	35.3	55.9	55.6	0.02s
(Croitoru et al., 2019) (MultiSelect-Net$_{iter2}$)*(ijcv,2019)*	81.7	51.5	54.1	62.5	49.7	68.8	55.9	50.4	33.3	57.0	56.5	0.15s
CAM (Zhou et al., 2016) *(cvpr,2016)*	52.3	66.4	25.0	66.4	39.7	**87.8**	34.7	53.6	45.4	43.7	51.5	0.2ms
GradCAM (Selvaraju et al., 2017) *(iccv,2017)*	44.1	68.4	50.0	61.1	51.8	79.3	56.0	47.0	44.8	42.4	54.5	27.8ms
GradCAM++ (Chattopadhyay et al., 2018) *(wacv,2018)*	74.7	78.1	38.2	69.7	56.7	84.3	61.6	61.9	43.0	44.3	61.2	28.0ms
Smooth-GradCAM++ (Omeiza et al., 2019) *(corr,2019)*	74.1	83.2	38.2	64.2	49.6	82.1	57.3	52.0	51.1	42.4	59.5	136.2ms
XGradCAM (Fu et al., 2020) *(bmvc,2020)*	68.2	44.5	45.8	64.0	46.8	86.4	44.0	57.0	44.9	45.0	54.6	14.2ms
LayerCAM (Jiang et al., 2021) *(ieee,2021)*	80.0	84.5	47.2	**73.5**	55.3	83.6	71.3	60.8	55.7	48.1	66.0	17.9ms
TCAM (ours)	79.4	**94.9**	**75.7**	61.7	**68.8**	87.1	**75.0**	**62.4**	**72.1**	45.0	**72.2**	18.5ms

CAM methods

- Standard CAM methods yield decent results → discriminative models are powerful for localization
- Leveraging temporal information during training yielded new state-of-the-art results

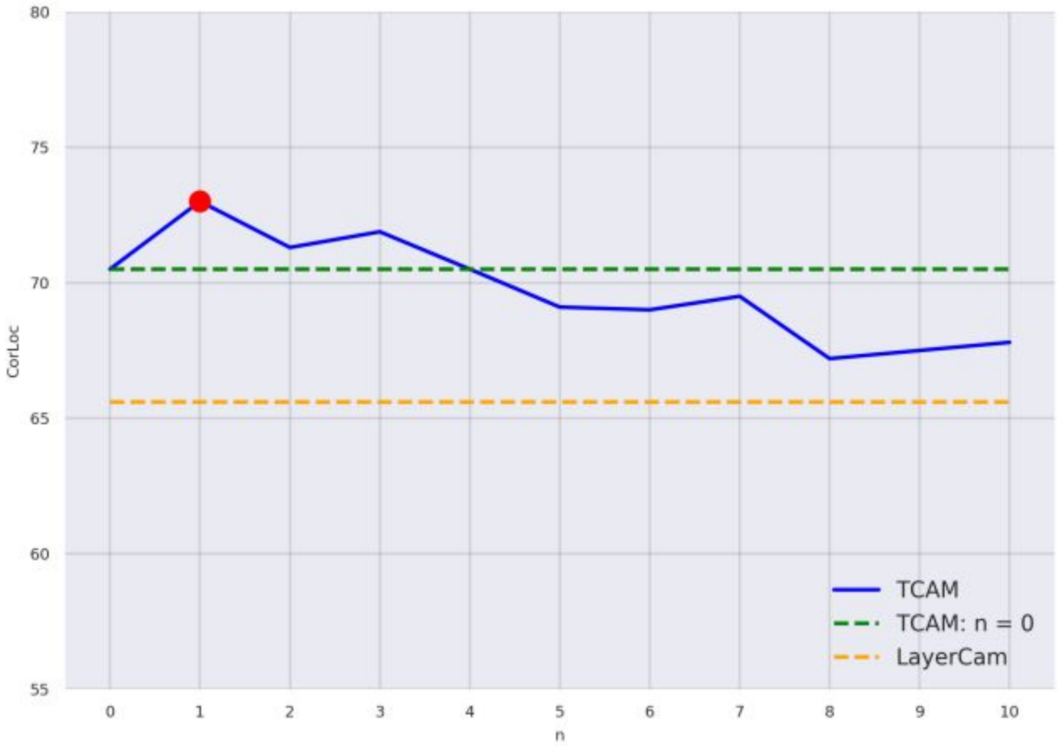
3. Experimental results: YouTube-Object v1.0 and v2.2 datasets

5. Experimental results: ablations

Methods		CorLoc
Layer-CAM (Jiang et al., 2021) *(ieee,2021)*		65.6
$n = 0$	Ours + $\mathbb{C}_t^+$ + $\mathbb{C}_t^-$	68.5
	Ours + $\mathbb{C}_t^+$ + $\mathbb{C}_t^-$ + CRF	69.6
	Ours + $\mathbb{C}_t^+$ + $\mathbb{C}_t^-$ + ASC	66.2
	Ours + $\mathbb{C}_t^+$ + $\mathbb{C}_t^-$ + CRF + ASC	70.5
$n > 0$	Ours + $\mathbb{C}_t^+$ + $\mathbb{C}_t^-$ + CRF + ASC + CAM-TMP	73.0
Improvement		+7.4

Time dependency: ***n* (number of frames)**

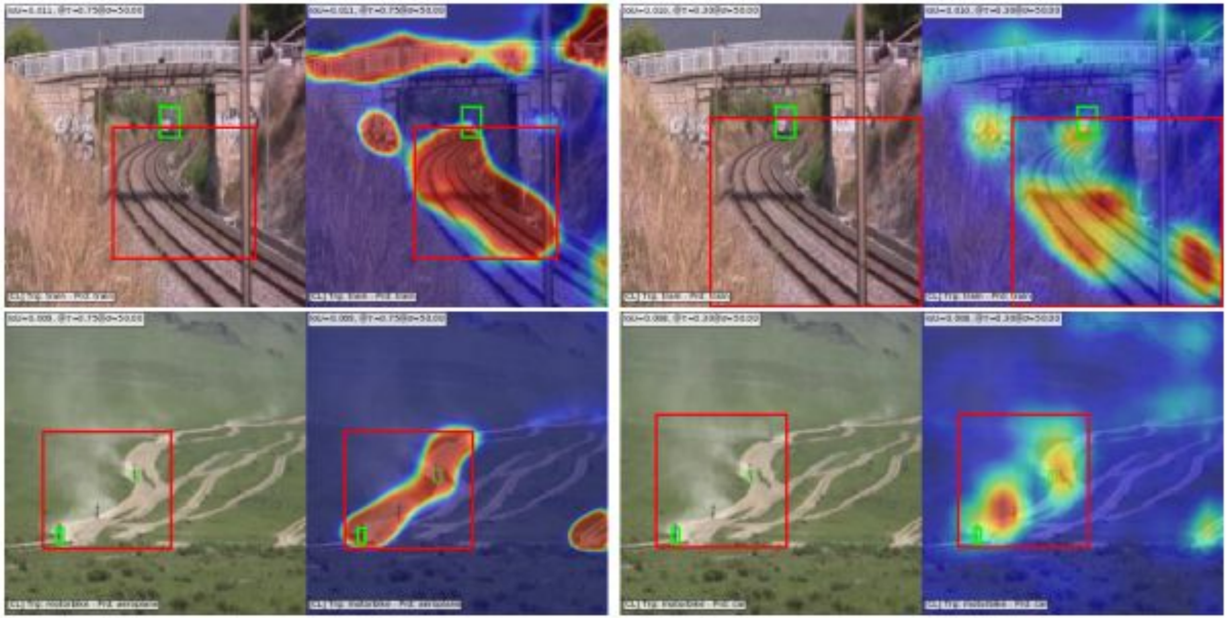
5. Experimental results: ablations



Time dependency: ***n* (number of frames)**

3. Experimental results: Failure cases

- Due to the co-occurrence of objects, and to small objects
- Caused by strong dependency to the pseudo-labels quality



TCAM Base-CAM

For questions and more details, please visit our poster #1703

Resources:

- Code: https://github.com/sbelharbi/tcam-wsol-video
- Demo videos: https://drive.google.com/drive/folders/1D8DgOdjT35Vf5Tqej3K5ZWFqz3LhgeQt?usp=sharing

ÉTS Le génie pour l'industrie

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

Digital Research Alliance of Canada | Alliance de recherche numérique du Canada

McGill