Explore Spatio-temporal Aggregation for Insubstantial Object Detection: Benchmark Dataset and Baseline

Computational Imaging Lab @ Nanjing University

Project Page Github Code CITE Lab

Kailai Zhou, Yibo Wang, Tao Lv, Yunqian Li, Linsen Chen, Qiu Shen,* Xun Cao*

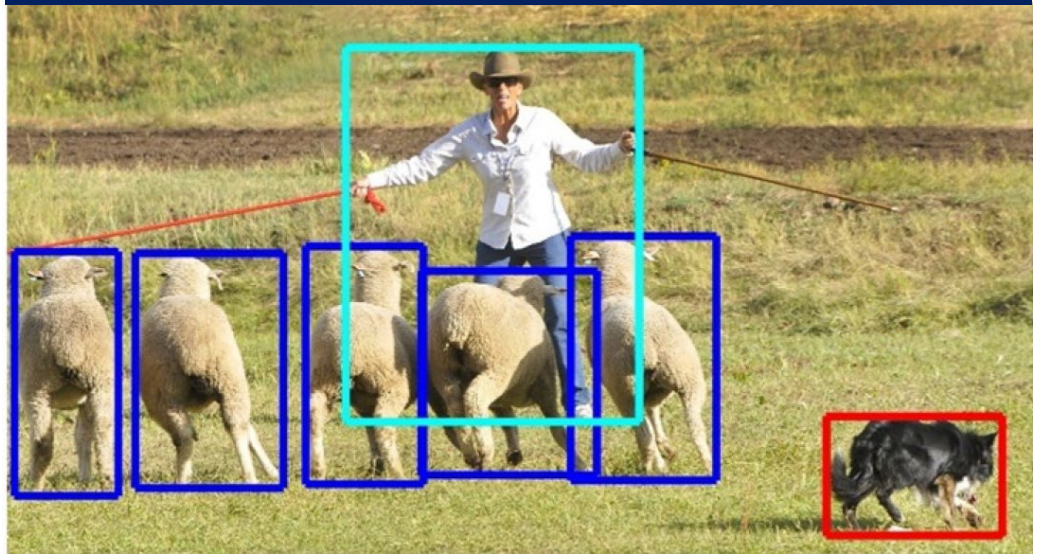
{calayzhou, ybwang, lvtao, lyq, linsen chen}smail.nju.edu.cn {shenqiu, caoxun}@nju.edu.cn

Background

Solid Substance

(salient visual features)

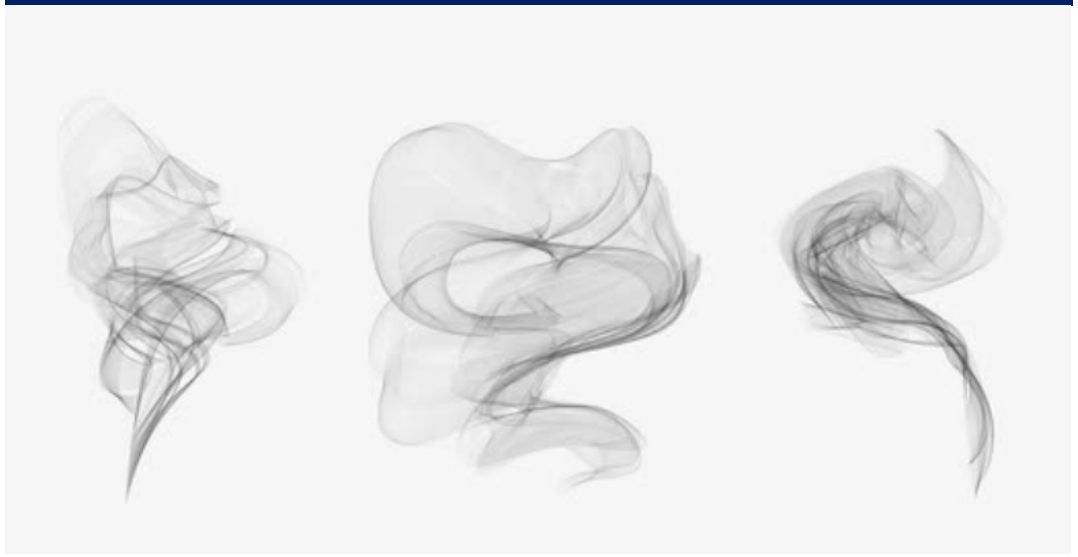
Person, Car, Animal...



Gaseous Substance

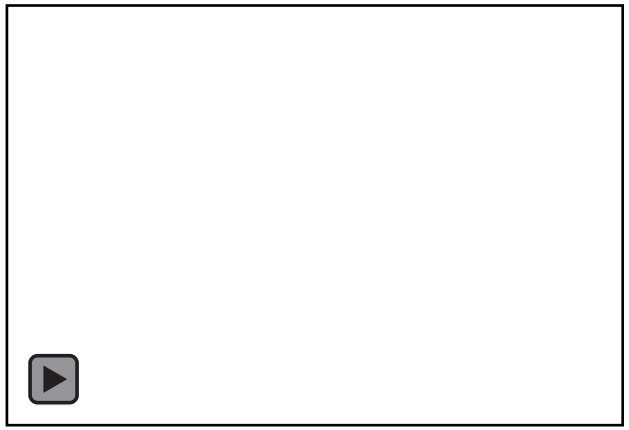
(cannot be seen by human)

CH4, C2H4, NH3...



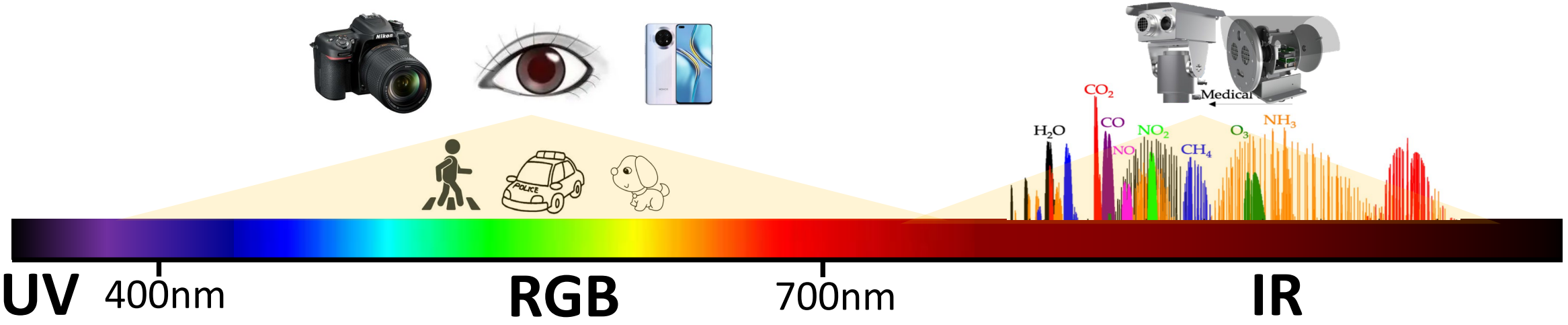
Q: Can we detect the gaseous substance with CV methods

Background



Human eyes/RGB camera

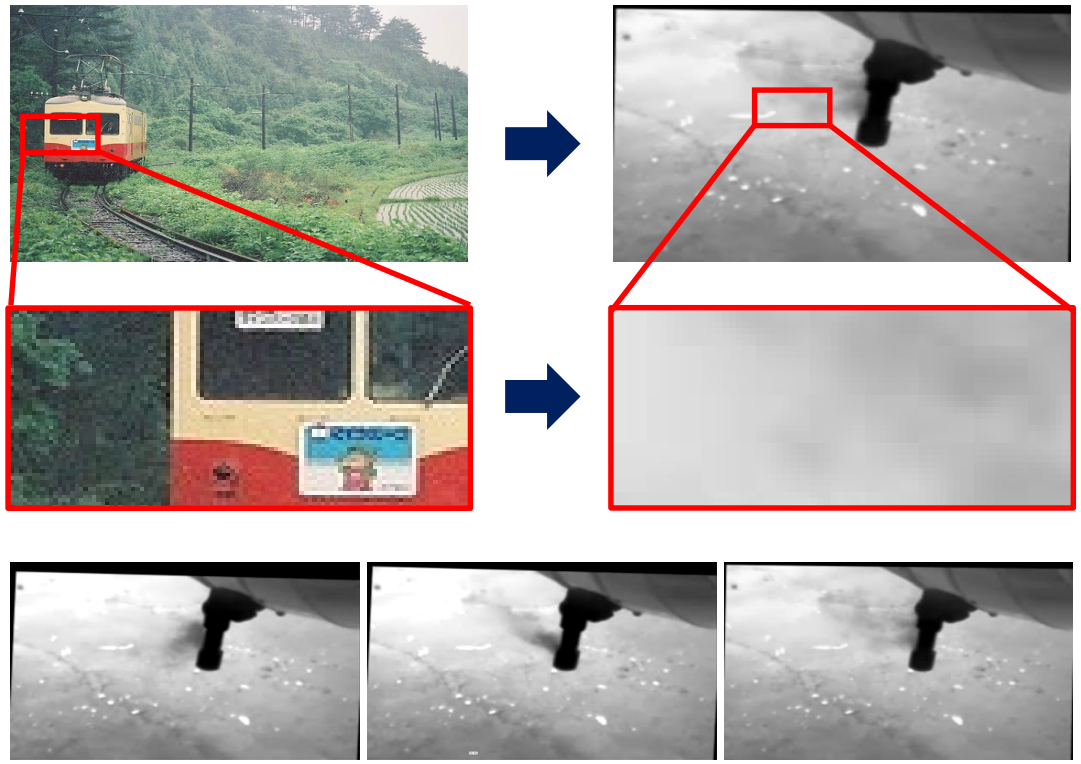
Infrared multispectral camera



Gas can be captured in the mid-infrared band

IOD task

“Insubstantial”

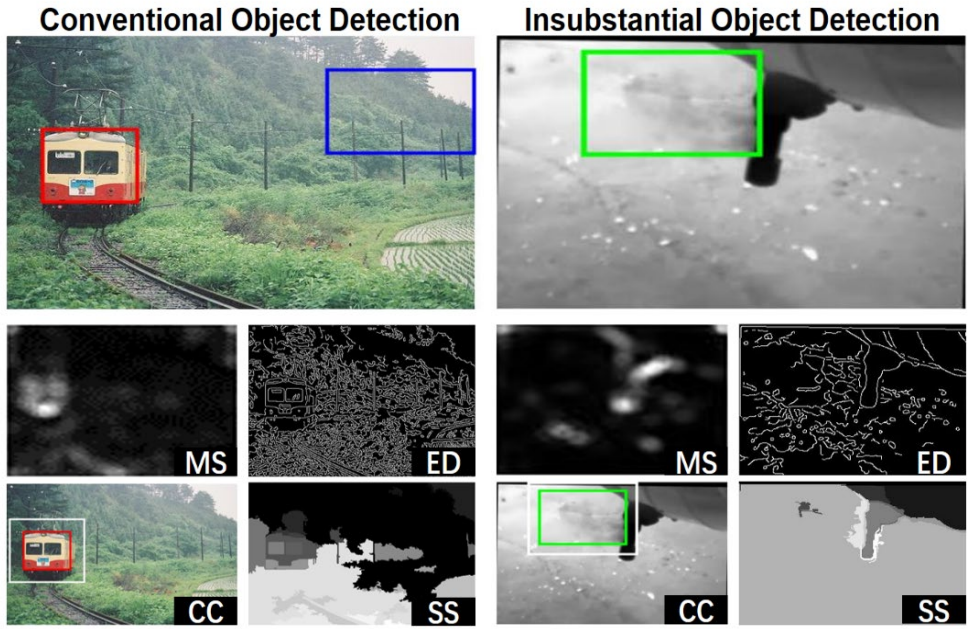


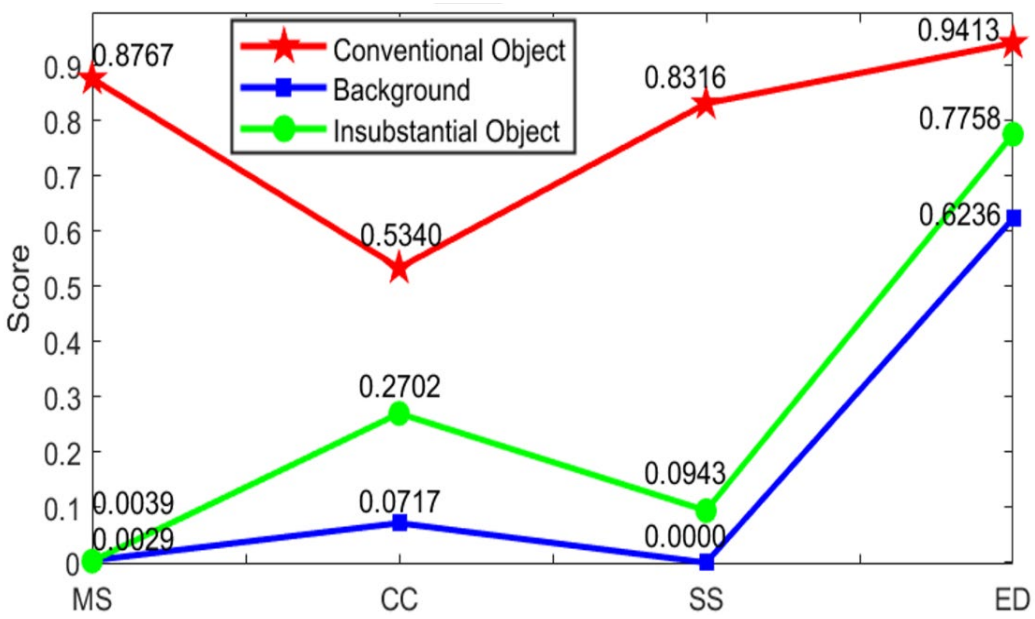
- lack color information
- indistinct boundary
- shape can be arbitrary

It’s of great difficulties to localize the insubstantial object

IOD task

What is an object?[1]





MS: Multi-scale Saliency
CC: Color Contrast

ED: Edge Density
SS: Superpixels Straddling

Insubstantial object is more similar to the background

[1] Alexe, Bogdan, Thomas Deselaers, and Vittorio Ferrari. "What is an object?." CVPR2010.

IOD task

Insubstantial Object Detection (IOD)

- indistinct boundary and amorphous shape
- the similarity to the background surroundings
- absence of color information and saliency

Smoke

Gas leak

Steam

IOD-Video Dataset

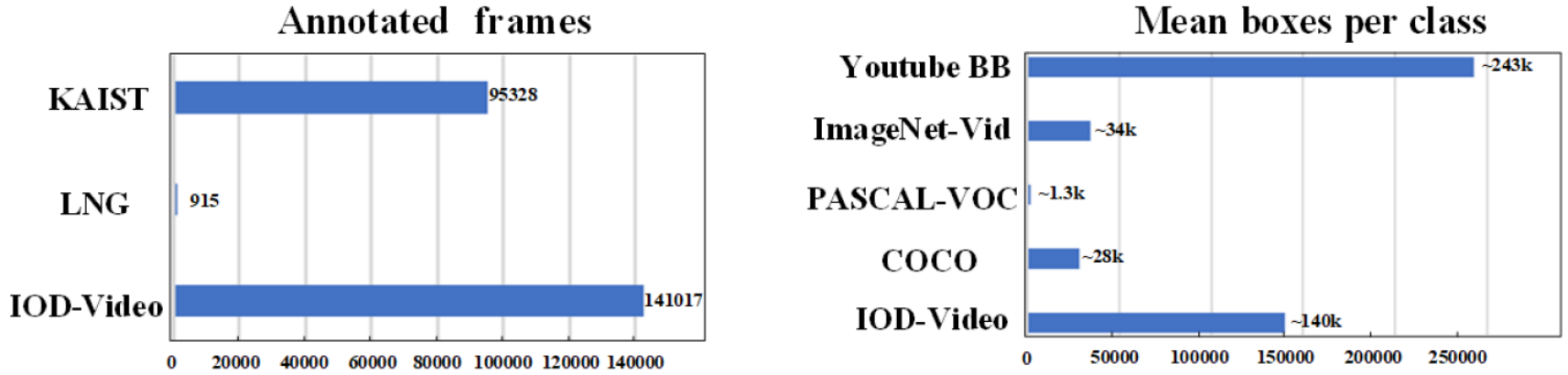
- we construct an IOD-Video dataset comprised of 600 videos (141,017 frames)
- It covers various distances (0~100m), sizes, visibility, and scenes
- all insubstantial objects (gas, smoke, steam, etc) are integrated into one category
- the carefully labeled frame-level annotations are provided

600 Videos

141K Frames

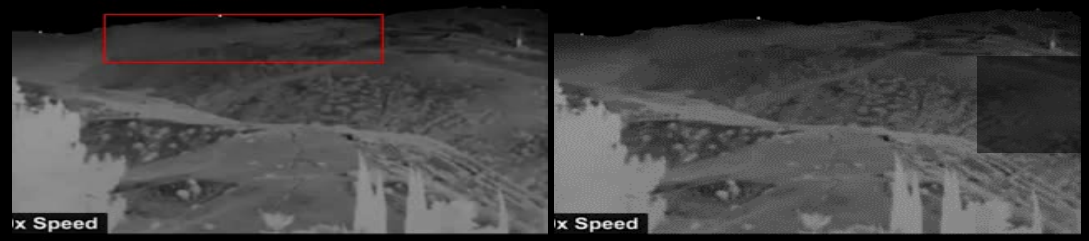
1 Category

8+ Scenes

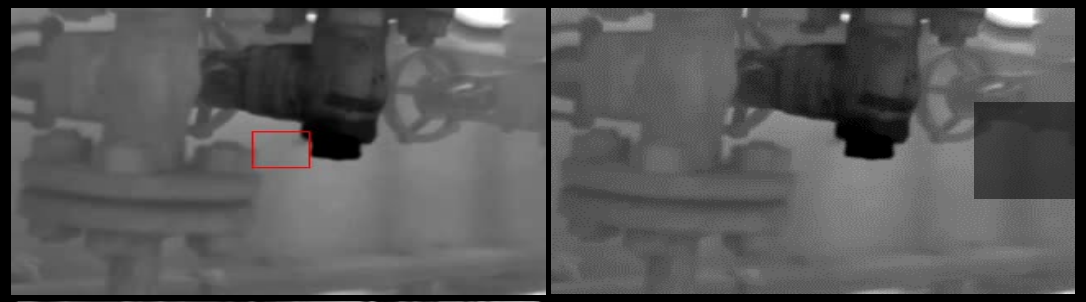


(a) Comparison with infrared detection dataset (b) Comparison with mainstream detection dataset

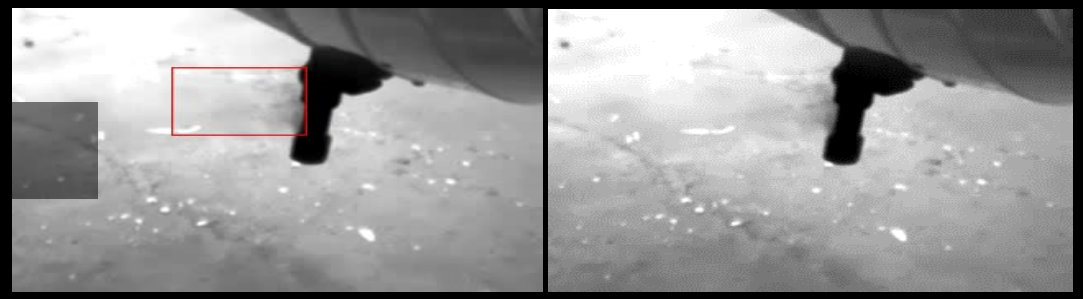
IOD-Video Dataset

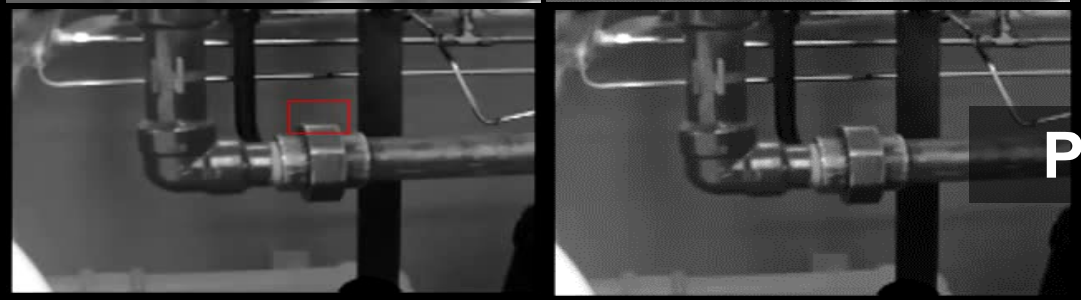


Wild

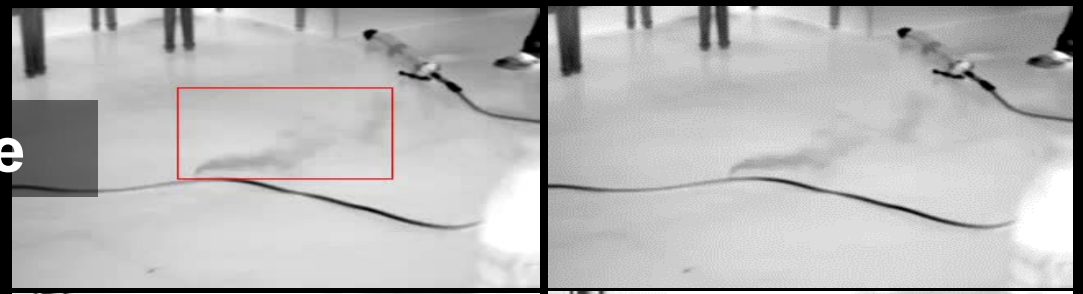


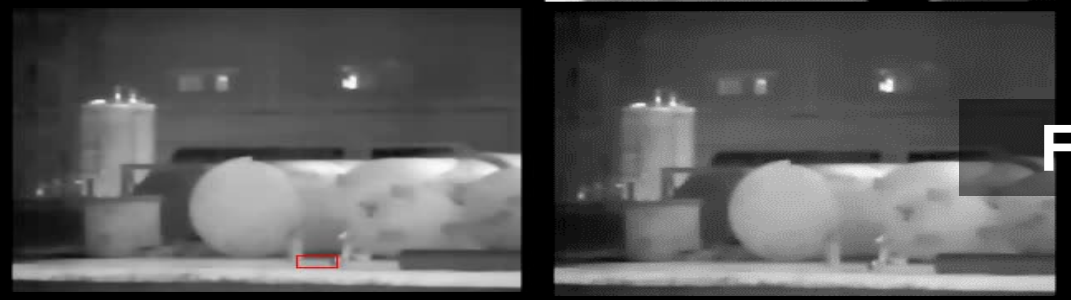
Valve



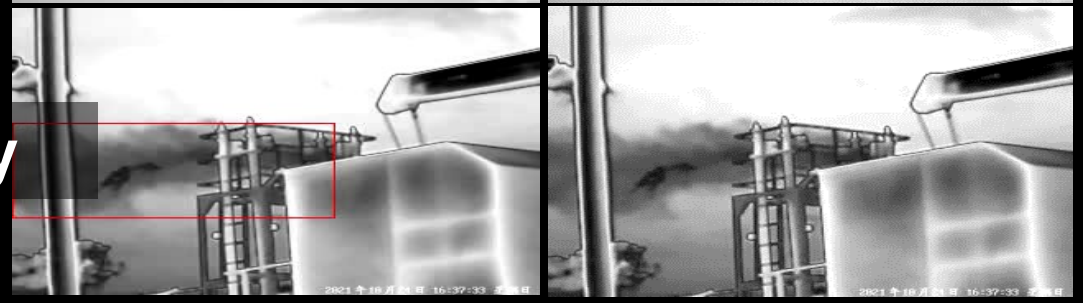


Pipeline

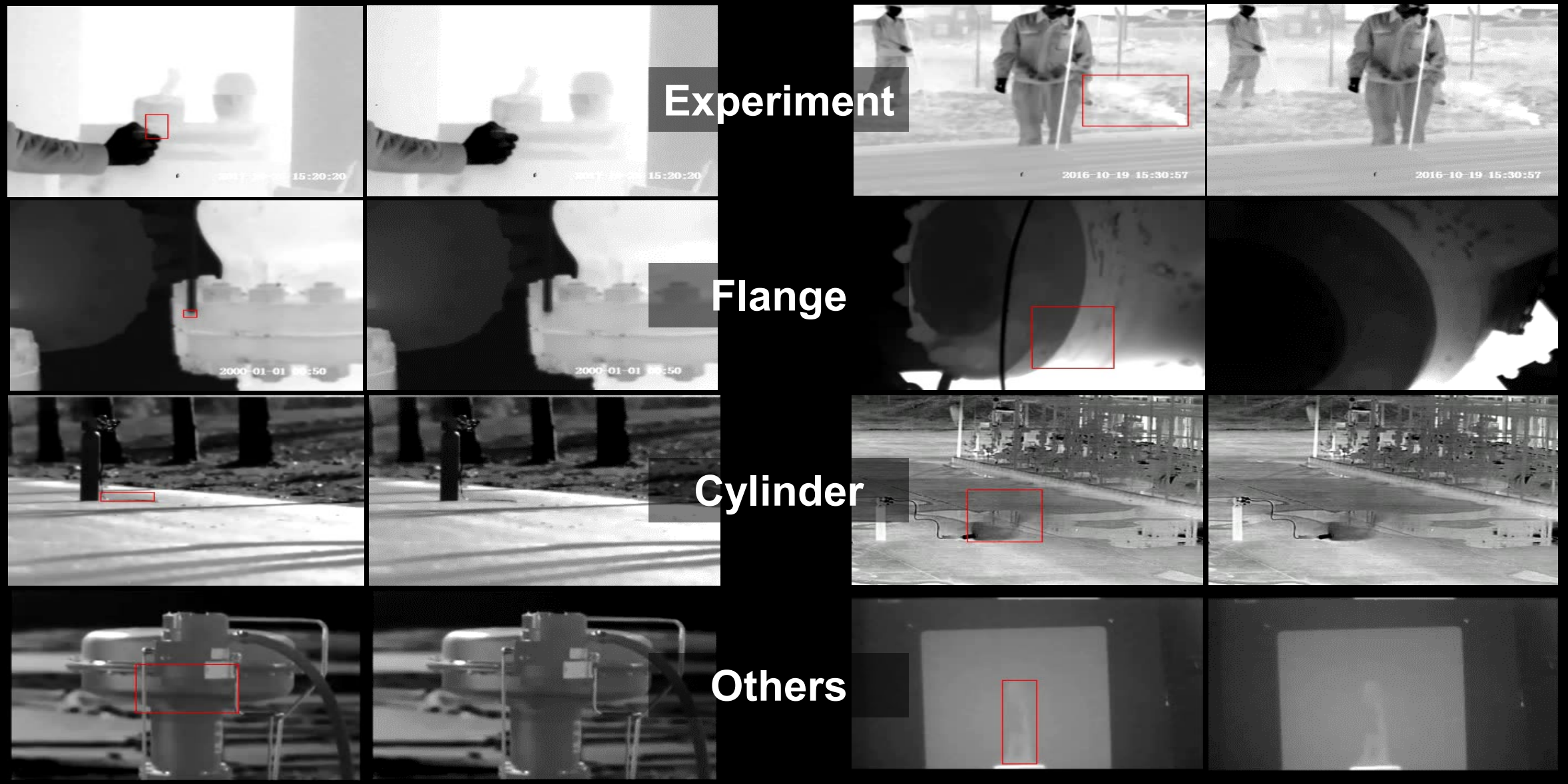




Factory



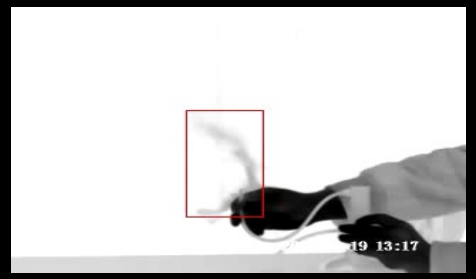
IOD-Video Dataset



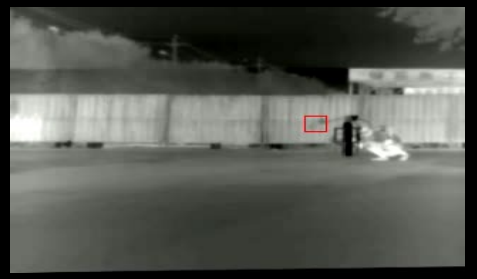
IOD-Video Dataset

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

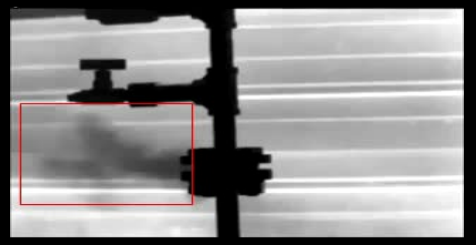
Various Distances



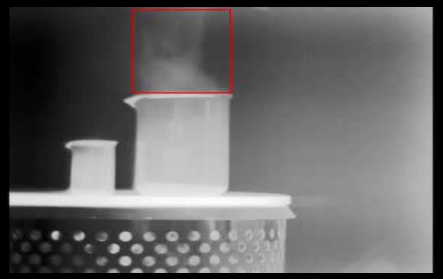

Various Sizes




Various Visibility

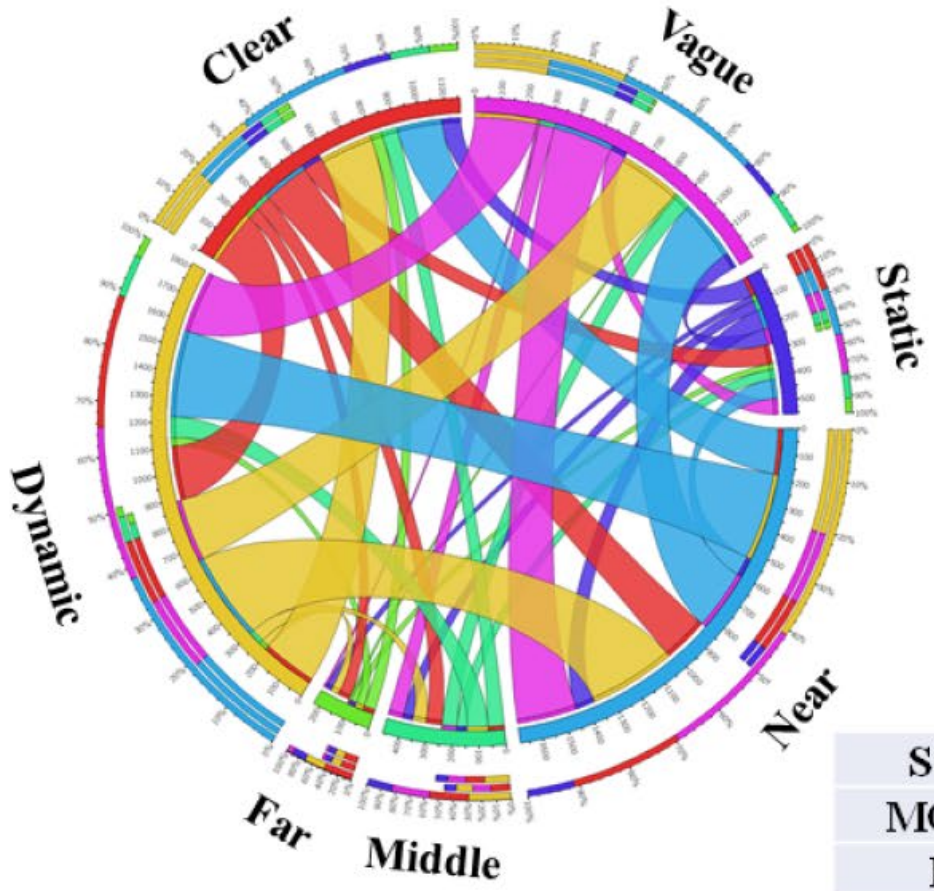
Static & Dynamic



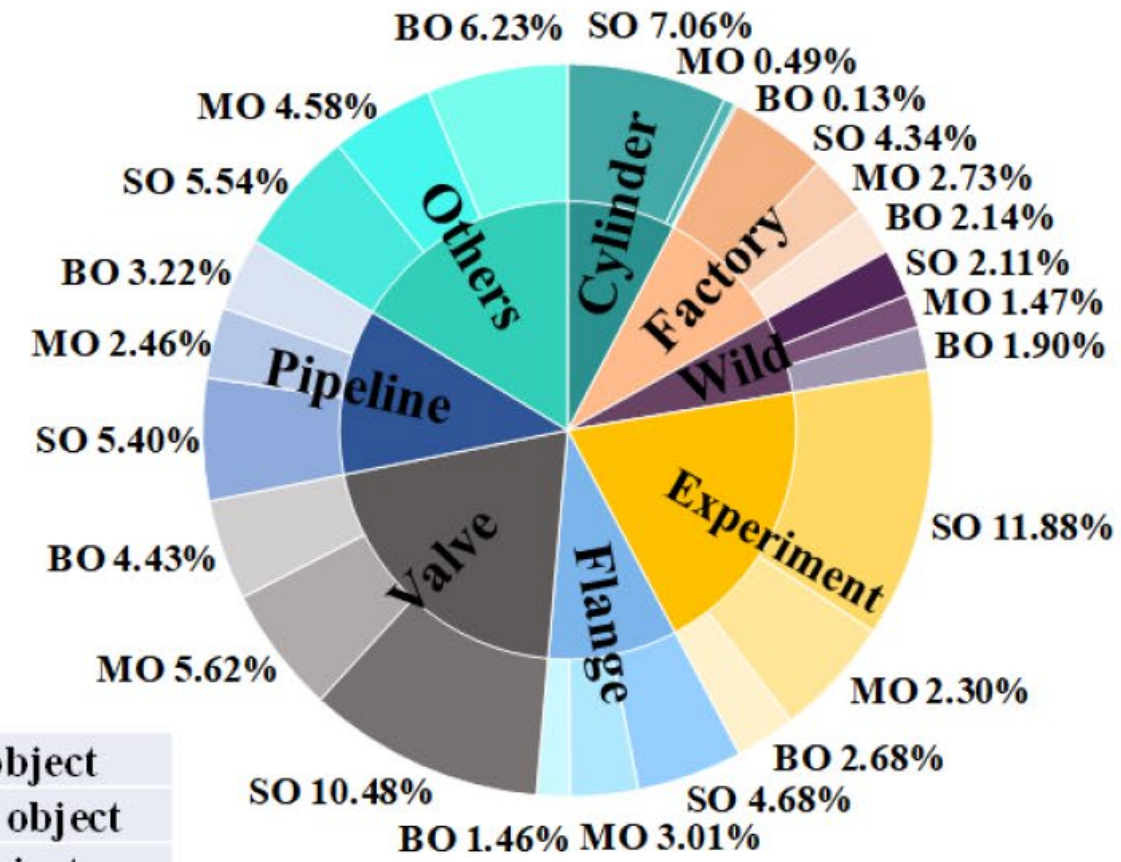

CVPR JUNE 19-24 2022
NEW ORLEANS • LOUISIANA

IOD-Video Dataset

IOD-Video Dataset

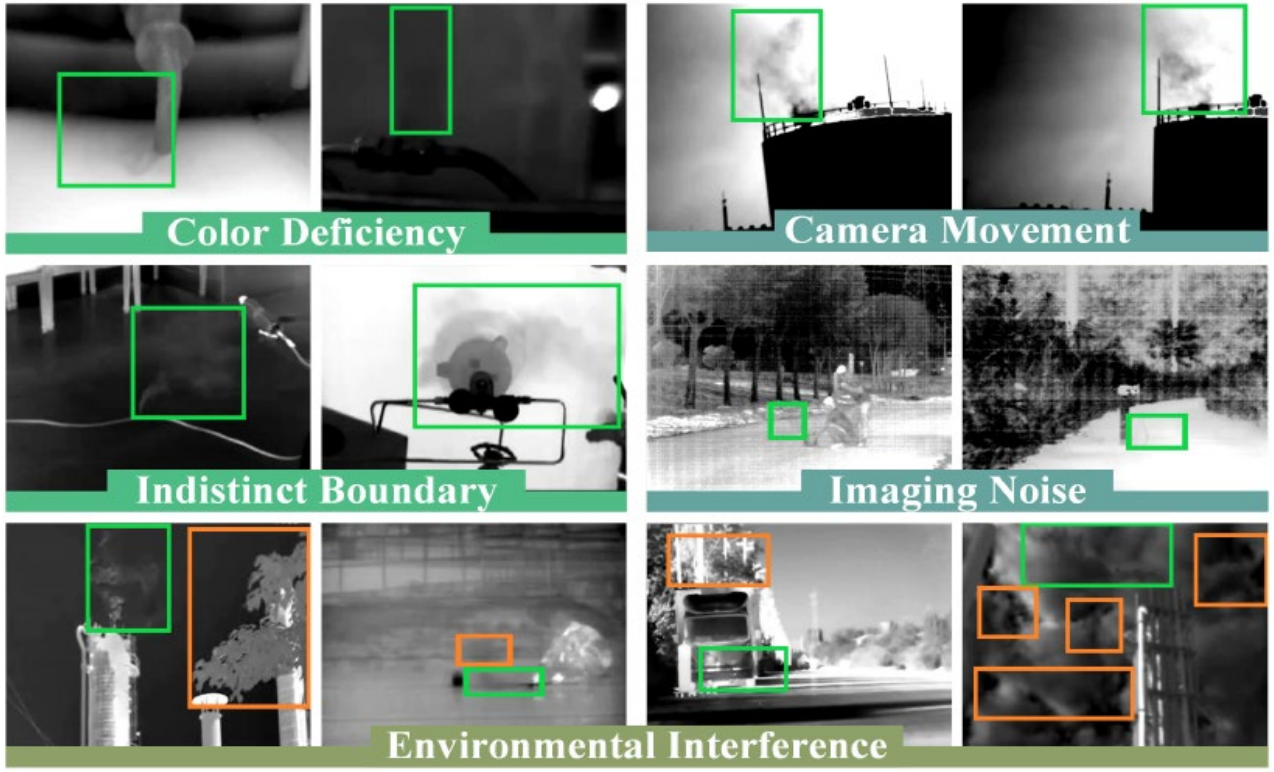


SO: small object
MO: middle object
BO: big object

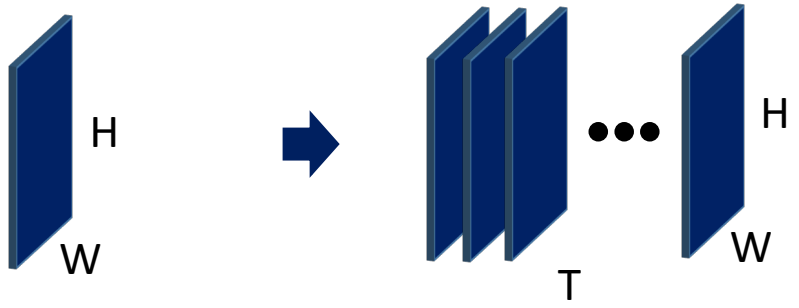


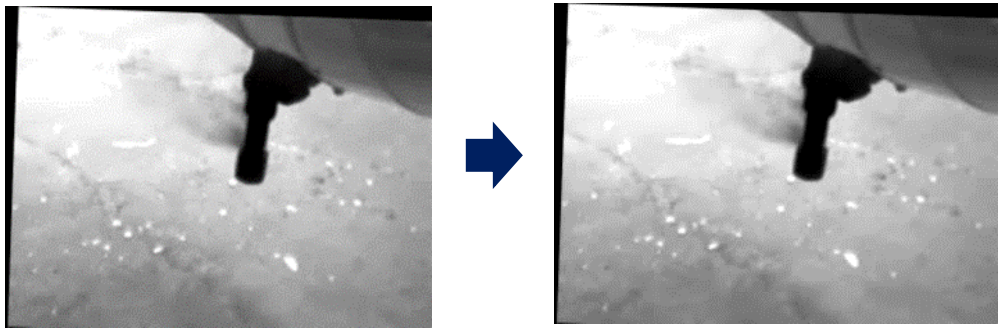
Multi-dependencies among IOD-Video attributes

Challenge



Solution





- insubstantial characteristics: color absence, indistinct boundary
- photography restrictions: camera movement, imaging noise
- environmental interference

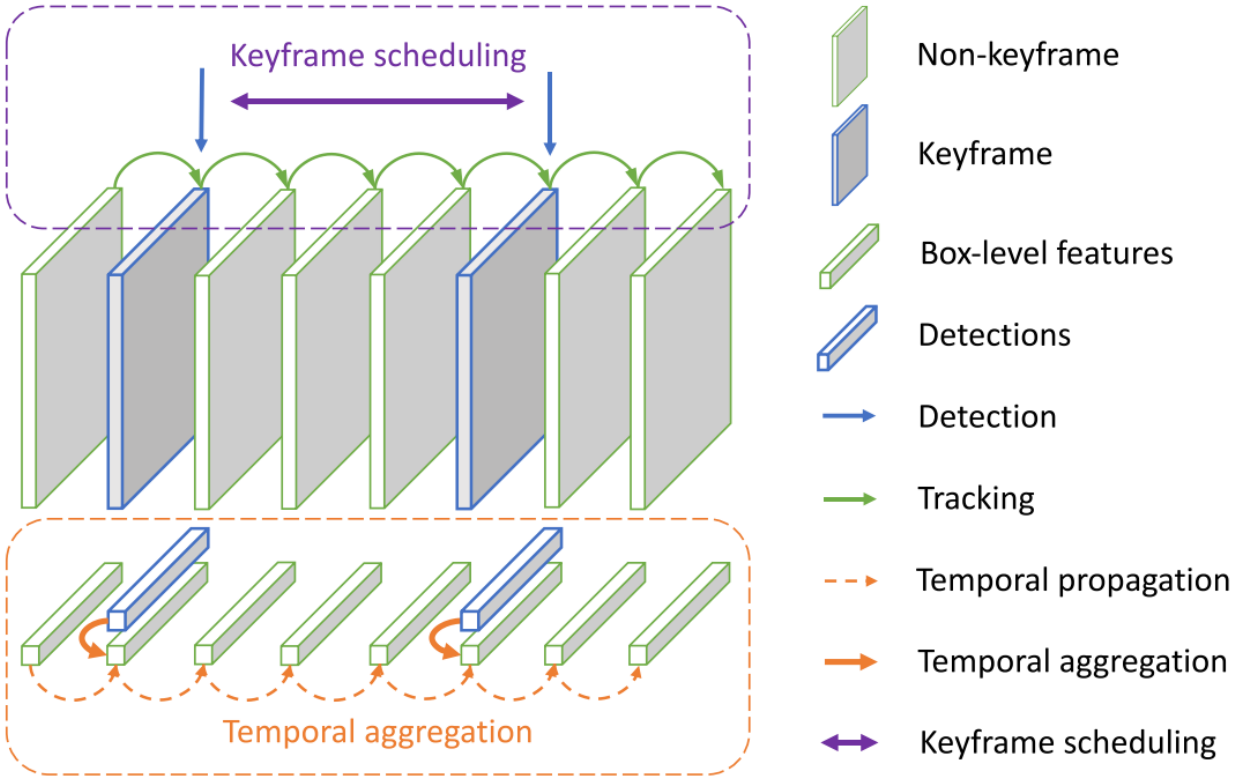
Deficiency in spatial features->Spatio-temporal Aggregation

Relate Work

Video Object Detection

Aim to address feature degradation
e.g., motion blur, occlusion, and defocus

- Track-based post-processing
- Flow-guided feature aggregation
- Memory-guided feature aggregation
- Object level feature aggregation

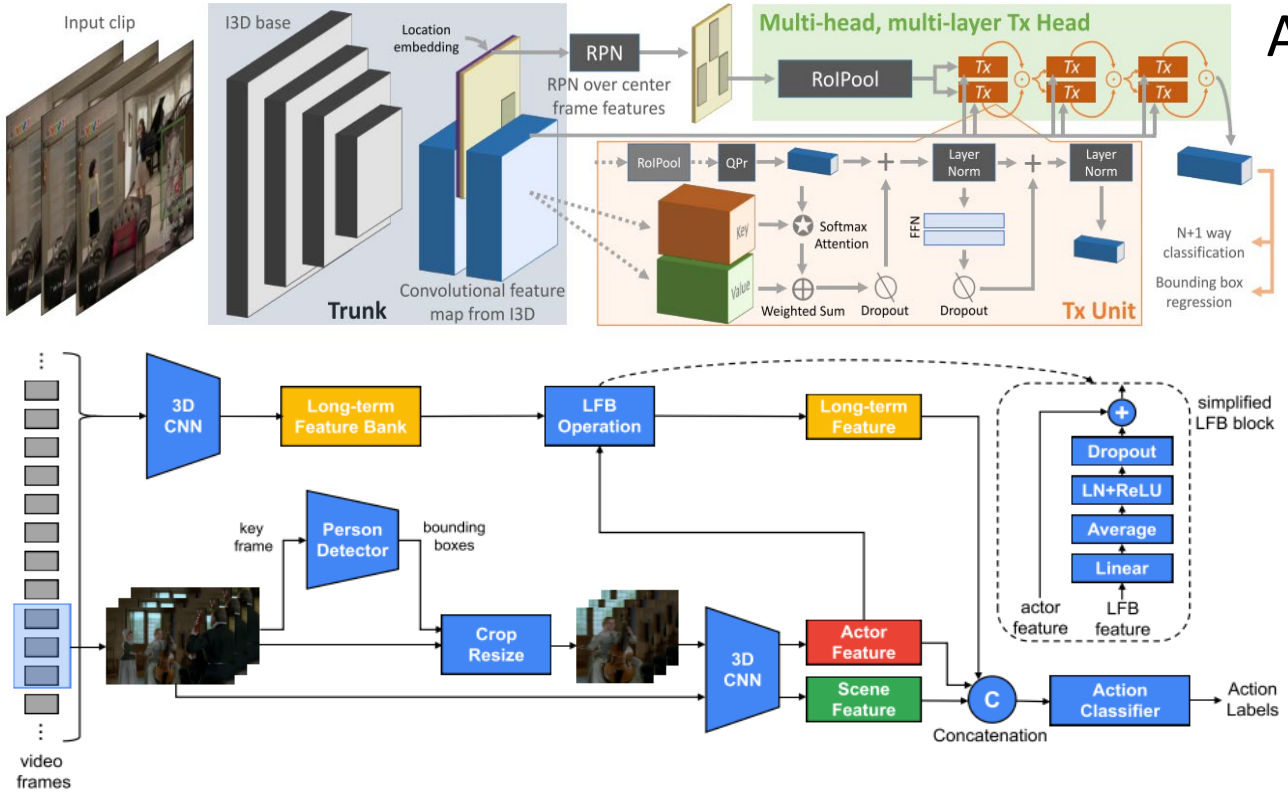


Propagate the rich information from key to non-key frame

[1] Yao, Chun-Han, et al. "Video object detection via object-level temporal aggregation." ECCV2020.

Relate Work

Spatio-temporal Action Detection



Aim to localize and recognize human actions

Action Transformer Network [1]

- Feature extraction: I3D
- Recognition: Action Transformer units

Context – aware RCNN [2]

- Feature extraction: 2D-CNN
- Recognition: I3D

Lack analysis of different 3D-CNN's detection capability

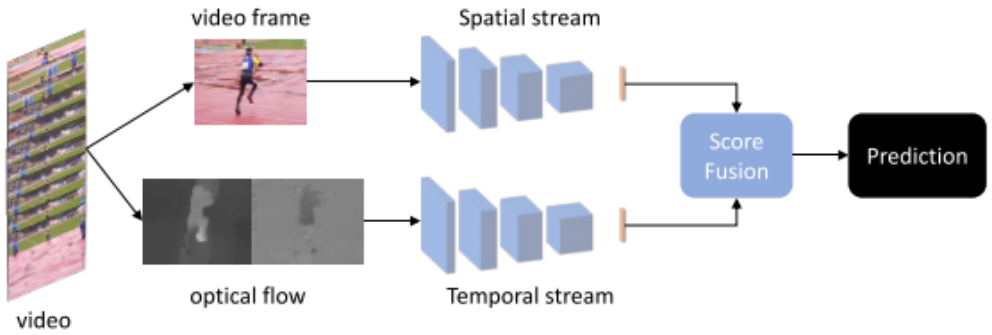
[1] Girdhar, Rohit, et al. "Video action transformer network." ICCV2019.

[2] Wu, Jianchao, et al. "Context-aware rcnn: A baseline for action detection in videos." ECCV2020.

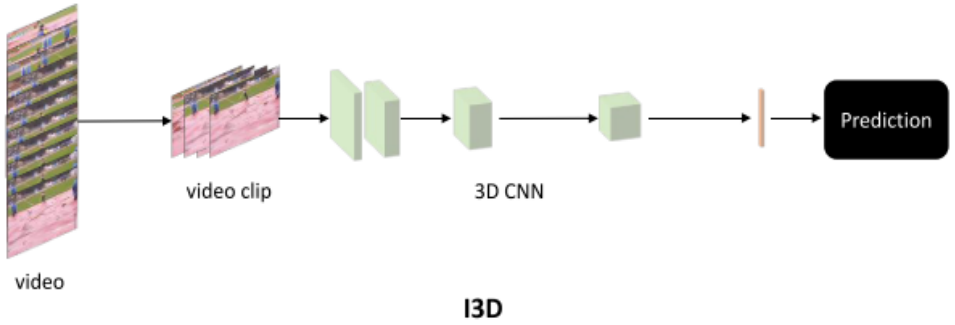
Relate Work

Action Recognition

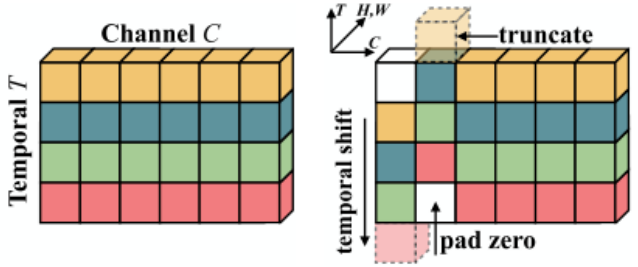
- Two-stream networks



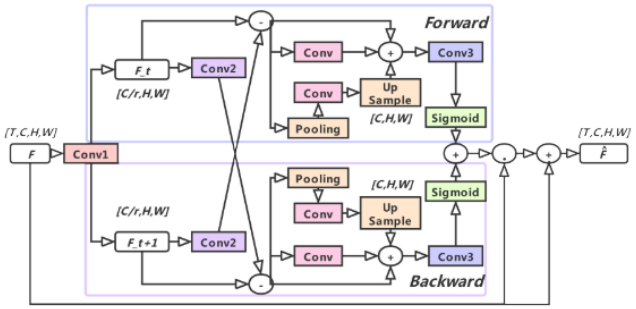
- 3D Convolution



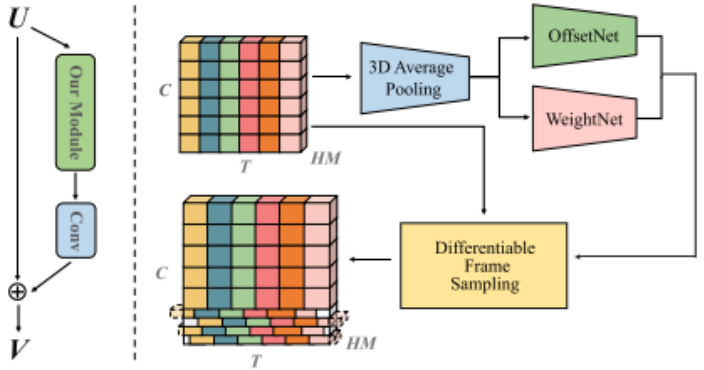
- Compute-effective 2D-CNNs



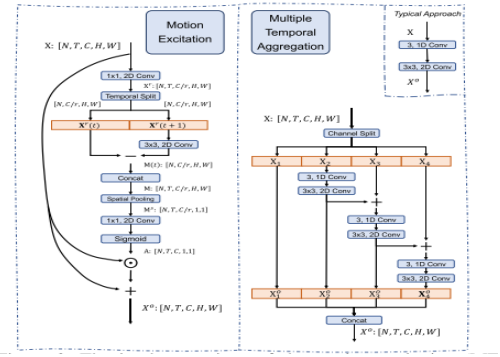
TSM



TDN



TIN



TEA

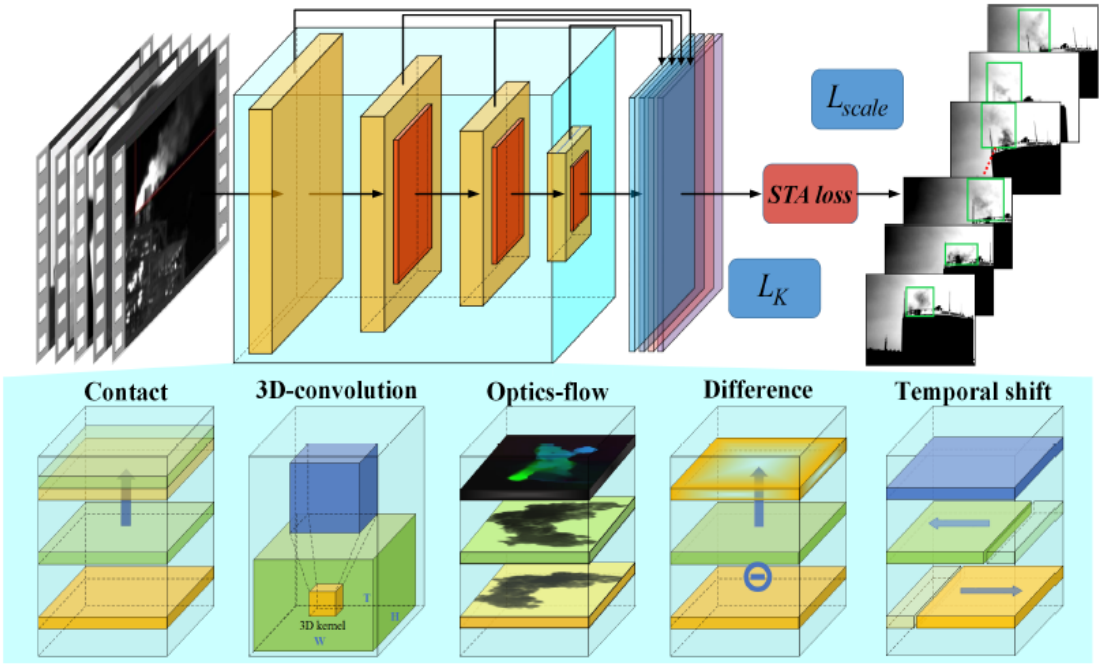
Study of 3D-CNN design focuses on video classification

STA Framework

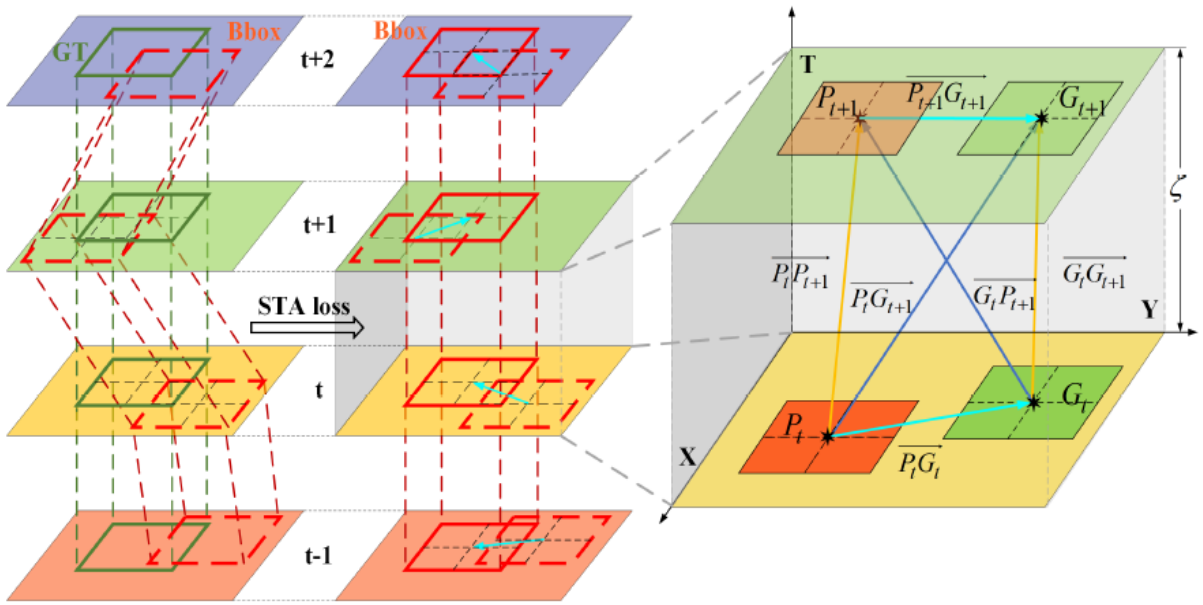
Spatio-temporal Aggregation (STA) Framework

We construct the STA Framework based on CenterNet from two aspects:

- backbone: different action recognition methods are inserted as backbones
- loss: the STAloss is specifically designed to leverage the temporal consistency



(a) Spatio-temporal Aggregation framework



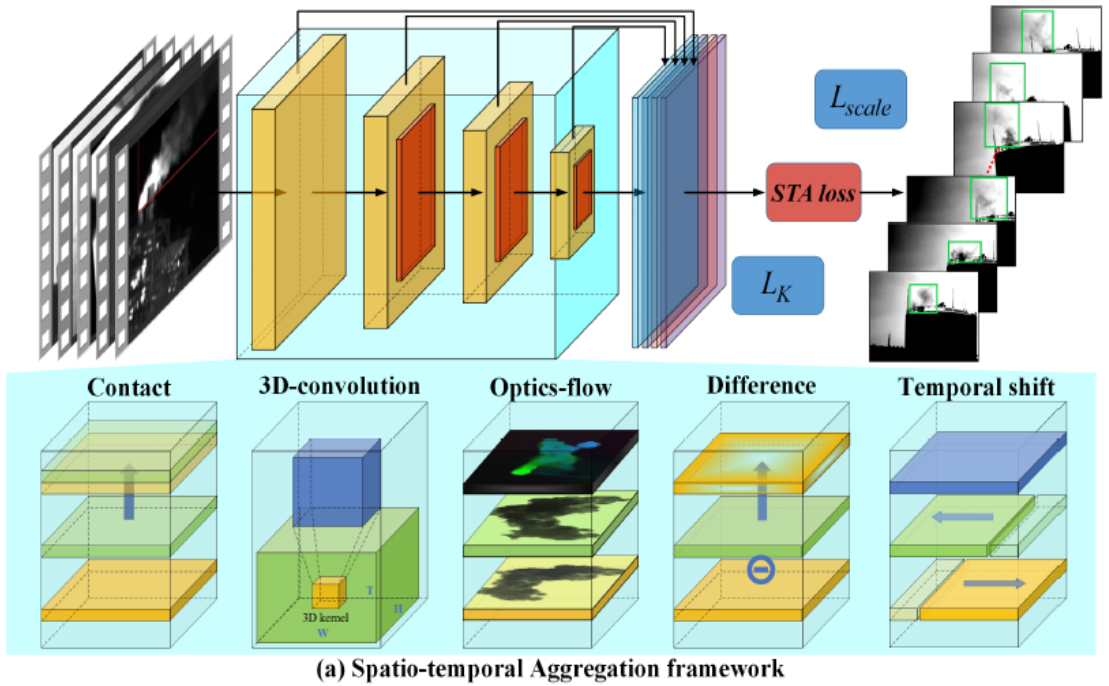
(b) Spatio-temporal aggregation loss (STA loss)

STA Framework

Fair evaluation of action recognition model's detection capability

Method	Pre-train	Backbone	Frames×Views	Venue	V1 Top1	V2 Top1
TSN [218]	I	BN-Inception	8×1	ECCV 2016	19.7	-
I3D [14]	I,K	ResNet50-like	32×6	CVPR 2017	41.6	-
NL I3D [219]	I,K	ResNet50-like	32×6	CVPR 2018	44.4	-
NL I3D + GCN [220]	I,K	ResNet50-like	32×6	ECCV 2018	46.1	-
ECO [283]	K	BNIncep+ResNet18	16×1	ECCV 2018	41.4	-
TRN [269]	I	BN-Inception	8×1	ECCV 2018	42.0	48.8
STM [92]	I	ResNet50-like	8×30	ICCV 2019	49.2	-
STM [92]	I	ResNet50-like	16×30	ICCV 2019	50.7	-
TSM [128]	K	ResNet50	8×1	ICCV 2019	45.6	59.1
TSM [128]	K	ResNet50	16×1	ICCV 2019	47.2	63.4
bLVNet-TAM [43]	I	BLNet-like	8×2	NeurIPS 2019	46.4	59.1
bLVNet-TAM [43]	I	BLNet-like	16×2	NeurIPS 2019	48.4	61.7
TEA [122]	I	ResNet50-like	8×1	CVPR 2020	48.9	-
TEA [122]	I	ResNet50-like	16×1	CVPR 2020	51.9	-
TSM + TPN [248]	K	ResNet50-like	8×1	CVPR 2020	49.0	62.0
MSNet [110]	I	ResNet50-like	8×1	ECCV 2020	50.9	63.0
MSNet [110]	I	ResNet50-like	16×1	ECCV 2020	**52.1**	**64.7**
TIN [182]	K	ResNet50-like	16×1	AAAI 2020	47.0	60.1
TEINet [132]	I	ResNet50-like	8×1	AAAI 2020	47.4	61.3
TEINet [132]	I	ResNet50-like	16×1	AAAI 2020	49.9	62.1

Table 3. **Results of widely adopted methods on Something-Something V1 and V2 datasets**. We only report numbers without using



(a) Spatio-temporal Aggregation framework

SOTA models on the **motion-focus** dataset (Something-Something V1 & V2) are inserted into the STA framework as spatio-temporal backbones

Video classification -> Video level detection evaluation

[1] Zhu, Yi, et al. "A comprehensive study of deep video action recognition." arXiv preprint arXiv:2012.06567

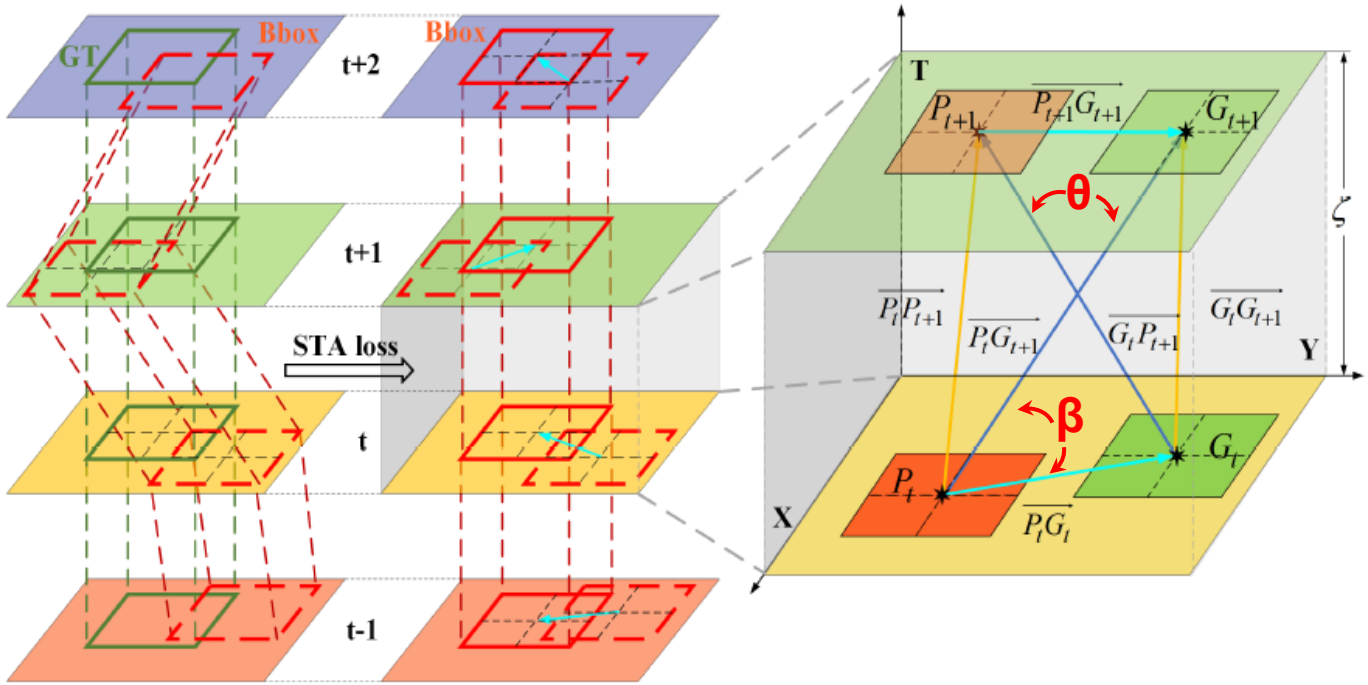
STA Framework

STAloss motivation: four predicted boxes has the same IoU with GT boxes, nevertheless, they are staggered in the time axis.

Gt boxes

Adjusted boxes

Predicted boxes



(b) Spatio-temporal aggregation loss (STA loss)

STAloss goal:
min θ ->0° max β-> 90°

$$L_{STA\cos\theta}^{cross} = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{\overrightarrow{G_tP_{t+1}}\cdot\overrightarrow{P_tG_{t+1}}}{\left|\overrightarrow{G_tP_{t+1}}\right|\left|\overrightarrow{P_tG_{t+1}}\right|}$$

$$L_{STA\cos\theta}^{self} = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{\overrightarrow{P_tP_{t+1}}\cdot\overrightarrow{G_tG_{t+1}}}{\left|\overrightarrow{P_tP_{t+1}}\right|\left|\overrightarrow{G_tG_{t+1}}\right|}$$

$$L_{STA\sin\beta}^{pre} = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{\left|\overrightarrow{P_tG_t}\right|}{\left|\overrightarrow{P_tG_{t+1}}\right|}$$

$$L_{STA\sin\beta}^{next} = \frac{1}{T-1}\sum_{t=1}^{T-1}\frac{\left|\overrightarrow{P_{t+1}G_{t+1}}\right|}{\left|\overrightarrow{P_{t+1}G_t}\right|}$$

Effect: pull the predicted boxes of each frame along the temporal axis together.

Impose constraints in the spatio-temporal space (2D->3D)

Experiment

	Method	Base Backbone	Pretrained Model	Frame AP (%)				
				@0.5	@0.75	@clear	@vague	0.5:0.95
Frame-based Detector	Faster RCNN [58]	ResNet-50	ImageNet	33.49	6.76	16.52	8.82	12.31
	SSD [50]	ResNet-50	ImageNet	30.21	3.95	12.78	7.82	9.99
	CenterNet [87]	ResNet-50	ImageNet	24.80	4.32	11.21	6.65	8.50
Video-based Detector	CRCNN [76]	ResNet101&FPN	ImageNet	36.15	7.46	18.84	9.14	13.52
	MOC [47]	DLA-34	COCO	36.81	**8.94**	19.96	9.62	10.96
	MOC + Flow [47]	DLA-34	COCO	34.50	7.28	18.18	8.49	9.41
Ours	TEA [46]	ResNet-50	ImageNet	**42.19**	8.66	**22.97**	**9.95**	**15.69**

Table 1. Comparisons with previous frame-based detectors and video-based detectors. Our basic spatio-temporal aggregation framework simply replaces the backbone of CenterNet with TEA [46] without any other complex design.

- Our STA Framework simply replaces the backbone of CenterNet with TEA
- It achieves overall better performances which suggest the spatio-temporal aggregation is crucial

Experiment

	Spatio-temporal Backbone	Base Backbone	Frame AP (%)				
			@0.5	@0.75	@clear	@vague	0.5:0.95
Concat	Concat [47]	ResNet-50	27.41	4.45	12.38	7.32	9.32
3D-Convolution	S3D [78]	ResNet-50	35.82	6.73	17.72	8.81	12.72
	I3D [7]	ResNet-50	36.83	7.39	18.78	9.29	13.43
Flow-based	MSNet [43]	ResNet-50	41.19	7.91	21.35	10.01	14.90
Difference	TDN [70]	ResNet-50	41.69	8.48	21.42	10.46	15.40
Temporal Shift	TSM [48]	ResNet-50	42.13	8.20	21.98	10.28	15.38
	TAM [18]	ResNet-50	41.95	8.53	21.49	10.61	15.50
	TIN [61]	ResNet-50	42.77	8.01	22.35	10.51	15.73
	TEA [46]	Res2Net-50	42.19	8.66	22.97	9.95	15.69
+STAloss	TIN [61]	ResNet-50	43.72	9.26	23.81	10.35	16.27
	TEA [46]	Res2Net-50	**45.08**	**9.50**	**24.43**	**10.91**	**16.99**

Table 2. The representative action recognition models are selected from the state-of-the-arts on Sth-Sth dataset [23], and we present the video-level detection performance of different spatio-temporal backbones. STAloss replaces the original offset loss L_{off} of CenterNet.

- Video-level detection capability of different action recognition models are fairly evaluated.
- Temporal shift models perform best which may preserve the spatial feature-level integrity.
- The STAloss provides a feasible way for further improvement for IOD task.

Demo

CenterNet

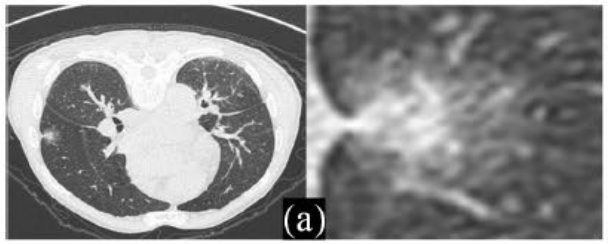
TEA+STAloss

Application

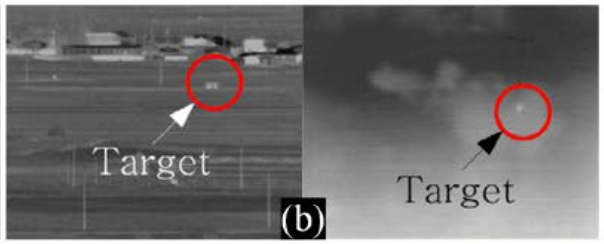
Similar Insubstantial Characteristic

Video Camouflaged Object Detection

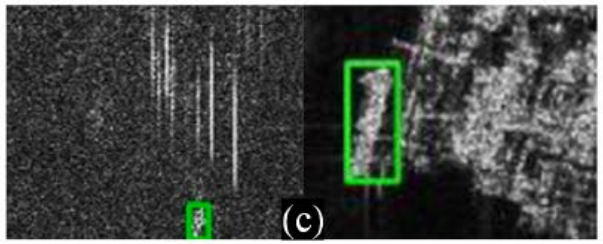
Video Smoke Detection



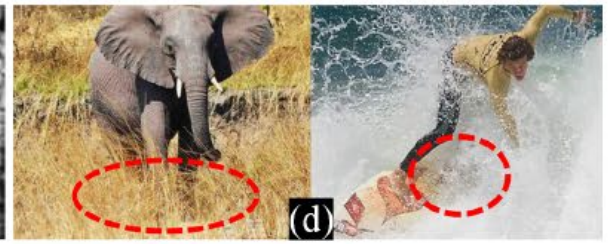
Pulmonary Nodule Detection



Infrared Dim Small Object Detection



Synthetic Aperture Radar (SAR) detection



Partially-occluded Targets in RGB Images

Conclusion

- We propose an IOD-Video dataset for insubstantial object detection to promote research on this challenging task.

- We develop a spatio-temporal aggregation framework in which the video-level detection capability of representative action recognition backbones can be fairly evaluated.

- Based on the temporal shift backbone which achieves best performance, the STAloss is specifically designed to leverage the temporal consistency for further improvement.

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

南京大学 NANJING UNIVERSITY

Thanks for your attention!

Computational Imaging Lab @ Nanjing University

Project Page **Github Code** **CITE Lab**

Acknowledgements: we thank the ***ZHIPUTECH*** for the device supports and data collection.