Object Detection with Self-Supervised Scene Adaptation

Zekun Zhang[1], Minh Hoai[1,2]

[1]Stony Brook University

[2]VinAI Artificial Intelligence Application and Research

Stony Brook University

VinAI RESEARCH

Poster # THU-PM-094

JUNE 18-22, 2023

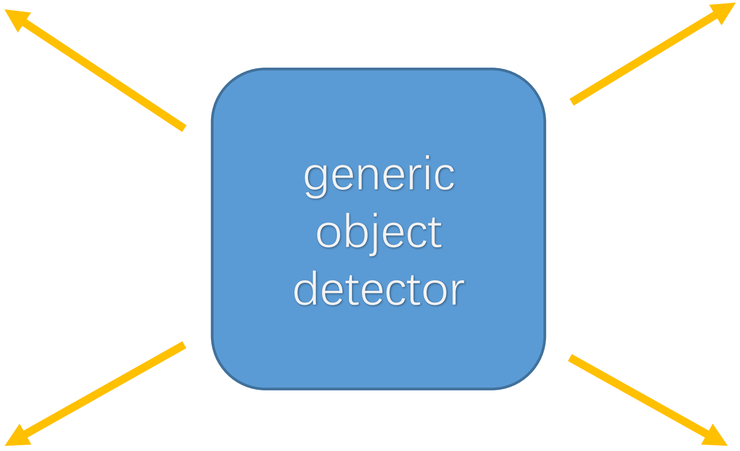
CVPR

VANCOUVER, CANADA

Object Detection with Scene Adaptation

Jackson Hole Town Square @ Pizzeria Caldera ©SeeJH.com 09/10/2020 03:06:21 PM

traffic monitoring


generic object detector

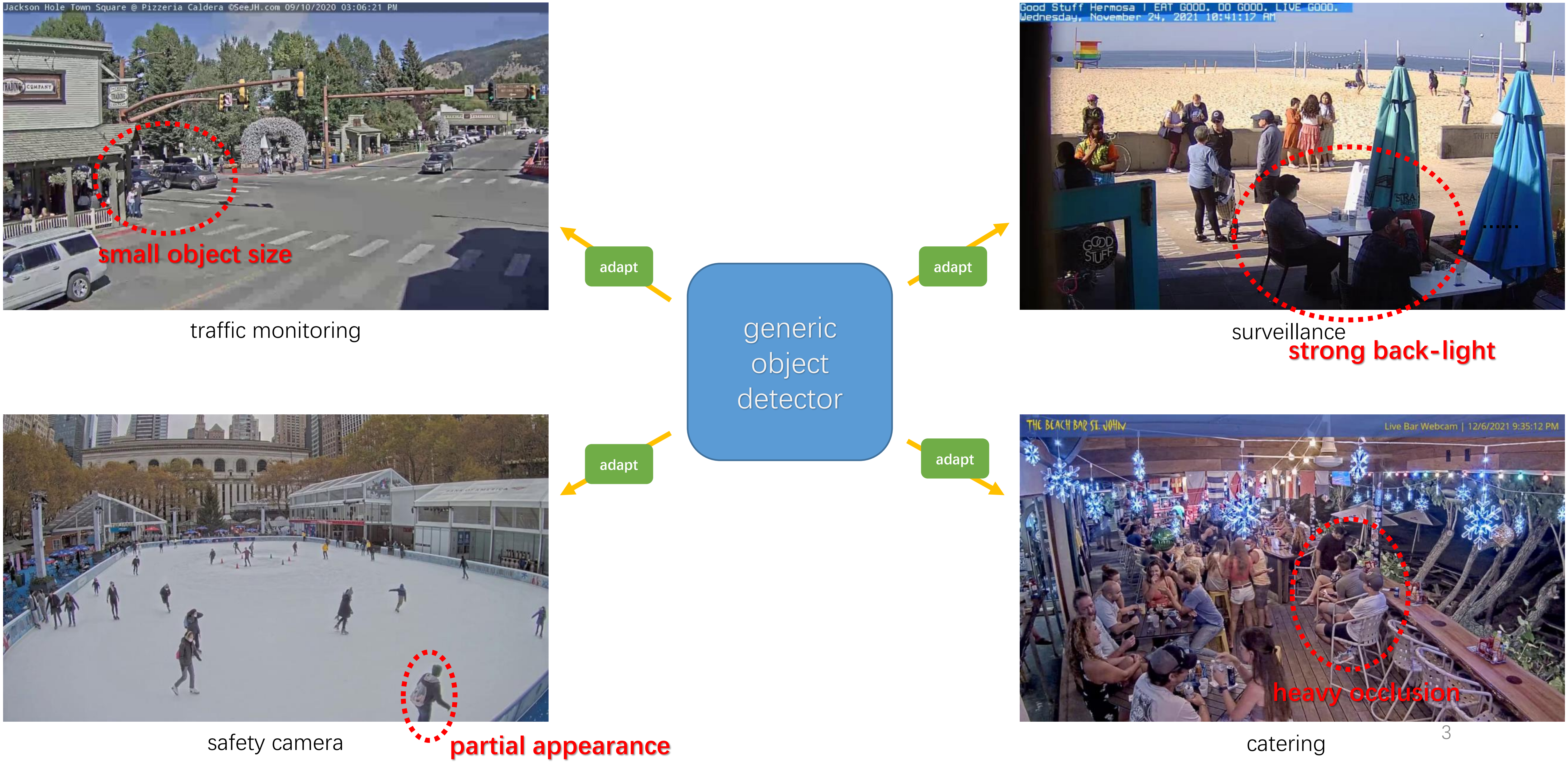
Good Stuff Hermosa | EAT GOOD. DO GOOD. LIVE GOOD.
Wednesday, November 24, 2021 10:41:17 AM
GOOD STUFF

surveillance

safety camera

catering

Object Detection with Scene Adaptation



Jackson Hole Town Square @ Pizzeria Caldera ©SeeJH.com 09/10/2020 03:06:21 PM
Good Stuff Hermosa | EAT GOOD. DO GOOD. LIVE GOOD.
Wednesday, November 24, 2021 10:41:17 AM
THE BEACH BAR ST. JOHN
Live Bar Webcam | 12/6/2021 9:35:12 PM
adapt
adapt
adapt
adapt

generic object detector

small object size

traffic monitoring

strong back-light

surveillance

partial appearance

safety camera

heavy occlusion

catering

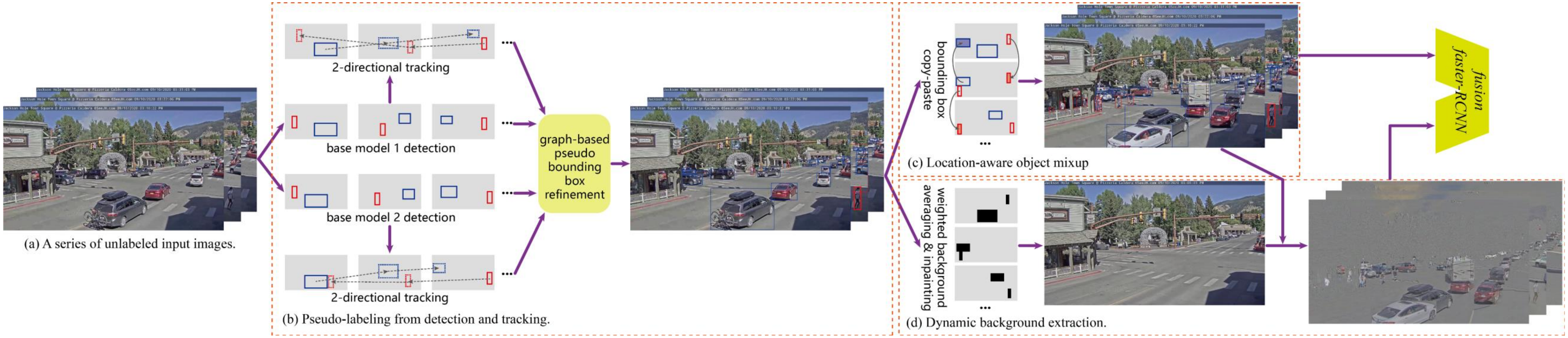
Contributions

- An object detection framework that
  - adapts to new scenes requiring no human annotation
  - utilizes stationary background and temporal correlation
- First scene adaptation object detection dataset: *Scenes100*
  - large-scale & diverse
  - long videos for training
  - annotation for evaluation

Object Detection with Scene Adaptation

- Scene adaptation is a special case of domain adaptation
  - Distribution shift causes performance drop
  - Lack of human annotation on target domain
  - Getting more attention in recent years: [RoyChowdhury *et al.* CVPR 2019], [Sohn *et al.* arXiv 2020], [Li *et al.* CVPR 2022], [Xu *et al.* CVPR 2022], [Zhao & Wang CVPR 2022], [Li *et al.* CVPR 2022], ⋯
- Uniqueness
  - Fixed camera gives stationary background
  - Temporal correlation among images
  - Less data variance harms generalization

Scene-Adaptive Object Detection Framework



(a) A series of unlabeled input images.

(b) Pseudo-labeling from detection and tracking.

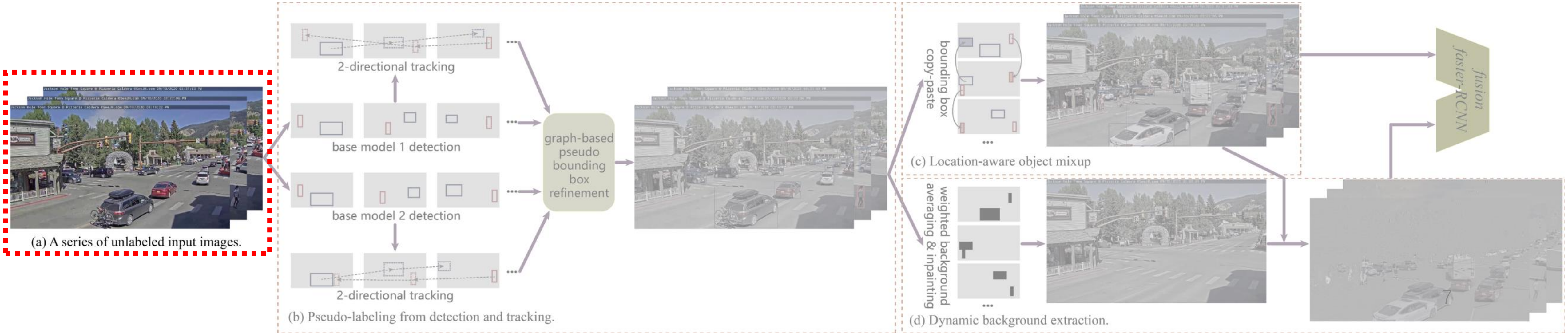
(c) Location-aware object mixup

(d) Dynamic background extraction.

Scene-Adaptive Object Detection Framework

Input

- Unlabeled video stream with fixed camera
- Multiple trained base object detectors



(a) A series of unlabeled input images.

(b) Pseudo-labeling from detection and tracking.

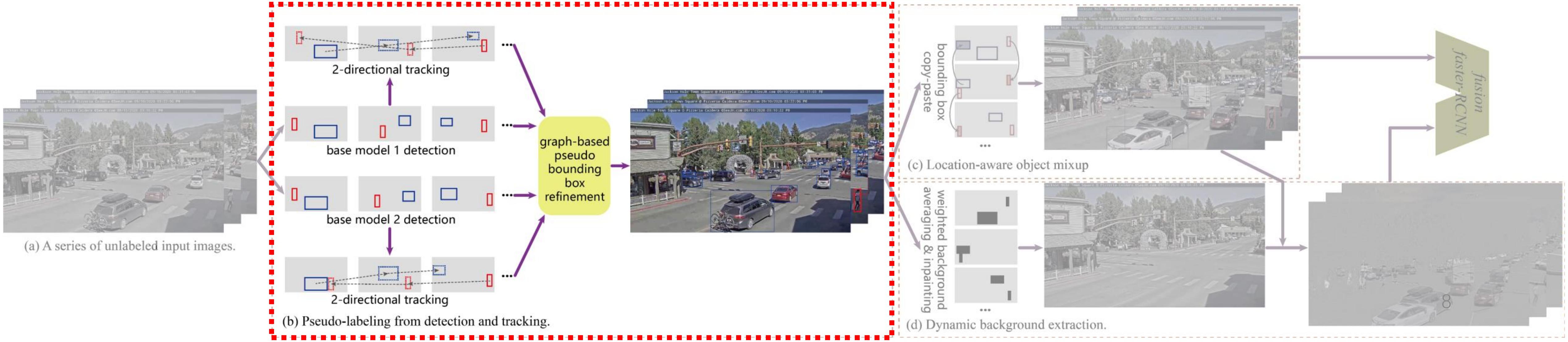
(c) Location-aware object mixup

(d) Dynamic background extraction.

Scene-Adaptive Object Detection Framework

Step 1: pseudo-labeling

- Base detectors generate bounding boxes
- 2-directional tracking initialized from detections
- Score thresholding and duplication removal



(a) A series of unlabeled input images.

(b) Pseudo-labeling from detection and tracking.

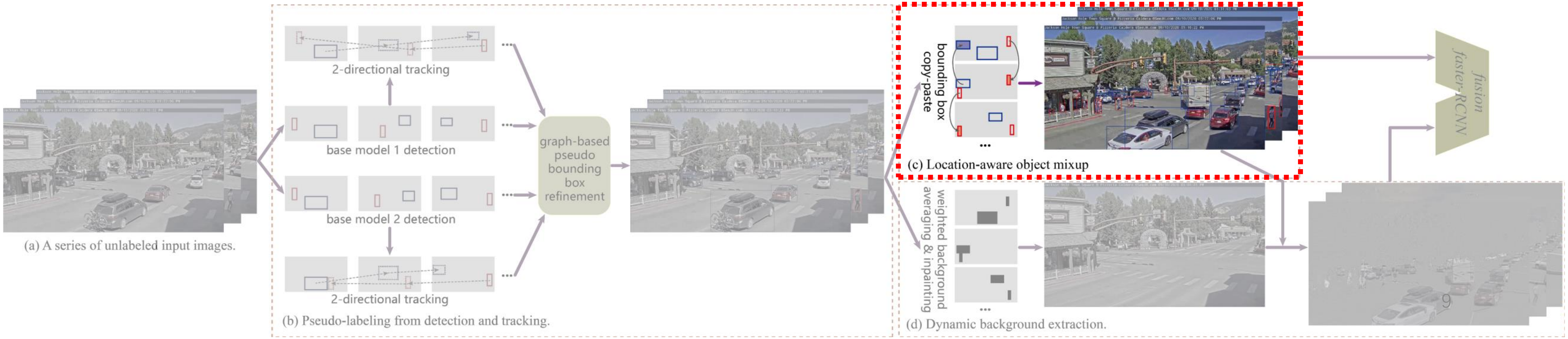
(c) Location-aware object mixup

(d) Dynamic background extraction.

Scene-Adaptive Object Detection Framework

Step 2: location-aware object mixup

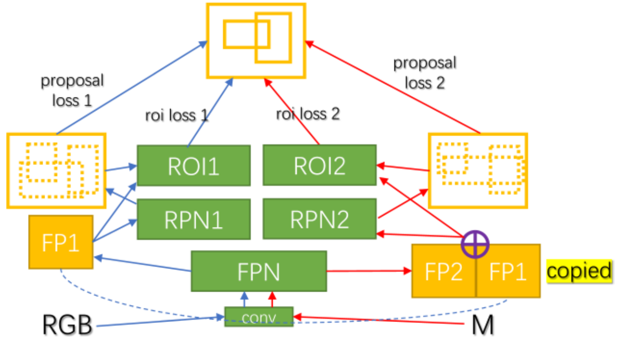
- Copy-paste detected objects while retaining positions
- Artifact-free mixed-up images improve generalization


(a) A series of unlabeled input images.
2-directional tracking
base model 1 detection
base model 2 detection
2-directional tracking
graph-based pseudo bounding box refinement
(b) Pseudo-labeling from detection and tracking.
bounding box copy-paste
...
(c) Location-aware object mixup
weighted background averaging & inpainting
...
(d) Dynamic background extraction.
fusion faster-RCNN

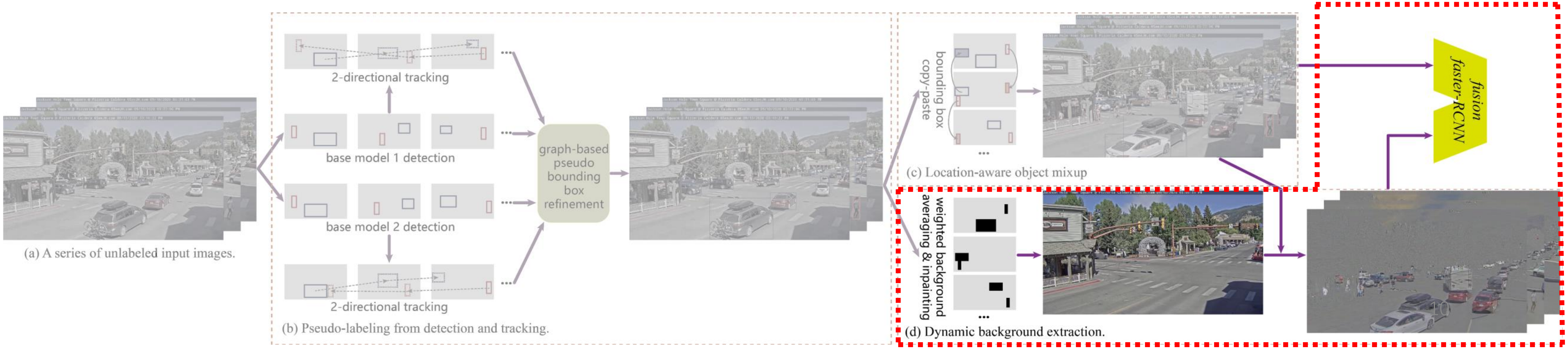
Scene-Adaptive Object Detection Framework

Step 3: dynamic background extraction & object mask fusion

- Moving average of background pixels
- Use image with background subtracted as additional input to fusion faster-RCNN



mid-fusion faster-RCNN



(a) A series of unlabeled input images.

(b) Pseudo-labeling from detection and tracking.

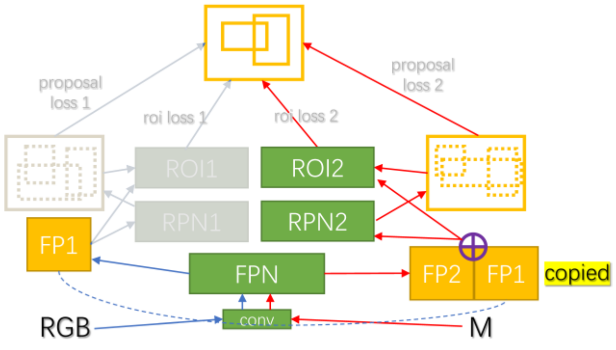
(c) Location-aware object mixup

(d) Dynamic background extraction.

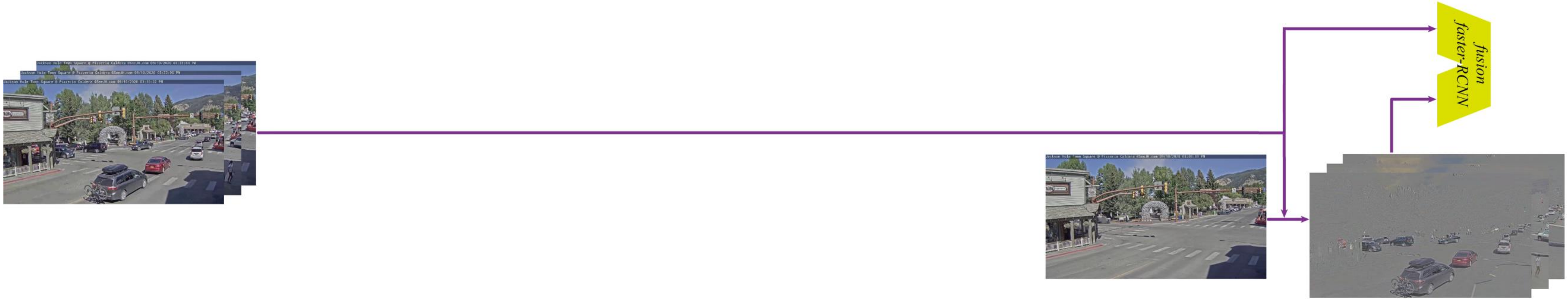
Scene-Adaptive Object Detection Framework

After training process

- Trained fusion faster-RCNN
- Latest background image



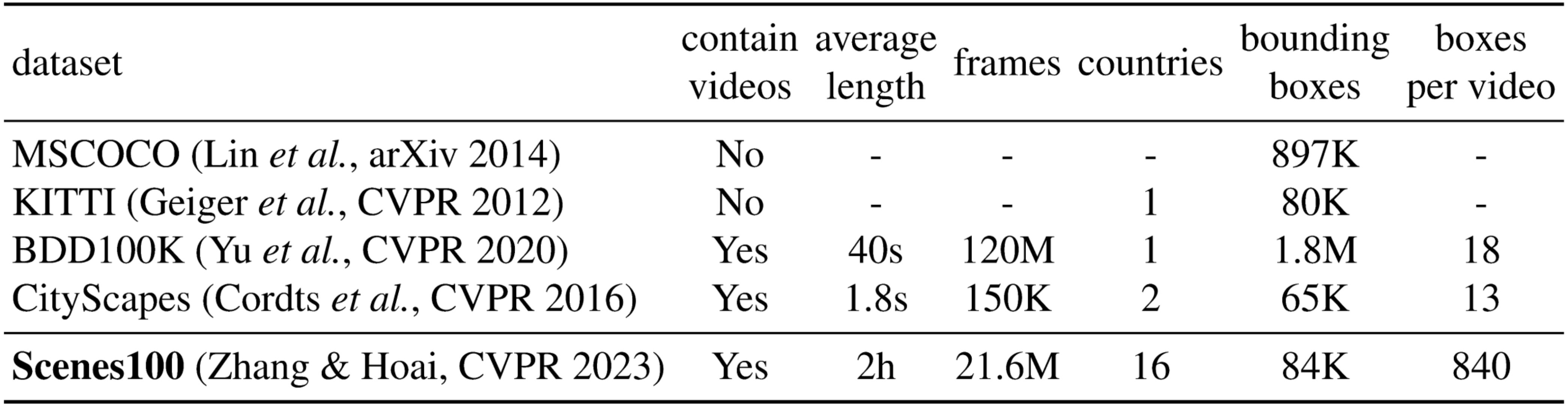
mid-fusion faster-RCNN



Scenes100 Object Detection Dataset

Scenes100 Object Detection Dataset

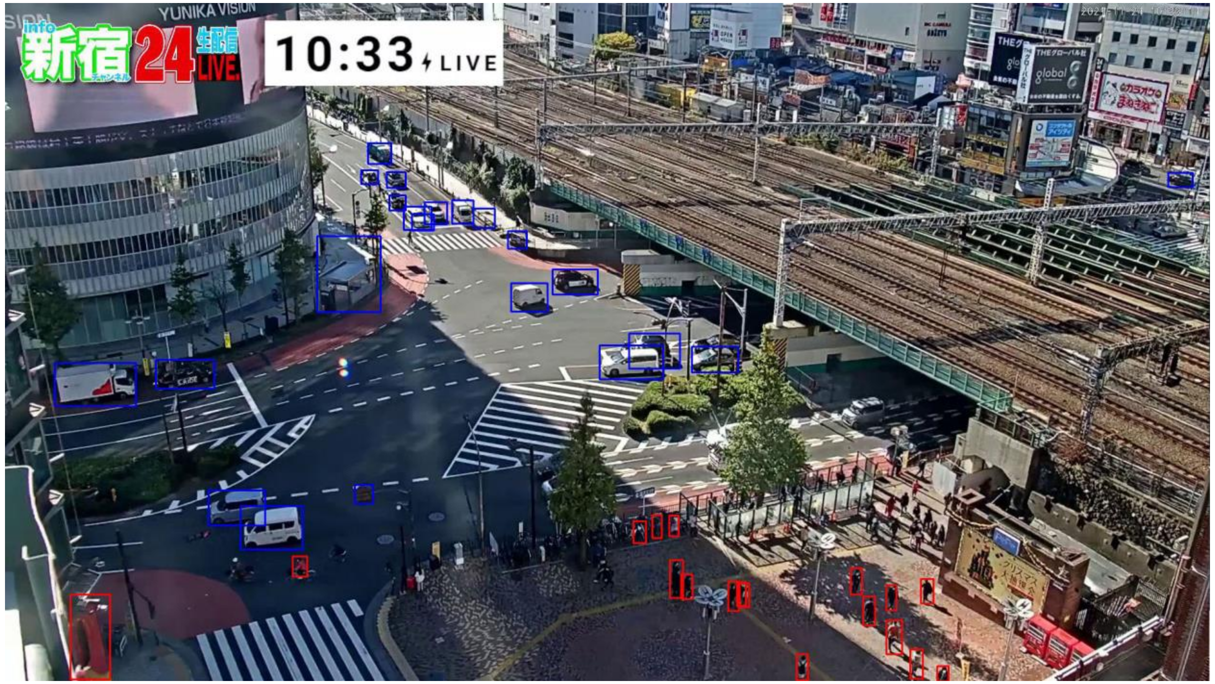
- First video object detection dataset for scene adaptation
- Large scale, diverse, long video, fixed camera
- High quality annotation for reliable performance evaluation



dataset	contain videos	average length	frames	countries	bounding boxes	boxes per video
MSCOCO (Lin *et al.*, arXiv 2014)	No	-	-	-	897K	-
KITTI (Geiger *et al.*, CVPR 2012)	No	-	-	1	80K	-
BDD100K (Yu *et al.*, CVPR 2020)	Yes	40s	120M	1	1.8M	18
CityScapes (Cordts *et al.*, CVPR 2016)	Yes	1.8s	150K	2	65K	13
Scenes100 (Zhang & Hoai, CVPR 2023)	Yes	2h	21.6M	16	84K	840

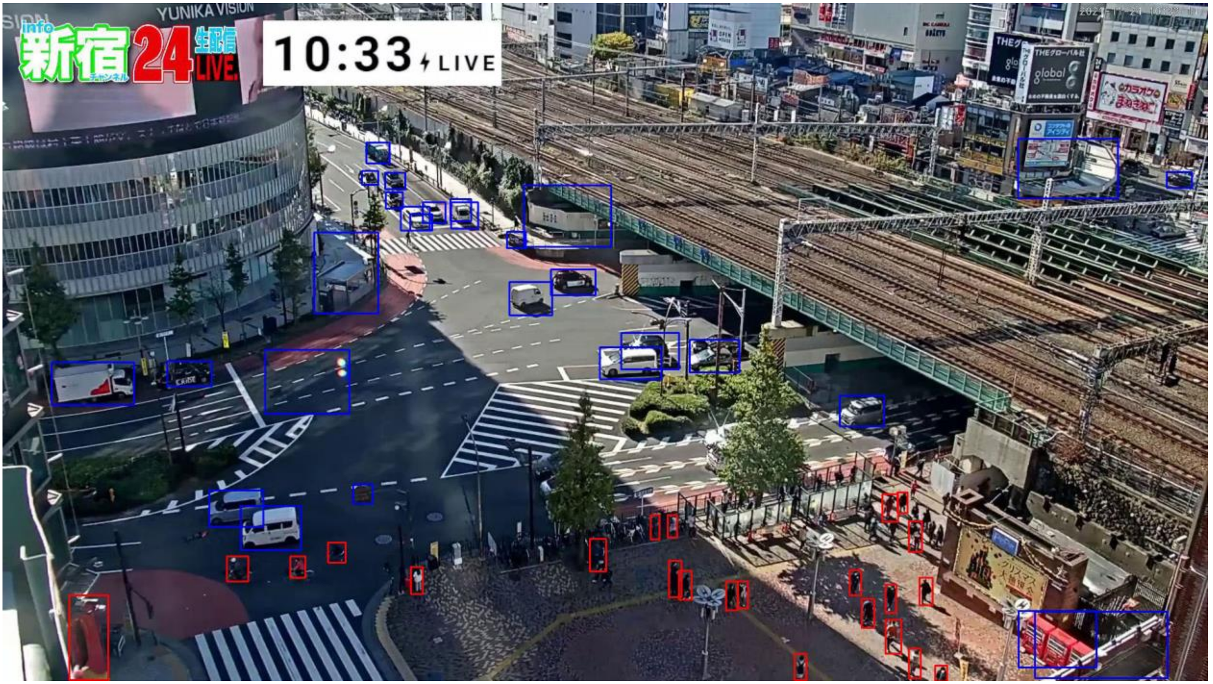
Qualitative Results

base model detection



mAP=25.13

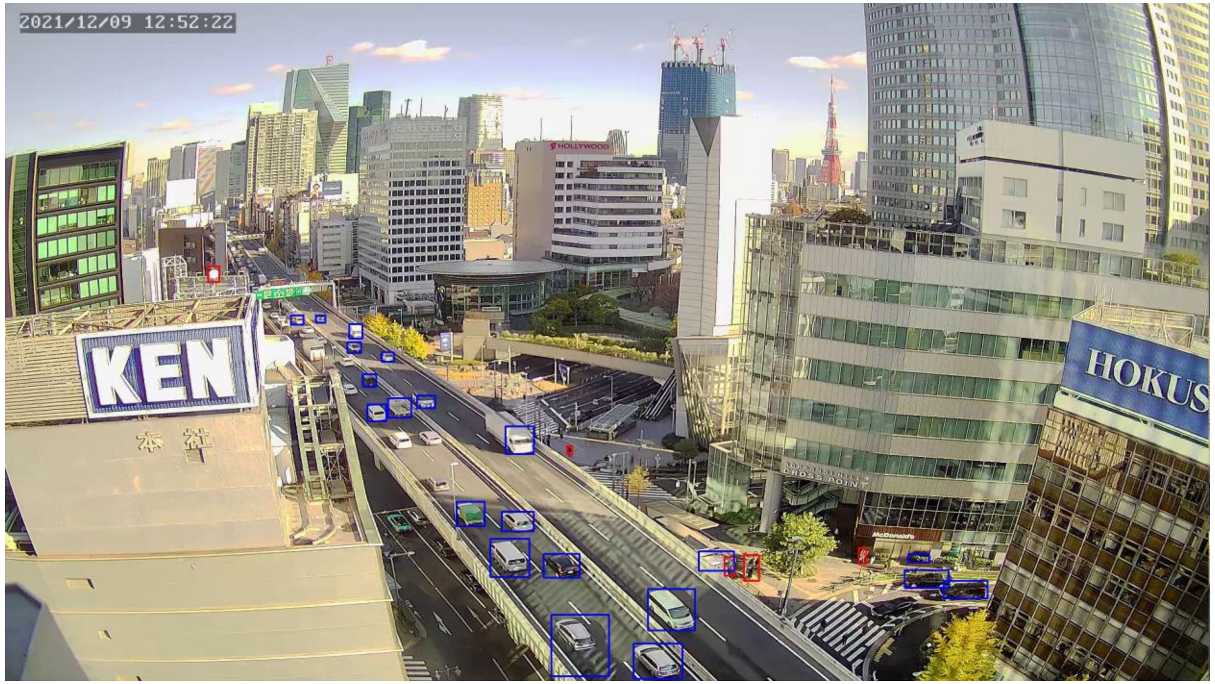
adapted model detection



mAP=33.36

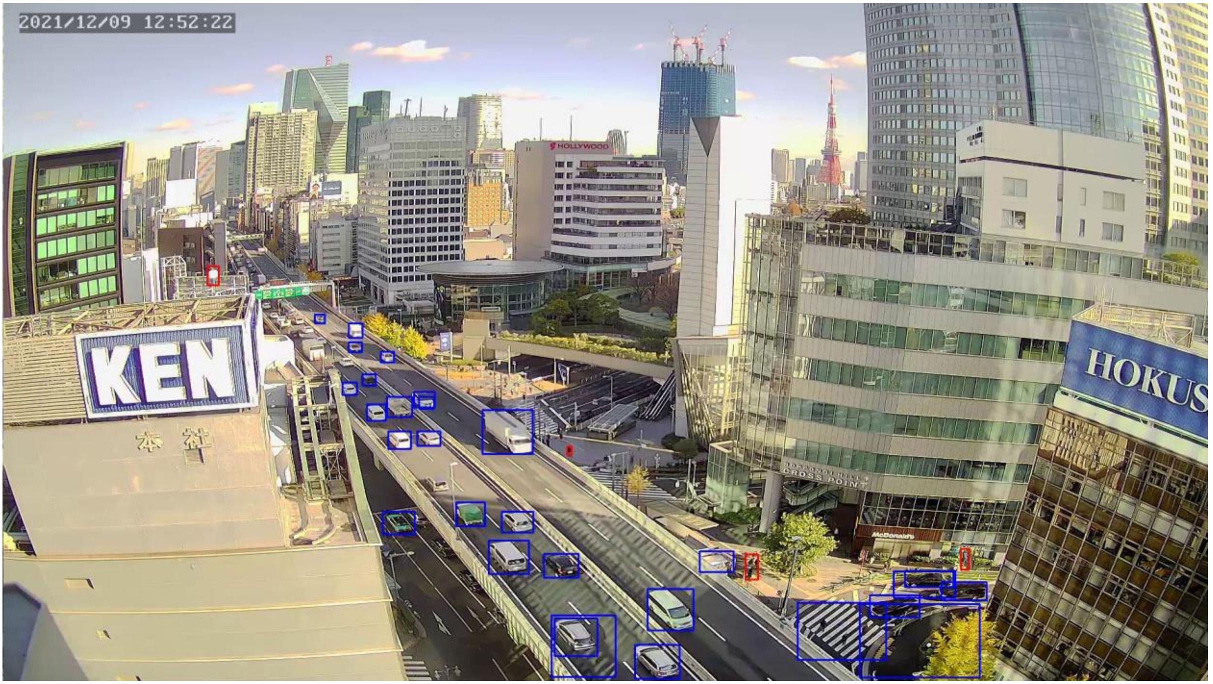
Qualitative Results

base model detection



mAP=17.59

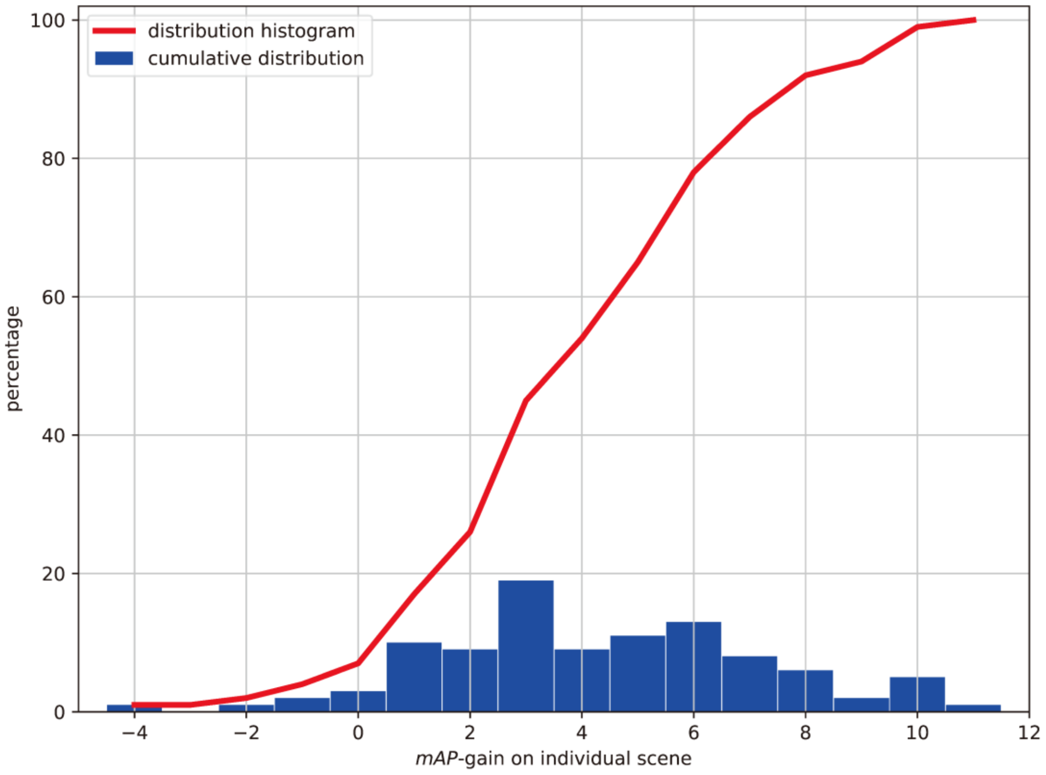
adapted model detection

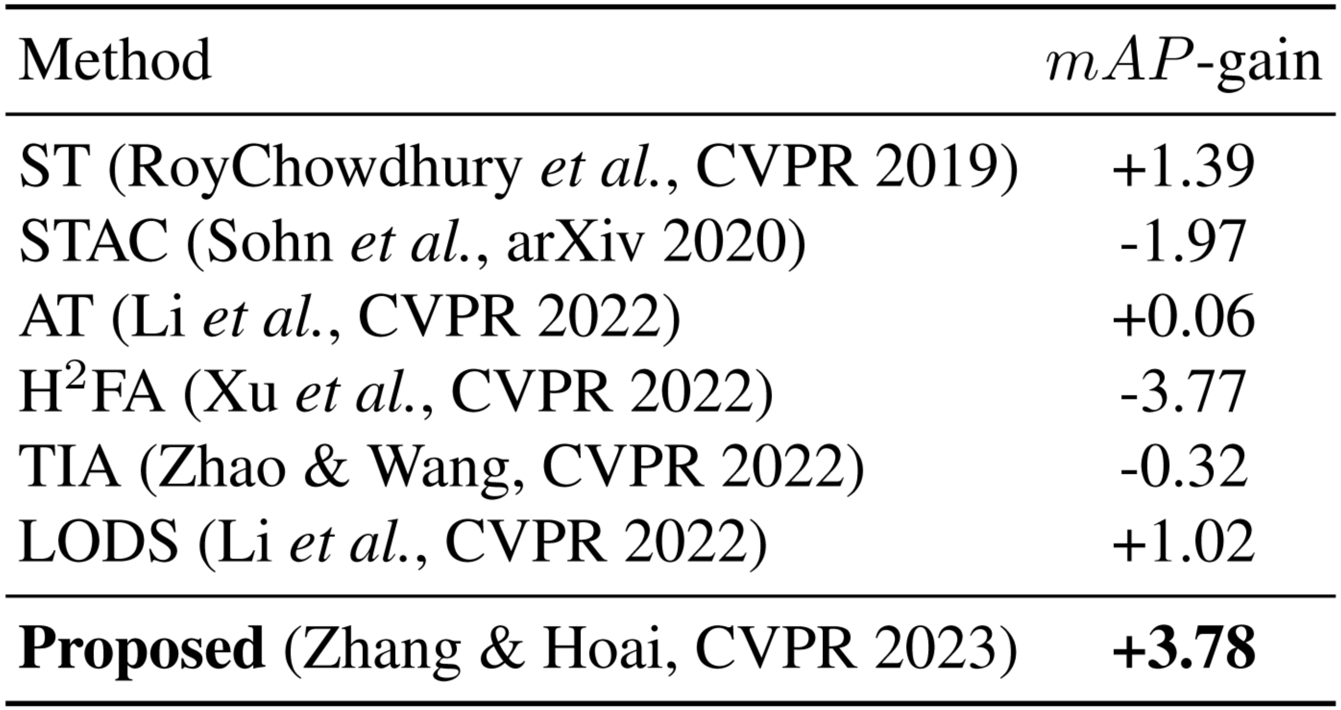


mAP=27.19

Quantitative Results

- Consistently reduces performance drop
- Significantly outperforms general domain adaptation methods



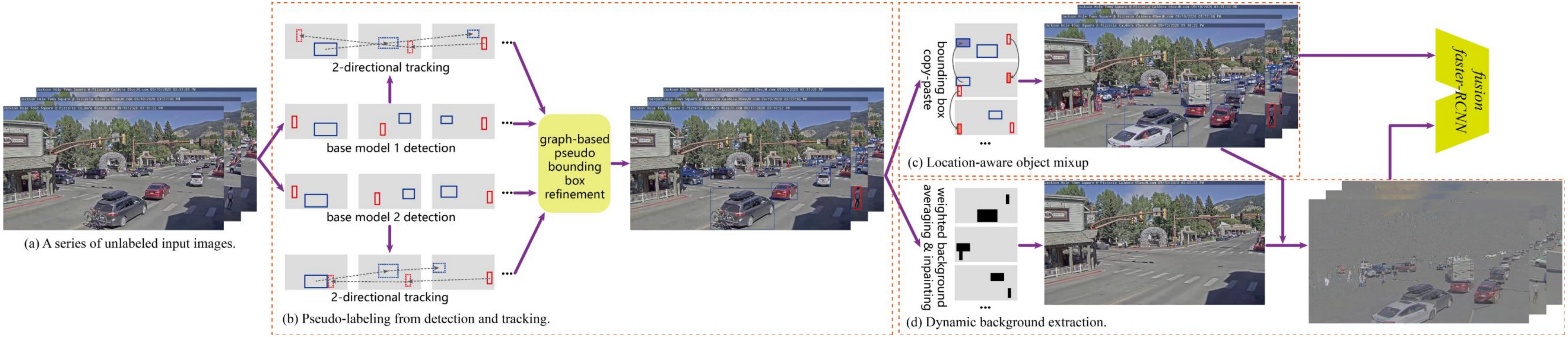


Method	mAP-gain
ST (RoyChowdhury *et al.*, CVPR 2019)	+1.39
STAC (Sohn *et al.*, arXiv 2020)	-1.97
AT (Li *et al.*, CVPR 2022)	+0.06
H^{2}FA (Xu *et al.*, CVPR 2022)	-3.77
TIA (Zhao & Wang, CVPR 2022)	-0.32
LODS (Li *et al.*, CVPR 2022)	+1.02
Proposed (Zhang & Hoai, CVPR 2023)	**+3.78**

Ablation Study

- Pseudo-labeling is essential
    - Tracking & model ensemble improve performance
- Location-aware mixup outperforms random mixup
- Object mask fusion improves performance greatly

Summary



(a) A series of unlabeled input images.

(b) Pseudo-labeling from detection and tracking.

(c) Location-aware object mixup

(d) Dynamic background extraction.

Code

&

Dataset

Talk to us on Thursday PM

at west building exhibit halls ABC 094 !