ZBS: Zero-shot Background Subtraction via Instance-level Background Modeling and Foreground Selection

by Yongqi An et al., CVPR 2023

Aniket Das

Indian Statistical Institute, Kolkata

November 2025

Course: Computer Vision
Instructor: Pradipta Maji

Agenda

1. Motivation
2. Background
3. Method
4. Experiments
5. Discussion
6. Conclusion

Why Background Subtraction?

- Extract moving objects from video frames into binary foreground masks.
- Core for surveillance, traffic analytics, robotics, and anomaly detection.
- Real-world difficulties: Shadows, illumination change, camera jitter, dynamic background, unseen object categories.

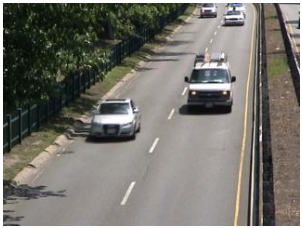


Figure: CDNet frame (left) & BGS mask (right)

Limitations of Traditional Methods

- Pixel-level BGS (SuBSENSE): Sensitive to shadows, jitter, and dynamic backgrounds.
- Supervised Deep BGS (BSUV-Net): Good accuracy, but often requires scene-specific training and fails on unseen categories.
- Unsupervised BGS (RT-SBS-v2): Misclassifies shadows and night lights as foreground.
- Need: A generalized robust method that handles unseen objects and complex scenes.

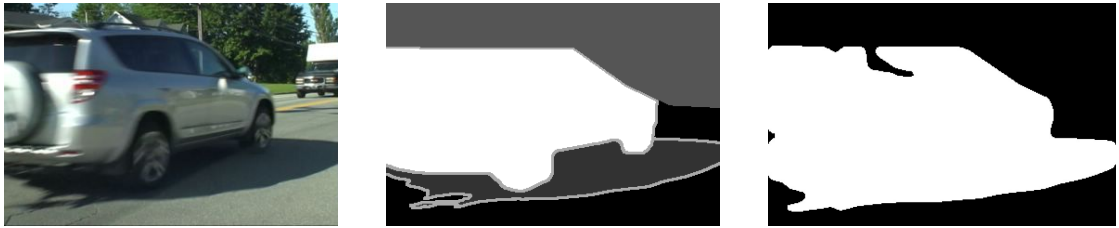
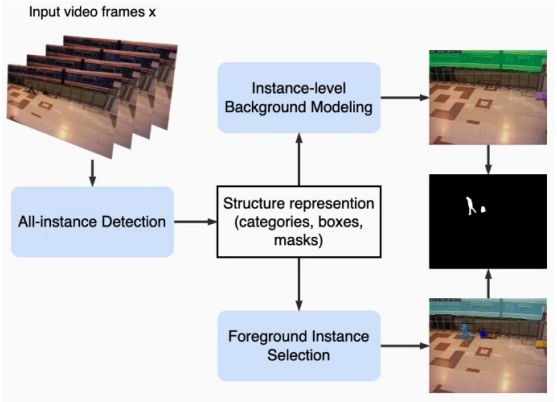


Figure: CDNet frame (left), Ground Truth (middle) & SuBSENSE (right)

Zero-shot Background Subtraction

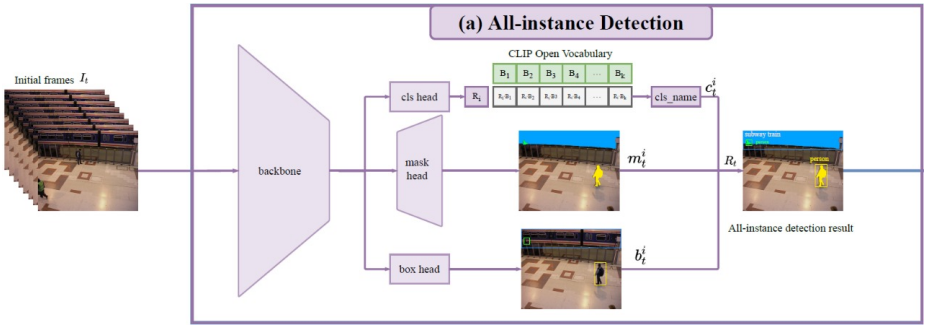
- Formulates background subtraction as an **instance-level motion detection** problem.
- The object is used as the minimum unit of the background model.
- **Algorithm stages:**
  - All-instance detection
  - Instance-level background modeling
  - Foreground instance selection
- Key idea is to shift from pixel-centric to **instance-centric** modeling using a zero-shot detector, Detic [Zhou 2022].



[Zhou 2022] Detecting Twenty-thousand Classes using Image-level Supervision, Xingyi Zhou et al., 2022

All-instance Detection

- **Goal:** Detect and segment all objects in the image as much as possible using Detic.
- **Input:** A frame from the video.
- **Output:** A structured instance-level representation $R_t = \{(c_t^i, m_t^i, b_t^i)\}$, including categories, masks, and boxes.



Instance-level Background Modeling

- **Goal:** Distinguish which detected objects are moving and which remain static, forming an **instance-level background model**.
- **Definition:** The background model is the collection of stationary object instances:

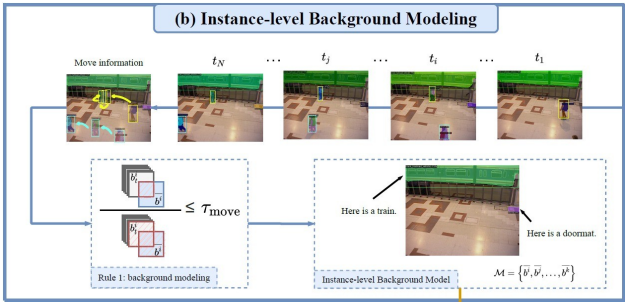
$$\mathcal{M} = \{\bar{b}^i, \bar{b}^j, \ldots, \bar{b}^k\},$$

  where, each $\bar{b}^i$ represents the average (or median) box of instance i across its trajectory.
- **Tracking:** SORT [Bewley 2016] is used to track and maintain consistent instance IDs between frames.

[Bewley 2016] Simple Online and Realtime Tracking, Alex Bewley et al., 2016

Instance-level Background Modeling

- Compute motion stability for each instance using the minimum IoU $\mathrm{IoU}_{\min}(b_t^i, \bar{b}^i)$, which measures how much the instance moves across its trajectory.
- **Update Strategy:** If an instance remains stationary ($\mathrm{IoU}_{\min} \geq \tau_{\text{move}}$), it is added to the background model $\mathcal{M}$; otherwise, it is removed.



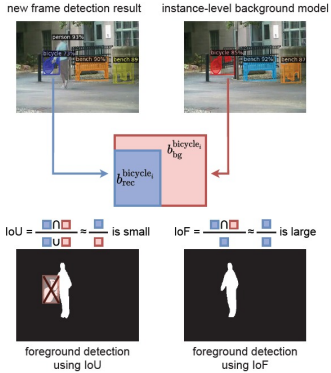
$$\mathcal{M} = \begin{cases} \mathcal{M} \cup \bar{b}^i, & \text{if } \mathrm{IoU}_{\min}(b_t^i, \bar{b}^i) \geq \tau_{\text{move}}, \\ \mathcal{M} \setminus (\mathcal{M} \cap \bar{b}^i), & \text{otherwise.} \end{cases}$$

Foreground Instance Selection

- **Selection Strategy:** If an instance in the current frame has a large overlap (IoU) with one in the background model, it is considered **background**; otherwise, it is **foreground**.
- **Challenge:** Occlusion can cause IoU to drop even for static or partially covered objects.
- **Improvement:** Introduce **IoF** (Intersection over Foreground) to complement IoU and handle occlusion robustly:

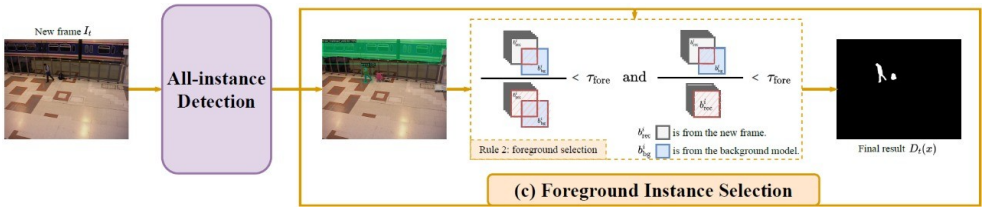
$$\mathrm{IoF}(b^i_{\mathrm{rec}}, b^i_{\mathrm{bg}}) = \frac{|b^i_{\mathrm{rec}} \cap b^i_{\mathrm{bg}}|}{|b^i_{\mathrm{rec}}|}.$$

- If IoU is small but IoF remains large, the instance is likely occluded but still static.



Foreground Instance Selection

- Use both metrics (**IoU** and **IoF**) to decide if an instance belongs to the foreground.
- **Rule:** An instance b^i_{rec} is classified as **foreground**, only if both IoU and IoF are below the threshold τ_{fore}.



$$D_t(x) = \begin{cases} \text{FG}, & \text{if } \text{IoU}(b^i_{\text{rec}}, b^i_{\text{bg}}) < \tau_{\text{fore}} \ \wedge \ \text{IoF}(b^i_{\text{rec}}, b^i_{\text{bg}}) < \tau_{\text{fore}}, \\ \text{BG}, & \text{otherwise.} \end{cases}$$

The ZBS Algorithm

Initialize the zero-shot detector as $\mathcal{Z}$ and the background model as $\mathcal{M}$
while current frame I_t is valid **do**
  Stage 1: All-instance detection
  output the result $R_t \leftarrow \mathcal{Z}(I_t)$
  Stage 2: Instance-level background model
  get the track of each instance from b_t^i (part of R_t)
  calculate the $\mathrm{IoU}_{\mathsf{min}}$ of $b_t^i, \bar{b}^i$
  update $\mathcal{M}$ based on $\mathrm{IoU}_{\mathsf{min}}$ and τ_{move}
  Stage 3: Foreground instance selection
  separate $\mathcal{M}$ and b_t by instance ID
  calculate the IoU and IoF of $\mathcal{M}$ and b_t^i
  get a binary mask $D_t(x)$ based on IoU & IoF and τ_{fore}
  current frame $\leftarrow$ next frame
end while

Experimental Setup

- **Dataset:** CDNet 2014 (53 sequences, 11 categories).
- **Input:** 50 consecutive frames from the same sequence.
- **Labeling protocol:** Static and shadow pixels are treated as *background* (negative), moving pixels as *foreground* (positive), and non-ROI regions are ignored.
- **Universal parameters:** $\tau_{\text{conf}} = 0.6$, $\tau_{\text{move}} = 0.5$, $\tau_{\text{fore}} = 0.8$ in all experiments.
- **Evaluation metrics:** Precision (P), Recall (R), and F-Measure (F1).
- **Comparisons:** ZBS results are compared with the well-known SubSENSE BGS [St-Charles 2015] results and the consecutive Ground Truths.

[St-Charles 2015] SuBSENSE: A Universal Change Detection Method With Local Adaptive Sensitivity, Pierre-Luc St-Charles et al., 2015

Visualization - ZBS

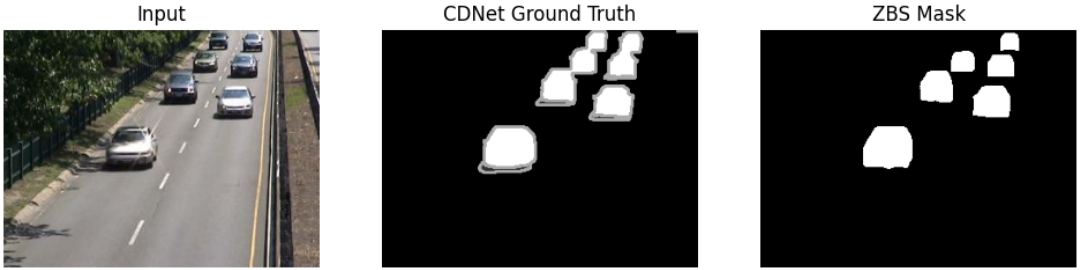
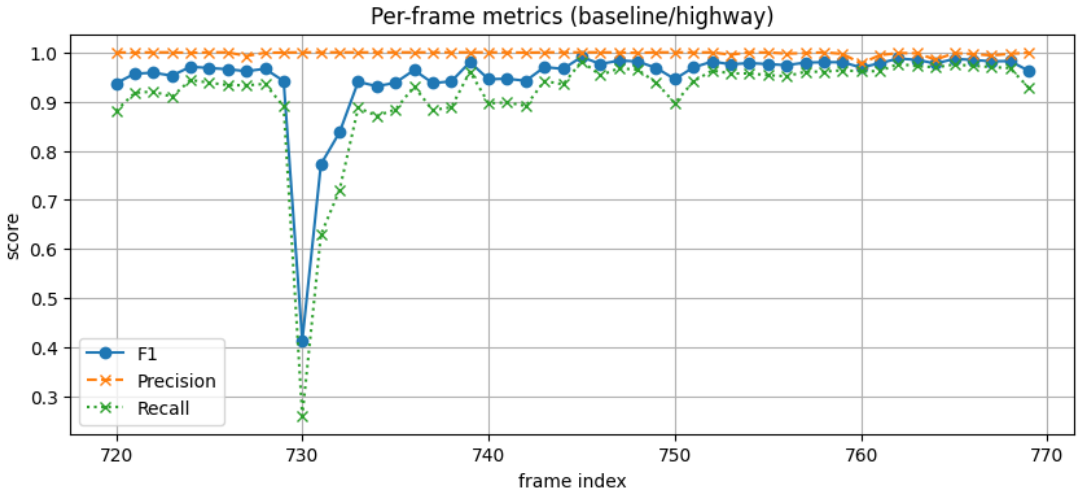


Figure: Result of ZBS on CDNet frames, Category: baseline/highway

F1-Measure - ZBS



Visualization - SuBSENSE vs ZBS

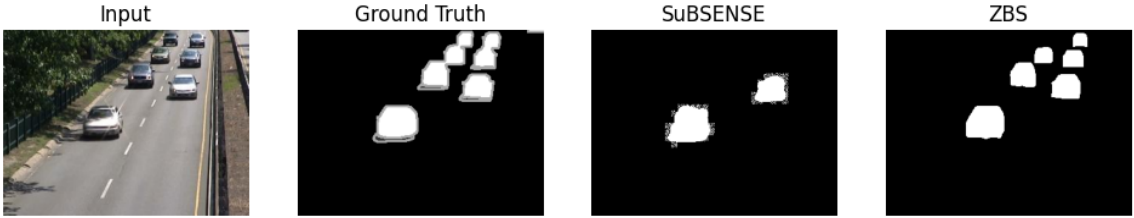
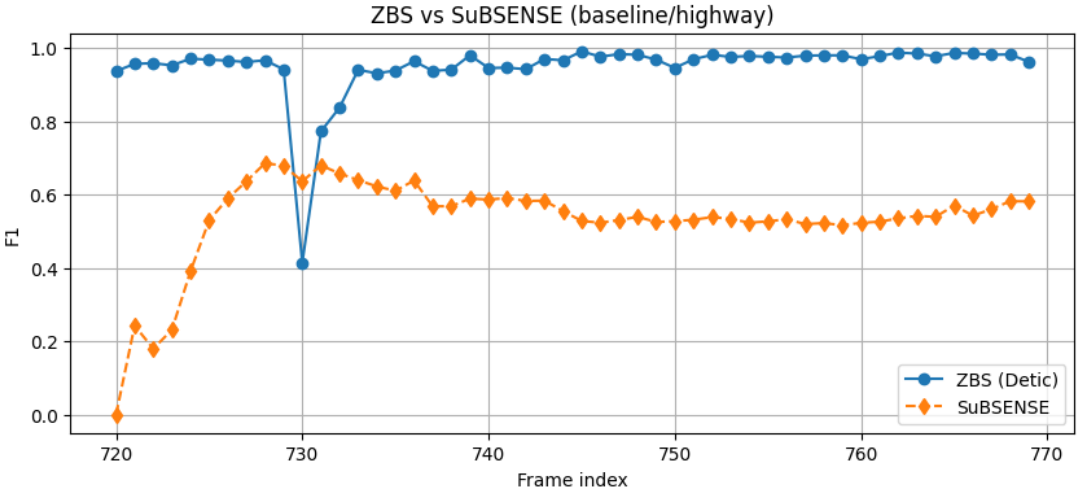


Figure: Results of SuBSENSE & ZBS on CDNet frames, Category: baseline/highway

F1-Measure - SuBSENSE vs ZBS



Visualization - ZBS across different categories

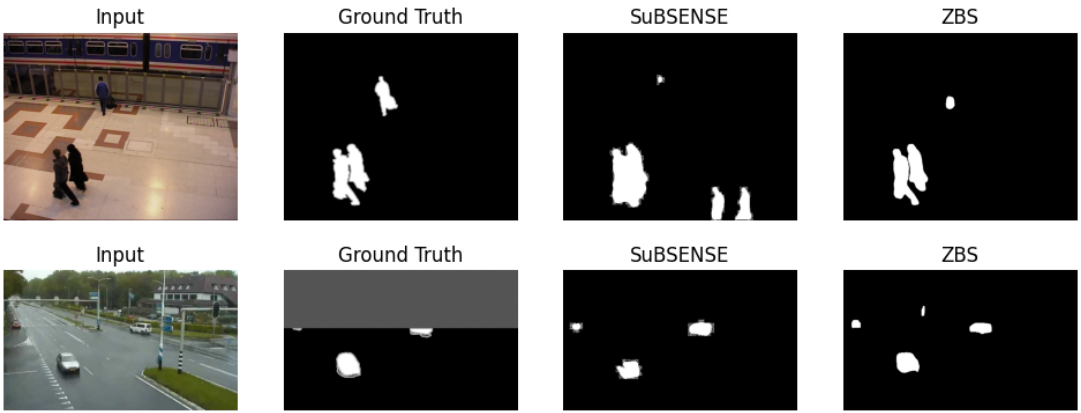


Figure: Results, CDNet, Categories: baseline/PETS2006 (top) & PTZ/twoPositionPTZCam (bottom)

Visualization - ZBS across different categories

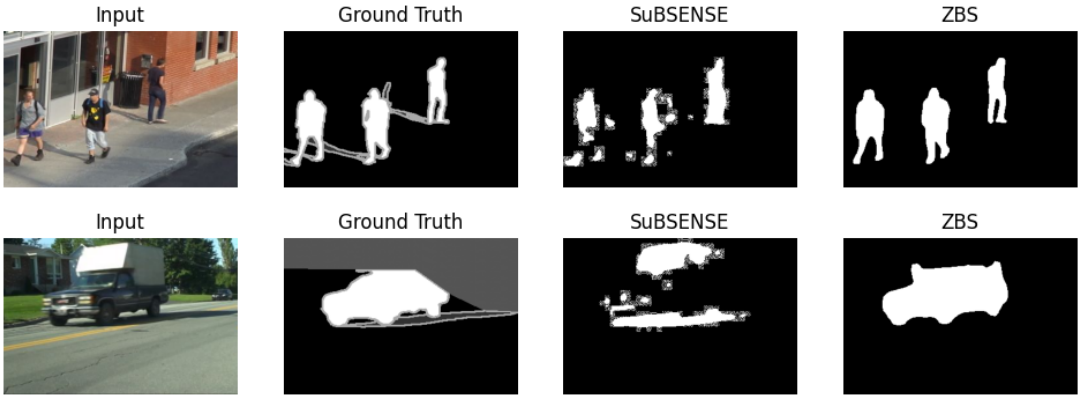


Figure: Results, CDNet, Categories: shadow/busStation (top) & shadow/bungalows (bottom)

Results

- **ZBS** achieves the highest overall **F-Measure of 0.8515** on **CDNet-2014** bechmark, outperforming all unsupervised BGS methods.
- Outperforms the previous SOTA unsupervised method by **+4.7%** in overall F-Measure.
- Shows strong performance in categories such as **PTZ**, **Shadow**, and **CameraJitter**.
- In our experimental comparison of **ZBS** vs **SuBSENSE** on selected frame intervals from multiple CDNet categories, we observed an average improvement of $\approx$ **+20%** in overall F-Measure, confirming ZBS's robustness on limited test segments.

Limitations

- **High computational cost:** The heavy backbone detector (Detic) requires a GPU for real-time or near real-time performance.
- **Adverse weather sensitivity:** Turbulence, heavy fog, or strong illumination changes can degrade detection quality, lowering recall.
- **Challenging cases:** Small, fast-moving, or heavily occluded objects remain difficult to detect and segment reliably.

Conclusion

- **ZBS** is the first *instance-level*, zero-shot background subtraction method.
- Leverages pre-trained zero-shot detectors to generalize across unseen categories without scene-specific training.
- Improves robustness against shadows, jitter, and complex backgrounds by modeling objects instead of pixels.
- Achieves state-of-the-art performance among unsupervised BGS methods on the CDNet 2014 benchmark.

Thank You!