Structure landmark-based self-supervised tracking

Jianjie Zhao[1] Shu Liu[2]

College of Engineering and Computer Science, Australian National University

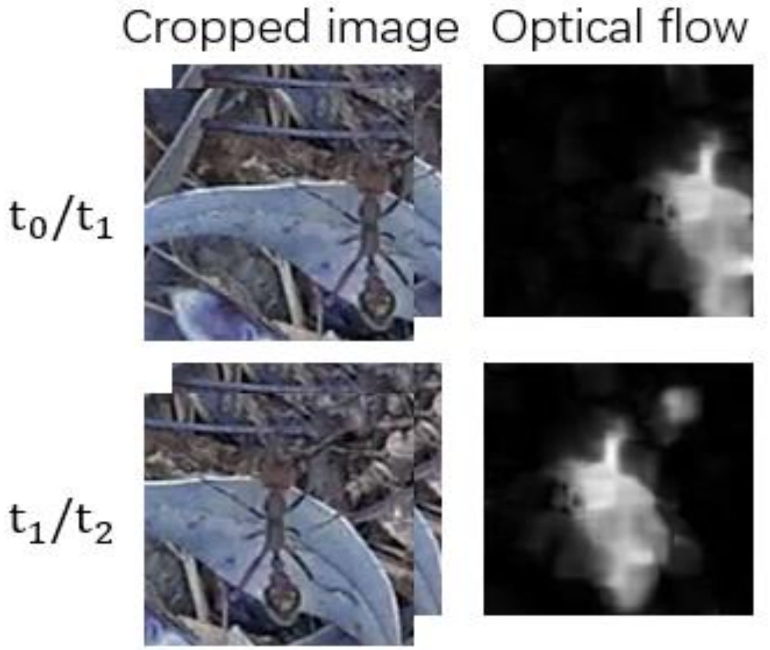
[1]u5491675@anu.edu.au [2]u5764321@anu.edu.au

1. Background

- Tracking in computer vision is popular topic. With the help of the application of machine learning, a tracking model can capture the predominating spatial features with no annotation.

2. Related Work

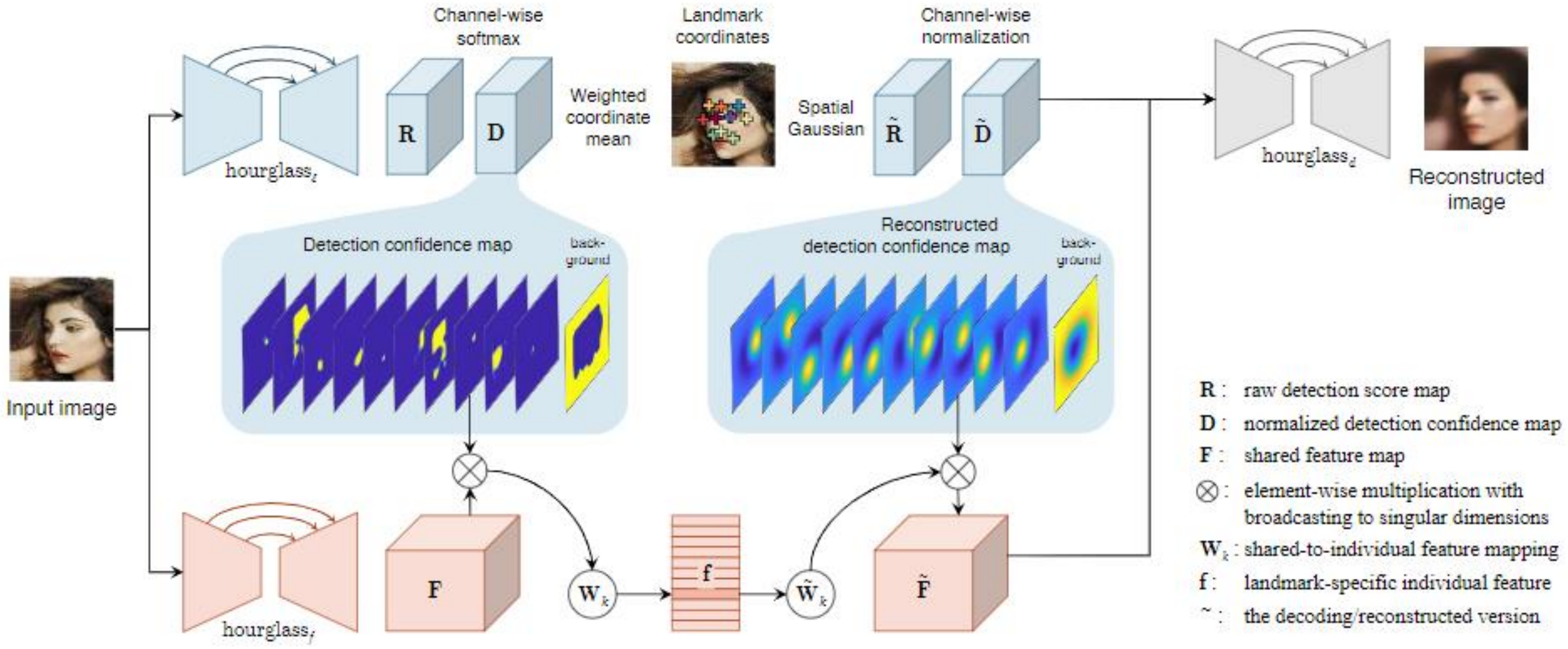
- Optical flow provides a motion field which can be taken as an auxiliary information for tracking.



2. Related Work

- Unsupervised structure representation provides a method for extracting the critical landmarks from a image.

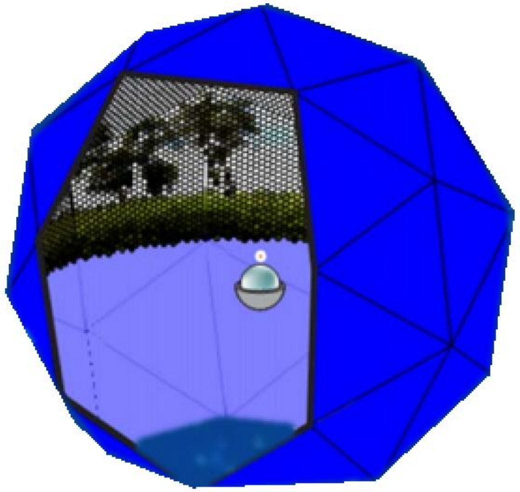
Zhang Y, Guo Y, Jin Y, et al. Unsupervised discovery of object landmarks as structural representations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2694-2703.



3. Method

- Ant-in-wild dataset (video)

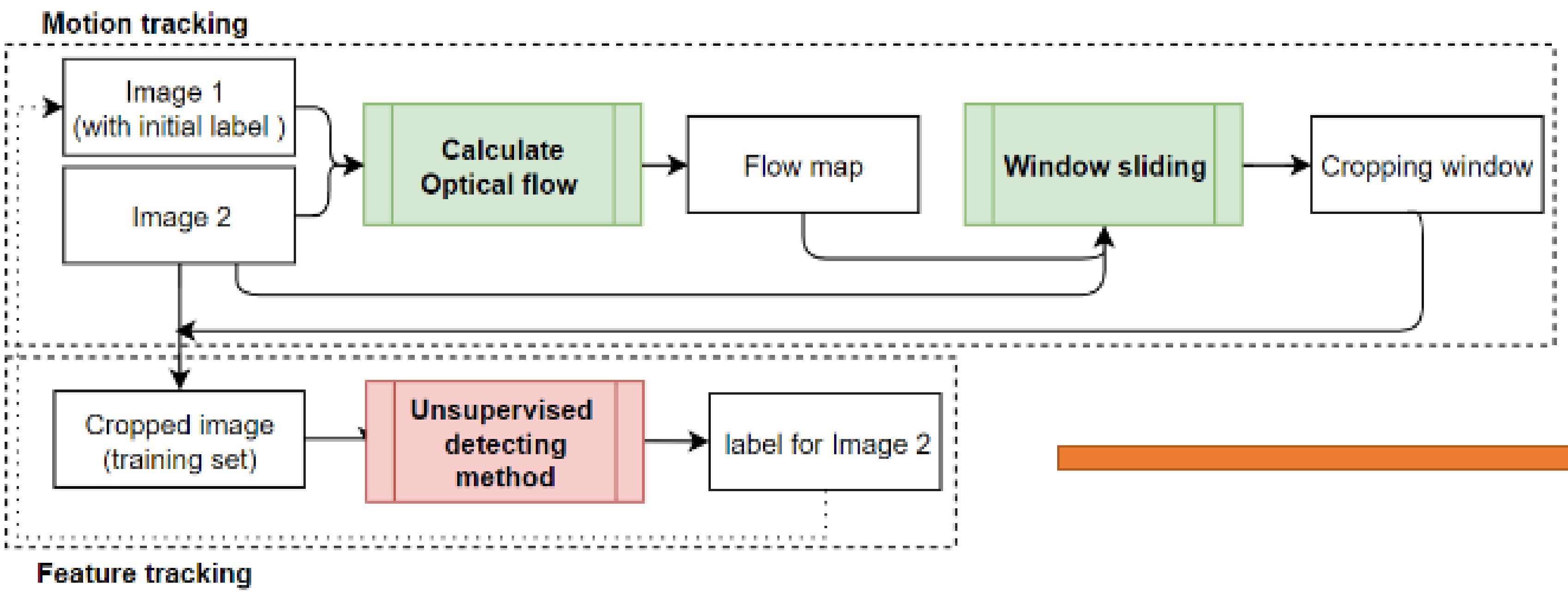
- Ant-on-ball dataset

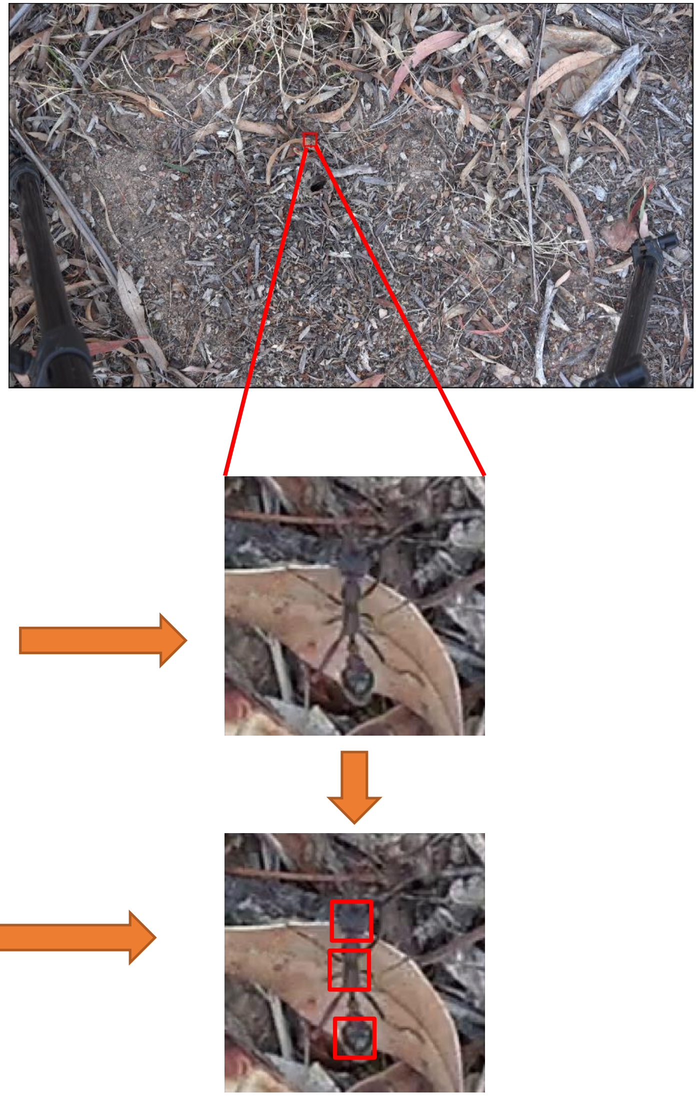


	Dimension	Dataset size	Background	Target moving	Annotation
Ant-in-wild	2160x3840	1425	Noised	Unrestricted	None
Ant-on-ball	971x736	5403	Non-noised	Restricted	None

3. Method

- Adaptable system architecture for tracking





3. Method

- Constricts
  - The target should have a relatively obvious motion compared with the background.
  - The target should occupy a large proportion area in the cropped images.
- Advantages:
  - The optical flow produces a cropping window which shrinks the searching region of the target.
  - The self-supervised method minimizes the annotation requirement.
  - The motion tracking section rejects the static background, which enhances the performance of the system in a complex scene.

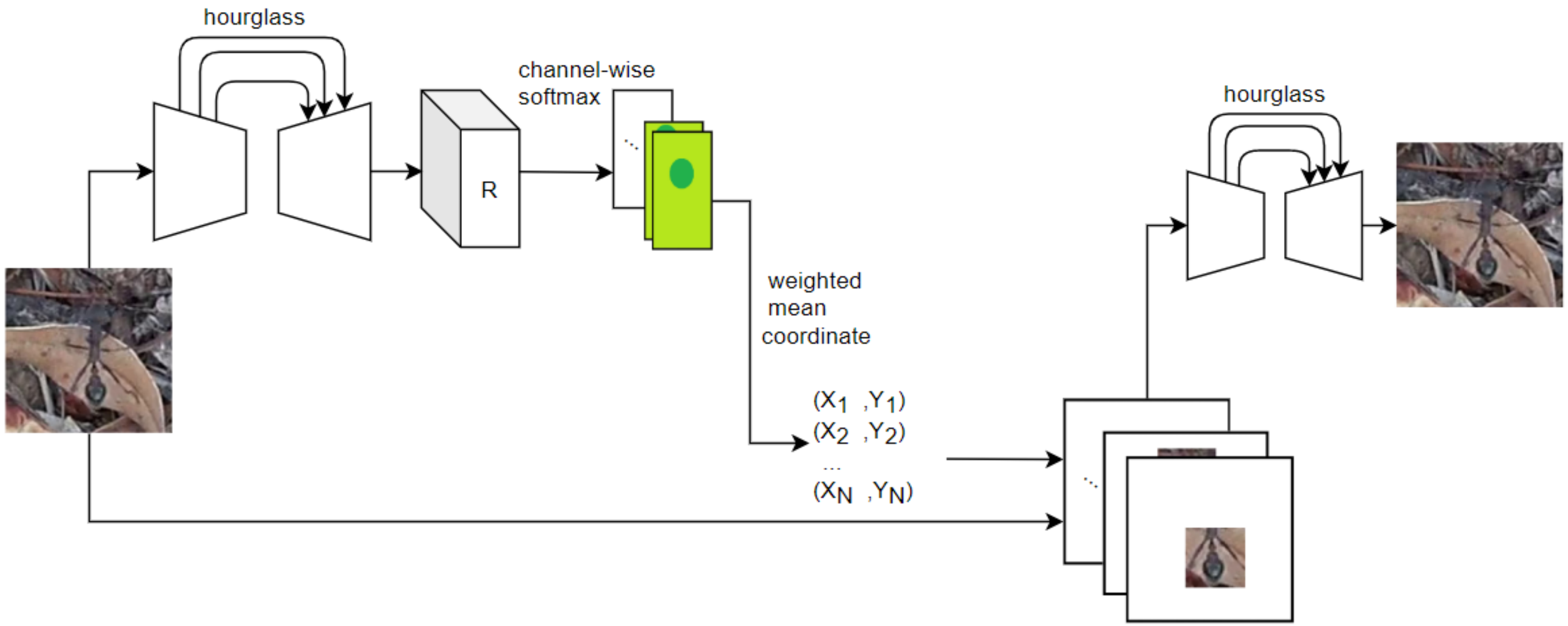
3. Method

Algorithm1 Training set generating
Define maximal frame $\boldsymbol{n}$ **Define** frame index $\boldsymbol{i}$ = 0 **Define** training set Λ as an empty list **Input** Image sequence I_n from a video, where n is the index **Input** Initial position of target in the i^{th} frame (x_i, y_i) **Input** maximal displacement d_x, d_y While i < n: Calculate the optical flow map K based on I_i and I_{i+1}. Capture the most concentrated motion on K with a window sliding (d_x, d_y) far from the position(x_i, y_i), the centre position and side length of the window denotes as $(x_w, y_w) = (d_x + x_i, d_y + y_i)$ and α. Cropping the squared patch I_i' in the region of $(x_w \pm \frac{a}{2}, y_w \pm \frac{a}{2})$ from the I_i. Append I_i' into Λ as a training sample. Update (x_i, y_i) with the values of (x_w, y_w)

- The generating of training set of the unsupervised method requires a initial annotation for determining the tracking target.
- For each annotation, n frames after the annotated frame will be put into training set.
- With m annotations, we acquire a training set with mn images in m different scene.

3. Method

- Our unsupervised for extracting the predominating features.



- 1. Generate K coordinate pairs (X_K, Y_K) with a hourglass network.
- 2. Generate K attention mask with the coordinate pairs. Then cropping the input image with the attention masks.
- 3. Reconstruct the input image with the cropped image patches.

3. Method

- Landmark constraint → regularization terms
  - Concentration constraint

    It is defined by the squared sum of variance of the value of all pixels along x-axis and y-axis.

    $\rightarrow L_c = \left(\sigma_x^2 + \sigma_y^2\right)^2$
  - Distance constraint

    It is defined by the sum of the distance between all coordinate pairs.

    $\rightarrow L_D = \sum_{k \neq k'}^{1,\ldots,K} \exp(-\left\|\left(x_{k'}, y_{k'}\right) - (x_k, y_k)\right\|_2^2)$
- Loss function

  $\rightarrow L = E_{mse} + \lambda_c L_c + \lambda_d L_d$

4. Training

- Dataset

  Ant-in-wild (with both motion tracking & feature tracking)

  Ant-on-ball (with only feature tracking as the ant motion is restricted)

- Hardware

  CPU: i7-8700

  GPU: NVIDIA GTX 1060 (6GB)

- Setting for motion tracking

  Window size: 128x128
  Number of annotation: 2
  Maximal length of valid frames: 250

- Setting for feature tracking

  Number of channel for training: 1

  Number of landmark: 4

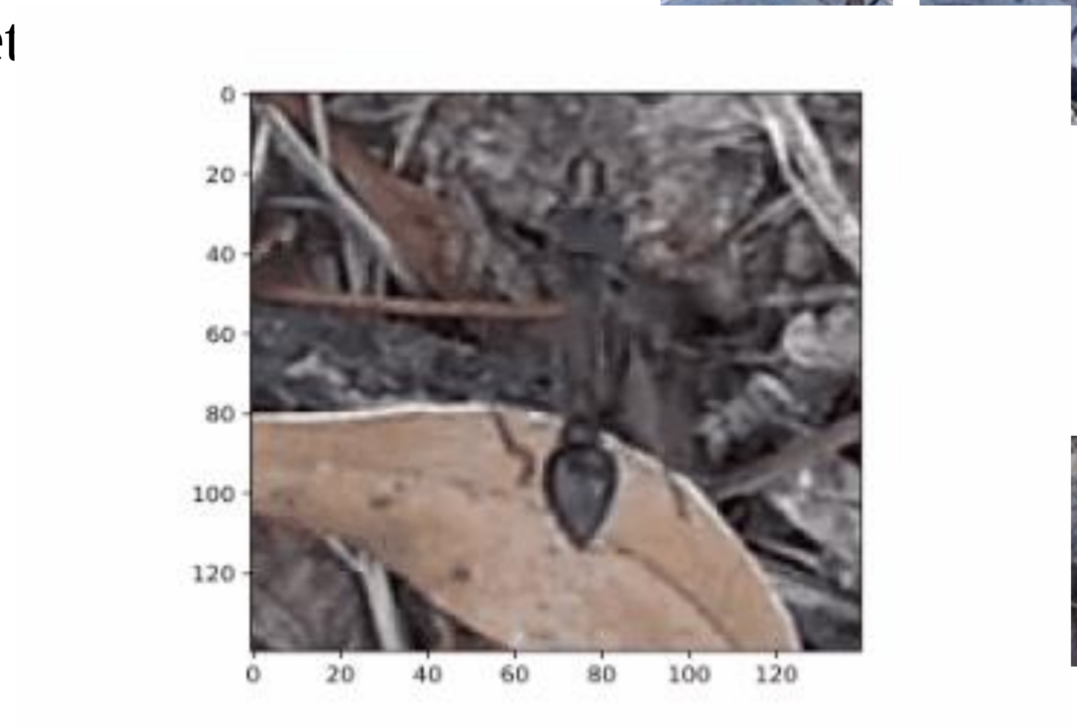
  Weight for regularization: $\lambda_c = 0.01$, $\lambda_d = 0.001$

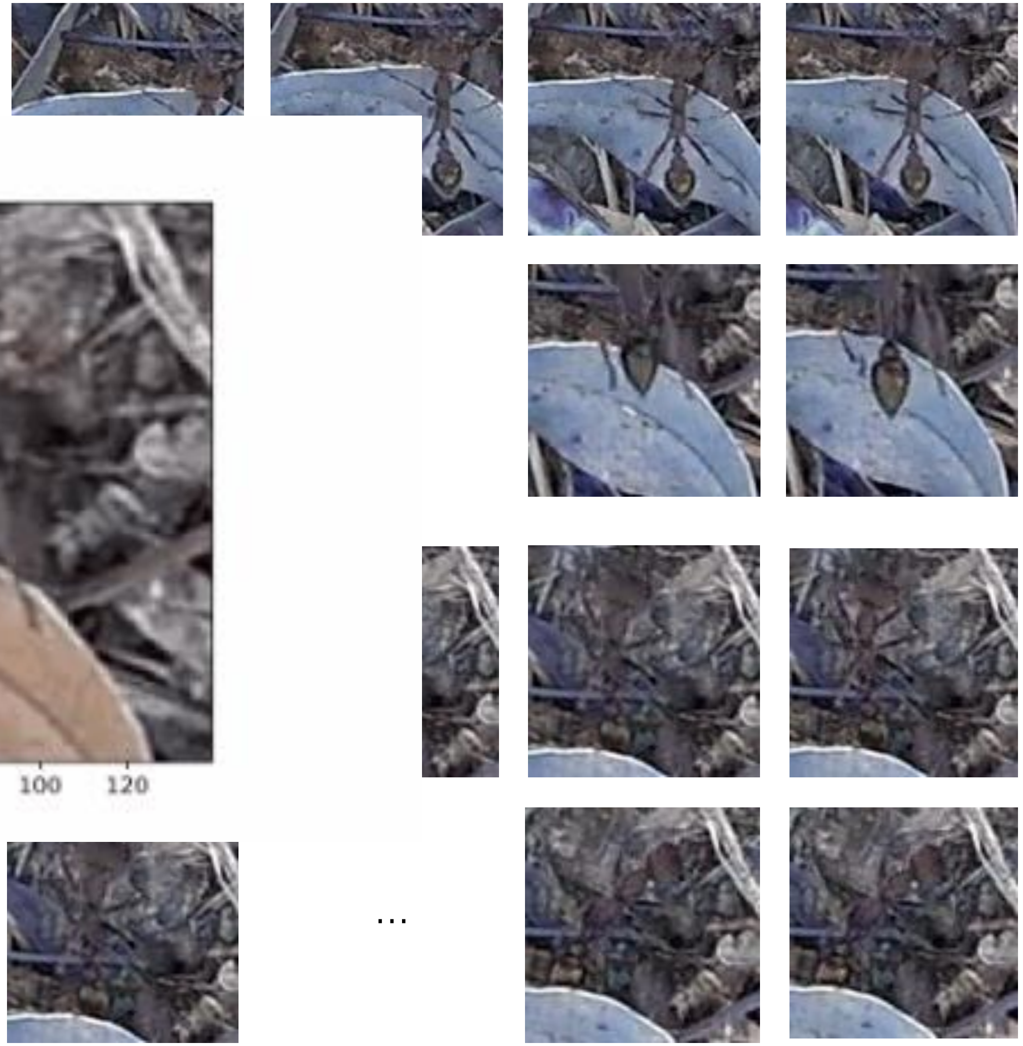
  Learning rate: 0.01 (before 100 epochs), 0.001 (after 100 epochs)

  Data argumentation: Rotation, Flipping

5. Result – motion tracking

- Ant-in-wild dataset

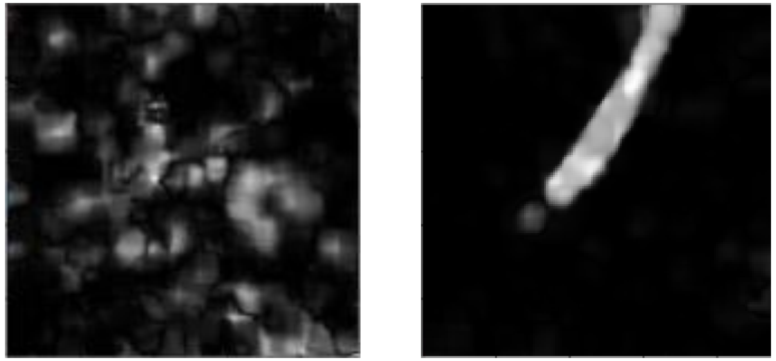




5. Result – motion tracking

- Failure on Ant-in-wild dataset

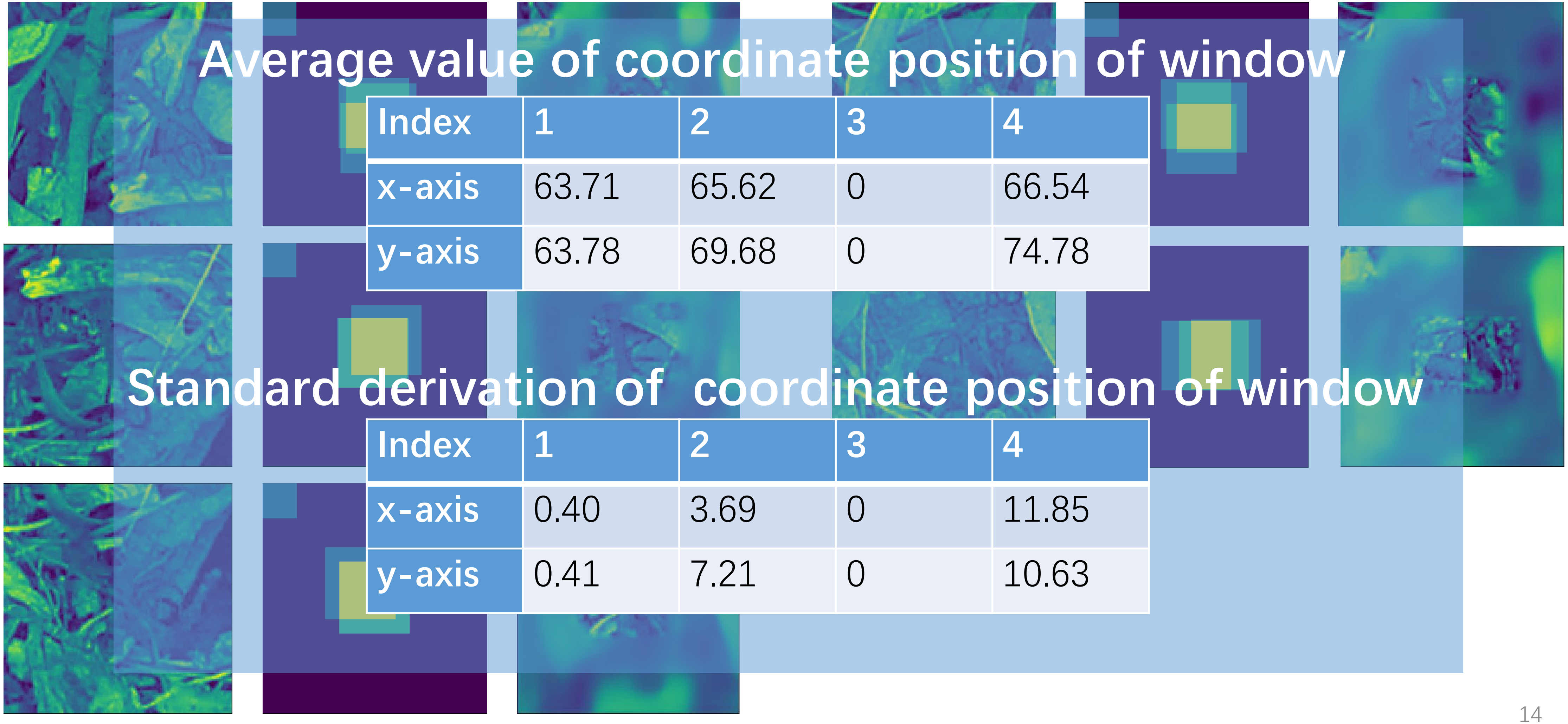
Unexpected motion



Object overlapping

5. Result – feature tracking

- Ant-in-wild dataset



Average value of coordinate position of window

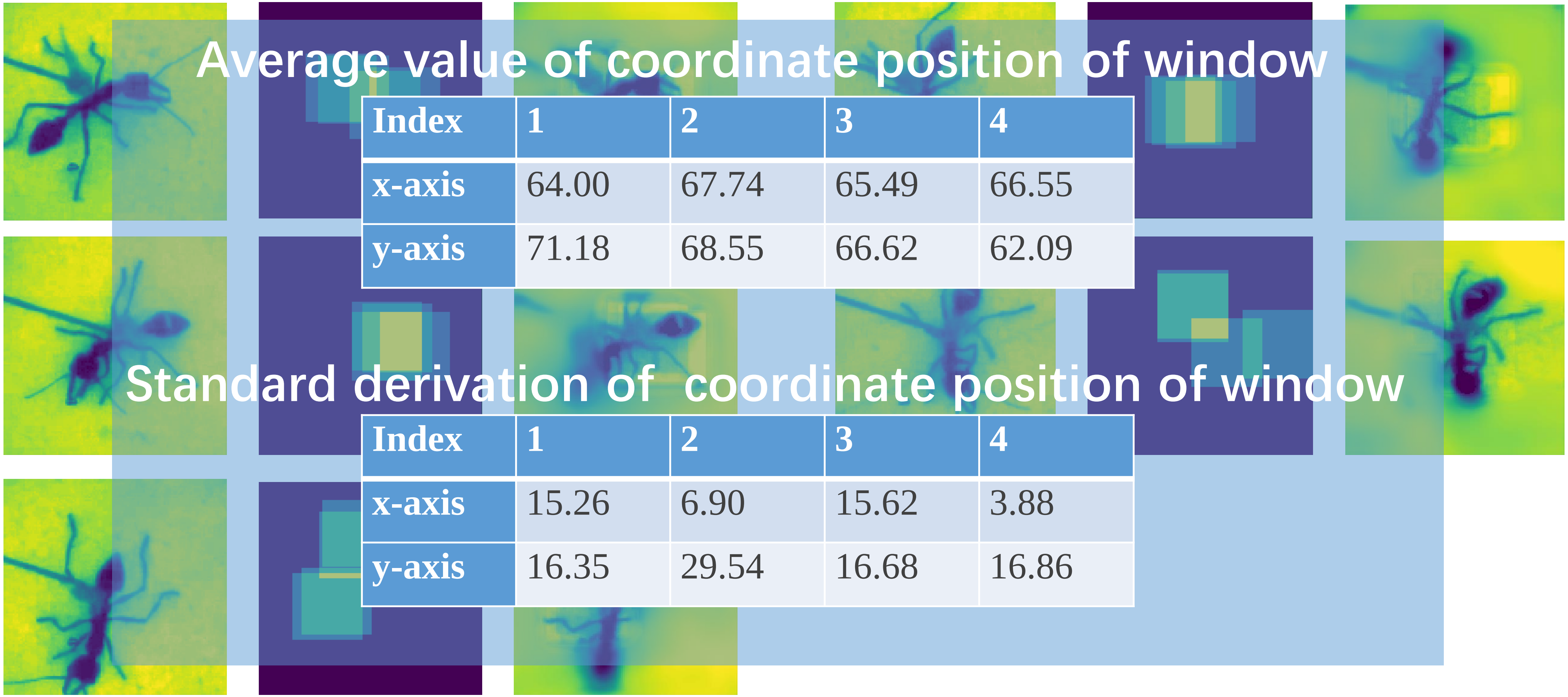
Index	1	2	3	4
x-axis	63.71	65.62	0	66.54
y-axis	63.78	69.68	0	74.78

Standard derivation of coordinate position of window

Index	1	2	3	4
x-axis	0.40	3.69	0	11.85
y-axis	0.41	7.21	0	10.63

5. Result – feature tracking

- Ant-on-ball dataset



Average value of coordinate position of window

Index	1	2	3	4
x-axis	64.00	67.74	65.49	66.55
y-axis	71.18	68.55	66.62	62.09

Standard derivation of coordinate position of window

Index	1	2	3	4
x-axis	15.26	6.90	15.62	3.88
y-axis	16.35	29.54	16.68	16.86

6. Reference

[1] Zhang Y, Guo Y, Jin Y, et al. Unsupervised discovery of object landmarks as structural representations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2694-2703.