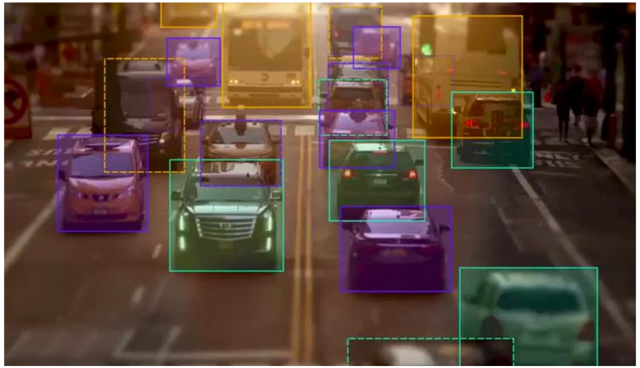
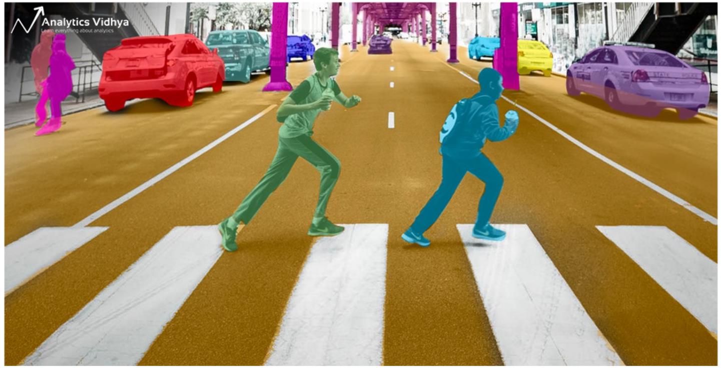
Fast Online Object Tracking and Segmentation: A Unifying Approach

Alessandro, Vish, and Mel

Table of content

Key terms





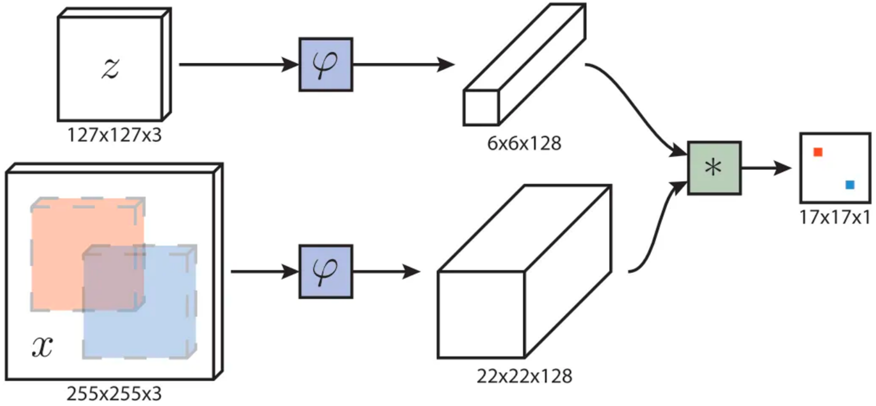
- Visual Object Tracking
  - Draw bounding box on object of interest in a scene
- Video Object Segmentation (VOS)
  - Draw a binary pixel mask over the scene indicating if the object is contained in the pixel
  - Historically more computationally expensive
    - Yeo *et al.* only manages 4fps and 0.1fps using CNN features

Motivations

- Success of fast-tracking approaches based on convolutional Siamese networks
- The availability of Youtube-VOS, a dataset of 4,000+ videos with pixel-level annotations on 70+ common objects

SiamFC - Bertinetto *et al.*

- Siamese network uses shared convolutional weights
- Compares features via cross-correlation function producing similarity “heat map”
- *x* is centered at last known location of object
- Maximum cross-correlation (response) per candidate window indicates match



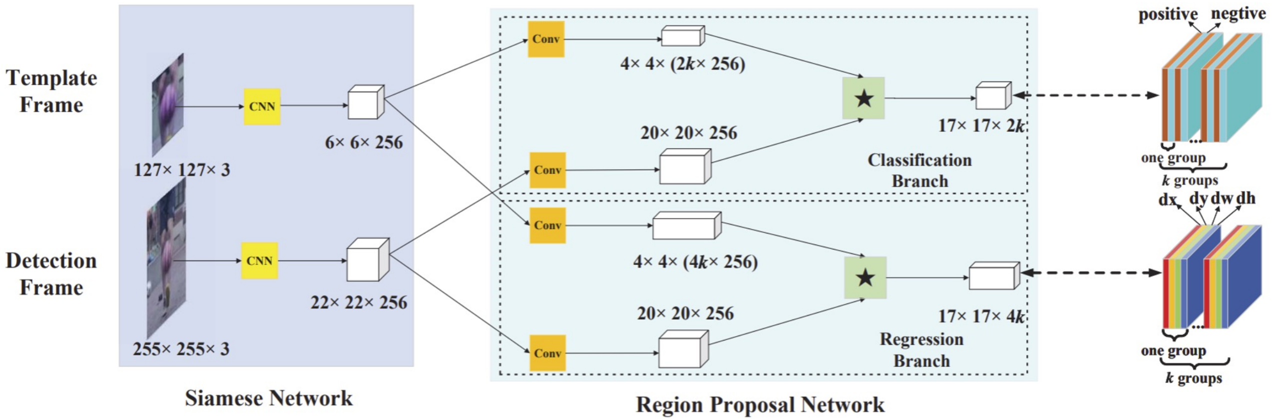
SiamFC - Continued

- Used fixed bounding box sizes
- Scale issues were addressed with rescaled exemplar images

How can we draw an appropriately shaped bounding box from this information?

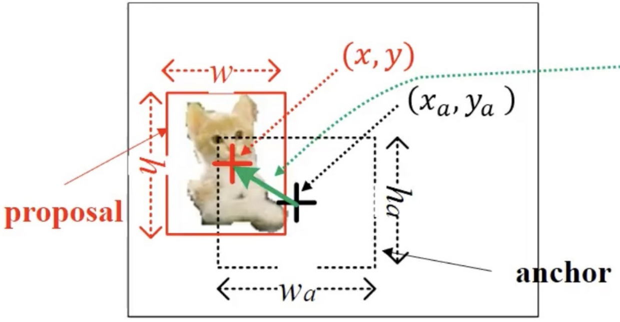
SiamRPN - Li *et al.*

- Also for object tracking
- Finds optimal aspect ratios of bounding boxes for *k* anchors
  - Aspect ratios are hyperparameters
- Classification and regression branches identify object presence and optimal bounding box in candidate window respectively



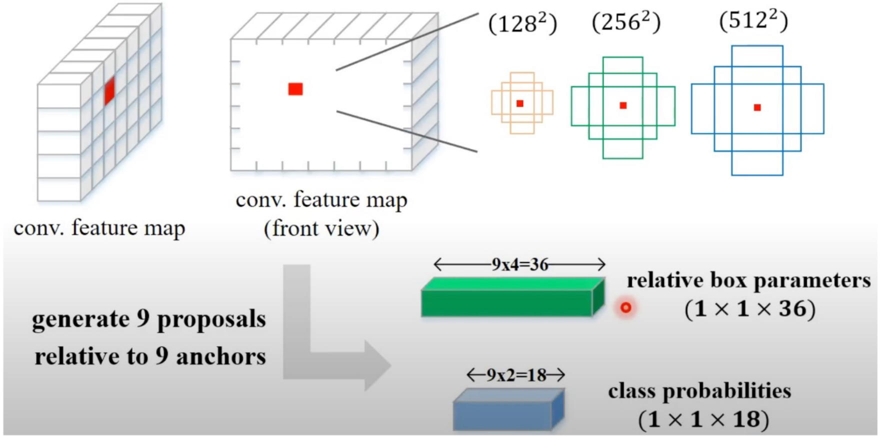
SiamRPN - continued

- Anchors are centered in candidate window
- For regression branch, several aspect ratios and scales are explored—a proposal
    - Each with corresponding offset (dx, dy, dw, dh)
- For classification branch, each RoW



relative encoding

$$d_x = (x - x_a)/w_a$$
$$d_y = (y - y_a)/h_a$$
$$d_w = w/w_a$$
$$d_h = h/h_a$$



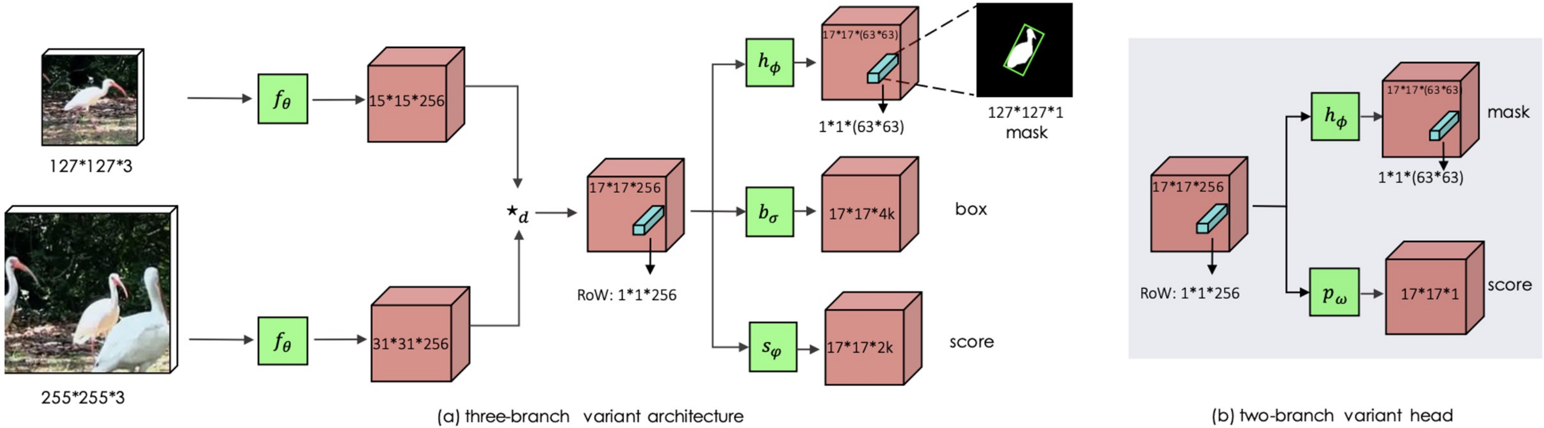
SiamRPN - Continued

- Only produced bounding box…

How can we draw a pixel mask from this information?

SiamMask Architecture elements

- Extra branch and loss is essential for encoding the information necessary to produce a pixel-wise binary mask.



Architecture elements

- It predicts w×h binary masks (one for each RoW) using a simple two-layers neural network hφ with learnable parameters φ. Let mn denote the predicted mask corresponding to the n-th RoW

$$m_n = h_\phi(g_\theta^n(z,\ x)).$$

Loss function

- each RoW is labelled with a ground-truth binary label yn ∈ {±1} and also associated with a pixel-wise ground-truth mask cn of size w×h
- Binary logistic regression loss over all RoWs

$$\mathcal{L}_{mask}(\theta,\ \phi) = \sum_n \Big(\frac{1+y_n}{2wh}\sum_{ij}\log(1+e^{-c_n^{ij} m_n^{ij}})\Big).$$

Architecture elements

- ResNet-50 until final convolutional layer of the 4th stage (stride 1 and dilated convolutions). No downsampling in conv4.
- Depth-wise cross correlated output features resulting in a feature map of size 17x17.

block	score	box	mask
conv5	$1 \times 1, 256$	$1 \times 1, 256$	$1 \times 1, 256$
conv6	$1 \times 1, 2k$	$1 \times 1, 4k$	$1 \times 1, (63 \times 63)$

Table 9. Architectural details of the *three-branch* head. k denotes the number of anchor boxes per RoW.

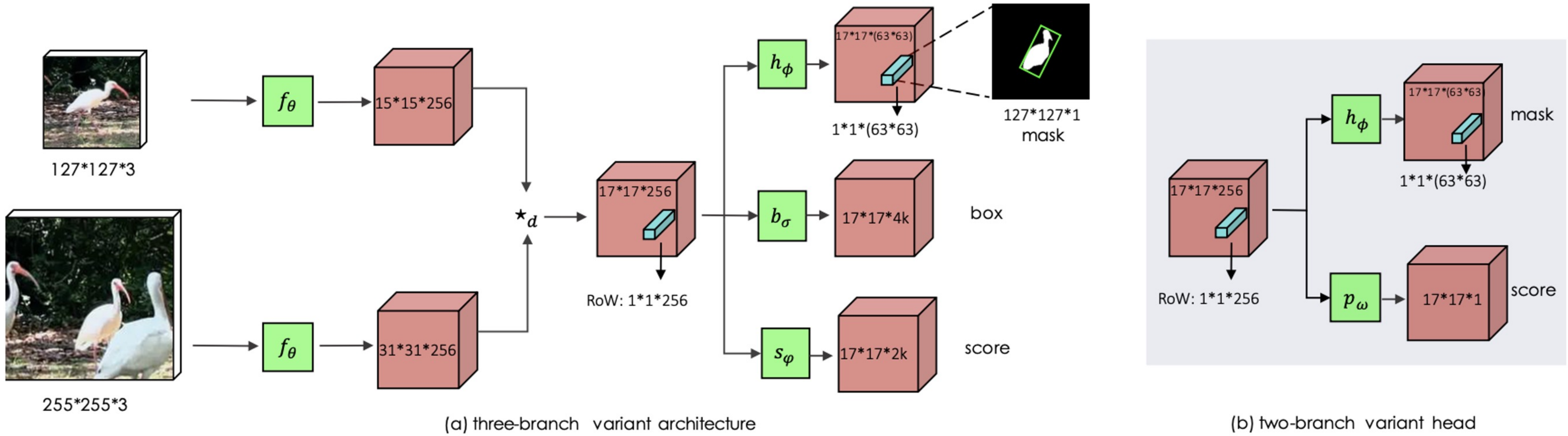
block	score	mask
conv5	$1 \times 1, 256$	$1 \times 1, 256$
conv6	$1 \times 1, 1$	$1 \times 1, (63 \times 63)$

Table 10. Architectural details of the *two-branch* head.

Training

- Examplar and search image patches of 127×127 and 255×255 pixels respectively.
- Pre-trained on the ImageNet-1k classification task.
- SGD with a first warmup phase in which the learning rate increases linearly from 10−3 to 5×10−3 for the first 5 epochs and then decreases logarithmically until 5×10−4 for 15 more epochs.
- Datasets: COCO, ImageNet-VID and YouTube-VOS.
- It selects the output mask using the location attaining the maximum score in the classification branch.

Architecture



Results - Visual Object Tracking

- VOT-2016 for representation types comparison
- VOT-2018 for state-of-the-art comparison
- How much does object representation matter?

	mIOU (%)	mAP@0.5 IOU	mAP@0.7 IOU
Fixed a.r. Oracle	73.43	90.15	62.52
Min-max Oracle	77.70	88.84	65.16
MBR Oracle	84.07	97.77	80.68
SiamFC [3]	50.48	56.42	9.28
SiamRPN [63]	60.02	76.20	32.47
SiamMask-*Min-max*	65.05	82.99	43.09
SiamMask-*MBR*	67.15	85.42	50.86
SiamMask-*Opt*	**71.68**	**90.77**	**60.47**

Results - Visual Object Tracking

- Results on VOT-2018 and VOT-2016

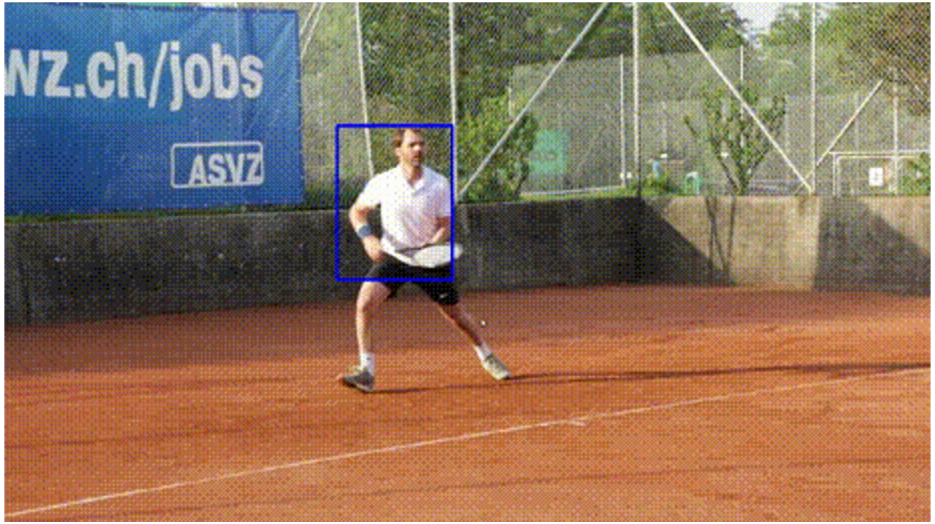
	SiamMask-*Opt*	SiamMask	SiamMask-2B	DaSiamRPN [63]	SiamRPN [28]	SA_Siam_R [15]	CSRDCF [33]	STRCF [29]
EAO ↑	**0.387**	**0.380**	0.334	0.326	0.244	0.337	0.263	0.345
Accuracy ↑	**0.642**	**0.609**	0.575	0.569	0.490	0.566	0.466	0.523
Robustness ↓	0.295	0.276	0.304	0.337	0.460	0.258	0.318	**0.215**
Speed (fps) ↑	5	55	60	160	**200**	32.4	48.9	2.9

Table 2. Comparison with the state-of-the-art under the EAO, Accuracy, and Robustness metrics on VOT-2018.

	VOT-2018			VOT-2016			
	EAO ↑	A ↑	R ↓	EAO ↑	A ↑	R ↓	Speed
SiamMask-box	0.363	0.584	0.300	0.412	0.623	0.233	**76**
SiamMask	0.380	0.609	**0.276**	0.433	0.639	**0.214**	55
SiamMask-*Opt*	**0.387**	**0.642**	0.295	**0.442**	**0.670**	0.233	5

Results - Video Object Segmentation

Can operate online, runs in real-time, and only requires a simple bounding box initialisation



Results - Video Object Segmentation

FT - Finetuned
M - Mask
J - Jaccard index
F - F-measure

- Test on DAVIS-2016, DAVIS-2017, and Youtube-VOS
- Extract axis-aligned bounding box from the mask

	FT	M	$\mathcal{J}_{\mathcal{M}}\uparrow$	$\mathcal{J}_{\mathcal{O}}\uparrow$	$\mathcal{J}_{\mathcal{D}}\downarrow$	$\mathcal{F}_{\mathcal{M}}\uparrow$	$\mathcal{F}_{\mathcal{O}}\uparrow$	$\mathcal{F}_{\mathcal{D}}\downarrow$	Speed
OnAVOS [53]	✔	✔	**86.1**	**96.1**	5.2	**84.9**	**89.7**	5.8	0.08
MSK [39]	✔	✔	79.7	93.1	8.9	75.4	87.1	9.0	0.1
MSK_b [39]	✔	✘	69.6	-	-	-	-	-	0.1
SFL [9]	✔	✔	76.1	90.6	12.1	76.0	85.5	10.4	0.1
FAVOS [8]	✘	✔	82.4	96.5	4.5	79.5	89.4	5.5	0.8
RGMP [57]	✘	✔	81.5	91.7	10.9	82.0	90.8	10.1	8
PML [7]	✘	✔	75.5	89.6	8.5	79.3	93.4	7.8	3.6
OSMN [59]	✘	✔	74.0	87.6	9.0	72.9	84.0	10.6	8.0
PLM [62]	✘	✔	70.2	86.3	11.2	62.5	73.2	14.7	6.7
VPN [22]	✘	✔	70.2	82.3	12.4	65.5	69.0	14.4	1.6
SiamMask	✘	✘	71.7	86.8	**3.0**	67.8	79.8	**2.1**	**55**

Table 4. Results on DAVIS 2016 (validation set). FT and M respectively denote if the method requires fine-tuning and whether it is initialised with a mask (✔) or a bounding box (✘).

	FT	M	$\mathcal{J}_{\mathcal{M}}\uparrow$	$\mathcal{J}_{\mathcal{O}}\uparrow$	$\mathcal{J}_{\mathcal{D}}\downarrow$	$\mathcal{F}_{\mathcal{M}}\uparrow$	$\mathcal{F}_{\mathcal{O}}\uparrow$	$\mathcal{F}_{\mathcal{D}}\downarrow$	Speed
OnAVOS [53]	✔	✔	**61.6**	**67.4**	27.9	**69.1**	**75.4**	26.6	0.1
OSVOS [5]	✔	✔	56.6	63.8	26.1	63.9	73.8	27.0	0.1
FAVOS [8]	✘	✔	54.6	61.1	**14.1**	61.8	72.3	**18.0**	0.8
OSMN [59]	✘	✔	52.5	60.9	21.5	57.1	66.1	24.3	8.0
SiamMask	✘	✘	54.3	62.8	19.3	58.5	67.5	20.9	**55**

Table 5. Results on DAVIS 2017 (validation set).

	FT	M	$\mathcal{J}_{\mathcal{S}}\uparrow$	$\mathcal{J}_{\mathcal{U}}\uparrow$	$\mathcal{F}_{\mathcal{S}}\uparrow$	$\mathcal{F}_{\mathcal{U}}\uparrow$	$\mathcal{O}\uparrow$	Speed
OnAVOS [53]	✔	✔	60.1	46.6	**62.7**	51.4	55.2	0.1
OSVOS [5]	✔	✔	59.8	**54.2**	60.5	**60.7**	**58.8**	0.1
OSMN [59]	✘	✔	60.0	40.6	60.1	44.0	51.2	8.0
SiamMask	✘	✘	**60.2**	45.1	58.2	47.7	52.8	**55**

Table 6. Results on YouTube-VOS (validation set).

Ablation studies

AN = Alex Net

RN = ResNet-50 proposed

w/o R = without final refinement

	AN	RN	EAO ↑	$\mathcal{J}_{\mathcal{M}}\uparrow$	$\mathcal{F}_{\mathcal{M}}\uparrow$	Speed
SiamFC	✔		0.188	-	-	86
SiamFC		✔	0.251	-	-	40
SiamRPN	✔		0.243	-	-	**200**
SiamRPN		✔	0.359	-	-	76
SiamMask-2B w/o R		✔	0.326	62.3	55.6	43
SiamMask w/o R		✔	0.375	68.6	57.8	58
SiamMask-2B-score		✔	0.265	-	-	40
SiamMask-box		✔	0.363	-	-	76
SiamMask-2B		✔	0.334	67.4	63.5	60
SiamMask		✔	**0.380**	**71.7**	**67.8**	55

Table 7. Ablation studies on VOT-2018 and DAVIS-2016.

Failure cases

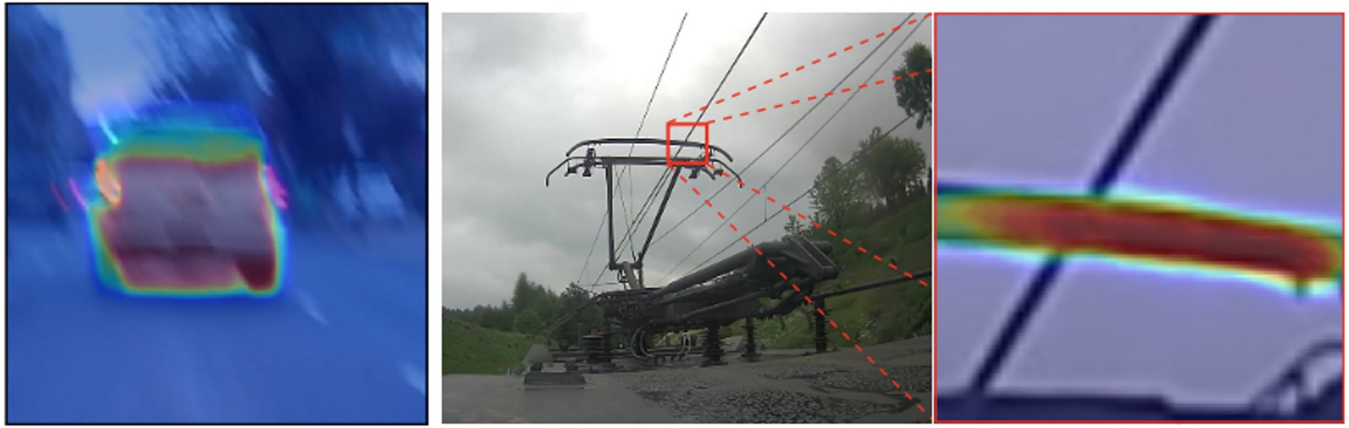


Figure 5. Failure cases: motion blur and “non-object” instance.