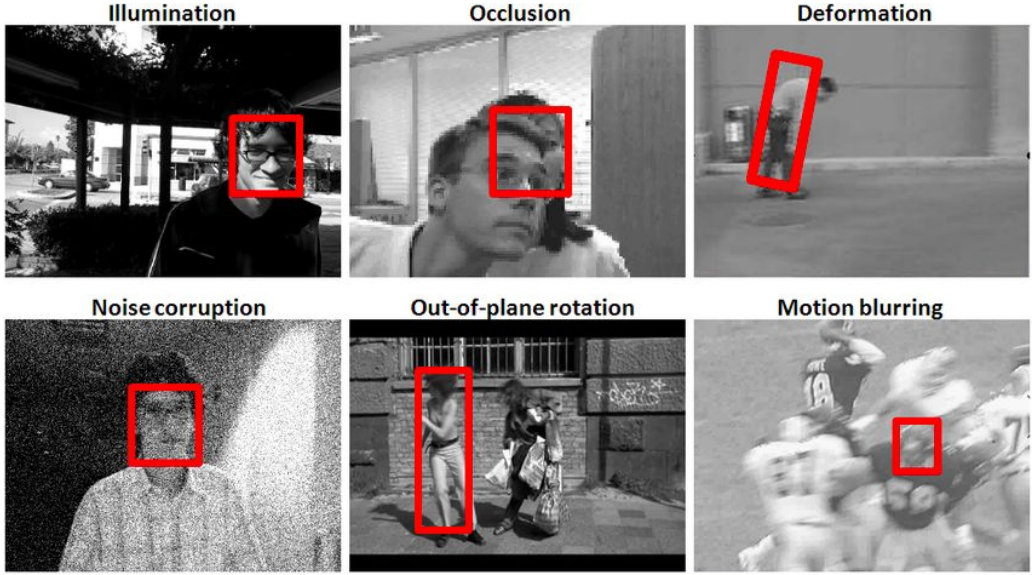
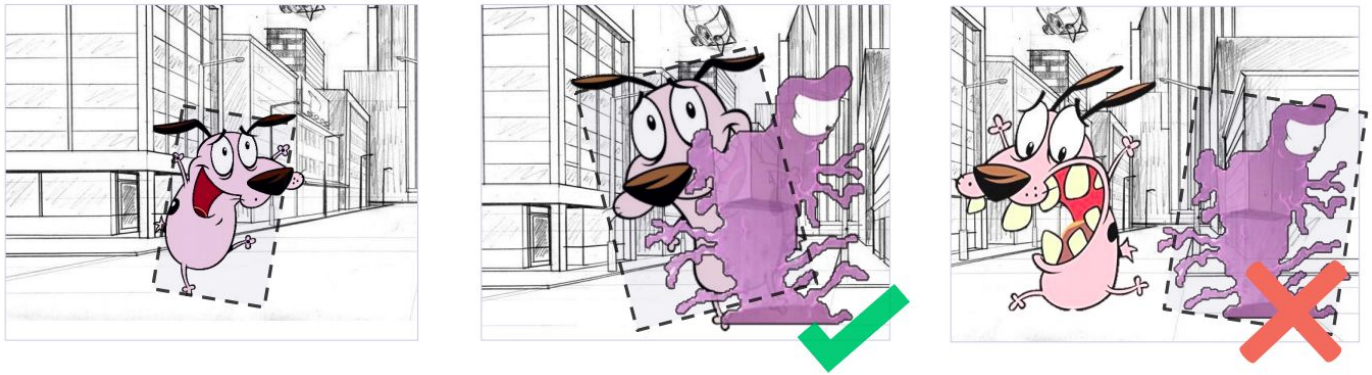
Visual object tracking

Беляев Владислав

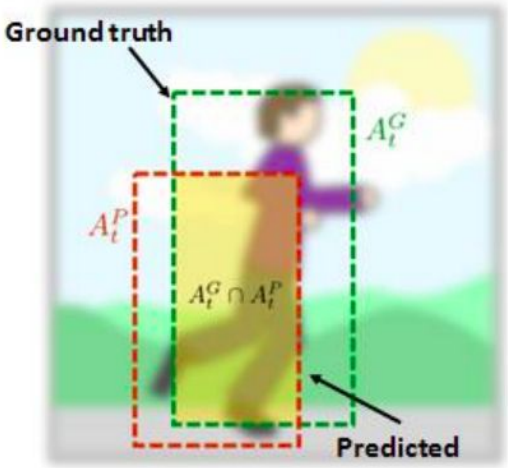
Трекинг объектов

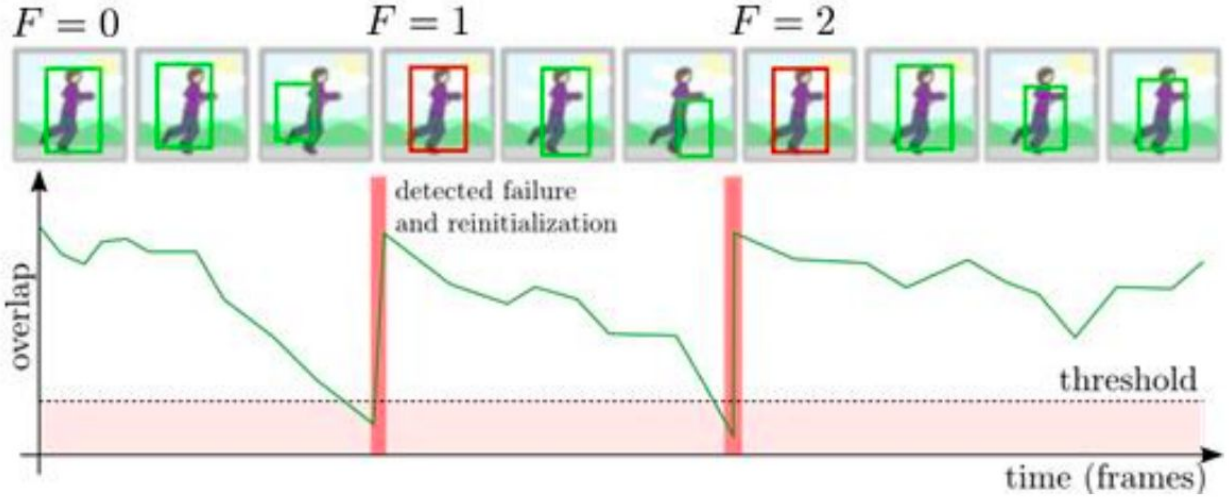
- Задача: отследить перемещение объекта на последовательности кадров, т.е. предсказать его местоположение для каждого кадра
- Положение задается прямоугольной вращающейся рамкой
- Начальное положение объекта дано





Performance measures



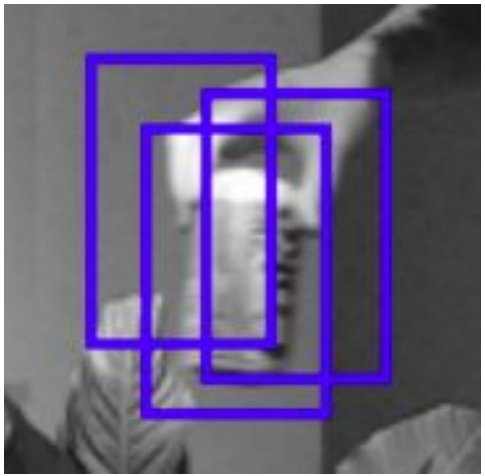


Две наименее коррелирующие между собой метрики:

- Robustness, количество раз когда трекер потерял объект
- Accuracy, среднее перекрытие ограничивающих рамок
- Expected Average Overlap (EAO), комбинация базовых метрик как критерий определения абсолютного победителя в VOT challenge

SotA 2018

Tracking approaches in top 10:



Winner of the realtime challenge:

SiamRPN (Bo Li et al. 2018)

Aim to discriminate foreground
from the non-semantic background.

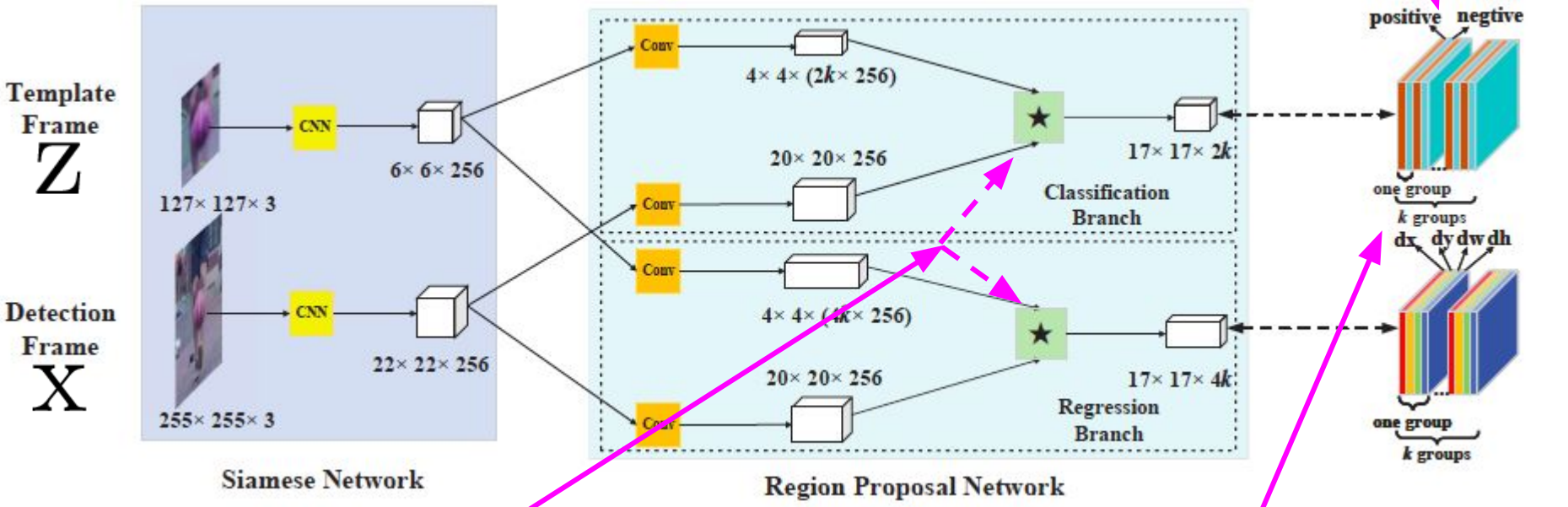
Winner of the VOT2018 challenge:

MFT (Martin Danelljan et al. 2018)

Train classifier to incorporate information
from all subwindows in frequency domain.

Siamese Region Proposal Network

$\mathrm{IoU} > \mathbf{0.6}$

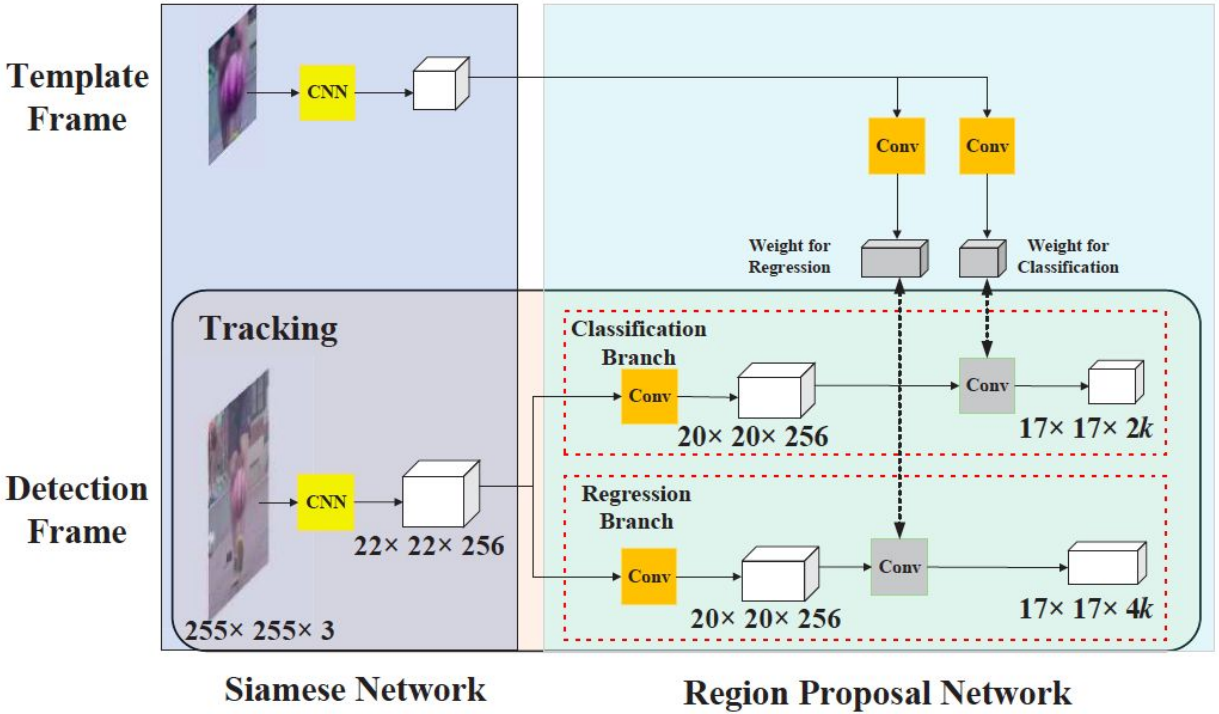


$$A^{cls}_{w\times h\times 2k} = [\varphi(x)]_{cls} \star [\varphi(z)]_{cls}$$

$$A^{reg}_{w\times h\times 4k} = [\varphi(x)]_{reg} \star [\varphi(z)]_{reg}$$

$$loss = L_{cls} + \lambda L_{reg}$$

Tracking as one-shot detection

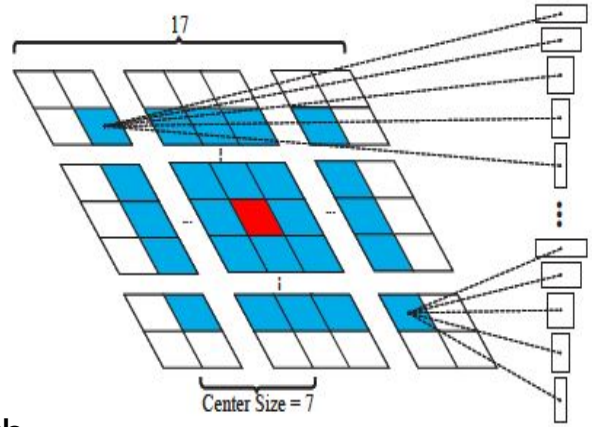


Meta-learning: we can now reinterpret the template branch in Siamese subnetwork as training parameters to predict the kernel of the local detection task

Proposal selection

Inference phase:

- Collect the top K bounding boxes (anchor + refinement from regression) according to positive activations
- Discarding the bounding boxes generated by the anchors too far away from the center (more than 7 pixels)
- The top K proposals are re-ranked after multiply the classification score by the temporal penalty
- Non-maximum-suppression
- Select final bounding box



$$penalty = e^{k*max(\frac{r}{r'},\frac{r'}{r})*max(\frac{s}{s'},\frac{s'}{s})}$$

r and r’ - proposal’s and last frame ratio of height and width
s and s’ - proposal’s and last frame scale

DCF paradigm

A linear classifier:

$$\min_{\mathbf{w},b} \sum_{i=1}^{m} L\left(y_i,\ f(\mathbf{x}_i)\right) + \lambda \left\|\mathbf{w}\right\|^2$$

Kernel Trick perform feature transformation:

$$\mathbf{x} \rightarrow \phi(\mathbf{x}) \quad < \mathbf{x}, \mathbf{x}' > \rightarrow < \phi(\mathbf{x}), \phi(\mathbf{x}') > = \kappa(\mathbf{x}, \mathbf{x}')$$

Solution can be expanded as a linear combination of the inputs:

$$\mathbf{w} = \sum_i \alpha_i \varphi(\mathbf{x}_i) \quad \boldsymbol{\alpha} = \left(K + \lambda I\right)^{-1} \mathbf{y}$$

$$\boldsymbol{\alpha} = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\mathbf{y})}{\mathcal{F}(\mathbf{k}) + \lambda}\right) \qquad k_i = \kappa(\mathbf{x},\ P^i\mathbf{x}), \quad \forall i = 0, \ldots, n-1$$

Given a single input z:

$$y' = \sum_i \alpha_i \kappa(\mathbf{x}_i, \mathbf{z}).$$

$$\bar{k}_i = \kappa(\mathbf{z},\ P^i\mathbf{x})$$

$$\hat{\mathbf{y}} = \mathcal{F}^{-1}\left(\mathcal{F}(\bar{\mathbf{k}}) \odot \mathcal{F}(\boldsymbol{\alpha})\right)$$

MFT

- **Tracking:**

Dense sampling: classifiers that are trained with all subwindows (of fixed size) of an image. The circulant matrix for this sample is generated by collecting its full cyclic shifts:

$$\boldsymbol{X}^\top = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n^2}]^\top \in \mathbb{R}^{n^2 \times n^2} \qquad n \times n \text{ image patch } \boldsymbol{x} \in \mathbb{R}^{n^2 \times 1}$$

Our goal is to learn a discriminative function:

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \boldsymbol{x}_i$$

$$f(\boldsymbol{X}; \theta) = \boldsymbol{\theta}^\top \boldsymbol{X} = \boldsymbol{\theta} \circledast \boldsymbol{x} = \mathcal{F}^{-1}\left(\hat{\boldsymbol{\theta}} \odot \hat{\boldsymbol{x}}^*\right) \qquad \boldsymbol{x}_* = \arg\max_{\boldsymbol{x}_i} f(\boldsymbol{x}_i; \boldsymbol{\theta}_{\text{model}})$$

- **Learning:**

$$\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta}} \mathcal{E}(\boldsymbol{\theta}, \boldsymbol{\mathcal{D}}) + \mathcal{R}(\boldsymbol{\theta})$$

- **Updating:**

$$\boldsymbol{\theta}_{\text{model}} = (1 - \alpha)\boldsymbol{\theta}_{\text{model}} + \alpha\boldsymbol{\theta}_*$$

Tricks and improvements

- Gaussian Shaped labels:

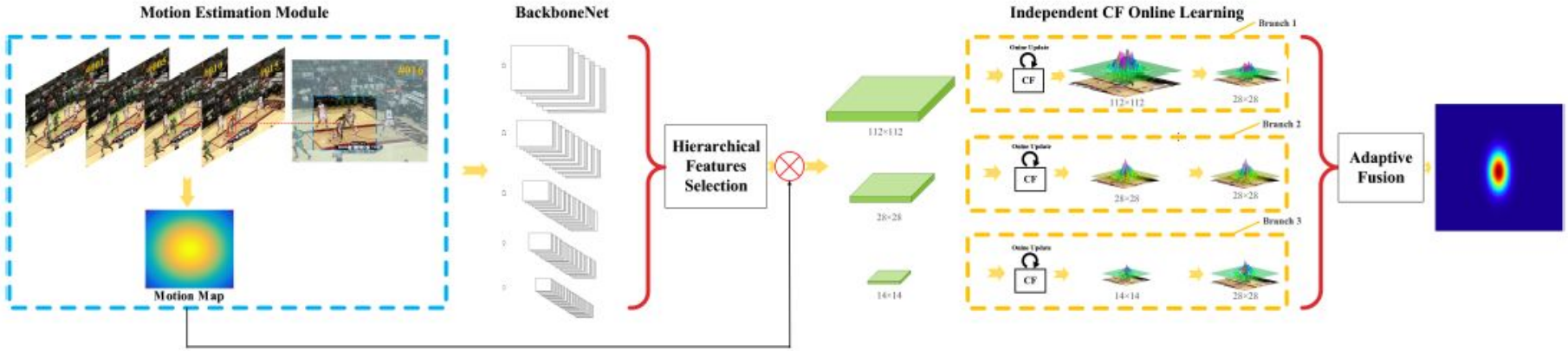
$$y_{ij} = \exp\left(-\left((i-i')^2 + (j-j')^2\right)/s^2\right), \quad \forall i,j = 0,\ldots,n-1$$

- Cosine window:

$$x_{ij} = \left(x_{ij}^{\text{raw}} - 0.5\right)\sin(\pi i/n)\sin(\pi j/n), \quad \forall i,j = 0,\ldots,n-1$$

- Motion Estimation Module:

Multi-hierarchical Independent Correlation Filters



- The motion estimation module previously locates the target by learning the law of movement of the target then determines the search area.
- Then hierarchical features are extracted by CNN (also HOG and Color Names) to represent the hierarchical semantic information.
- After that, hierarchical features independently fed into different CFs to online update the parameters.
- Adaptive multi-branch CF fusion is utilized to generate the final score map to locate the center point coordinates of the target.

Result and discussion

		Baseline			Realtime		
	Tracker	EAO	A	R	EAO	A	R
1.	LADCF	0.389 ①	0.503	0.159 ③	0.066	0.314	1.358
2.	MFT	0.385 ②	0.505	0.140 ①	0.060	0.337	1.592
3.	SiamRPN	0.383 ③	0.586 ①	0.276	0.383 ①	0.586 ①	0.276 ②
4.	UPDT	0.378	0.536	0.184	0.068	0.334	1.363
5.	RCO	0.376	0.507	0.155 ②	0.066	0.400	1.704
6.	DRT	0.356	0.519	0.201	0.062	0.321	1.503
7.	DeepSTRCF	0.345	0.523	0.215	0.063	0.418	1.817
8.	CPT	0.339	0.506	0.239	0.081	0.479	1.358
9.	SA_Siam_R	0.337	0.566 ②	0.258	0.337 ②	0.566 ②	0.258 ①
10.	DLSTpp	0.325	0.543	0.224	0.125	0.514	0.824

- Most failures due to: Occlusion
- Mostly affects accuracy: Occlusion + Scale change

Вопросы?