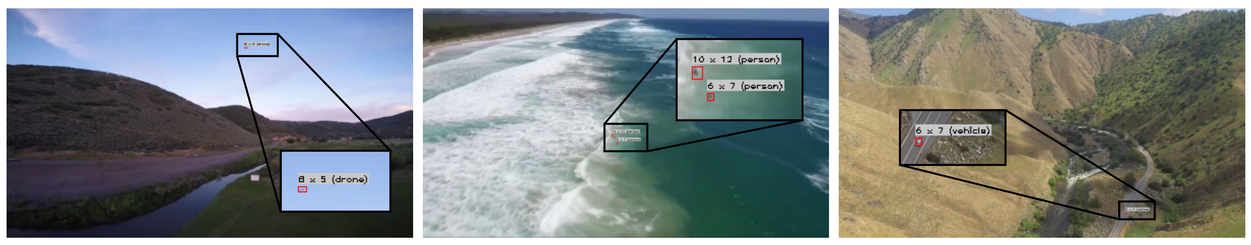
Correlation-based ConvNet for Small Object Detection in Videos

Brais Bosquet, Manuel Mucientes, Víctor M. Brea
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
University of Santiago de Compostela. Santiago de Compostela, Spain

OVERVIEW

Goal:

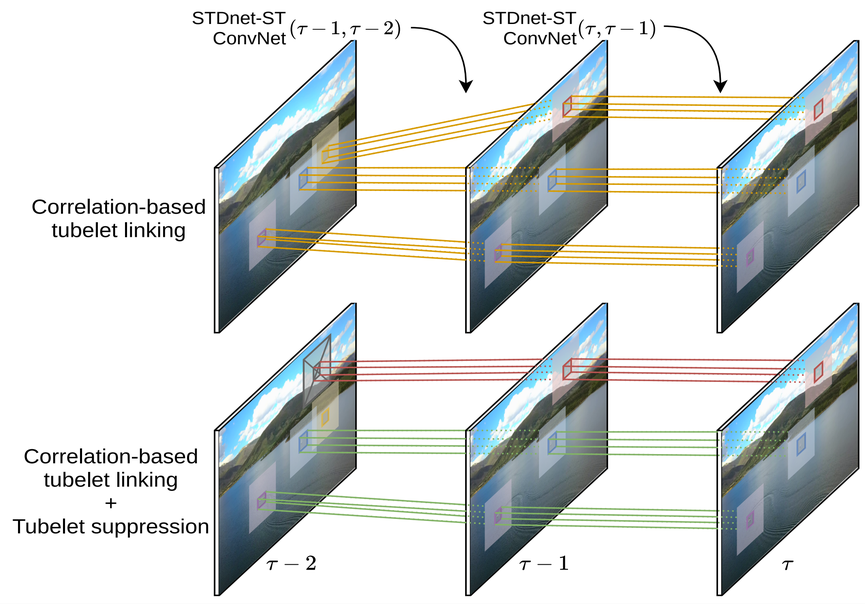
To detect small objects (under 16 x 16 px) in videos



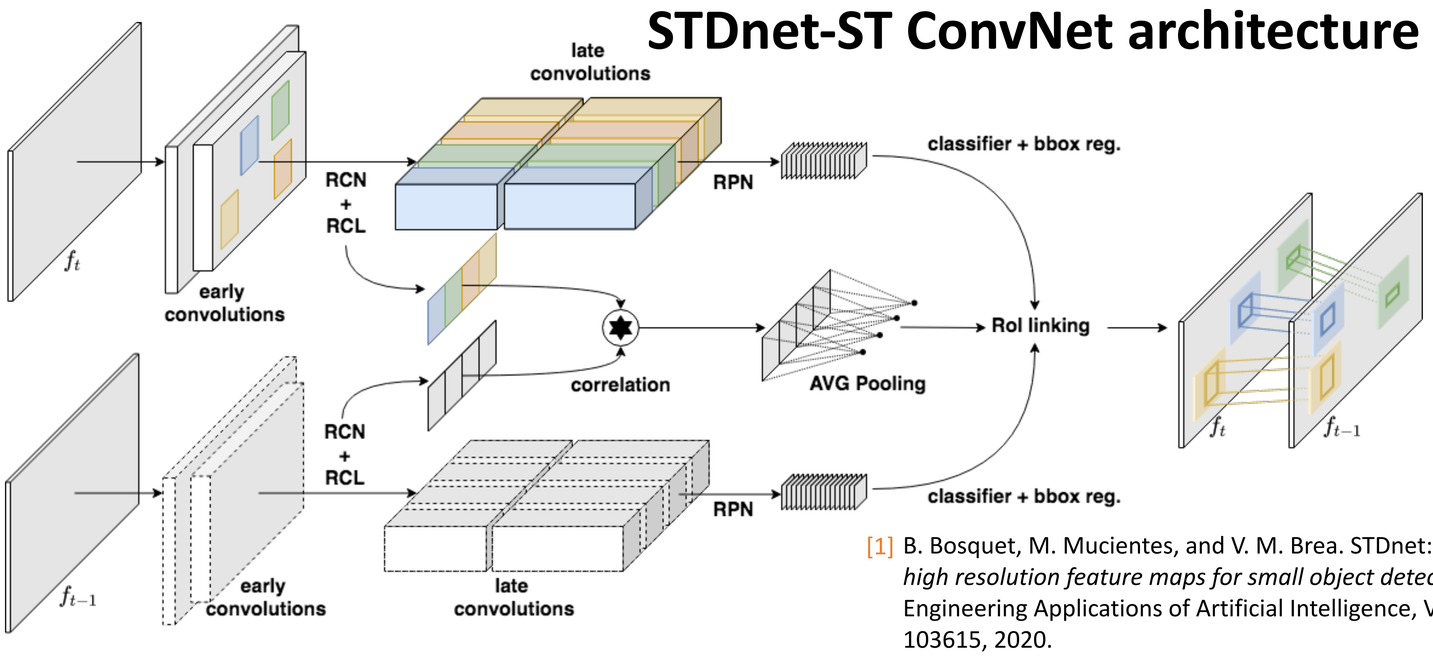
How it works:

Spatio-temporal STDnet (STDnet-ST) comprises two steps:

- **STDnet-ST ConvNet**: computes the detections of two consecutives frames and the correlation scores between them
- **STDnet-ST tubelet linking:**
  1. *Correlation-based tubelet linking*: generate object associations (tubelets) over time based on the Viterbi algorithm
  2. *Tubelet suppression algorithm*: identify and remove incorrect data associations



STDnet-ST ConvNet architecture



[1] B. Bosquet, M. Mucientes, and V. M. Brea. STDnet: *Exploiting high resolution feature maps for small object detection*. Engineering Applications of Artificial Intelligence, Vol. 91, No. 103615, 2020.

Two STDnet[1] branches

Correlation module

- *Correlation*: RCN regions[1]
- *RoI linking*: propagate correlation scores to final detections

STDnet-ST tubelet linking

a) Correlation-based tubelet linking

1. Compute score matrix for pair of frames
   - Replaces IoU with the correlation score

$$s_t^{ij} = p_{t-1}^i + p_t^j + \lambda \cdot c_t^{ij} \qquad c_t^{ij} = \rho\left(r_{t-1}^{k(i)}, r_t^{l(j)}\right)$$

2. Generate *tubelets* (Viterbi):

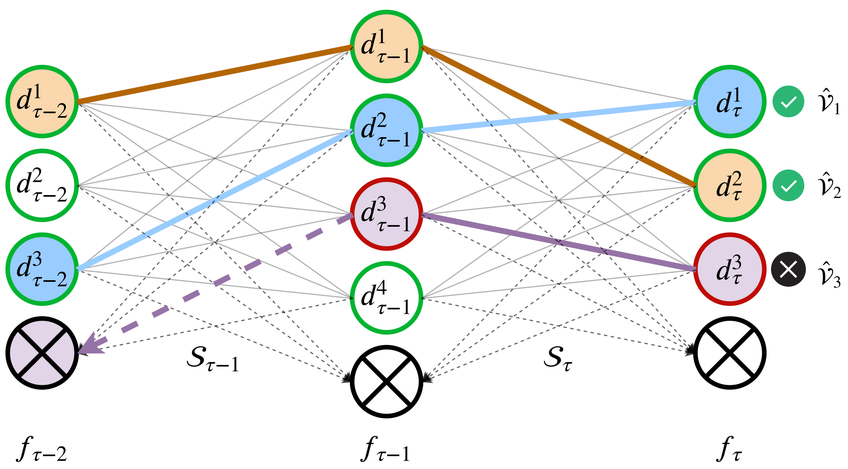
$$p_\tau^{i(\hat{v})} = \frac{1}{\tau}\sum_{t=1}^{\tau} p_t^{i(\hat{v})}$$

3. Confidence variability:

$$p_\tau^{i(\hat{v})} = \begin{cases} \max_{t=1}^{\tau} p_t^{i(\hat{v})} & \text{if } \sigma(\{p_t^{i(\hat{v})}\}_{t=1}^{\tau}) \leq \kappa \\ \frac{1}{\tau}\sum_{t=1}^{\tau} p_t^{i(\hat{v})} & \text{otherwise} \end{cases}$$

b) Tubelet suppression algorithm

- Discard unlikely tubelets using **dummy nodes** (⊗)



RESULTS & ACHIEVEMENTS

UAVDT

Method	$AP_{xs}^{@[.5,.95]}$	$AP_{xs}^{@.5}$
Faster R-CNN [27]	6.6	26.0
R-FCN [27]	9.2	32.5
RON [27]	3.7	19.7
SSD [27]	6.0	23.5
FGFA [154]	6.3	20.7
RDN [23]	9.3	27.9
MEGA [16]	9.2	26.6
FPN [78]	11.8	29.7
FPN-t	12.0	30.3
Cascade-FPN [13]	12.0	30.5
Cascade-FPN-t	12.3	31.2
STDnet [9]	12.5	35.1
STDnet++	12.6	35.4
STDnet-ST	13.1	36.0
STDnet-ST++	**13.3**	**36.4**

VisDrone2019- VID

Method	$AP_{xs}^{@[.5,.95]}$	$AP_{xs}^{@.5}$
FGFA [154]	3.8	16.8
RDN [23]	4.7	20.7
MEGA [16]	4.8	21.0
FPN [78]	6.2	19.9
Cascade-FPN [13]	6.1	20.2
FPN-t	6.3	20.2
Cascade-FPN-t	6.2	20.4
STDnet [9]	7.2	21.4
STDnet++	7.3	22.0
STDnet-ST	7.5	21.9
STDnet-ST++	**7.5**	**22.4**

USC-GRAD-STDdb

Method	$AP_{xs}^{@[.5,.95]}$	$AP_{xs}^{@.5}$
FGFA [154]	11.7	37.5
RDN [23]	15.5	48.6
MEGA [16]	17.4	53.1
FPN [78]	17.3	54.5
Cascade-FPN [13]	17.4	55.9
FPN-t	18.7	57.2
Cascade-FPN-t	19.1	58.9
STDnet [9]	18.3	57.8
STDnet++	18.9	59.1
STDnet-ST	20.1	62.1
STDnet-ST++	**21.4**	**63.4**

- **State-of-the-art** results:
  - UAVDT (*xs* subset)
  - VisDrone (*xs* subset)
  - USC-GRAD-STDdb[1]

- STDnet-ST generates object associations regardless the object temporal spatial distance. Critical for:
  - Small objects or fast motions
  - Low frame rate or skipping frames