Anomaly Detection in Video Sequence with Appearance-Motion Correspondence

Trong-Nguyen Nguyen and Jean Meunier

DIRO, University of Montreal, Montreal, QC, Canada

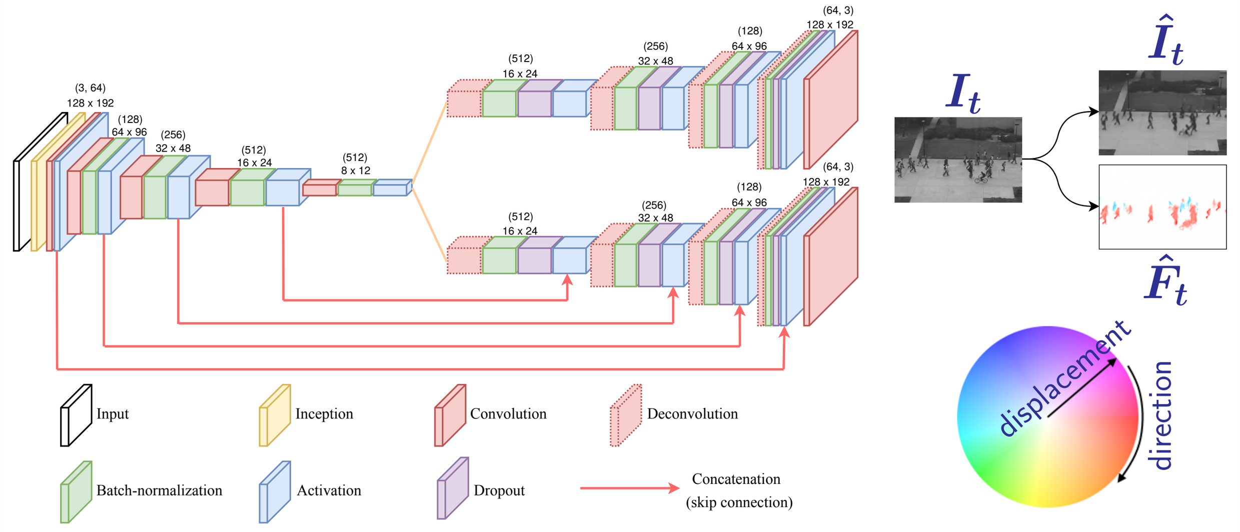
Université de Montréal

ICCV 2019 Seoul, Korea

Anomaly detection

- High diversity of possible anomalies
  - $\Rightarrow$ no general definition of anomaly
  - $\Rightarrow$ using only data of normality for training models

Proposed network



- Input: single frame of size $128 \times 192 \times 3$
- Groundtruth motion: FlowNet2 [Ilg *et al.*, CVPR2017]

Overall ideas

Considering common characteristics of normal events

- Learning regular appearance structures
  - $\Rightarrow$ using a convolutional auto-encoder
- Learning motions associated with these templates
  - $\Rightarrow$ using an U-Net translation model
- How to combine the two learnings?
  - $\Rightarrow$ sharing the encoding network
- Network depth for various camera distances?
  - $\Rightarrow$ let the network decide by itself using Inception
- And, how to estimate score of (ab)normality?
  - $\Rightarrow$ ~~computing on whole frame as other works~~
  - $\Rightarrow$ looking at the most unusual region

Learning appearance templates

- Typical problem of single frame reconstruction
- Objective function on intensity

$$\mathcal{L}_{int}(I, \hat{I}) = \|I - \hat{I}\|_2^2$$

- Constraint on gradient (reduce blur due to l_2 distance)

$$\mathcal{L}_{grad}(I, \hat{I}) = \sum_{d \in \{x,y\}} \left\| |g_d(I)| - |g_d(\hat{I})| \right\|_1$$

- Total loss for appearance stream

$$\mathcal{L}_{appe}(I, \hat{I}) = \mathcal{L}_{int}(I, \hat{I}) + \mathcal{L}_{grad}(I, \hat{I})$$

Learning associated motions

- Typical problem of image translation using U-Net
- Objective function on optical flow

$$\mathcal{L}_{flow}(F_t, \hat{F}_t) = \|F_t - \hat{F}_t\|_1$$

- Maybe an additional penalization would be better?
  - $\Rightarrow$ GANs worked well in related studies!
- We used conditional GAN
  - Condition: a video frame I_t
  - Input to classify: an optical flow F_t or $\hat{F}_t$

Frame-level normality score

- Considering small image patches

$$\begin{cases} \mathcal{S}_I(P) = \frac{1}{|P|} \sum_{i,j \in P} (I_{i,j} - \hat{I}_{i,j})^2 \\ \mathcal{S}_F(P) = \frac{1}{|P|} \sum_{i,j \in P} (F_{i,j} - \hat{F}_{i,j})^2 \end{cases}$$

$\mathcal{S}$: score function, P: patch (set of pixel positions)

- Patch location determined by motion stream

$$\tilde{P} \leftarrow \underset{P \text{ slides on frame}}{\operatorname{argmax}} \ \mathcal{S}_F(P)$$

- Score: weighted sum of 2 patches (for 2 streams)

$$\mathcal{S} = \log[w_F \mathcal{S}_F(\tilde{P})] + \lambda_{\mathcal{S}} \log[w_I \mathcal{S}_I(\tilde{P})]$$

$$\text{where} \begin{cases} w_F = \left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{F_i}(\tilde{P}_i)\right]^{-1} \\ w_I = \left[\frac{1}{n}\sum_{i=1}^{n} \mathcal{S}_{I_i}(\tilde{P}_i)\right]^{-1} \end{cases}$$

n: number of training frames

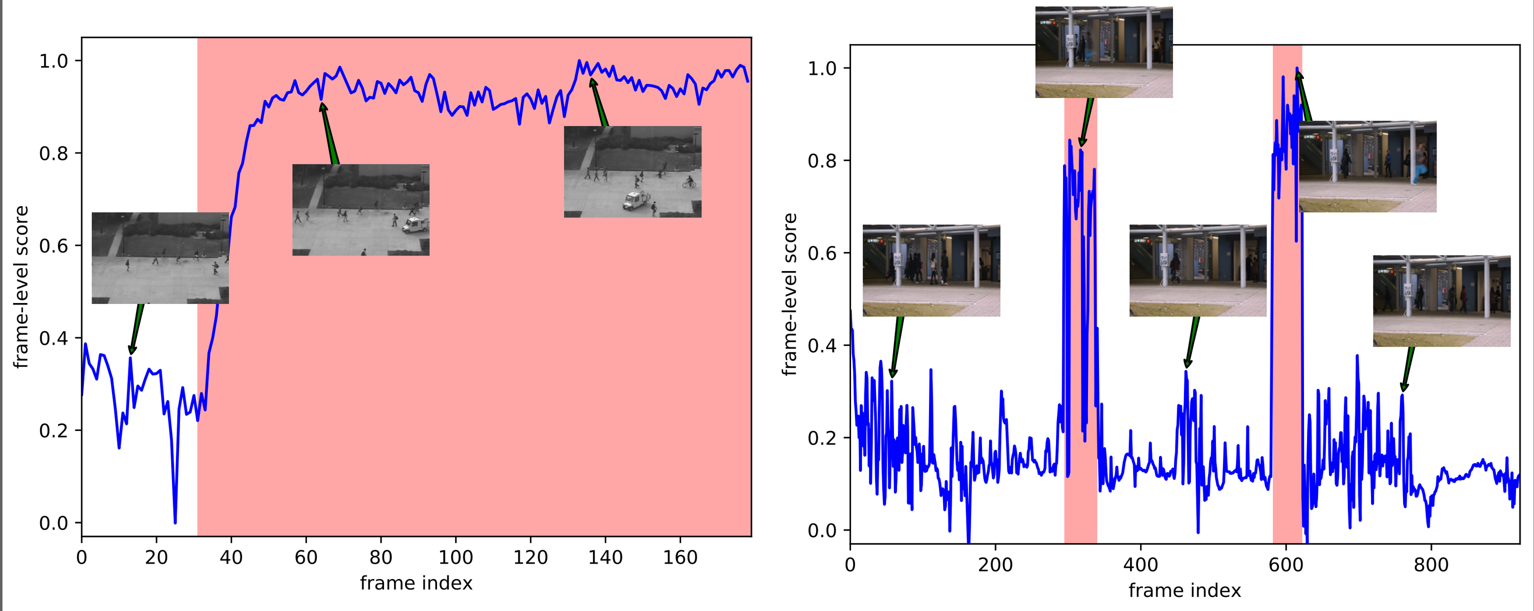
- (Optional) SSIM between I and $\hat{I}$
  - $\Rightarrow$ when we have problem with optical flow

Experimental results on frame-level anomaly detection

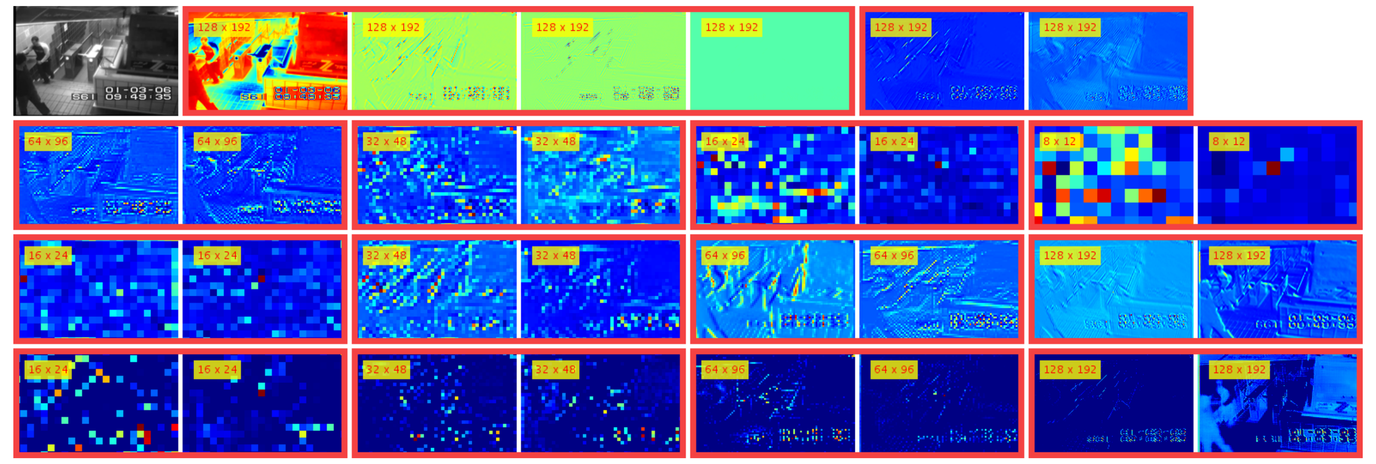
	Avenue†	Ped2†	Entrance	Exit	Belleview‡	Traffic-Train‡
Proposed architecture with motion stream						
Patch	0.869	0.962	61/18	17/5	0.751	0.490
SSIM	0.694	0.799	51/14	15/4	0.830	0.798
Architecture without motion stream						
Patch	0.702	0.773	58/16	14/7	0.838	0.380
SSIM	0.694	0.761	48/12	14/5	0.832	0.808

Note: True Positive / False Alarm for Entrance, Exit; †AUROC; ‡Avg Precision.

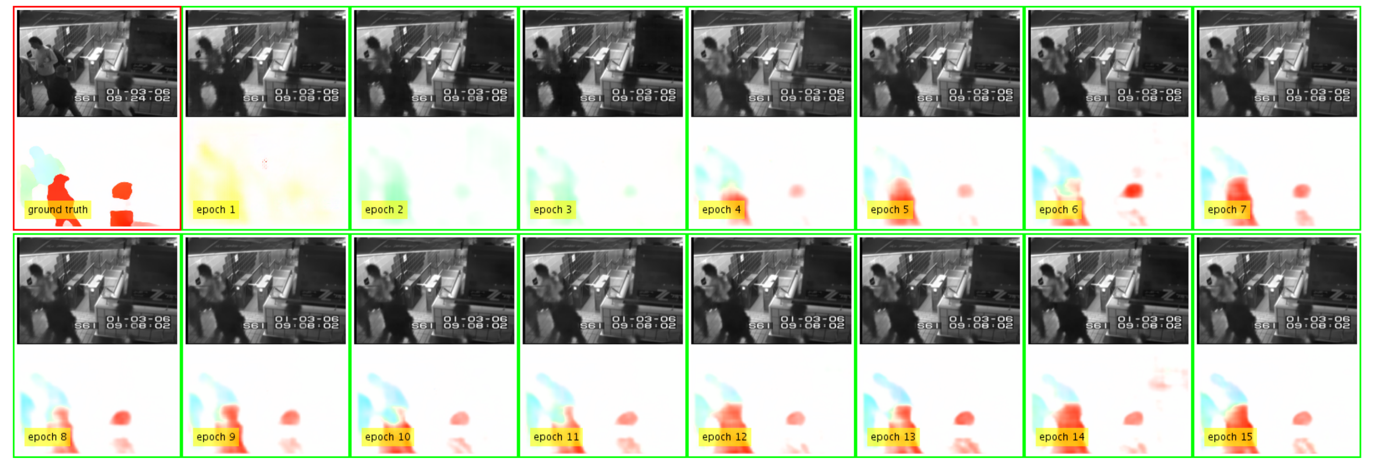
Demonstration of score sequence



Feature maps

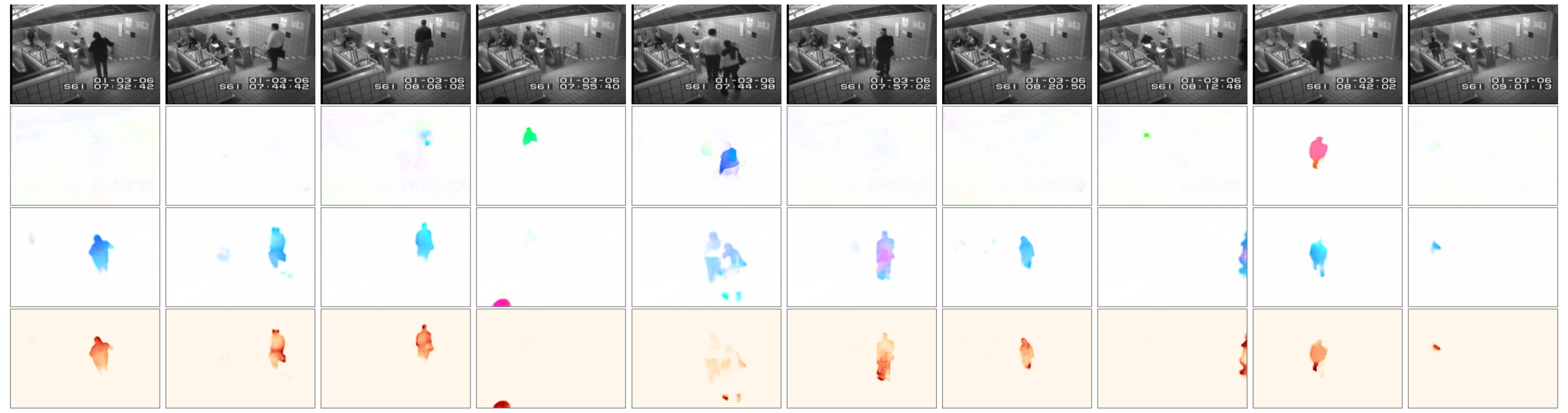


Outputs during optimization



UCSD Ped2 & CUHK Avenue

Subway datasets



Traffic-Train & Belleview

nguyetn@iro.umontreal.ca