# Anomaly Detection in Video Sequence with Appearance-Motion Correspondence
## — Supplementary Material —

Trong-Nguyen Nguyen, Jean Meunier
DIRO, University of Montreal
{nguyetn, meunier}@iro.umontreal.ca

## Abstract

*This supplementary material provides these contents:*

- *ROC curves of our frame-level scores on the CUHK Avenue and UCSD Ped2 datasets, and Precision-Recall (PR) curves on the traffic datasets.*
- *Experimental results of using either appearance reconstruction stream or motion prediction stream for score estimation.*
- *Impact of integrating motion stream and patch-based score estimation.*
- *Visualization of some feature maps in different blocks obtained in our experiments.*
- *Reconstructed frames and predicted motions after some training epochs.*

## 1. Flow field color coding

Figure 1 shows the color coding used in visualization of our optical flow in the main paper. This color coding is similar to [5] where the color indicates motion direction and the saturation corresponds to the pixel displacement.

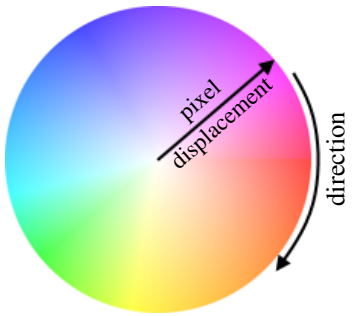



Figure 1: The color coding used for visualizing our optical flow in the main paper.

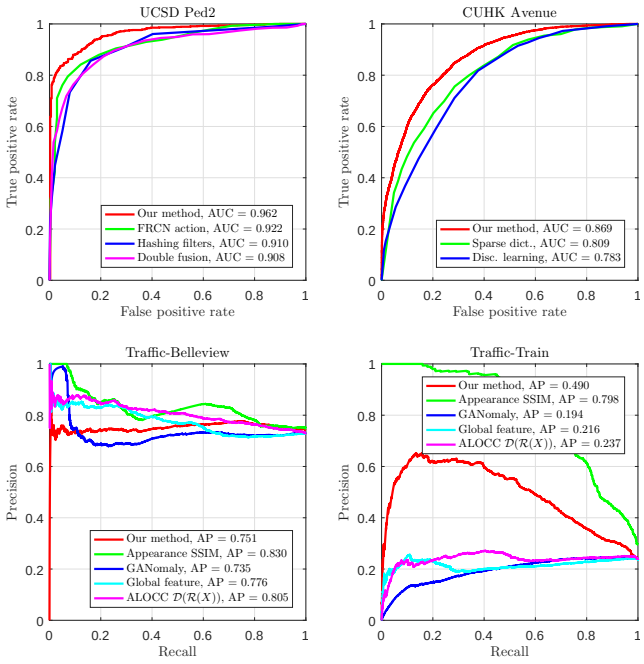



Figure 2: Top: ROC curves on the Ped2 and Avenue datasets. Bottom: PR curves on the Belleview and Train datasets. The corresponding Area Under Curve (AUC) and Average Precision (AP) are also provided. Best viewed in color.

## 2. Evaluation curves on 4 datasets

Figure 2 displays ROC and PR curves of our frame-level scores obtained in the experiments. Some state-of-the-art methods are also added into the figure to provide a visual comparison. These methods consist of FRCN action [4], hashing filters [10], AMDN double fusion [9], sparse dictionary [6], discriminative learning [3], GANomaly [1], autoencoder with global features [7] and ALOCC [8]. The ROC curves of the first 5 mentioned studies are provided in their original papers.

## 3. Experimental results on single streams

As indicated in the main paper, our frame-level score is estimated as a weighted combination of two partial scores

$$\mathcal{S} = \log[w_F \mathcal{S}_F(\tilde{P})] + \lambda_{\mathcal{S}} \log[w_I \mathcal{S}_I(\tilde{P})] \quad (1)$$

where $\mathcal{S}_F(\tilde{P})$ and $\mathcal{S}_I(\tilde{P})$ are respectively partial scores calculated from the motion and appearance streams, $w_F$ and $w_I$ are corresponding weights computed from the training data, $\lambda_{\mathcal{S}}$ is a hyperparameter controlling the contribution of partial scores to the summation, and $\tilde{P}$ is the patch providing the highest value of $\mathcal{S}_F$ in the considering frame.

In this section, we present the evaluation results in the cases of using only one of the two partial scores as the frame-level score indicator (see Figure 3). Both AUC and average precision (AP) measures are also provided for a convenient comparison with other studies. Note that the AUC and AP values are not comparable though there is a connection between ROC and PR spaces, and they are both affected by the balance of the two classes in each dataset [2].

Figure 3 shows that the combination of the two partial scores improved the detection ability since its AUC and AP increased compared with individual measures. For the Subway datasets, this combination reduced the risk of false detection, but the number of detected anomalous events was also slightly decreased (Subway Entrance).

## 4. Impact of motion stream and patch-based score estimation for anomaly detection

Table 1 shows the experimental results obtained on the 6 benchmark datasets using patch-based normality assessment and SSIM on appearance stream. We also remove the motion stream and the motion-oriented discriminator (Sections 3.3 & 3.4 in main paper) for the assessment of motion impact. SSIM was suggested due to the errors in optical flow measurement (camera jitter in Traffic-Train and low-quality frames in Belleview). Without motion stream, the model becomes a reconstruction auto-encoder of *single* frame, and the results on the first 5 datasets still demonstrate the efficiency of the proposed patch-based normality score. Using motion significantly improved results of the first 4 datasets while SSIM on appearance stream was just slightly reduced for the others (*i.e*. 0.830 vs. 0.832 for Belleview, and 0.798 vs. 0.808 for Traffic-Train).

| | Avenue† | Ped2† | Entran. | Exit | Belle.‡ | 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; ‡AP.

Table 1: Experimental results using patch-based normality assessment and SSIM on appearance stream.

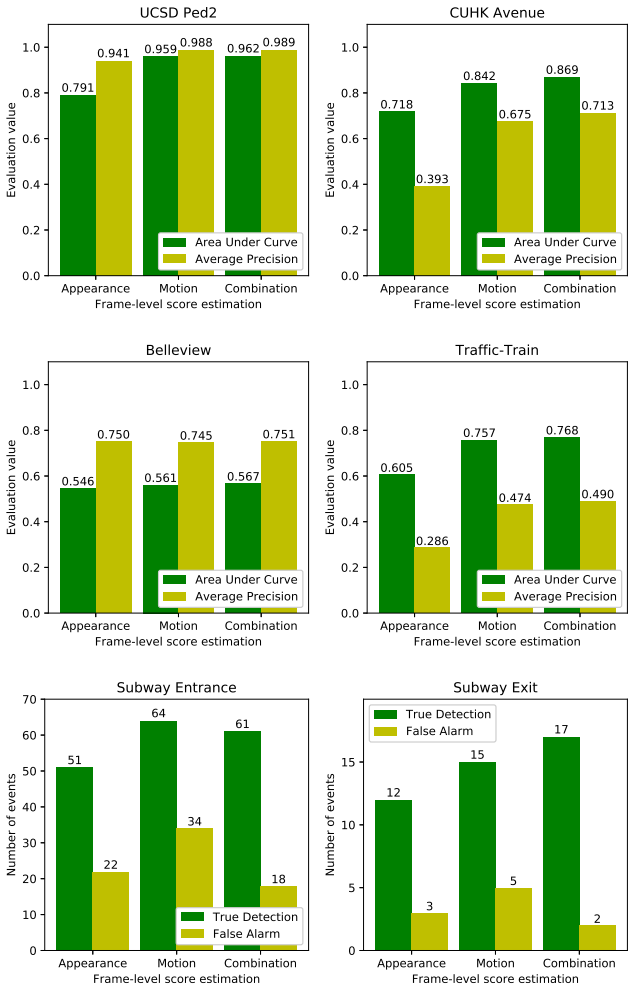



Figure 3: Evaluation results of our model using only the appearance reconstruction (Conv-AE), the motion prediction (U-Net) and their combination. The frame-level AUROC and Average Precision scores are provided for the Ped2, Avenue, Belleview and Train datasets. The numbers of true positive detections (*i.e*. true positive) and false alarms are presented for the Entrance and Exit datasets.

## 5. Feature maps

A visualization of some feature maps given an input frame for each dataset is shown in Figure 4. Each example is represented by 4 rows of images. We illustrate two feature maps (grouped in a red bounding box) for each layer block, except for the Inception module where 4 feature maps are shown for the $1 \times 1$, $3 \times 3$, $5 \times 5$ and $7 \times 7$ convolutional

| Dataset | Training epoch | Batch size |
|---|---|---|
| UCSD Ped2 | 15 | 16 |
| CUHK Avenue | 25 | 8 |
| Subway Entrance | 25 | 16 |
| Subway Exit | 15 | 8 |
| Traffic-Train | 25 | 16 |
| Traffic-Belleview | 120 | 8 |

Table 2: Number of training epochs and batch size in our experiments. These values were selected according to the number of training images in each dataset and the memory capacity of our hardware (Intel i7-7700K, 16 GB memory, GTX 1080).

filters. The first two rows include the input frame, activation maps resulting from the Inception module and subsequent blocks of the shared encoder. The third and fourth rows respectively consist of feature maps in the decoder of motion and appearance streams. The value of units in each map was normalized to provide a good visualization.

Figure 4 shows that our motion stream attempts to emphasize the image edges to provide a smooth optical flow (because FlowNet2 [5] was used as the ground truth motion estimator) while the other one tends to reconstruct appearance textures. By observing all feature maps provided by the Inception module, we found that $7 \times 7$ convolutional filters extracted informative details only on the CUHK Avenue, Subway Entrance and Traffic-Belleview datasets (best viewed when the feature map is enlarged). It demonstrated the reasonable use of Inception module right after the input layer to let the network automatically decides its appropriate low-level filter sizes.

## 6. Model optimization during training phase

In this section, we show the outputs of the proposed model after some training epochs given the same input for each dataset. The number of training epochs and batch size are presented in Table 2.

In Figure 5, the correspondence between a reconstructed frame and its predicted motion can be clearly observed. A sharper frame would be obtained together with a motion with more details (*e.g.* epochs 2 vs. 4 in the UCSD Ped2 experiment) as the number of epochs increases. It also demonstrates that the model encountered difficulty in optimizing the two streams on the Traffic-Train dataset due to the sudden change of lighting and camera jitter. However, the overall structure of the acquired scene was still preserved (*e.g.* poles and seats). The use of SSIM on the input frame and its reconstruction hence improved the anomaly detection results (presented in the main paper).

## References


[1] Samet Akcay, Amir Atapour-Abarghouei, and Toby P. Breckon. Ganomaly: Semisupervised anomaly detection via adversarial training. In *Computer Vision – ACCV 2018*, Cham, 2018. Springer International Publishing.

[2] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

[3] Allison Del Giorno, J. Andrew Bagnell, and Martial Hebert. A discriminative framework for anomaly detection in large videos. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 334–349, Cham, 2016. Springer International Publishing.

[4] Ryota Hinami, Tao Mei, and Shin'ichi Satoh. Joint detection and recounting of abnormal events by learning deep generic knowledge. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3639–3647, Oct 2017.

[5] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

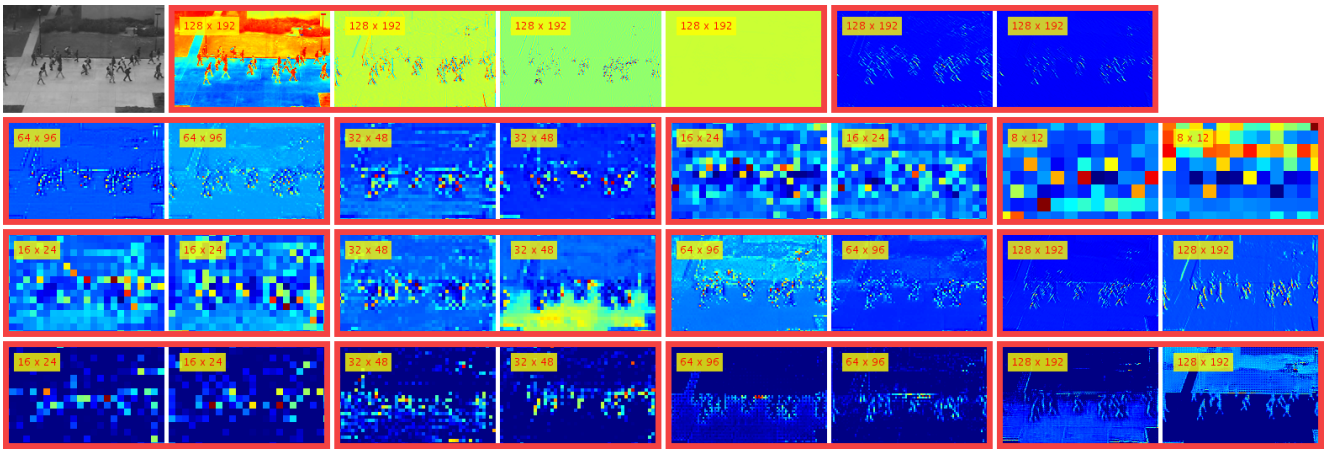
[6] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *2013 IEEE International Conference on Computer Vision*, pages 2720–2727, Dec 2013.

[7] Medhini G. Narasimhan and Sowmya Kamath S. Dynamic video anomaly detection and localization using sparse denoising autoencoders. *Multimedia Tools and Applications*, 77(11):13173–13195, Jun 2018.

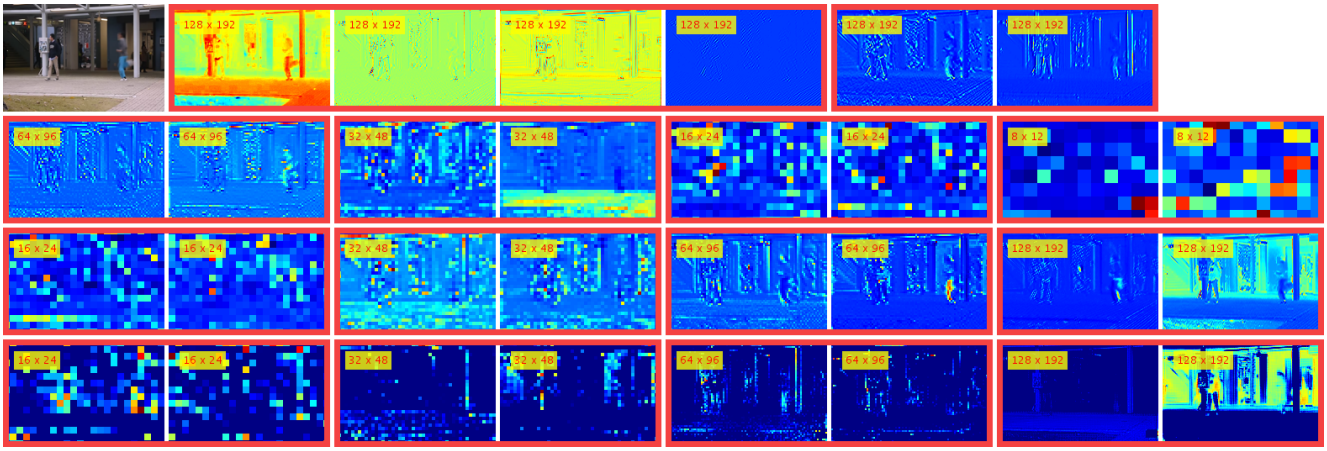
[8] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. Adversarially learned one-class classifier for novelty detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Dan Xu, Yan Yan, Elisa Ricci, and Nicu Sebe. Detecting anomalous events in videos by learning deep representations of appearance and motion. *Computer Vision and Image Understanding*, 156:117 – 127, 2017. Image and Video Understanding in Big Data.

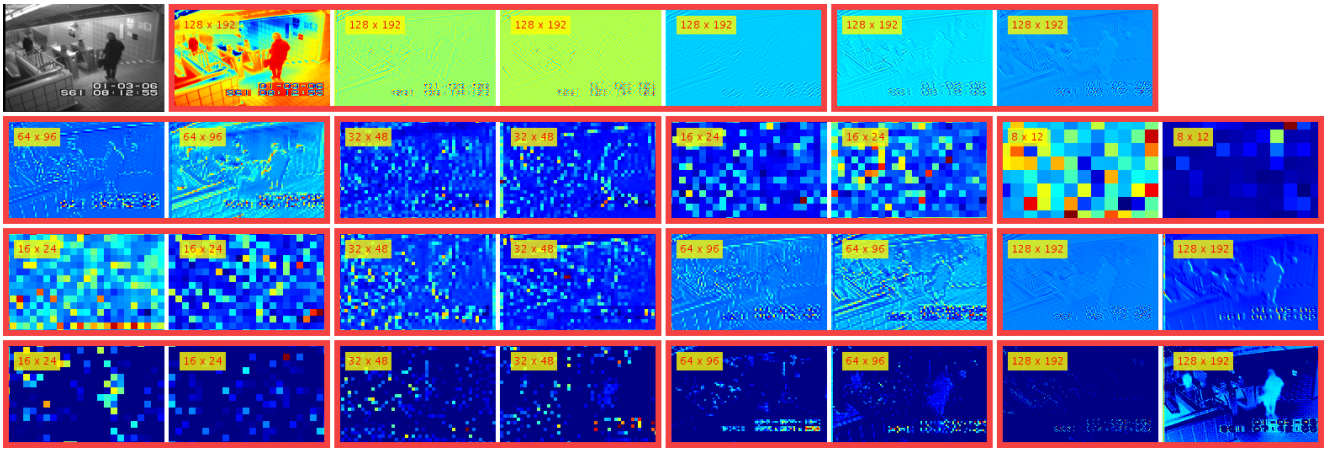
[10] Ying Zhang, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Shun Sakai. Video anomaly detection based on locality sensitive hashing filters. *Pattern Recognition*, 59:302 – 311, 2016. Compositional Models and Structured Learning for Visual Recognition.



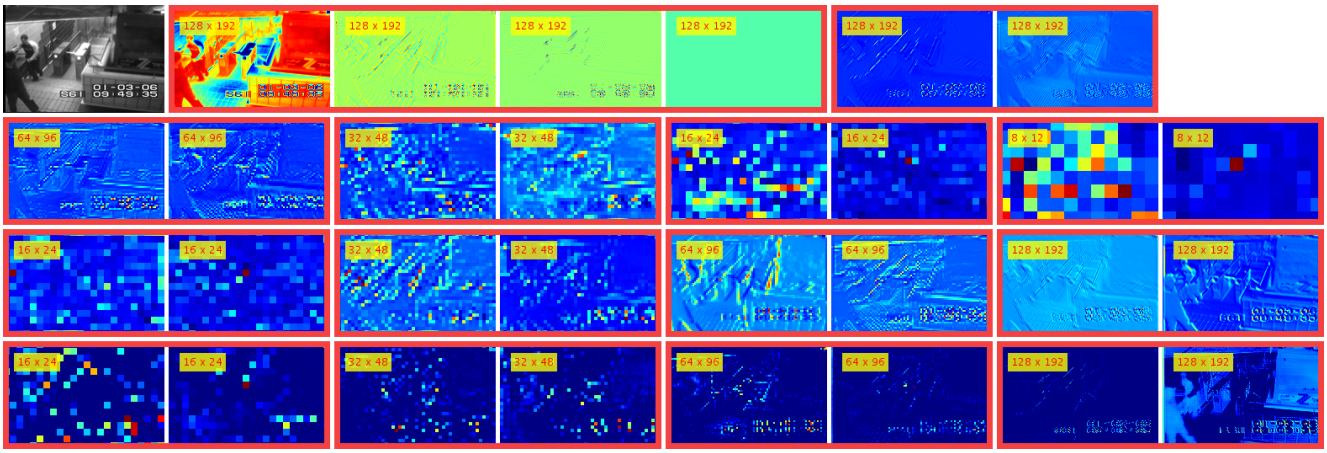

(a) UCSD Ped2



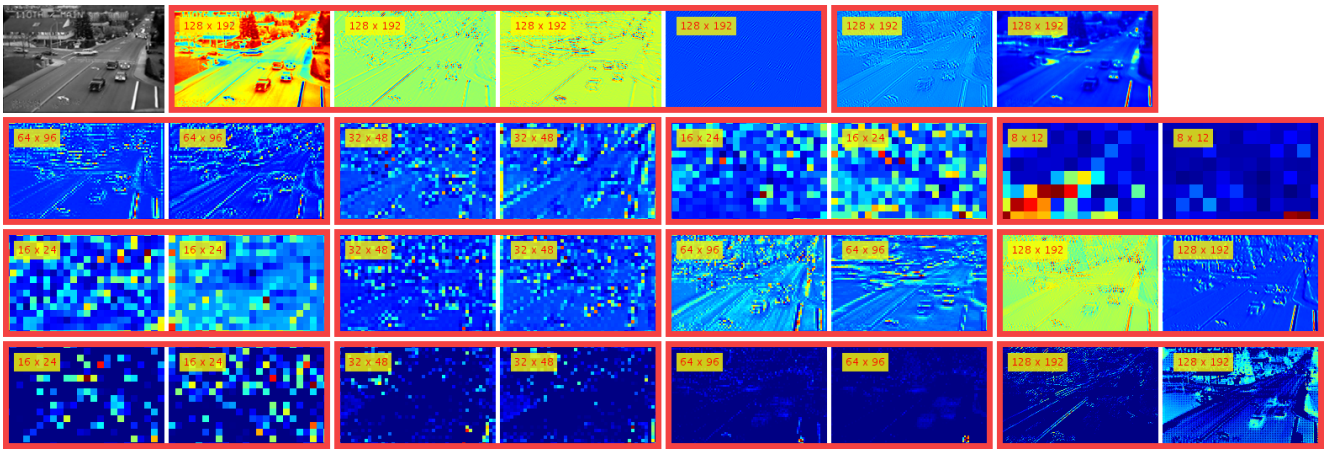

(b) CUHK Avenue



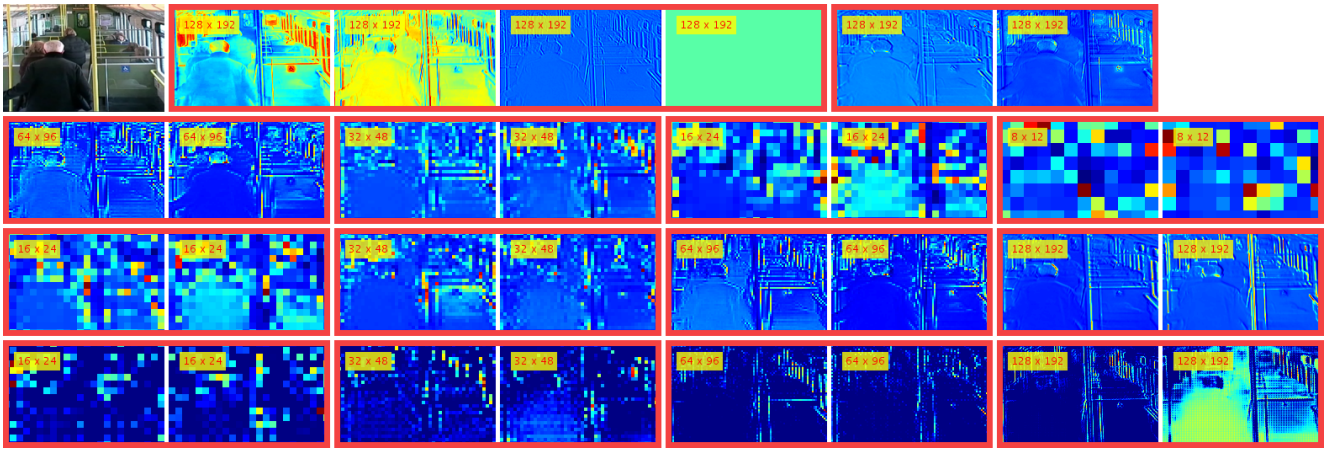

(c) Subway Entrance




(d) Subway Exit

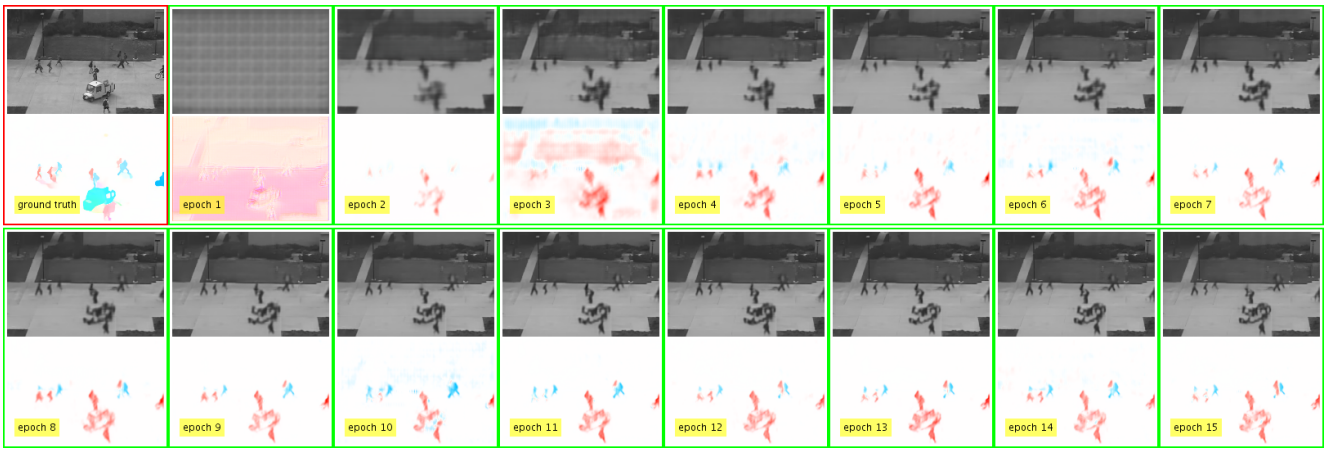



(e) Traffic-Belleview

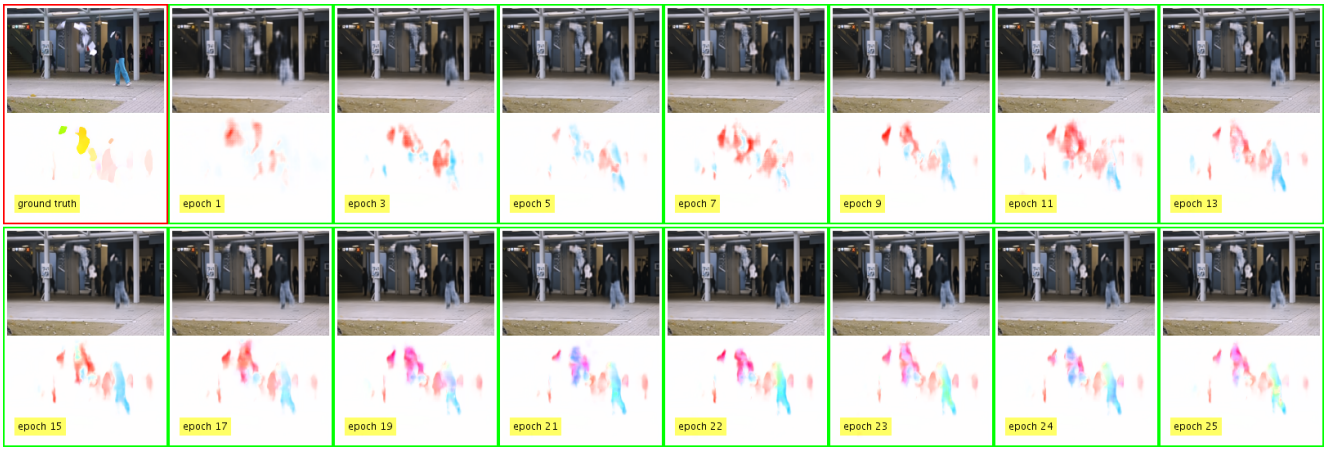



(f) Traffic-Train

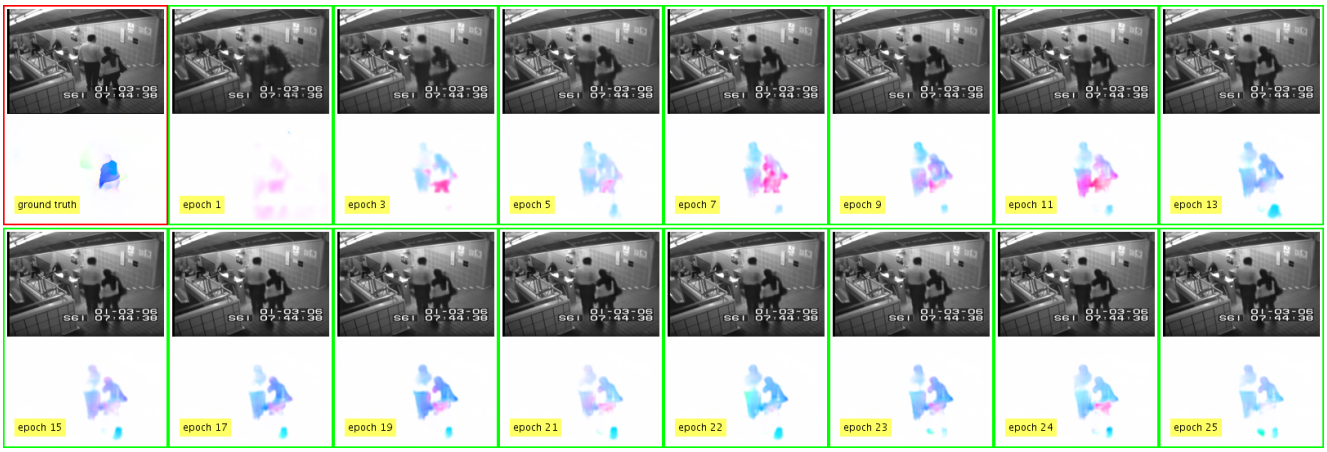
Figure 4: Visualization of some activation maps (together with their spatial resolution) given an input frame for each dataset. Channels sampled from the same block are grouped by a red bounding box. Best viewed in color.



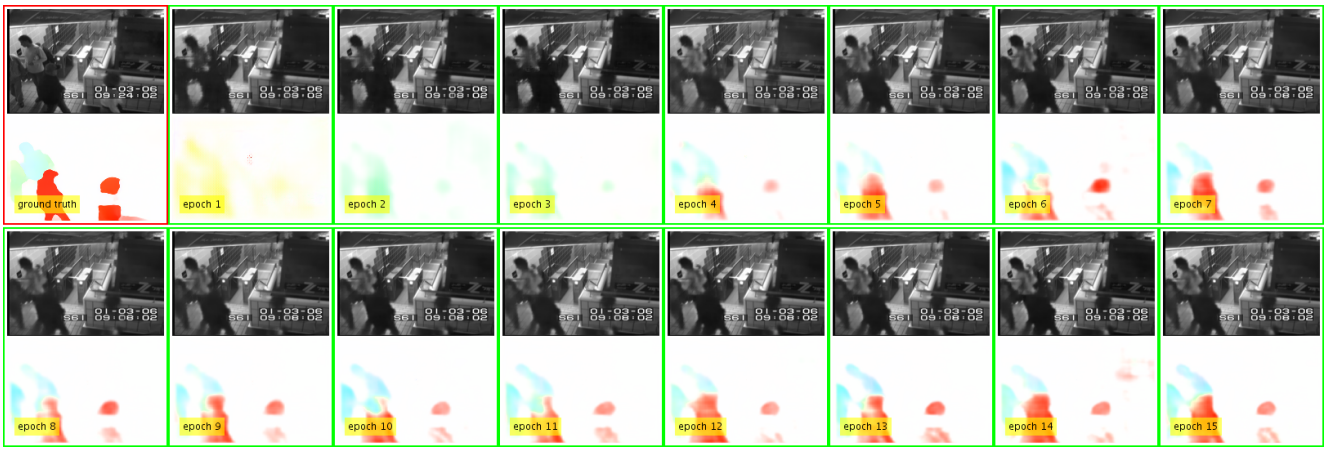

(a) UCSD Ped2



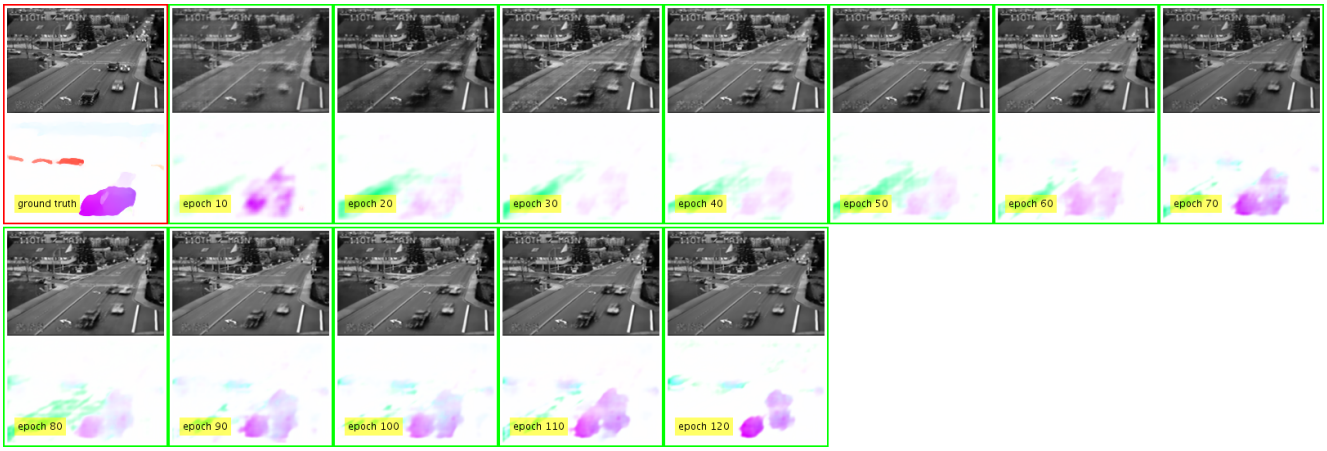

(b) CUHK Avenue



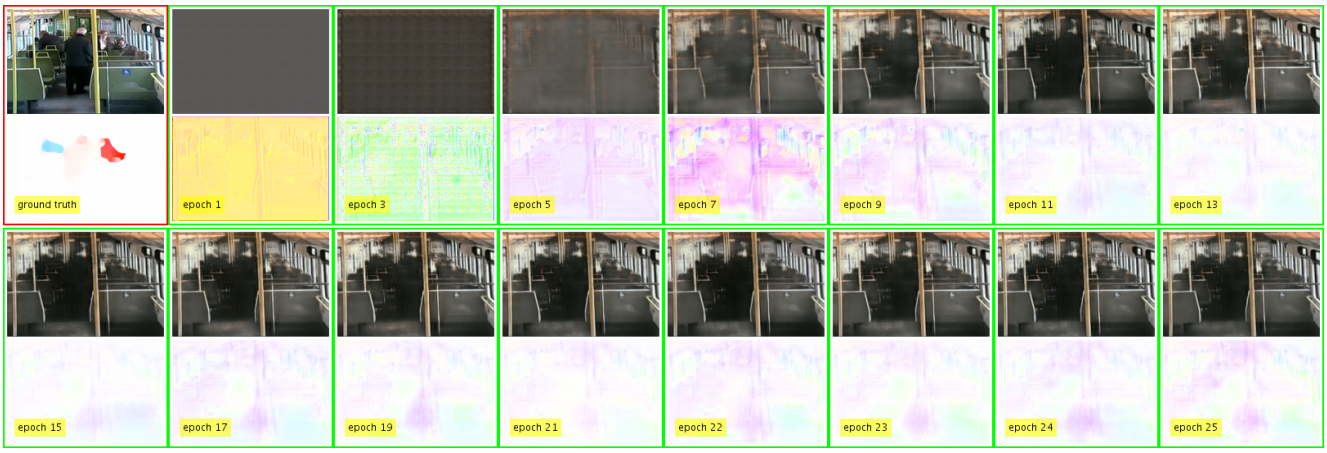

(c) Subway Entrance




(d) Subway Exit




(e) Traffic-Belleview




(f) Traffic-Train

Figure 5: Visualization of model outputs provided by the two streams after some training epochs. Note that these input frames were from the test set and were not employed for training the models. Best viewed in color.