Learning Temporal Regularity in Video Sequences

Supplementary Materials

Mahmudul Hasan Jonghyun Choi[†] Jan Neumann[†] Amit K. Roy-Chowdhury Larry S. Davis[‡]

UC Riverside Comcast Labs, DC[†] University of Maryland, College Park[‡]

{mhasa004@,amitrc@ece.}ucr.edu {jonghyun_choi,jan_neumann}@cable.comcast.com[†] lsd@umiacs.umd.edu[‡]

Table of Contents

1. Dataset Details

We use three challenging datasets to demonstrate our methods. They are curated for anomaly or abnormal event detection and are referred to as Avenue [1], UCSD pedestrian [2], and Subway [3] datasets. We describe the details of datasets in the supplementary material.

Avenue. There are total 16 training and 21 testing video sequences. Each of the sequences is short; about 1 to 2 minutes long. The total number of training frames is $15,328$ and testing frame is $15,324$. Resolution of each frame is 640×360 pixels.

UCSD Pedestrian. This dataset has two different scenes - Ped1 and Ped2.

UCSD-Ped1. It has 34 short clips for training, and another 36 clips for testing. All testing video clips have frame-level ground truth labels. Each clip has 200 frames, with a resolution of 238×158 pixels.

UCSD-Ped2. It has 16 short clips for training, and another 12 clips for testing. Each clip has 150 to 200 frames, with a resolution of 360×240 pixels.

Subway. The videos are taken from two surveillance cameras in a subway station. One monitors the exit and the other monitors the entrance. In both videos, there are roughly 10 people walking around in a frame. The resolution is 512×384 pixels.

Subway-Entrance. It is 1 hour 36 minutes long with $144,249$ frames in total. There are 66 unusual events of five different types: (a) walking in the wrong direction (WD); (b) no payment (NP); (c) loitering (LT); (d) irregular interactions between people (II) and (e) misc, including sudden stop, running fast.

Subway-Exit. It is 43 minutes long with $64,901$ frames. Three types of unusual events are defined in the subway exit video: (a) walking in the wrong direction (WD), (b) loitering near the exit (LT), and (c) miscellaneous, including suddenly stop and look around, janitor cleaning the wall, someone gets off the train and gets on again very soon. In total, 19 unusual events are defined as ground truth.

Go to Table of Contents

2. Learned Temporal Regularity

In Section 4.2 in the main paper, we visualize the temporal regularity by 1) synthesizing the regular frame and 2) visualizing accumulated regularity score within a video as a heat-map obtained by convolutional autoencoder (conv-autoencoder). Here, we present more examples per each dataset with a heat-map obtained by the improved trajectory based autoencoder (IT-autoencoder) for comparison. Compared to conv-autoencoder's regular score, the regular score by IT-autoencoder is up to patch precision and cannot capture the regularity well.

2.1. CUHK Avenue Dataset

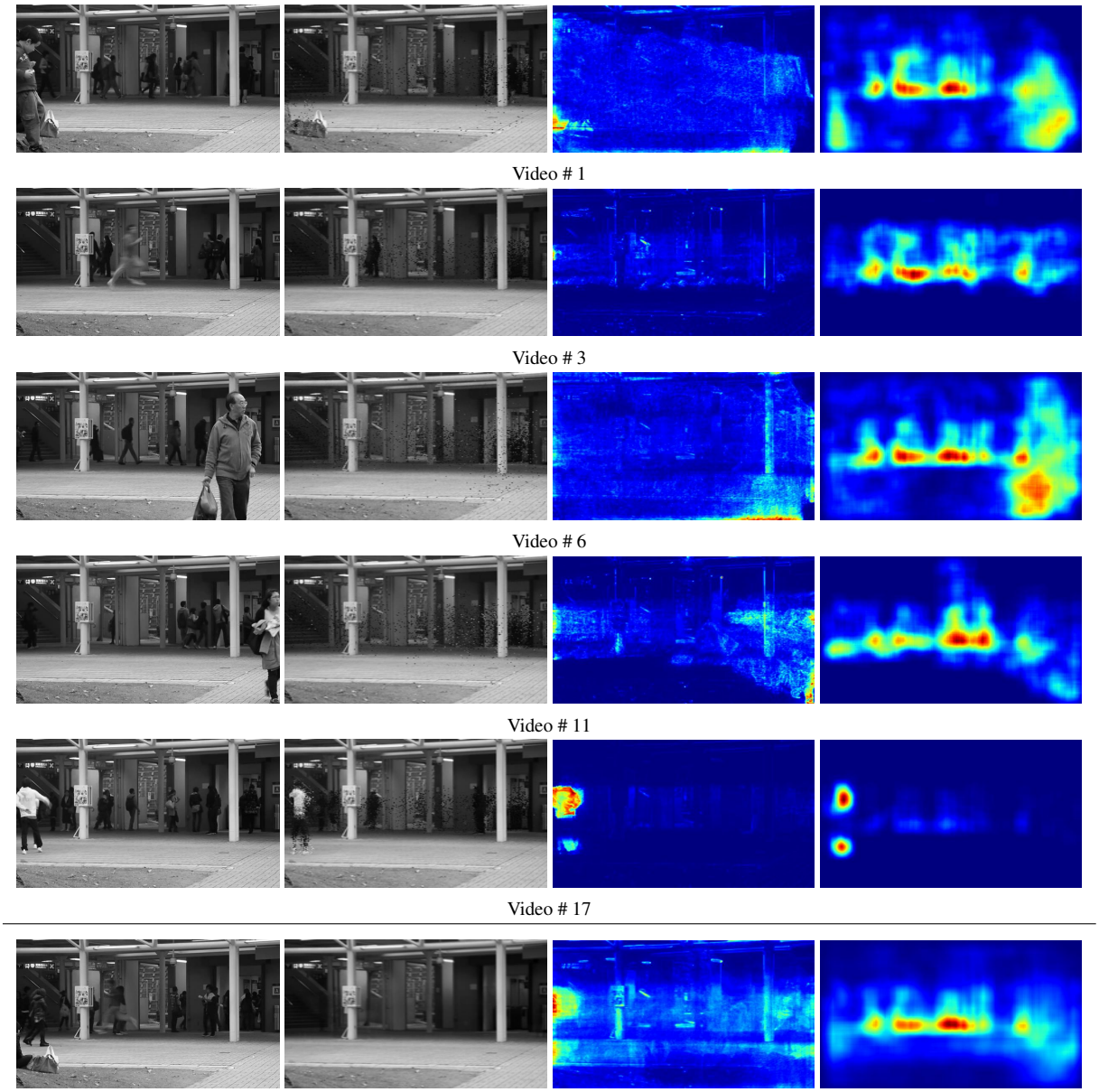


Figure 1. (Left) A sample irregular frame. (Second) A synthesized regular frame obtained by the pixel value of lowest reconstruction score across all frames of a video. (Third) Accumulated regularity score obtained by convolutional-autoencoder. (Fourth) Accumulated regularity score obtained by IT-autoencoder.

Go to Table of Contents

2.2. UCSD Ped1

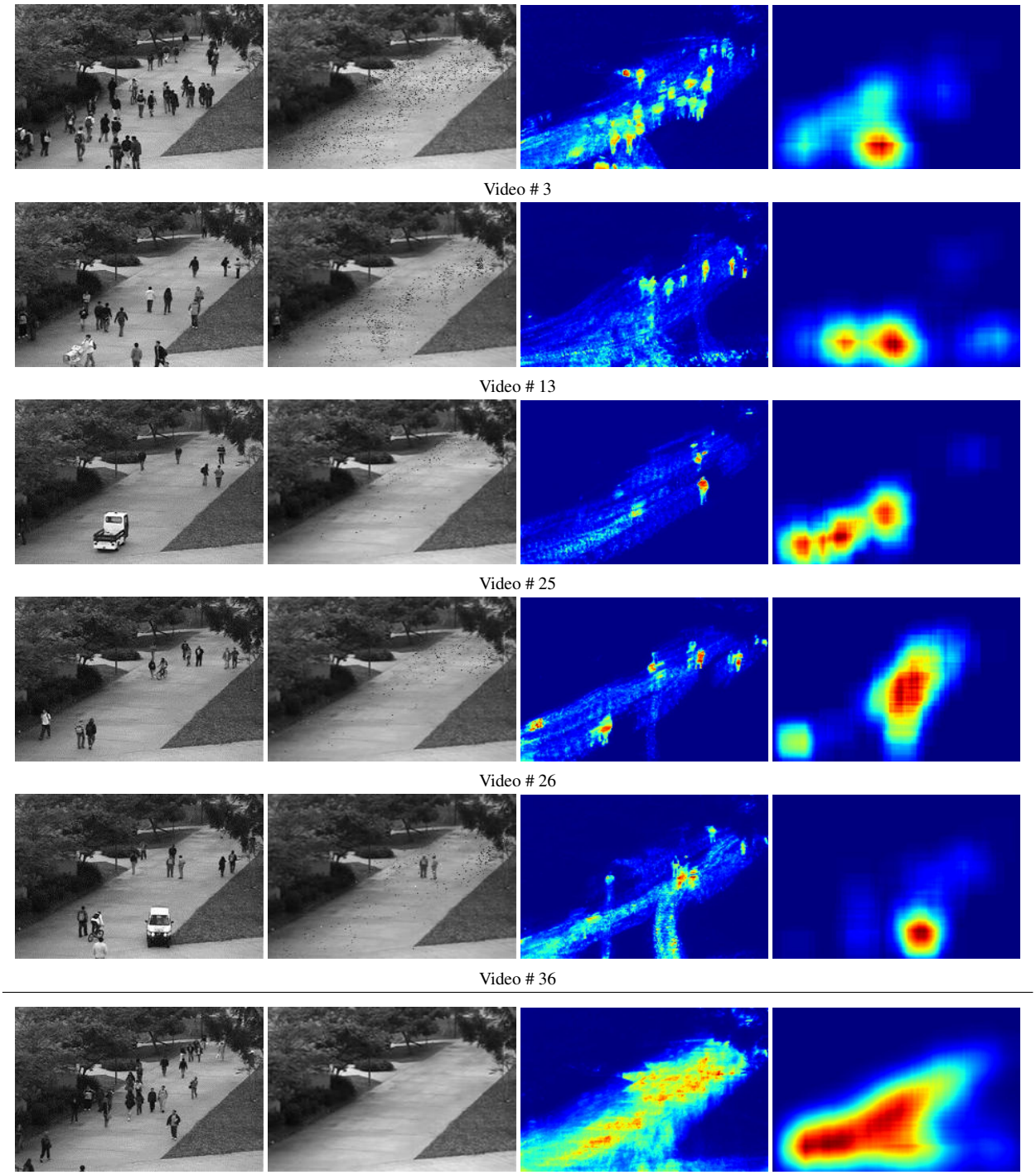


Figure 2. Same layout in all figures in Section 2.1. Especially, in video 36, we can observe the trajectory of a SUV in the heatmap of accumulated regular score (third).

Go to Table of Contents

2.3. UCSD Ped2

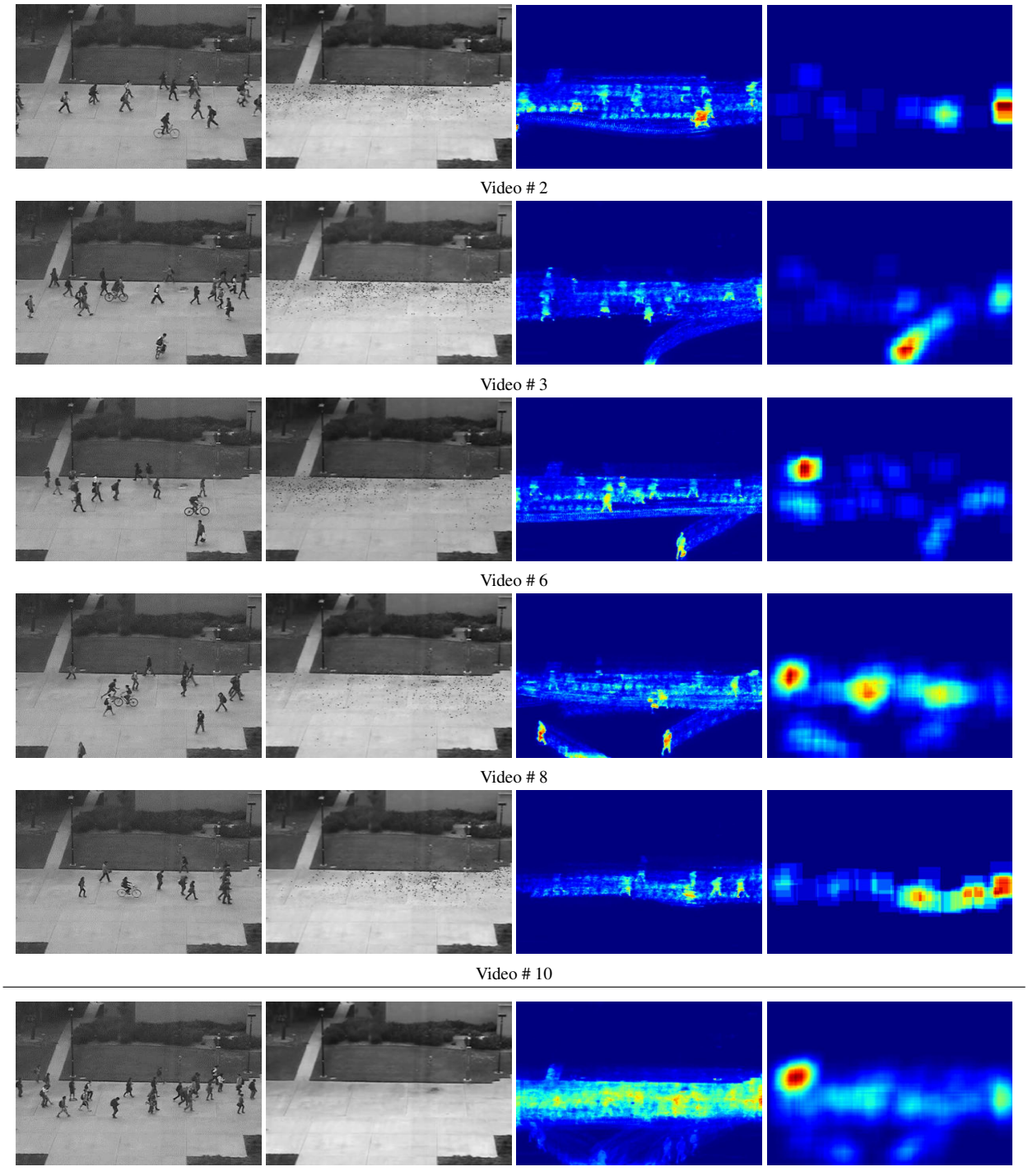


Figure 3. Same layout in all figures in Section 2.1. Especially, in video 8, we can clearly observe a trajectory of two people in the heatmap of accumulated regular score (third column). In the synthesized regular frame by conv-autoencoder (second), there are dots. Those dots are outliers in regularity score due to lack of data as there is no dots in the regular frame by all videos thanks to statistically significant amount of data.

Go to Table of Contents

2.4. Subway Enter

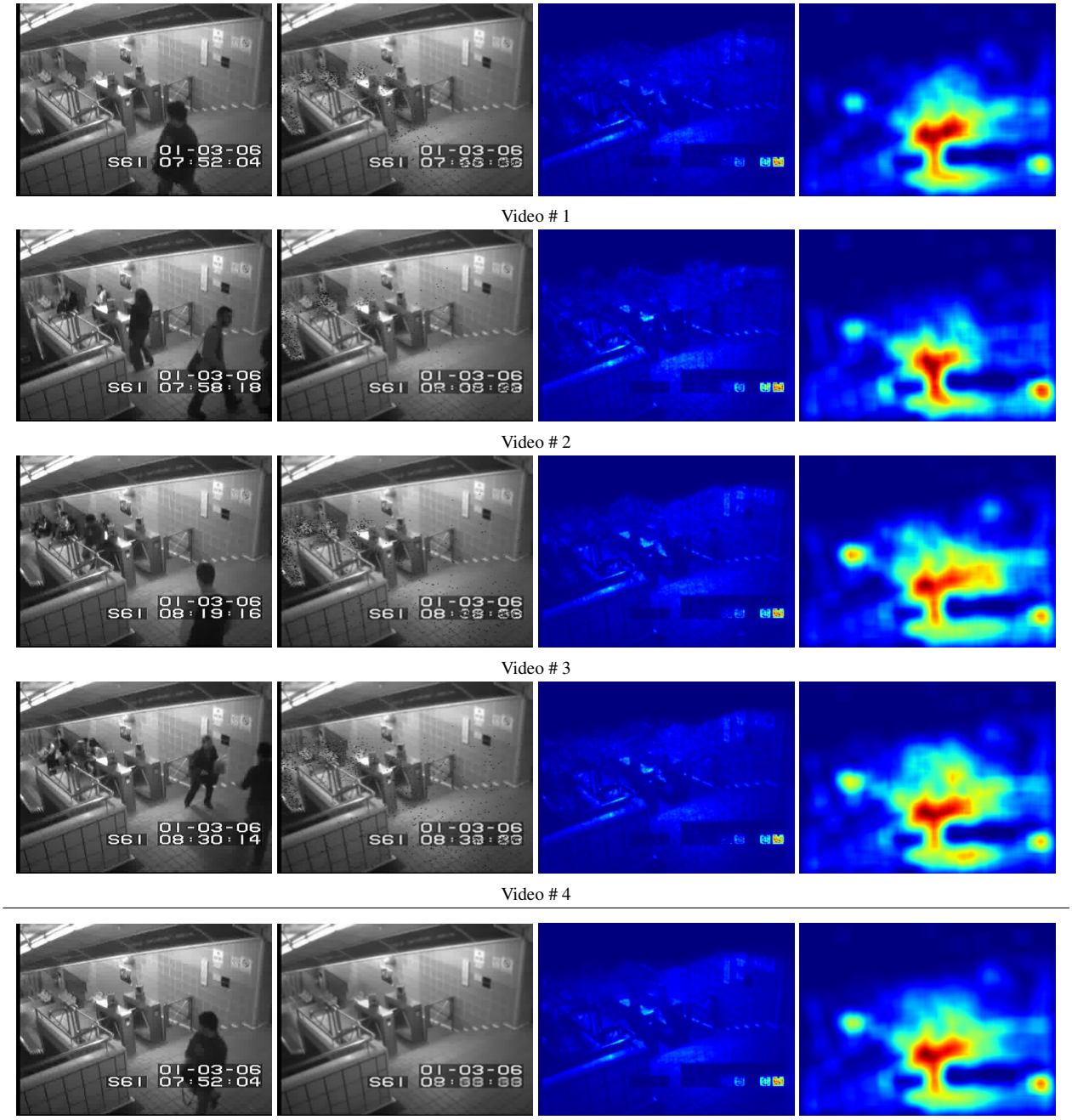


Figure 4. Same layout in all figures in Section 2.1. Compared to other datasets, we have relatively high regularity score (more blue). It is because length of the videos is long so that the irregular motion is averaged out in long minutes. Obviously, the clock ticking is not part of regular motions.

Go to Table of Contents

2.5. Subway Exit

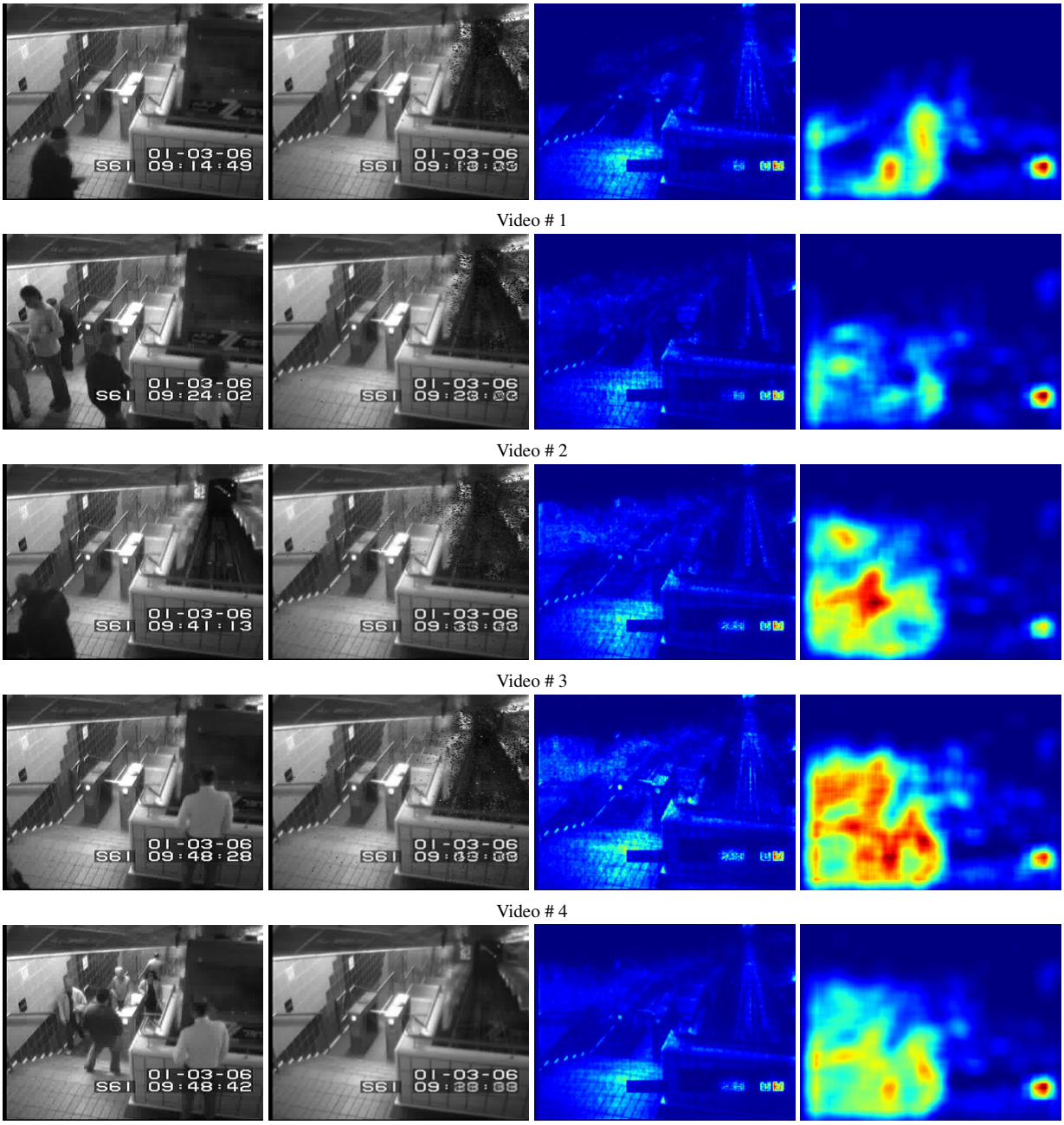


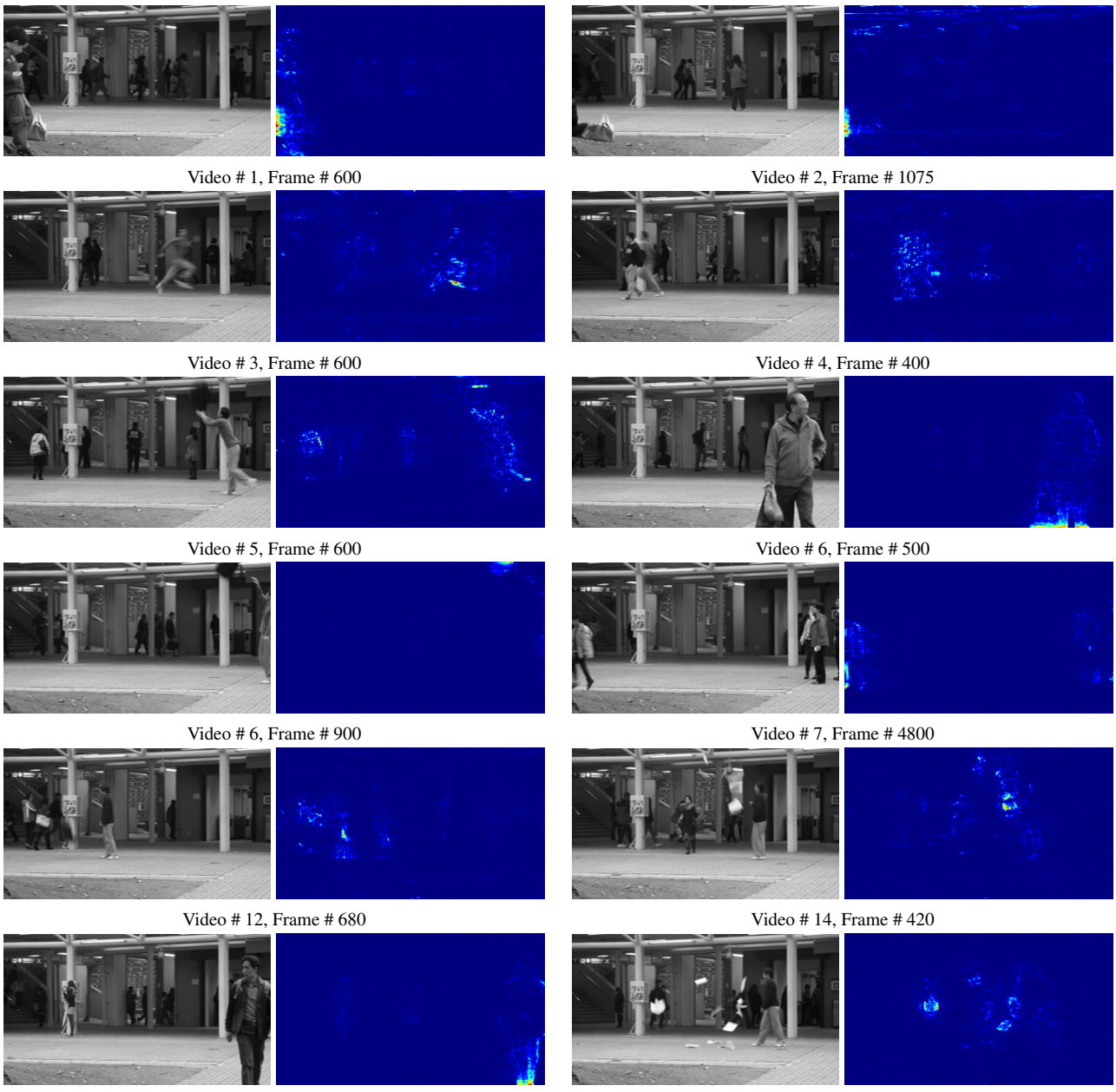
Figure 5. Same layout in all figures in Section 2.1. Similar to Subway Enter. But interestingly, IT-autoencoder has a very high accumulated irregular score in the stair regions.

Go to Table of Contents

3. Object Detection in Irregular Motion

Using the regularity score, we can obtain locations of objects involved in irregular motion in each frame, which is a usually the objects of interest. We present several frames with irregular motions for each dataset and its corresponding objects location in the frame. Note that we have high irregularity response at the edge of the objects where the motion changes most significantly. It is better presented in video: `reg_score_video.avi`

3.1. CUHK Avenue Dataset



Video # 1, Frame # 600 | Video # 2, Frame # 1075

Video # 3, Frame # 600 | Video # 4, Frame # 400

Video # 5, Frame # 600 | Video # 6, Frame # 500

Video # 6, Frame # 900 | Video # 7, Frame # 4800

Video # 12, Frame # 680 | Video # 14, Frame # 420

Video # 15, Frame # 550 | Video # 20, Frame # 100

Figure 6. (Left) a frame (Right) Object regularity score. In video 6 (frame # 500), the bottom part of legs, which are the most prominent object involved in a irregular motion, exhibits very high scores to other regions. In video 14 and 20, the flying papers are well captured.

Go to Table of Contents

3.2. UCSD Ped1

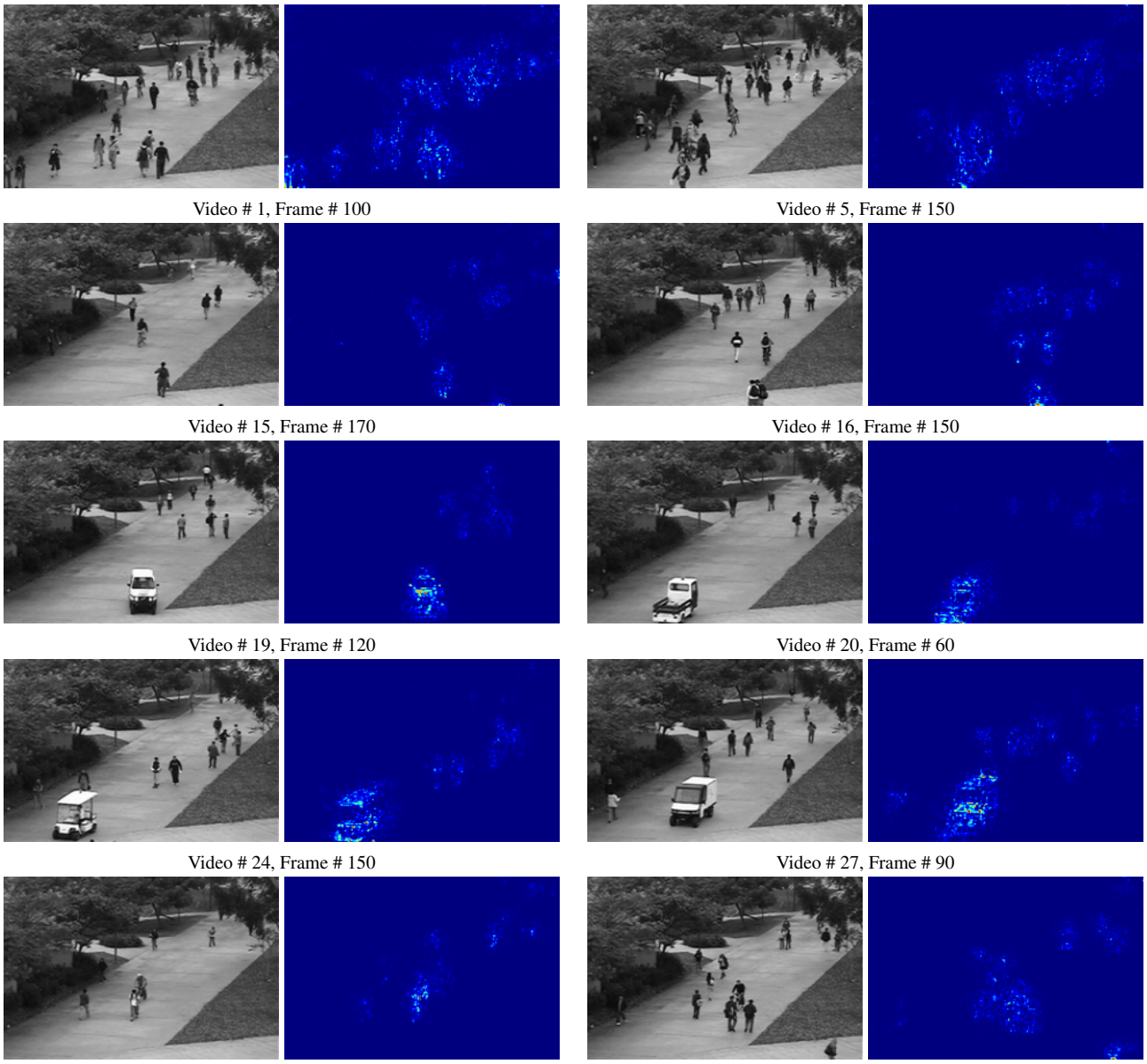
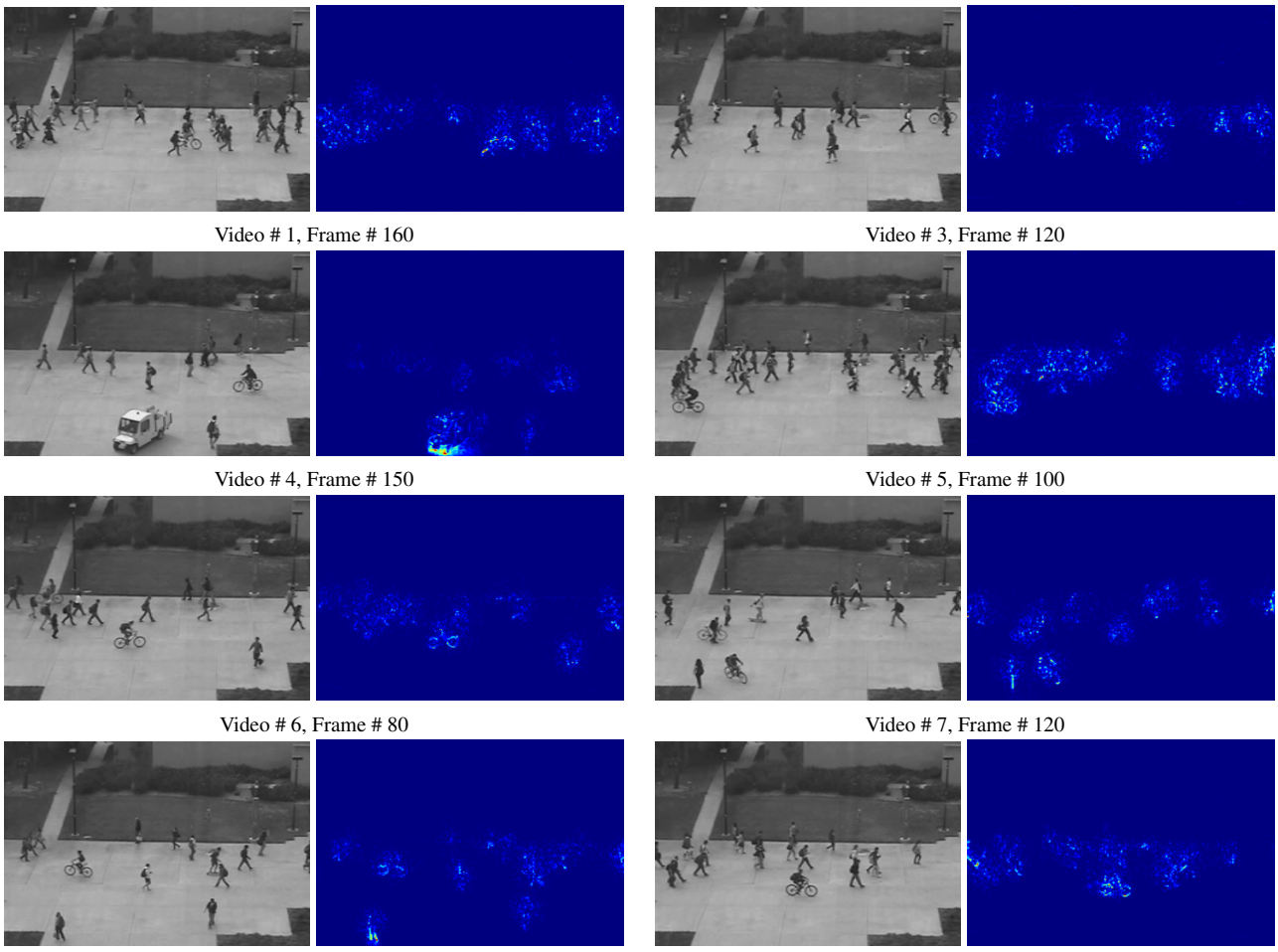


Figure 7. Same layout in all figures as in Section 3.1. The moving cars are easily identified in video #19, #20, #24, and #27 and fast moving persons in video #29.

Go to Table of Contents

3.3. UCSD Ped2



Video # 1, Frame # 160 | Video # 3, Frame # 120

Video # 4, Frame # 150 | Video # 5, Frame # 100

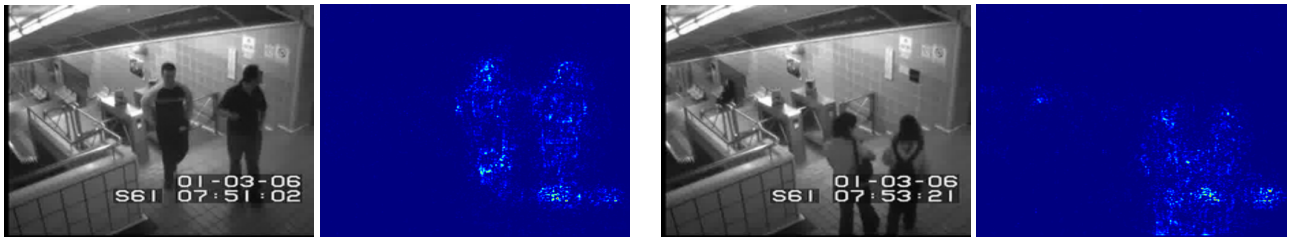
Video # 6, Frame # 80 | Video # 7, Frame # 120

Video # 8, Frame # 130 | Video # 9, Frame # 100

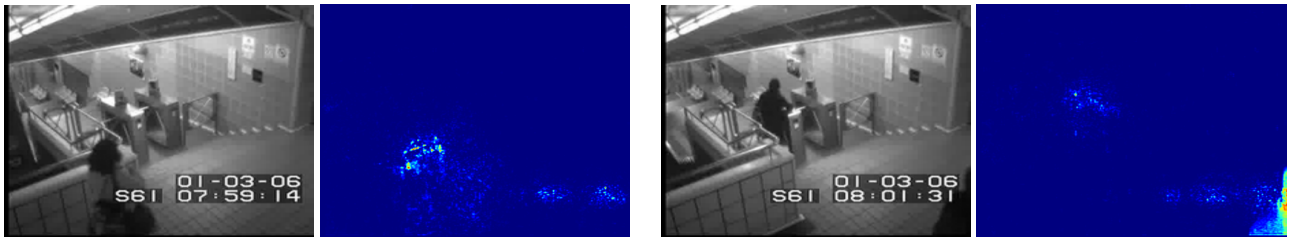
Figure 8. Same layout in all figures as in Section 3.1. The moving cars (frames in video 6) and fast moving persons are easily localized.

Go to Table of Contents

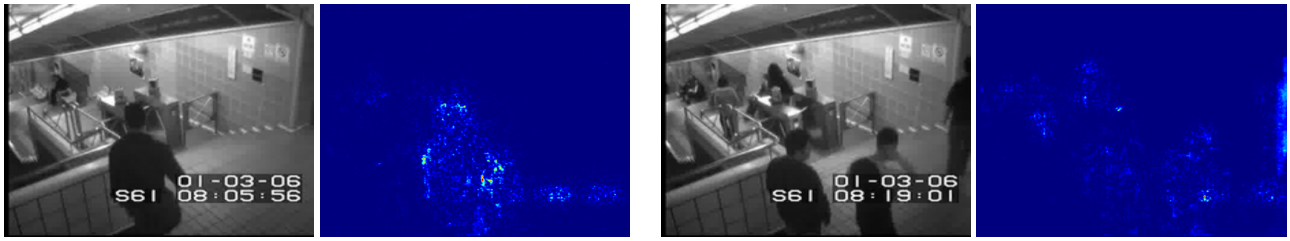
3.4. Subway Enter



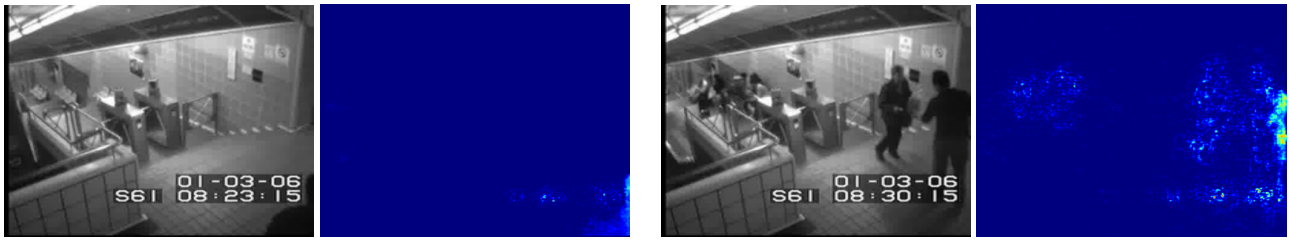
Video # 1, Frame # 9830 Video # 1, Frame # 13310



Video # 2, Frame # 2130 Video # 2, Frame # 5540



Video # 2, Frame # 12170 Video # 3, Frame # 11800

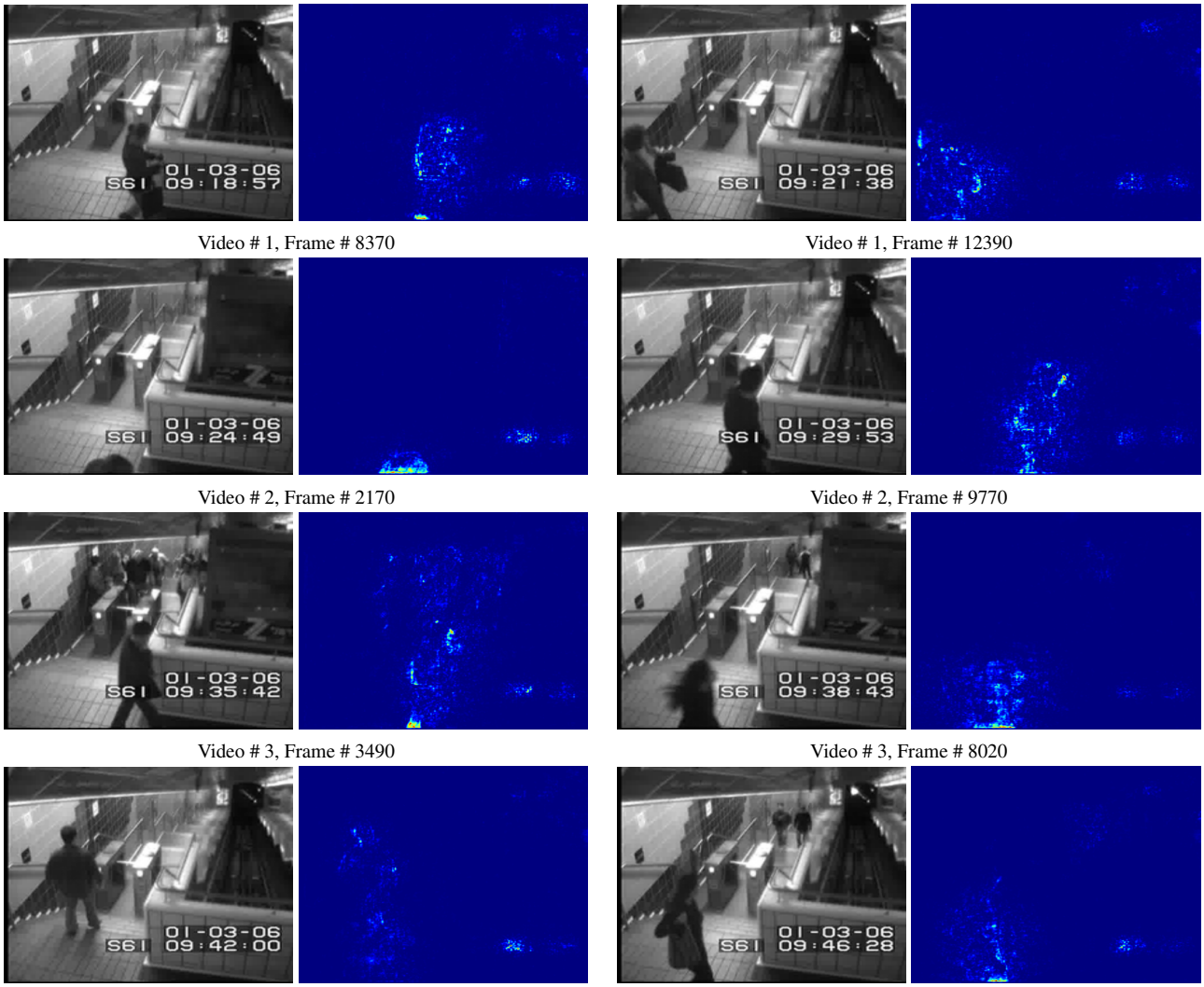


Video # 3, Frame # 18150 Video # 4, Frame # 8640

Figure 9. Same layout in all figures as in Section 3.1. Moving persons are easily localized.

Go to Table of Contents

3.5. Subway Exit



Video # 1, Frame # 8370 Video # 1, Frame # 12390

Video # 2, Frame # 2170 Video # 2, Frame # 9770

Video # 3, Frame # 3490 Video # 3, Frame # 8020

Video # 3, Frame # 12940 Video # 4, Frame # 4640

Figure 10. Same layout in all figures as in Section 3.1. Similar to Subway Enter dataset; moving persons are easily revealed.

Go to Table of Contents

4. Predicting Past and Future Regular Frames

As in Section 4.4 in the main paper, we present a predicted regular frames of the past and the future of a given single image. The left most column in each figure in this section shows the given single image from which we predict the past and the future regular frames. Second column presents the images of 0.1 second before the moment of the given image. Third column presents the reconstructed 'regular' frame of the moment of the given image. Fourth column presents the images of 0.1 second after the moment of the given image. Note that the objects involved in the irregular motions are gradually appearing from the past and gradually disappearing in the future.

4.1. CUHK Avenue Dataset

It is best viewed in a video form: `frame_pred_avenue.mp4`

In the video, we put the all twelve videos into one file for the ease of playing. We first show the single seed frame for a second and show the predicted video followed by a blank frames.

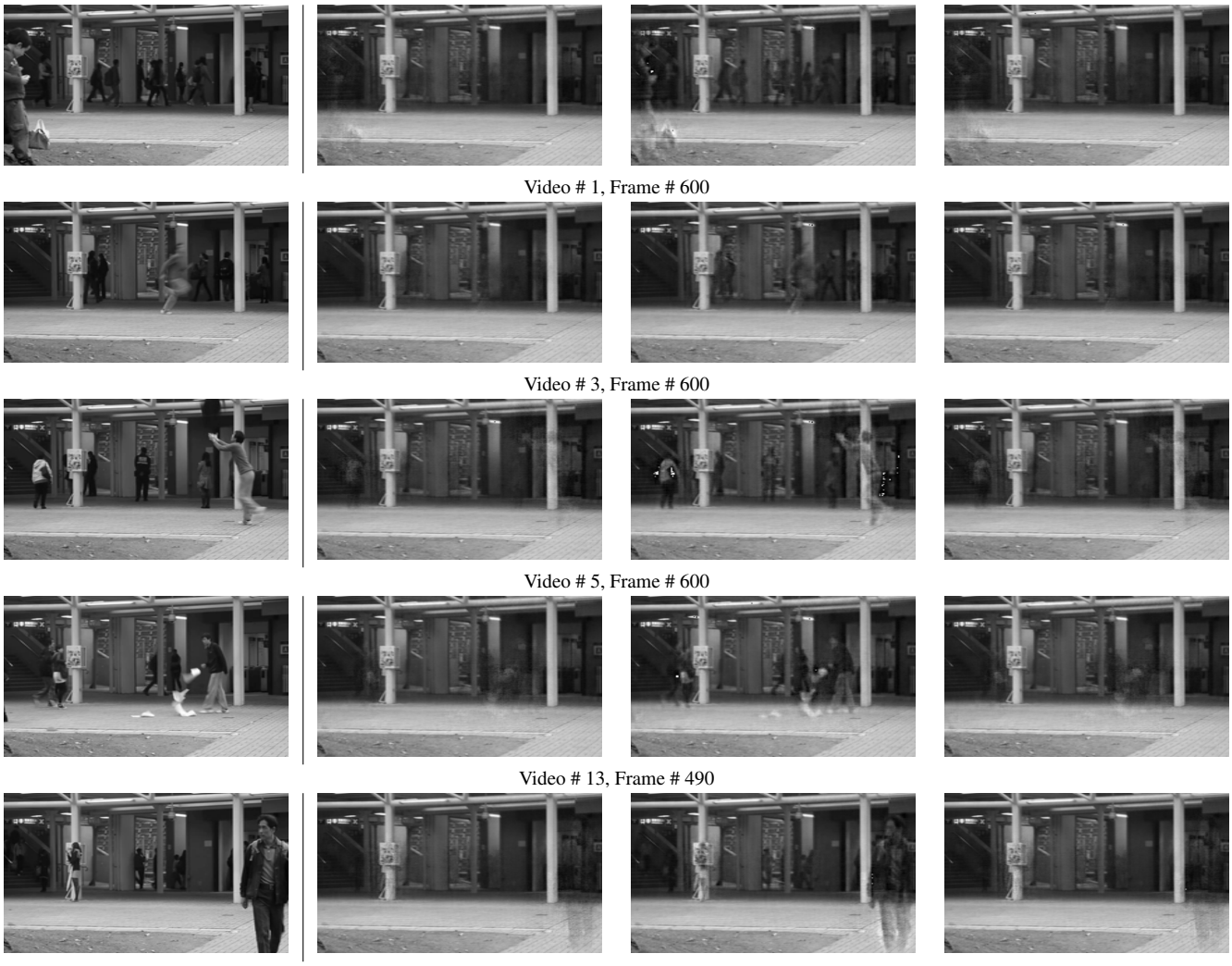
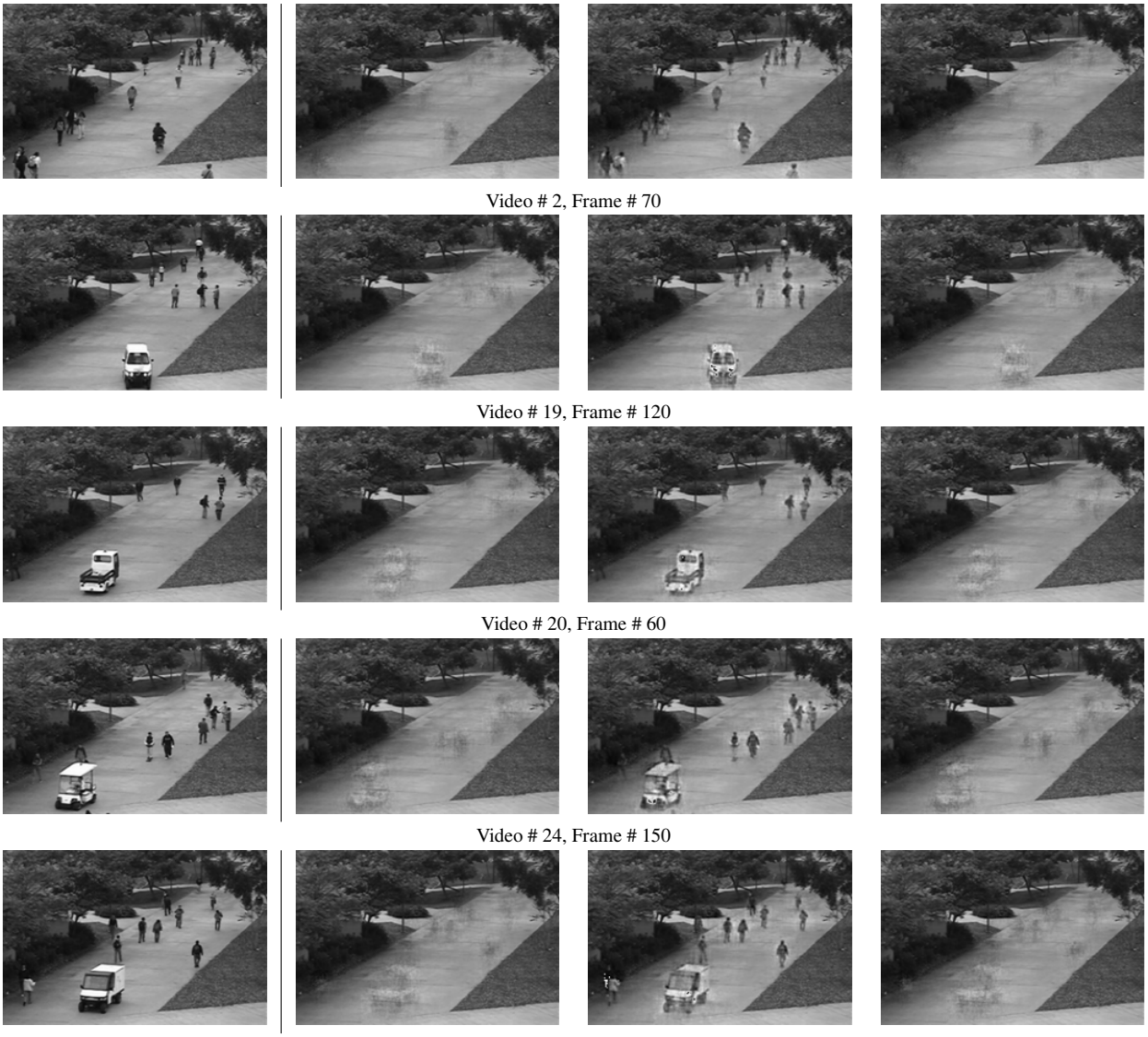


Figure 11. Same layout as discussed in Section 4. The regularity enforces that the objects involved in irregular motion gradually appearing and disappearing. In video 1 (first row), the crowd and the person in front are gradually appearing and disappearing. In video 13, in the future frame, the paper is closer to the ground compared to the past.

Go to Table of Contents

4.2. UCSD Ped1

We do not provide a video for this dataset but figure will explain it.



Video # 2, Frame # 70

Video # 19, Frame # 120

Video # 20, Frame # 60

Video # 24, Frame # 150

Video # 27, Frame # 90

Figure 12. Same layout as discussed in Section 4. In video 20, the car moves a little bit upwards in the future frame.

Go to Table of Contents

4.3. UCSD Ped2

We do not provide a video for this dataset but figure will explain it.

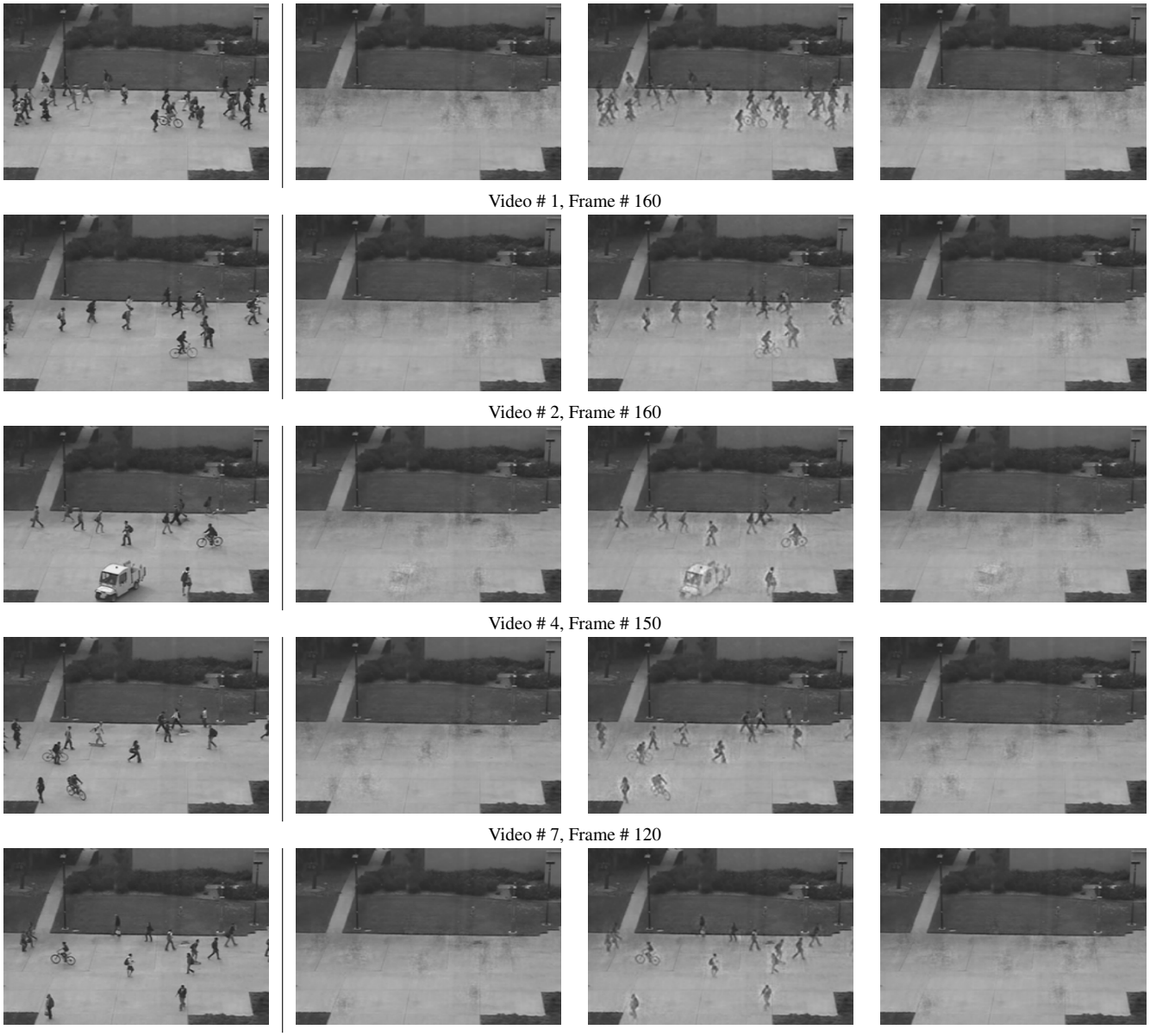


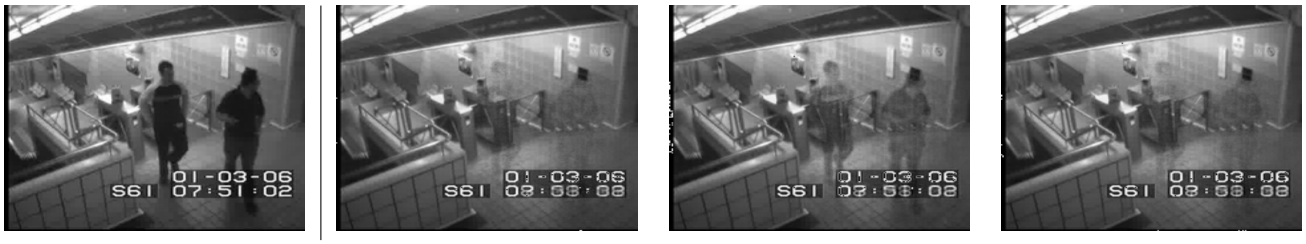
Figure 13. Same layout as discussed in Section 4. In video 4, the car appears clearer than the past frame predicted (second) and moves a bit more south than the past frames.

Go to Table of Contents

4.4. Subway Enter

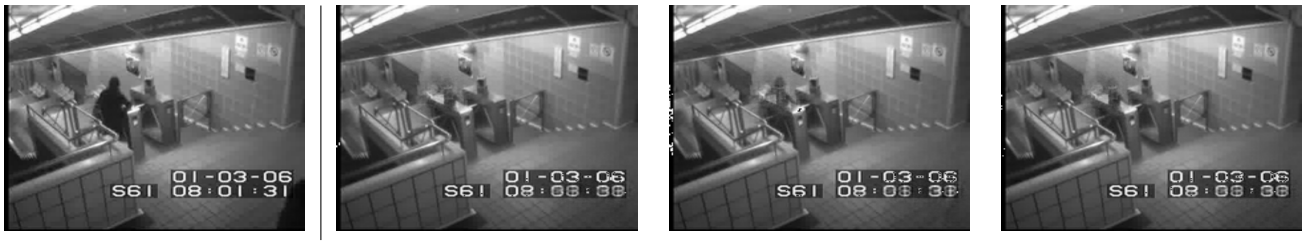
We do not provide a video for this dataset but figure will explain it.

Video # 1, Frame # 180



Video # 1, Frame # 9830

Video # 1, Frame # 13310



Video # 2, Frame # 5540

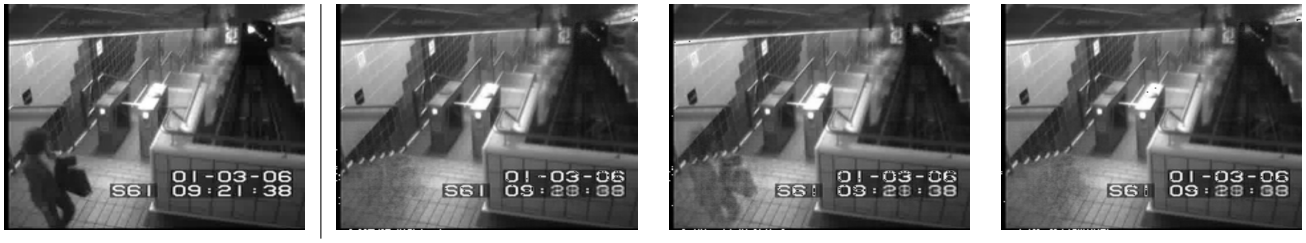
Video # 2, Frame # 12170

Figure 14. Same layout as discussed in Section 4.

Go to Table of Contents

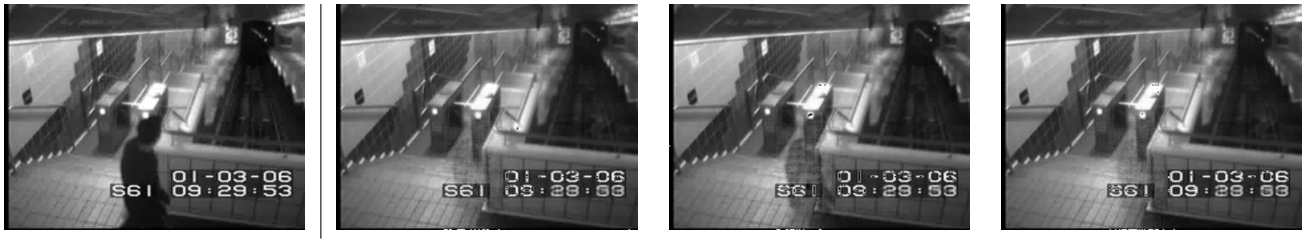
4.5. Subway Exit

We do not provide a video for this dataset but figure will explain it.

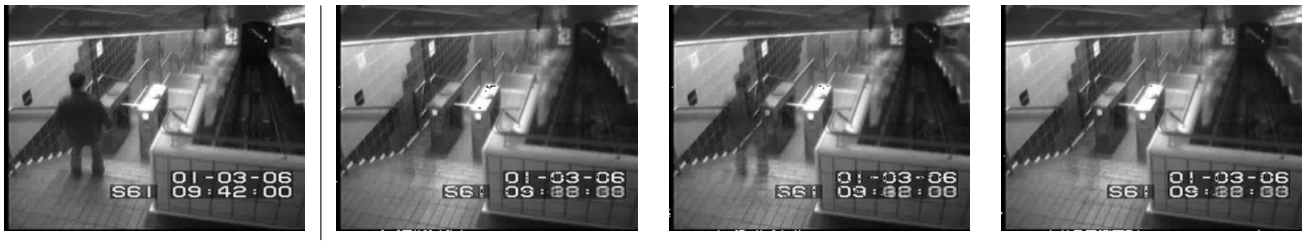


Video # 1, Frame # 12390

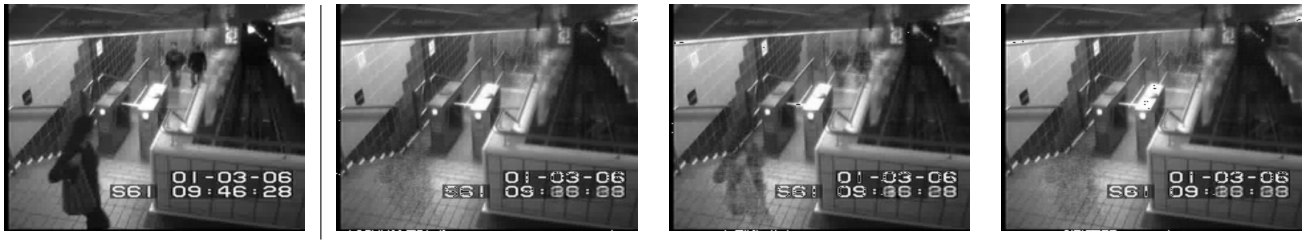
Video # 1, Frame # 1010



Video # 2, Frame # 9770



Video # 3, Frame # 12940



Video # 4, Frame # 4640

Figure 15. Same layout as discussed in Section 4.

Go to Table of Contents

5. Anomalous Event Detection and Generalization Analysis on Multiple Datasets

We visualize the regularity score (defined in Eq. (3) in the main paper) to detect anomalous events in video. When the regularity score is low in a local temporal window, the video segment is determined containing anomalous events. We additionally compare with the generalizability of the trained model using various training sets. Blue (conventional) represents the score obtained by a model trained on the *specific target* dataset. Red (generalized) represents the score obtained by a model trained on *all* datasets. (This is the model we use for all other experiments.) Yellow (transfer) represents the score obtained by a model trained on *all datasets except that specific target* datasets.

5.1. CUHK Avenue Dataset

The target dataset is CUHK Avenue. Thus, the 'conventional' represents the score obtained by a model trained only on the Avenue dataset. The 'generalized' represents the score obtained by a model trained on all datasets we used. The 'transfer' represents the score obtained by a model trained on all datasets except the Avenue dataset. Surprisingly, the generalized model performs very well same as the target model (conventional). And the transfer model also performs decently.

By comparing 'conventional' and 'generalized', we observe that the model is powerful enough not being harmed by other datasets. At the same time, by comparing 'transfer' and either 'generalized' or 'conventional', we observe that the model is not too much overfitting to the given dataset as it can generalized to *unseen* videos. Consequently, we believe that the proposed network structure is well balanced between overfitting and underfitting. Red shaded region represents the ground truth anomalous temporal segments defined by each data curators.

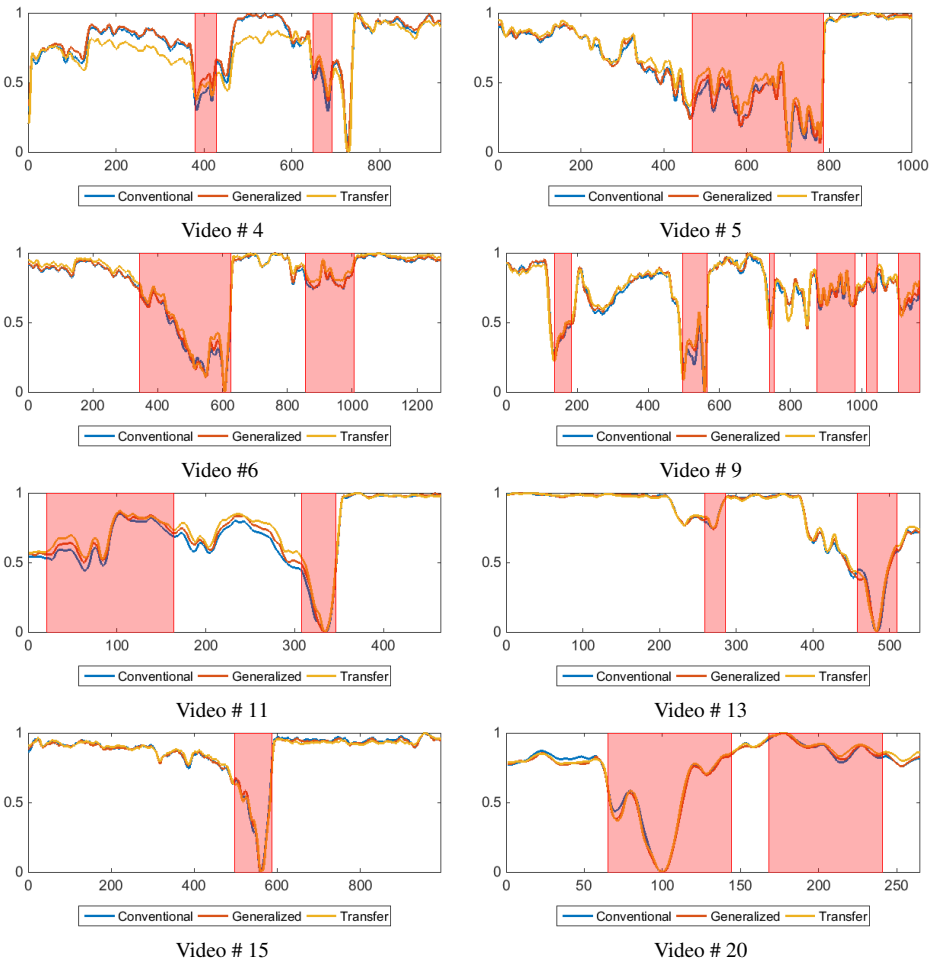


Figure 16. Our model captures anomalous regions as a form of local minima. In some of the ground truth anomalous region, however, the regularity score is not as much low as other regions. This is mainly due to the the anomalous action is happening in a small region or is well blended with the appearances of regular activity.

Go to Table of Contents

5.2. UCSD Ped1

Similar to Avenue dataset, the generalized model performs very well same as the target model (conventional) and the transfer model also performs very decently.

Figure 17. In video #5, at the end of the first region we have high regularity score even though it is in anomalous regions. This is mainly because the definition of anomalous event is different from the definition of regularity; regularity means temporally ordinary motions whereas the anomalous event can be defined as necessary - thus regular motion can be defined as anomaly.

Go to Table of Contents

5.3. UCSD Ped2

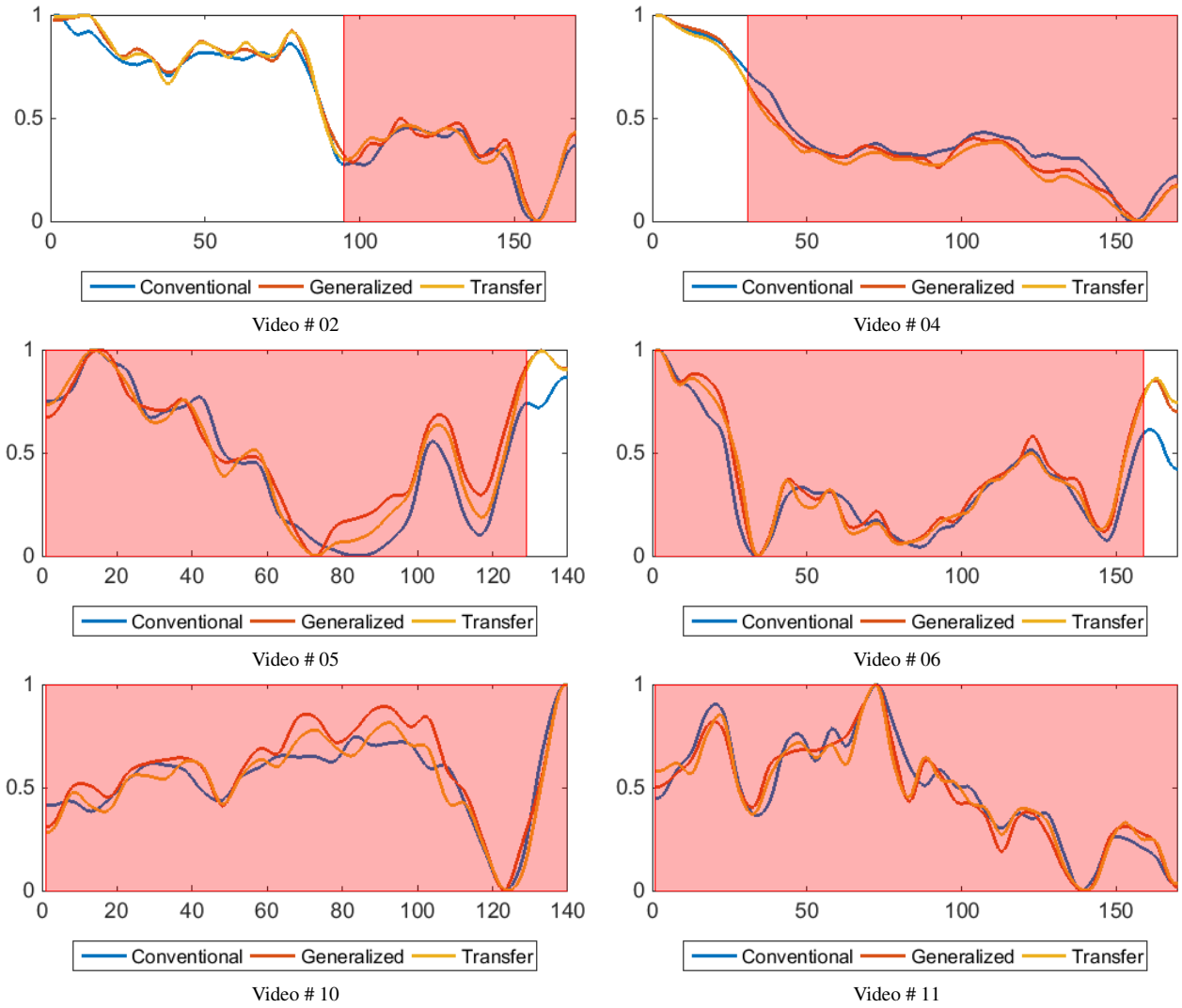
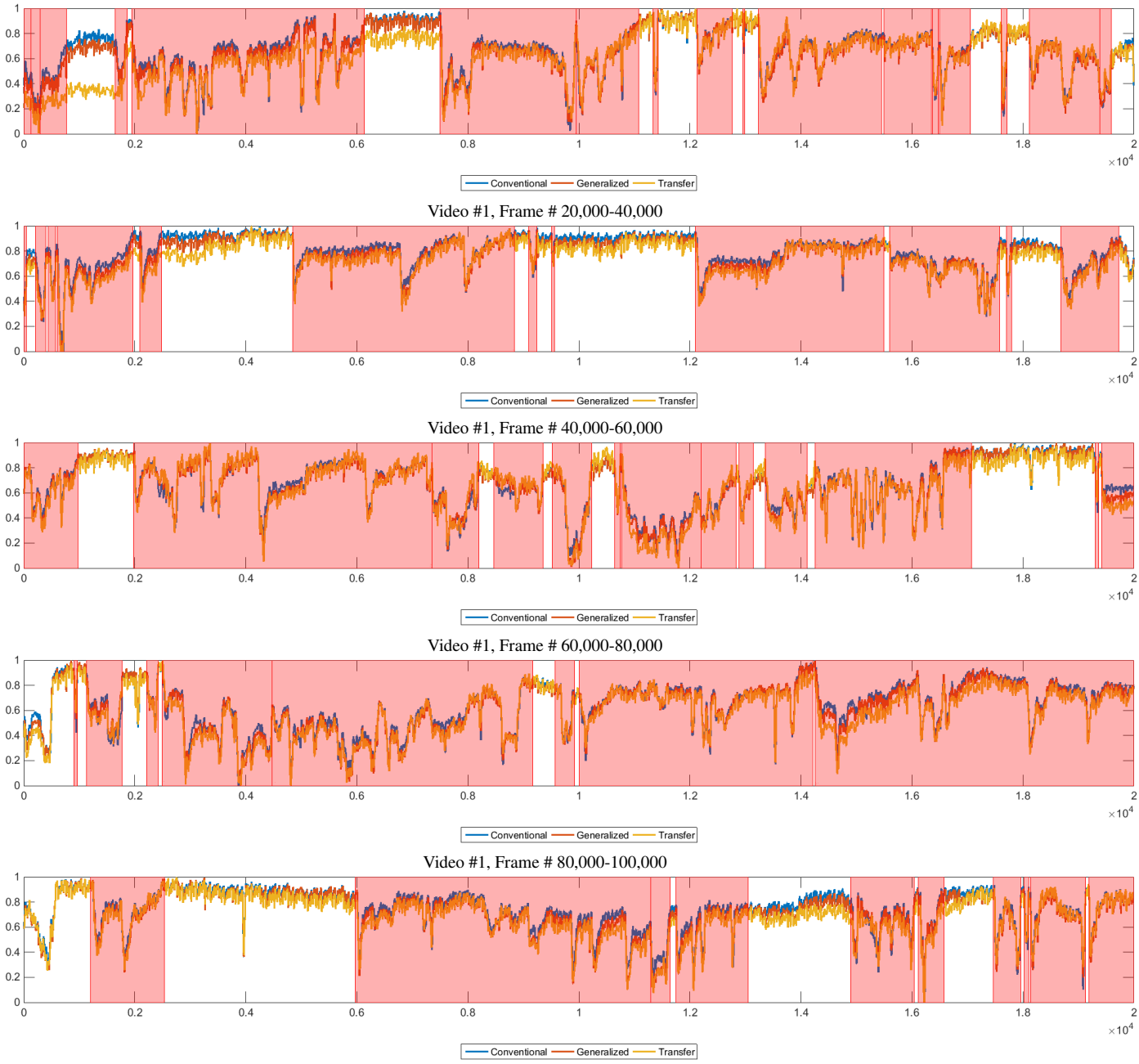


Figure 18. In video # 10 and 11, entire sequence is defined as an anomalous event. Some frames, however, shows regular motions as discussed in the previous section.

Go to Table of Contents

5.4. Subway Enter

The anomalous events are well captured by the regularity score as the definition of anomalous events in this dataset is similar to our definition of regularity - no presence of any abruption motions.



Video #1, Frame # 20,000-40,000

Video #1, Frame # 40,000-60,000

Video #1, Frame # 60,000-80,000

Video #1, Frame # 80,000-100,000

Video #1, Frame # 100,000-120,000

Figure 19. Low score regions are well aligned with the temporal regions of anomalous events.

Go to Table of Contents

5.5. Subway Exit

Similar to Subway-Enter, the definition of anomalous events in this dataset is similar to our definition of regularity.

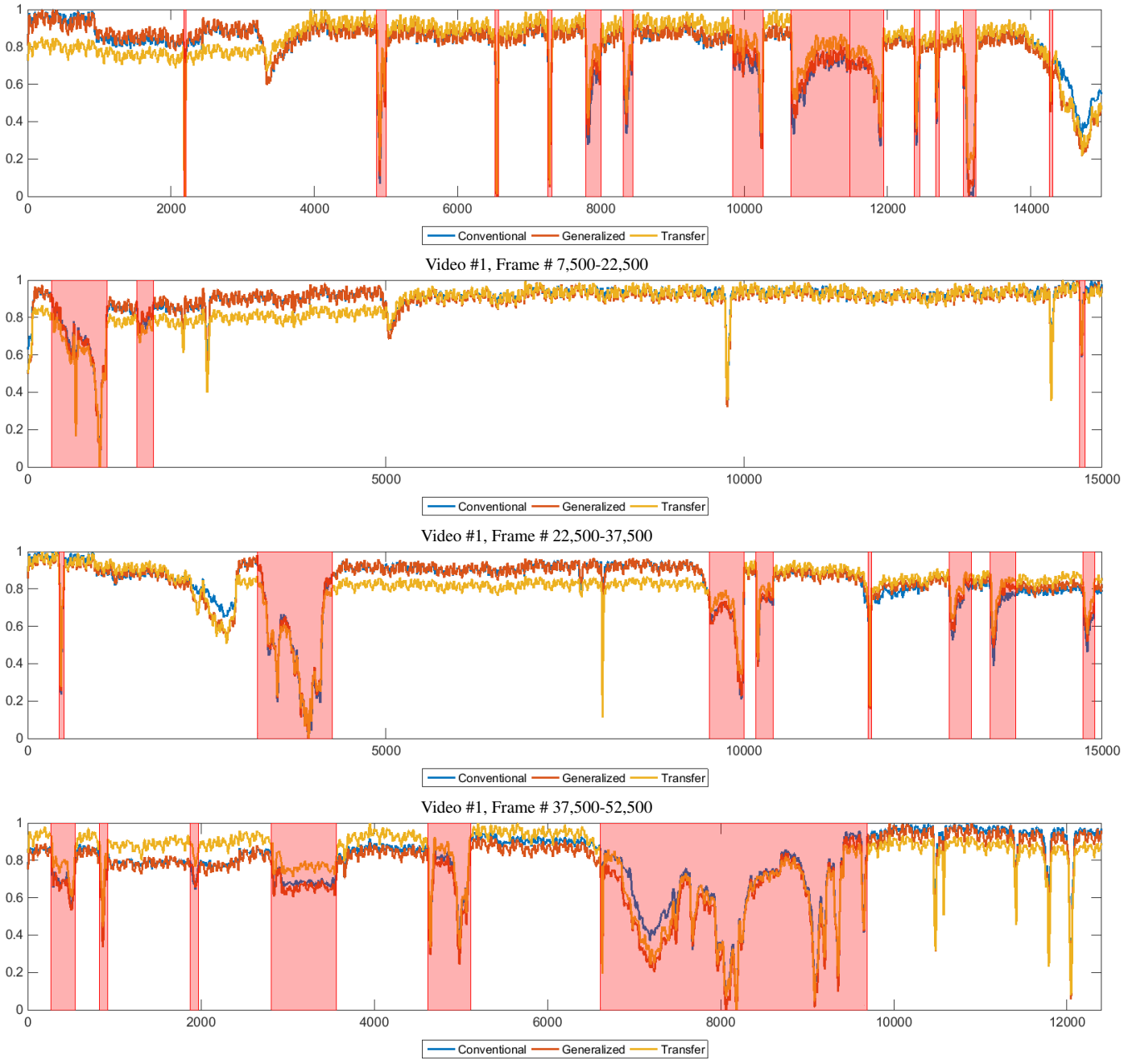


Figure 20. Low score regions are well aligned with the temporal regions of anomalous events.

Go to Table of Contents

6. Filter Response Visualization

We visualize the responses of learned convolutional filters in every layer. In the early convolutional layers, the filters capture various low level structural patches. Various learned filters capture complementary information as different filters shows very different responses on the same patch. As the layer goes deeper in convolution, the filters capture higher level structure in scale. The deconvolutional layers try to unpack the encoded (and noiseless) information in a hierarchical way in scale. Note that, for ease of visualization, we only show two frames of input and output.

6.1. CUHK Avenue Dataset

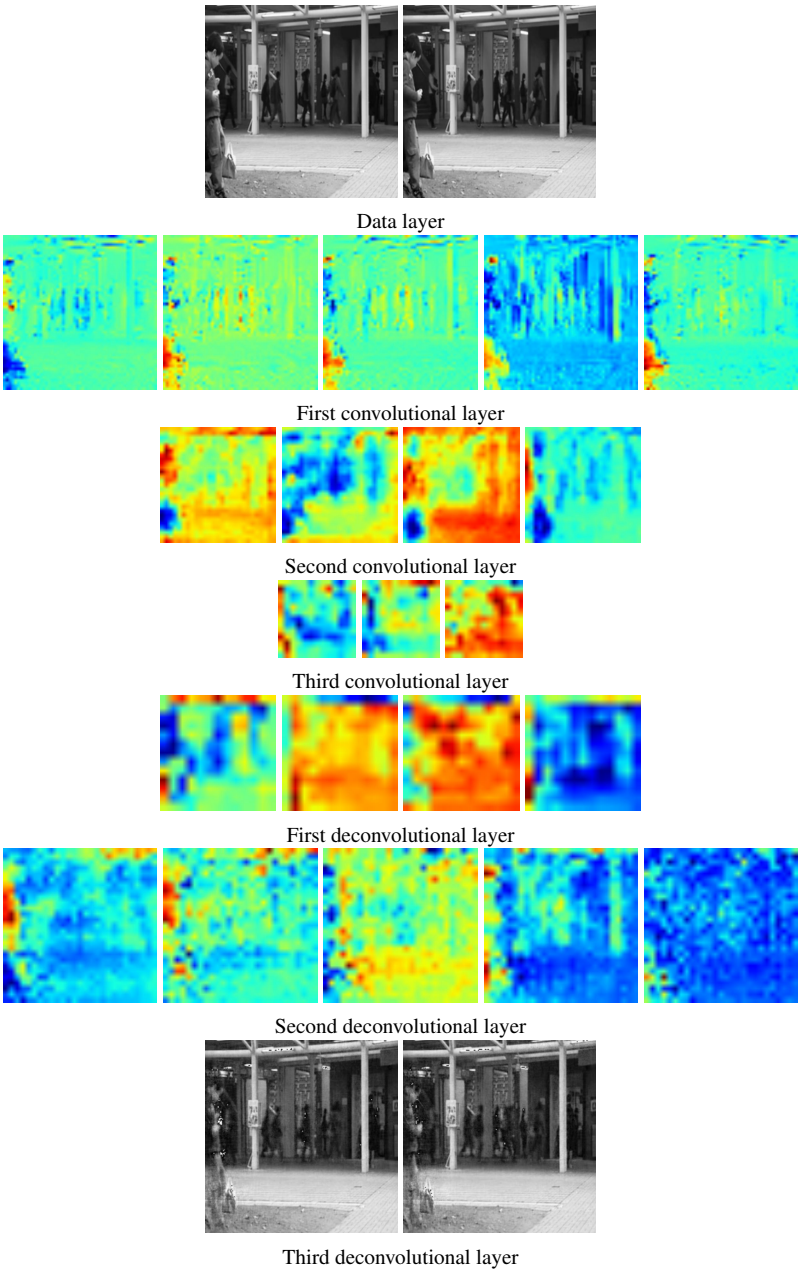


Figure 21. Responses of learned filters, evaluated on a video in CHUK Avenue dataset.

Go to Table of Contents

6.2. UCSD Ped1

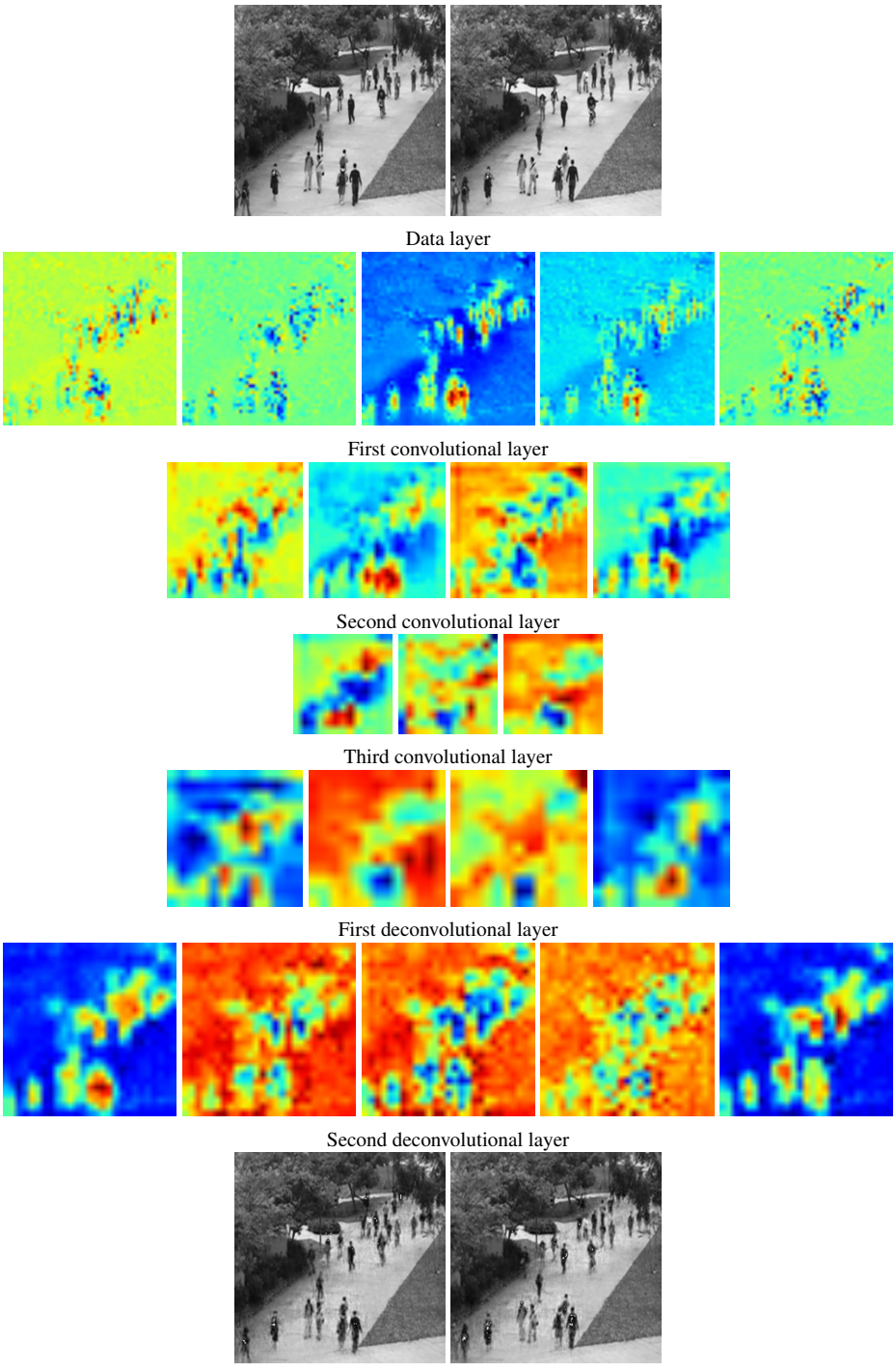


Figure 22. Responses of learned filters, evaluated on a video in UCSD-Ped1 dataset. Note that the filters captures various aspects of regularity as shown by various colored responses on the same region.

Go to Table of Contents

6.3. UCSD Ped2

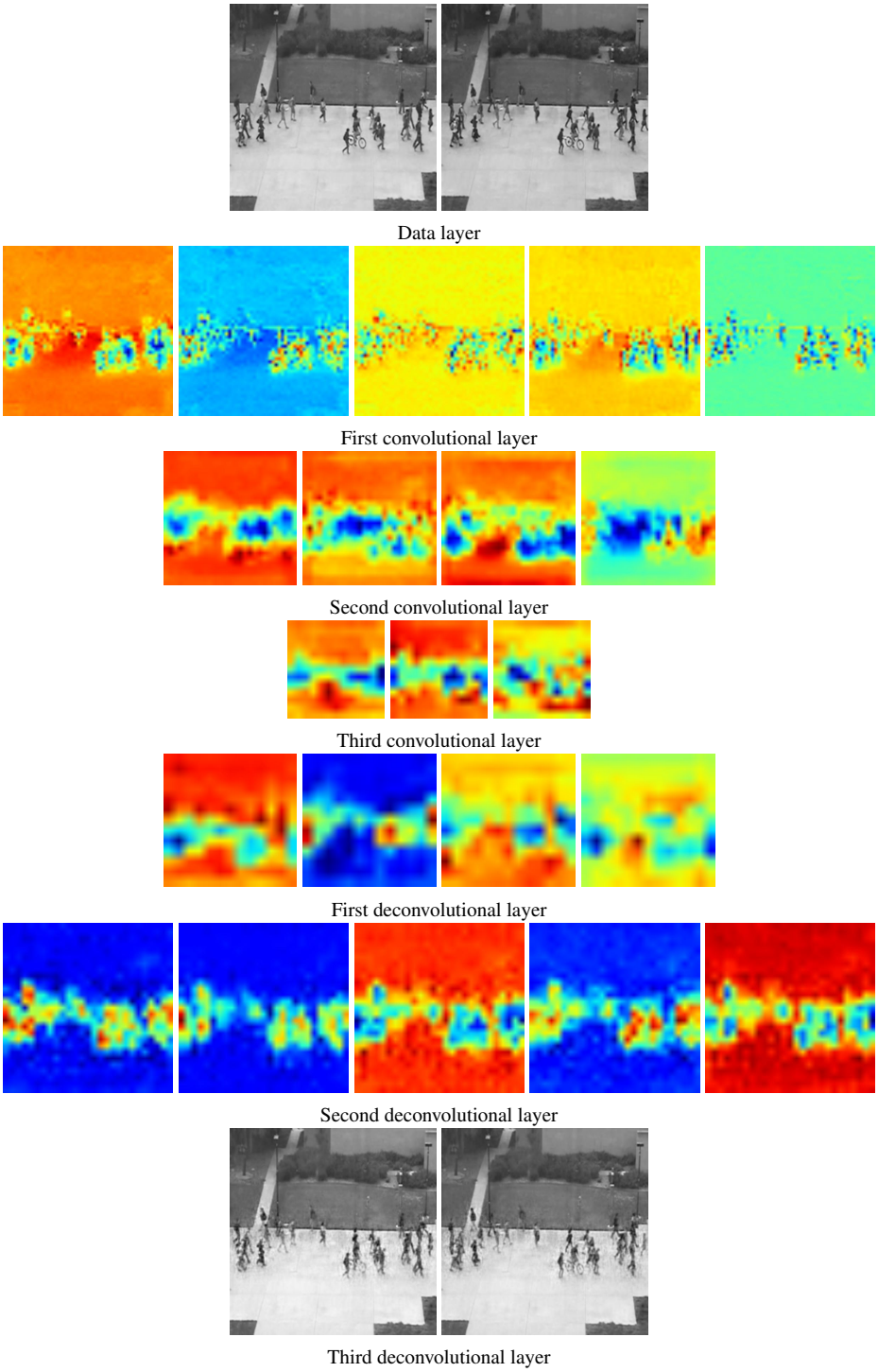


Figure 23. Responses of learned filters, evaluated on a video in UCSD-Ped2 dataset. Note that the filters captures various aspects of regularity as shown by various colored responses on the same region. Noticeably, the background of first convolutional layer outputs are in various colors.

Go to Table of Contents

6.4. Subway Enter

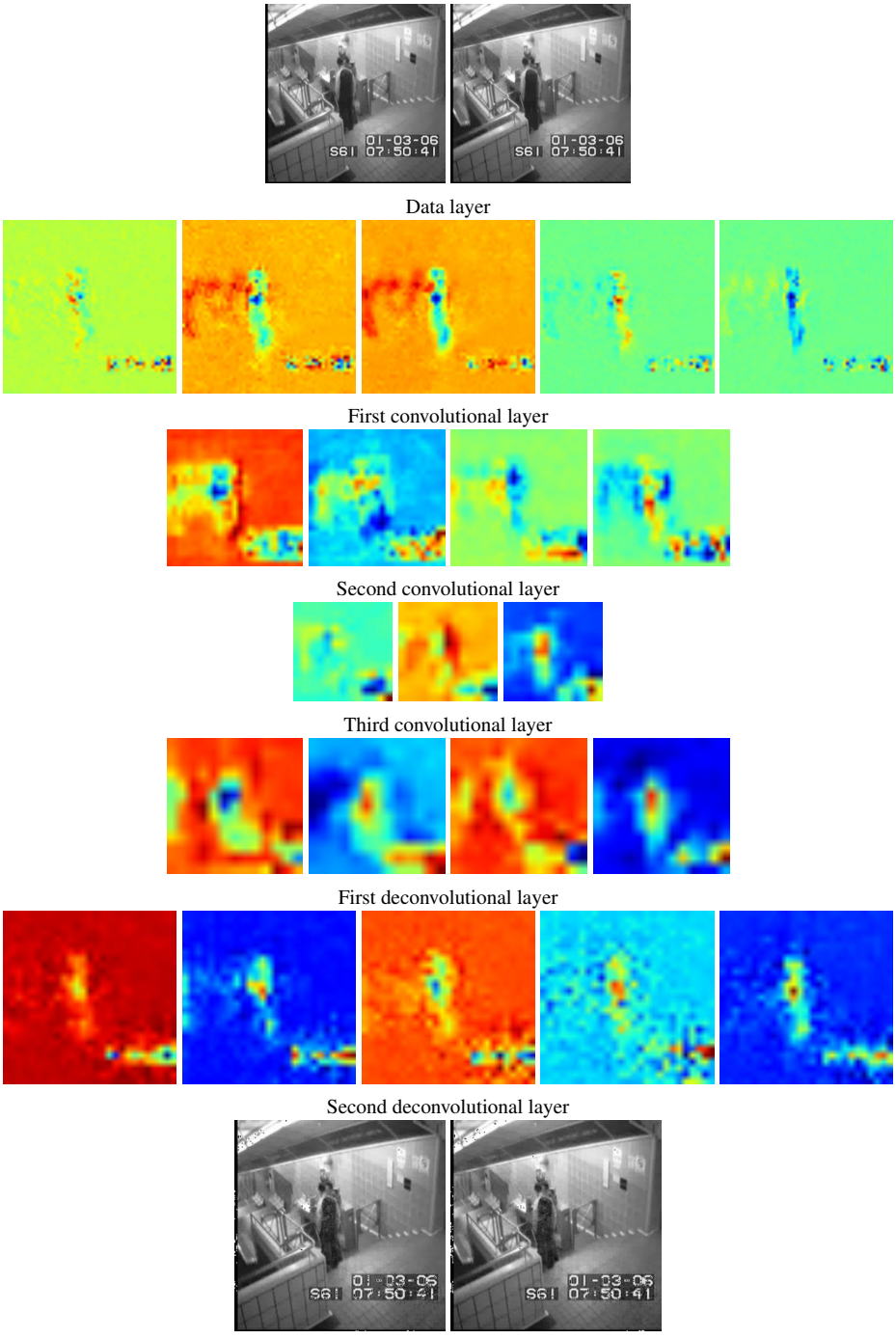


Figure 24. Responses of learned filters, evaluated on a video in Subway Enter dataset.

Go to Table of Contents

6.5. Subway Exit

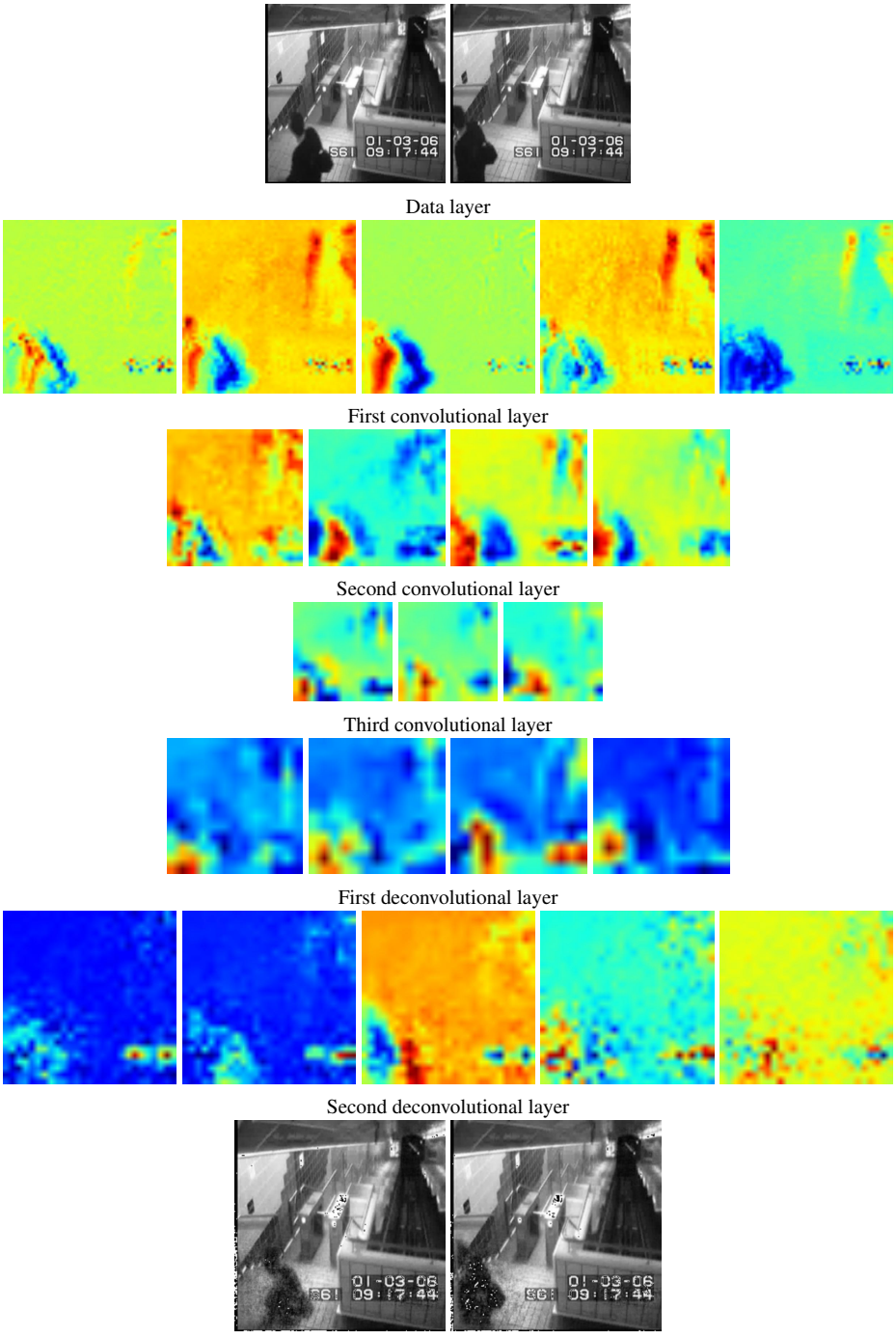
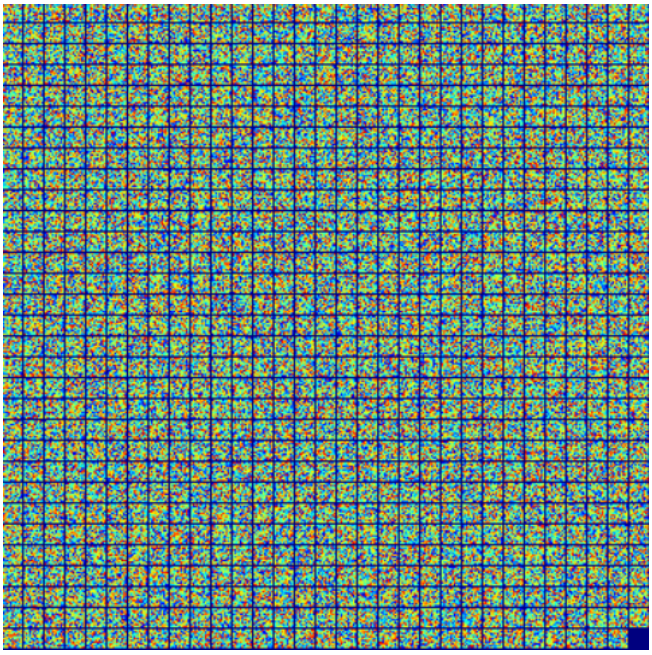


Figure 25. Responses of learned filters, evaluated on a video in Subway Exit dataset.

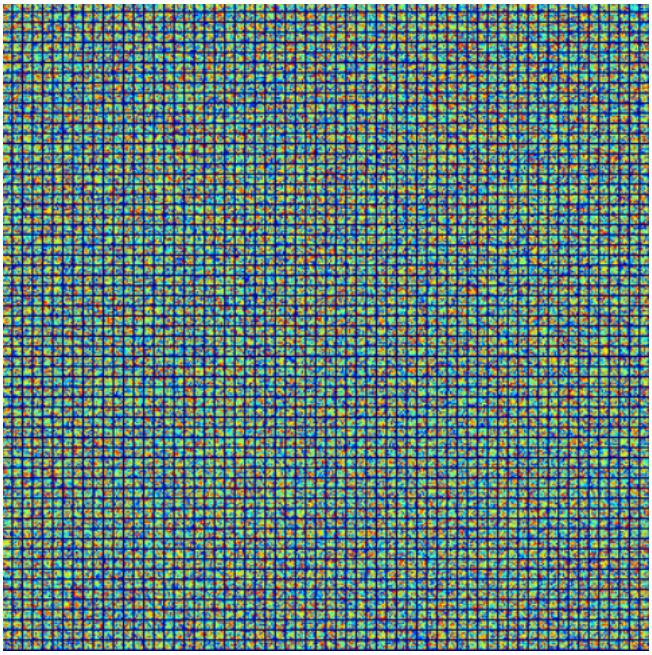
Go to Table of Contents

7. Filter Weights Visualization

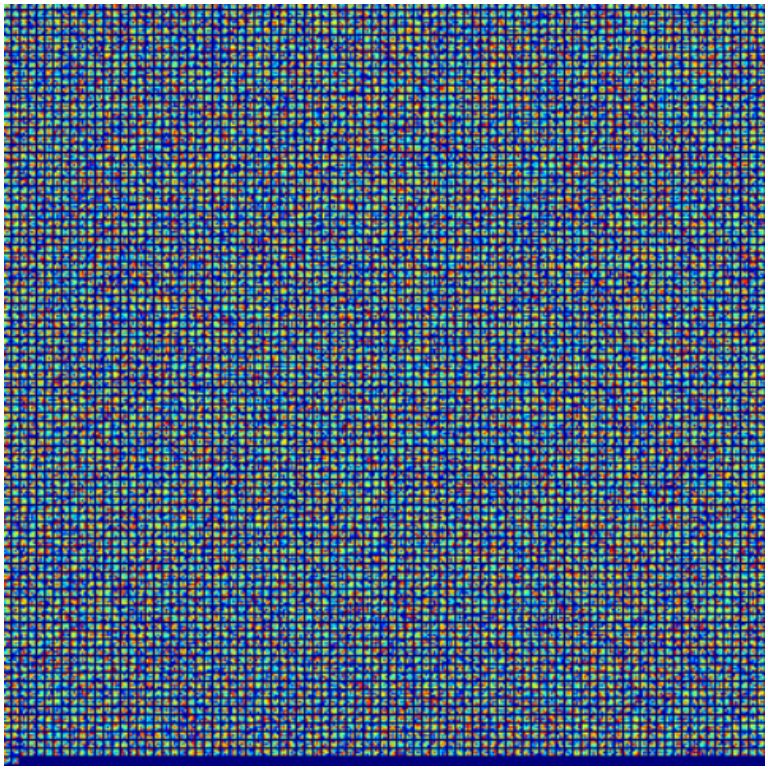
In addition to visualizing filter responses, we visualize the learned filters themselves. The learned filters are on a small spatial region and span in temporal dimensions up to 10 frames. Since ten frame cube is hard to visualize, we select the first 3 frames to visualize the filters of temporal regularity. Compared to the filters for object recognition that capture spatial structures [4], the temporal regular patterns do not have obvious spatial structure since they capture both spatial and temporal appearance thus look like random patterns. But they exhibit some forms of horizontal and vertical motions in a form of implicit horizontal and vertical lines of same colored pixels (best viewed in zoom-in).



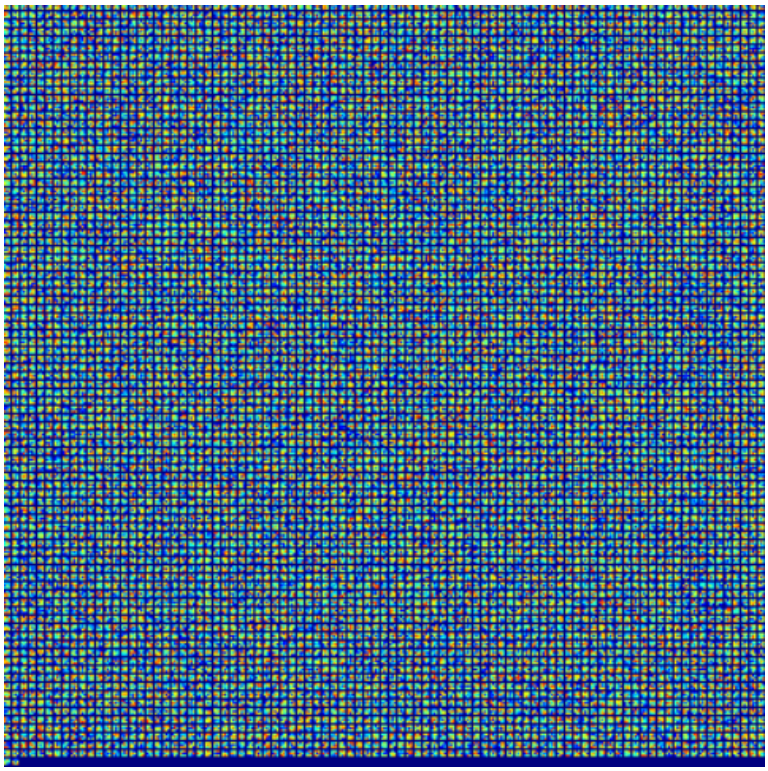
First convolutional layer



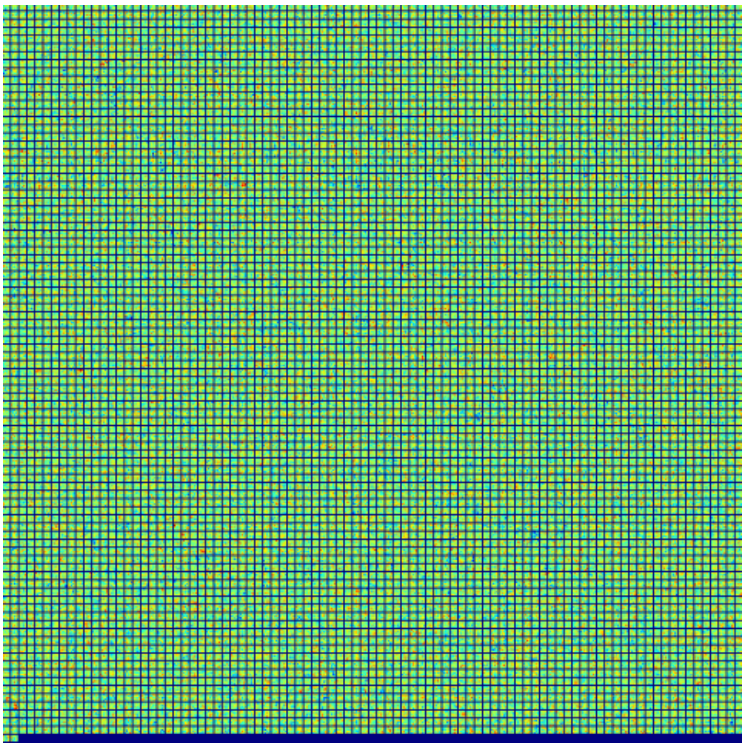
Second convolutional layer



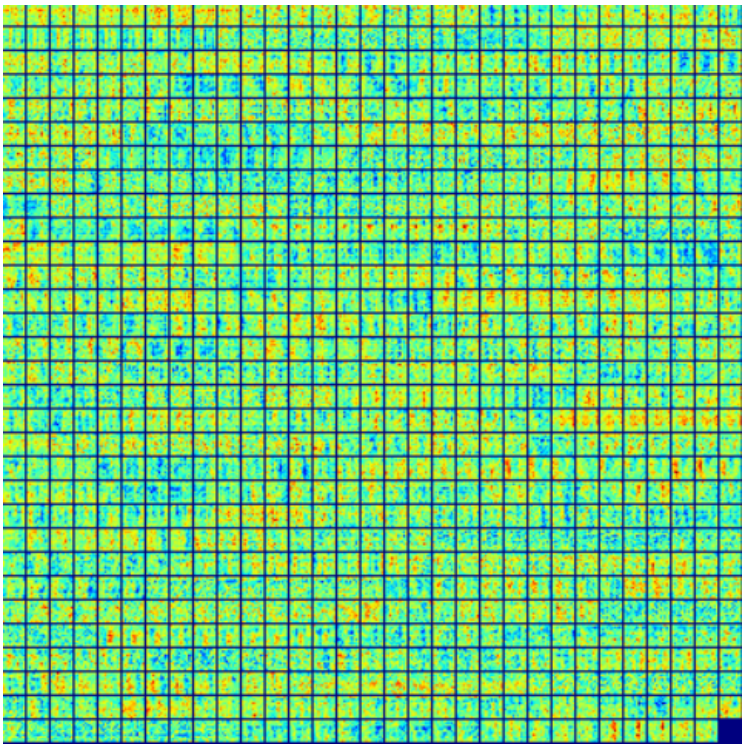
Third convolutional layer



First deconvolutional layer



Second deconvolutional layer



Third dconvolutional layer

Go to Table of Contents

References

[1] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *ICCV*, 2013, pp. 2720–2727. 2

[2] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *CVPR*, 2010, pp. 1975–1981. 2

[3] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 3, pp. 555–560, 2008. 2

[4] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Neural Networks," *CoRR*, 2013. 28