A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction
— Supplementary Material —

Zhian Liu[1], Yongwei Nie[1*], Chengjiang Long[2], Qing Zhang[3] and Guiqing Li[1]

[1]School of Computer Science and Engineering, South China University of Technology, China
[2]JD Finance America Corporation, Mountain View, CA, USA
[3]School of Computer Science and Engineering, Sun Yat-sen University, China

Abstract

In this supplementary material, we provide additional information that can not be included in the main manuscript due to the space limit.

1. Detailed Network Design

In Figure 1, we illustrate the detailed network architecture of the ML-MemAE-SC for flow reconstruction. Each cube in the network is the output feature maps for the corresponding layer. ML-MemAE-SC contains 4 levels in total. The kernel size of all convolutional layers in the network is fixed to 3×3. A basic convolution block contains a convolution layer, a batch-normalization layer and a ReLU activation layer sequentially. The downsampling and upsampling layers are implemented by stride-2 convolution and stride-2 deconvolution, respectively. The slot number of each memory module is fixed to 2K. Given an input flow with size (32,32,2), the feature maps sizes of each level are (32,32,32), (16,16,64), (8,8,128) and (4,4,256), respectively.

In Figure 2, we illustrate the detailed network architecture of the CVAE for flow-guided future frame prediction. Each cube in the network is the output feature maps for the corresponding layer. As shown, we have two encoders E_θ and F_ϕ that share similar architecture, and one decoder D_ψ. Inspired by the Variational UNet proposed in [1], we add skip connections between F_ϕ and D_ψ to help generating x_{t+1}. Following [1], the downsampling and upsampling layers are implemented by stride-2 convolution and subpixel convolution [6], respectively. And each Res-block follows a similar setting as in [2]. Our CVAE model also contains 4 levels in total, and the corresponding feature map sizes of each level are (32,32,64), (16,16,128), (8,8,128) and (4,4,128), respectively. We concatenate the sampled z with $E_\theta(\hat{y}_{1:t})$, which are sent to the decoder. Note that we utilize the last two bottleneck levels to estimate the distributions and sample data from them, and these two bottleneck levels share the same layer settings (please see the code for more details).

*Corresponding author: nieyongwei@scut.edu.cn

2. Sampling strategies during test time

Conditional variational autoencoder (CVAE), as a generative model, can produce different output results when given different latent code during testing. We test two sampling strategies: (1) **stochastical way**, *i.e.* sampling z from the posterior distribution $q(z|x_{1:t}, y_{1:t})$ randomly and (2) **deterministic way**, *i.e.* using the mean of the posterior distribution $q(z|x_{1:t}, y_{1:t})$ as the sampled z. For the UCSD Ped2 [5] dataset, the AUROC of the latter strategy is 99.3078%, while the former way gives performance ranging from 99.3065% to 99.3089%. This demonstrates that our model is robust though the predicted future frame is slightly different. But we still adopt the latter strategy to get statistically stable performance.

3. Number of reconstructed flows to CVAE

We have t previous frames and t corresponding optical flows ($t = 4$ in our setting). We explore the performance of our method when inputting different number of reconstructed flows into the CVAE based prediction module. For example, we can input all t reconstructed flows into CVAE, or just 1 reconstructed flow but $t - 1$ original flows into CVAE. As shown in Table 1, there are totally 4 variants. The results show that our method with all the four reconstructed flows achieves the best VAD performance.

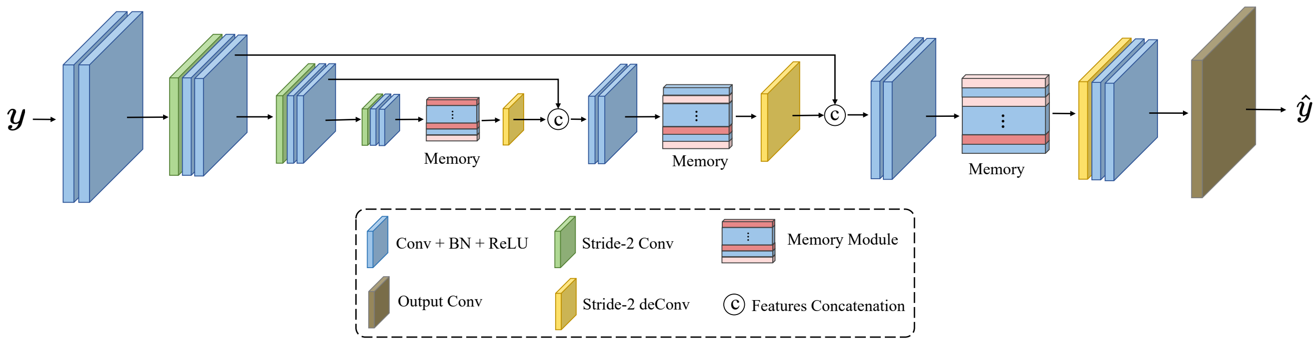


Figure 1: Detailed network architecture of the ML-MemAE-SC for flow reconstruction.

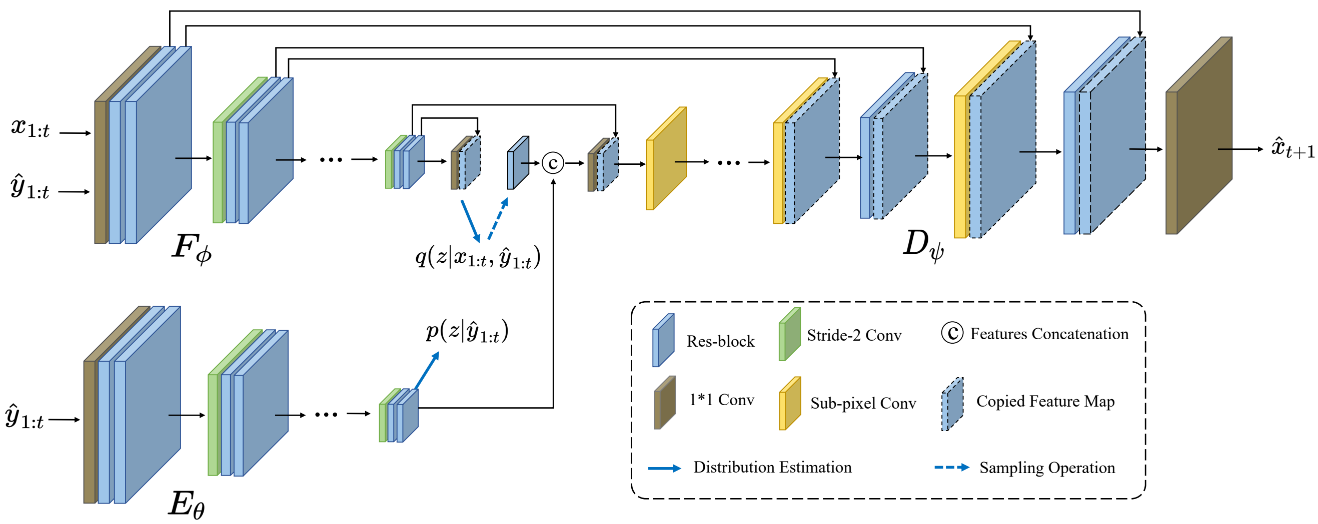


Figure 2: Detailed network architecture of the CVAE for flow-guided future frame prediction.

Table 1: Different number of reconstructed flows input into CVAE. As an example, $orig.\{1:t-1\}$ $recon.\{t\}$ means the flows from 1 to $t-1$ are original flows and the flow at time t is reconstructed, which are fed into CVAE for the future frame prediction. Results are obtained on Ped2.

	orig.{1:t-1} recon.{t}	orig.{1:t-2} recon.{t-1:t}	orig.{1:t-3} recon.{t-2:t}	orig.{1:t-4} recon.{t-3:t}
AUROC	98.70%	98.92%	99.25%	**99.31%**

4. Evaluation on UCF Crime

The three VAD datasets evaluated in the paper consist of surveillance videos with static backgrounds, for which anomalies come from dynamic foreground objects. Therefore, we extract STCs and process each foreground object separately. But our method can also be applied to the entire video frames. To show this, we conduct experiment on UCF-Crime dataset [7]. We select 10 videos for training and 6 for testing from UCF-Crime dataset. To be more specific, the training videos are *Normal_Videos165, 256, 267, 269, 279, 301, 355, 358, 489, 624*, and the test videos are *Arson011, Explosion004, Explosion008, Explosion013, Explosion021, Shooting008*. We train the proposed HF^2-VAD model on the entire frames and the AUROC result is 83.50% while that of VEC [8] is 81.12%.

5. Anomaly Detecting Cases

We visualize more anomaly detection examples of the proposed HF^2-VAD framework, showing some anomaly curves in Figure 3a, 3b-3c and 3d-3f for UCSD Ped2 [5], CUHK Avenue [3] and ShanghaiTech [4], respectively. In each subfigure, the red boxes in video frames denote the ground truth abnormal objects, and we plot the anomaly score of each frame over time. For a specific video, we calculate the AUROC under different model settings (higher AUROC means better anomaly detecting accuracy). We can observe that HF^2-VAD w/o FP or HF^2-VAD w/o FR can already detect most abnormal cases. Combining flow reconstruction and reconstructed-flow guided future frame prediction, the HF^2-VAD performs even better, producing relatively lower scores in the normal intervals and higher scores in the abnormal intervals.

6. More Qualitative Examples

We show more qualitative results of our proposed HF^2-VAD in Figure 4, demonstrating some flow reconstruction examples and frame prediction examples. As can be seen, given a video event (*i.e*., flow spatial-temporal cube and frame spatial-temporal cube), the output of ML-MemAE-SC are inclined to be reconstructed as a combination of some normal motion patterns. We can clearly see that the normal flow patches are reconstructed well while the abnormal ones are not, which is an apparent clue to detect anomaly. Using the reconstructed motion as condition, the predicted future frame for abnormal event is significantly different from the actual future, making it easier to be detected.

References

[1] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016. 1

[3] Cewu Lu, Jianping Shi, and Jiaya Jia. Abnormal event detection at 150 fps in matlab. In *Proceedings of the IEEE international conference on computer vision*, pages 2720–2727, 2013. 2, 5, 6

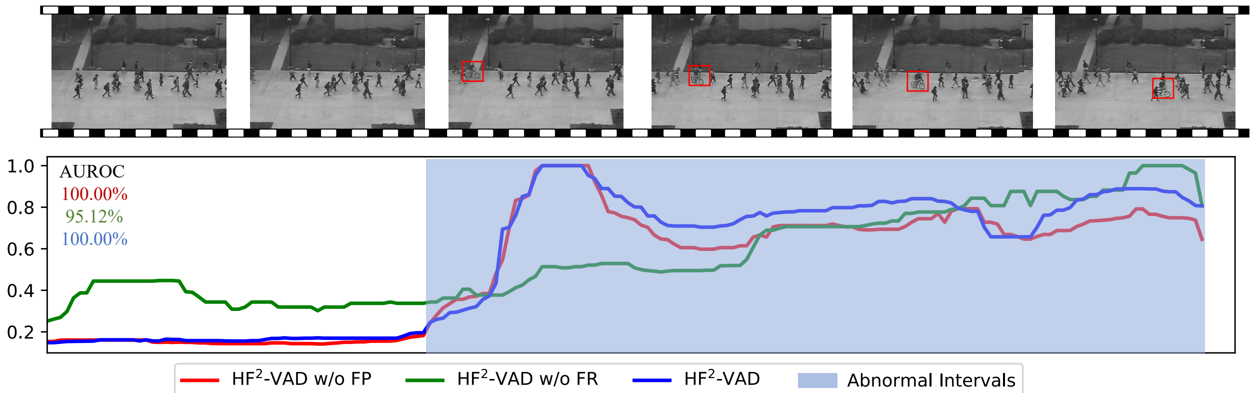
[4] Weixin Luo, Wen Liu, and Shenghua Gao. A revisit of sparse coding based anomaly detection in stacked rnn framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 341–349, 2017. 2, 5, 6

[5] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1975–1981. IEEE, 2010. 1, 2, 5, 6

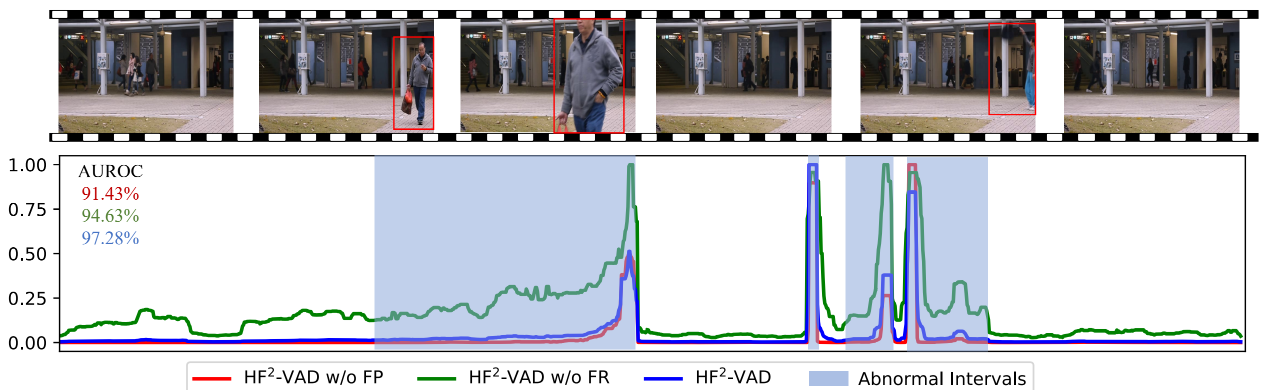
[6] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 1

[7] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *The IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 2

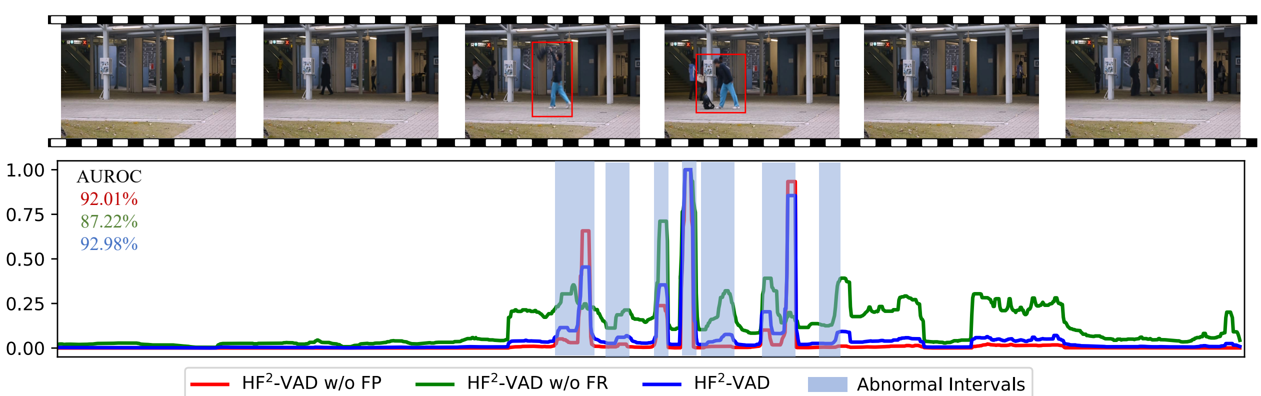
[8] Guang Yu, Siqi Wang, Zhiping Cai, En Zhu, Chuanfu Xu, Jianping Yin, and Marius Kloft. Cloze test helps: Effective video anomaly detection via learning to complete video events. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 583–591, 2020. 2



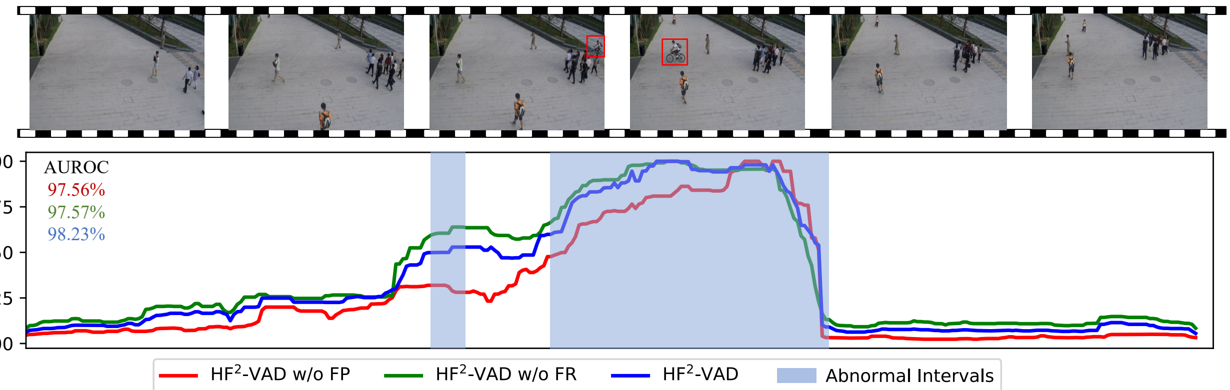
(a) Ped2 test video 01 with abnormal event: bicycle riding.



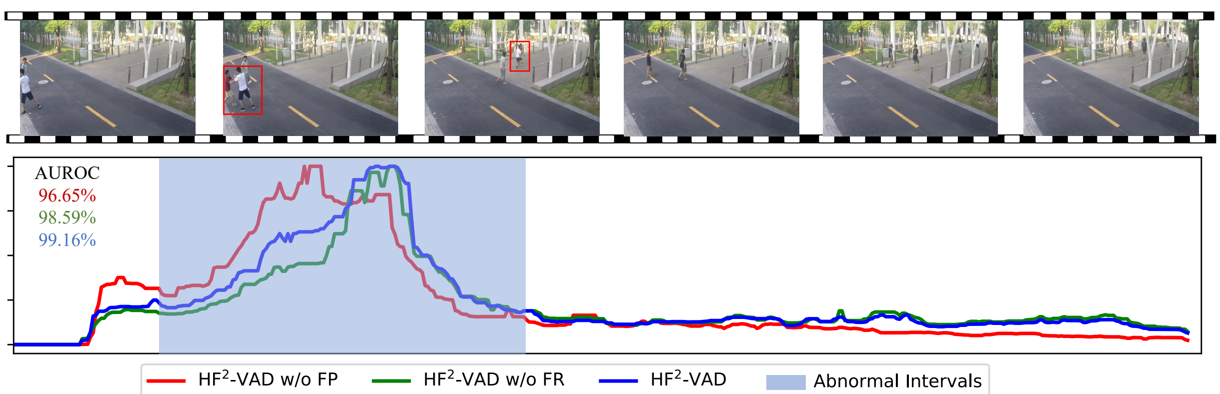
(b) Avenue test video 06 with abnormal events: wrong direction and throwing backpack.



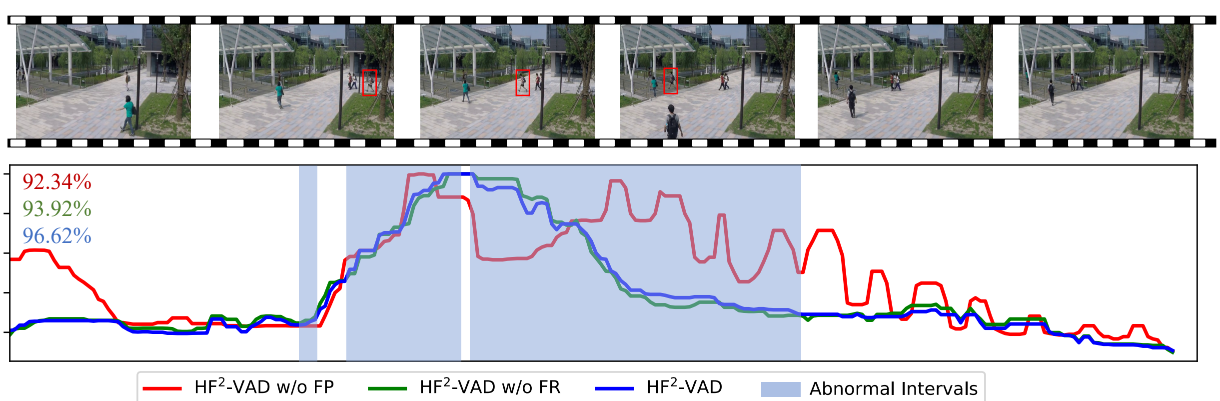
(c) Avenue test video 12 with abnormal event: throwing backpack.



(d) ShanghaiTech test video 00_0052 with abnormal event: bicycle riding.



(e) ShanghaiTech test video 05_0024 with abnormal event: fighting and chasing.



(f) ShanghaiTech test video 08_0079 with abnormal event: running.

Figure 3: Anomaly detecting examples on USCD Ped2 [5], CUHK Avenue [3] and ShanghaiTech [4]. The horizontal axis denotes time, while the vertical axis denotes anomaly score (higher value indicates more possible to be abnormal). The values in the upper left corner denote AUROCs under different model settings. Best viewed in color.

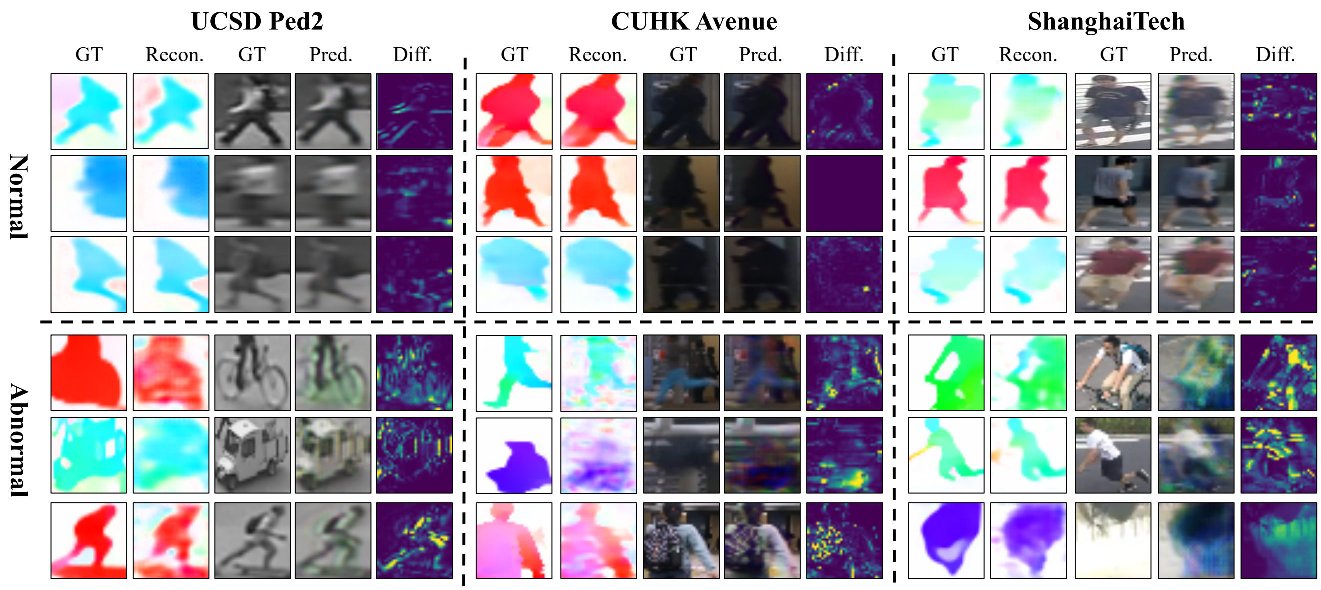


Figure 4: Visualization of some flow reconstruction and future frame prediction examples on UCSD Ped2 [5], CUHK Avenue [3] and ShanghaiTech [4] datasets. For each dataset, from left to right, we sequentially show the ground-truth flow, reconstructed flow, ground-truth future frame, predicted future frame and the prediction error map, respectively. The top and bottom regions show normal and abnormal samples respectively. The lighter color in the difference maps denotes larger prediction error. Best viewed in color.