1. Supplementary (Rebuttal)

We would like to highlight technique contribution as follows: that we customize a video prediction network more appropriately for anomaly detection purpose other than general purpose. Specifically, anomaly can be identified either from motion or appearance. For appearance, we adopt an UNet architecture because neighboring frames in normal events are similar, and UNet feeds features extracted from previous frames to predict future frames; For anomaly in appearance, UNet cannot predict future well. For motion, the optical flow consistence constraint actually enforces UNet to predict the next frame which agrees with motion of normal events. We do not take a separate network to leverage the motion information of observed frames for future prediction on the testing data, and it is not desirable for anomaly detection because even for abnormal events, sometimes a longer motion observation also helps the accurate prediction of next frame;

Both [4] (Decomposing motion and content for natural video sequence prediction) and [5](Generating the Future with Adversarial Transformers) achieve state-of-the-art performance. [5] uses a LSTM based motion encoder to encode ALL history motion for prediction. [5] proposes to learn transformers for future prediction. Both of them are designed for more general future prediction. We will include them in our paper. Even though motion constraint is also used in [4], it uses intensity difference as motion characterization while we use optical flow. Further, it uses history motion information of observed frames for future video prediction on the testing data, so it may fail the detection of anomaly in motion. But the motion constraint in our method only regularizes an UNet to better predict normal events. Further, in Table 1, we report optical flow PSNR for both normal and abnormal events and their gap. We can see that the gap of [4] between normal and abnormal events is smaller than ours in terms of motion. So our method is more suitable for anomaly detection.

Table 1. The normal and abnormal optical flow PSNR and its gap between these two PSNR. For each cell, the left one is [4] and the right one is our method.

	normal	abnormal	gap
Ped1	30.9 / 27.7	30.0 / 21.3	0.9 / **6.3**
Ped2	30.2 / 30.9	29.0 / 25.8	1.2 / **5.1**
Avenue	30.4 / 27.3	27.1 / 17.0	3.3 / **10.3**
ShanghaiTech	32.5 / 32.4	32.1 / 26.2	0.4 / **6.2**

We compare our work with [4] in Table 2 by using the codes provided from author[1]. We can see the promising results of [4] but our methods still achieves the best performance.

Table 2. We compare our method with [4] in Ped1, Ped2, Avenue and ShanghaiTech (ST) Datasets.

	Ped1	Ped2	Avenue	ST
[1]	75.7%	95.0%	83.8%	72.6%
our method	**83.1%**	**95.4%**	**84.9%**	**72.8%**

Q1:The paper does not include some recent papers [1][2][3] reporting better results.

A1: We compare our model with these advanced baselines in Table 3. Our method outperforms other baselines in Ped2 and Avenue. We will include the comparisons in our paper. In Ped1, objects become smaller as they move from near to far. Further, some anomalies are caused by small objects, like someone skateboarding. Though they are shaded by trees for more than 20 frames, these frames are still annotated with abnormal labels. Howeverour method uses only 4 previous frames to predict the next one, so we cannot handle these anomalies. We believe more frames used for future prediction can benefit such a case that anomalies can be found only by observing for a long period.

Table 3. We compare our method with [R1][R2][R3] in Ped1, Ped2 and Avenue Datasets.

	Ped1	Ped2	Avenue
AbnormalGAN [1]	**97.4%**	93.5%	N/A
DeepAppearance [2]	N/A	N/A	84.6%
GrowingGas [3]	93.8%	94.1%	N/A
Our method	83.1%	**95.4%**	**84.9%**

Q2: I do not follow the argument presented in the introduction / related work that all other works are based on the reconstruction error...

A2: We will claim that those Auto-Encoder based methods are not suitable for anomaly detection, which may learn an "Identity" mapping. Further, we have conducted some experiments to support this claim in Session 4.7. The words reconstruction error are not precise here and we will restate them and use Auto-Encoder reconstruction error later.

References

[1] M. Ravanbakhsh, M. Nabi, E. Sangineto, L. Marcenaro, C. S. Regazzoni, and N. Sebe. Abnormal event detection in videos using generative adversarial nets. In *ICIP*, pages 1577–1581, 2017.

[2] S. Smeureanu, R. T. Ionescu, M. Popescu, and B. Alexe. Deep appearance features for abnormal behavior detection in video. In *ICIAP*, pages 779–789, 2017.

[3] Q. Sun, H. Liu, and T. Harada. Online growing neural gas for anomaly detection in changing surveillance scenes. *Pattern Recognition*, 64:187–201, 2017.

[4] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee. Decomposing motion and content for natural video sequence prediction. *ICLR*, 2017.

[5] C. Vondrick and A. Torralba. Generating the future with adversarial transformers. In *CVPR*, pages 2992–3000, 2017.

[1] https://github.com/rubenvillegas/iclr2017mcnet