基于预测未来帧的异常事件检测Future Frame Prediction for Anomaly Detection – A New Baseline

摘要

视频异常事件检测的目标是识别出那些不服从期望行为的。在本文中，作者提出了一个基于预测未来帧（future frame prediction）的架构，此外，为了更好地预测未来的帧，除了通常使用的空间约束（spatial constrains），i.e. intensity constrain和gradient constrain，作者还加上了时序约束（temporal constrain），i.e.光流场约束（optical flow constrain）。作者分别在几个公开数据集和一个toy数据集上进行了实验，代码开源在github上:https://github.com/StevenLiuWen/ano_pred_cvpr2018

1. 介绍

视频异常事件检测的目标是识别出那些不遵从从期望行为的。这个任务十分困难因为异常事件通常是一个无界的概念，我们也无法列举出所有的异常事件类别，所以基于classification的策略是不合适的。并且，我们的数据集通常也大多都是一些正常的事件，异常事件在一个视频中也是少量的帧，人工对每帧去进行标注也是不现实的。

目前有许多这方面的研究，在正常的训练数据上进行特征重建是一个常用的策略。基于特征的使用选取，现有的方法可以大致分为两类：

1. 基于hand-crafted特征选取的：这种方法将视频用一些手工选取的特征进行代表，然后学习出一个正常事件的字典，从而那些异常事件在这个空间上会存在较大的reconstruction errors。
2. 基于DL方法的：采用auto-encoder的结果对正常事件的pattern进行自学习，然后同样，异常事件也在最后会存在较大的reconstruction loss.但是深度神经网络的泛化能力很强，因而，就算有异常事件发生，重建loss不一定会很大。

最近，生成对抗网络（GAN）的迅速发展也对视频预测任务产生了促进。综上，作者提出了基于预测未来帧的视频异常事件检测架构，具体来说，给定一些训练video clips,模型会学习生成一个predictor，它能根据现有的视频帧预测下一正常的帧，在测试阶段，输入一些帧，模型根据已经学习到的内容同样预测下一帧，如果和预测输出的帧满足，则判定为normal的，否则判定为abnormal的。总的来说，该文章的贡献如下：

1. 首次提出了这种基于future frame的架构
2. 引入了基于光流场的约束（optical flow constrain）
3. 在公开数据集和一个toy数据集上实验验证了该方法的有效性。

2. Related Work

2.1 基于手工特征选取的方法

基于手工特征选取的方法大都遵从以下三个模块：

1. 提取特征：例如low-level的轨迹特征（trajectory features）、HOG【7】、HOF【8】等。
2. 建立模型对regular patterns进行学习：Zhang等人【41】利用马尔科夫随机场，Mahadevan等人【25】利用高斯混合模型
3. 识别异常事件。

2.2 基于深度学习的方法

Xu等人【40】利用multi-layer的自动编码器完成特征学习，【14】中利用3D-Conv-AE对正常帧进行建模，【5】和【23】中利用ConvLSTM-AE对appearance和motion信息的patterns进行了建模。

2.3 视频帧预测

【27】中提出了一个基于对抗学习的multi-scale network来生成视频的未来帧，这也是本文参考的一个重要文章。

3. 基于预测未来帧的异常事件检测

总体的思想是，对正常的帧的模式进行学习，然后再基于已有的帧对未来帧进行一个预测，记为prediction，如果prediction和ground truth相差很大，就视为异常事件。举例来说，如果有一个视频，开始一部分都是人群在正常过马路，模型能实时地对下一帧进行预测，如果突然有一辆卡车出现，而模型预测的帧任然是人群过马路，这样产生出入较大，就视为异常事件。之前的大多工作也只考虑到了上述的appearance信息（一辆卡车突然冲入人群），但是作者任务motion信息同样重要（例如一个扒手经过一个人），所以提出了optical flow constant。总体的框架如下图所示：

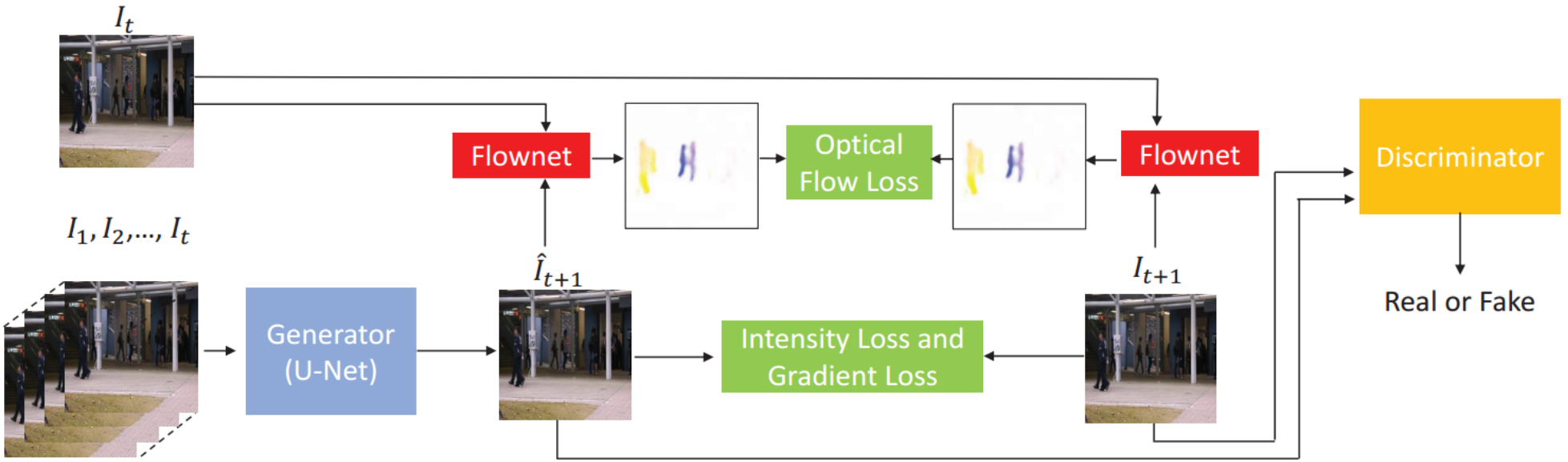

Figure 2. The pipeline of our video frame prediction network. Here we adopt U-Net as generator to predict next frame. To generate high quality image, we adopt the constraints in terms of appearance (intensity loss and gradient loss) and motion (optical flow loss). Here Flownet is a pretrained network used to calculate optical flow. We also leverage the adversarial training to discriminate whether the prediction is real or fake.

3.1 未来帧预测

Generator部分采用的是类似U-Net的结构，如下所示：

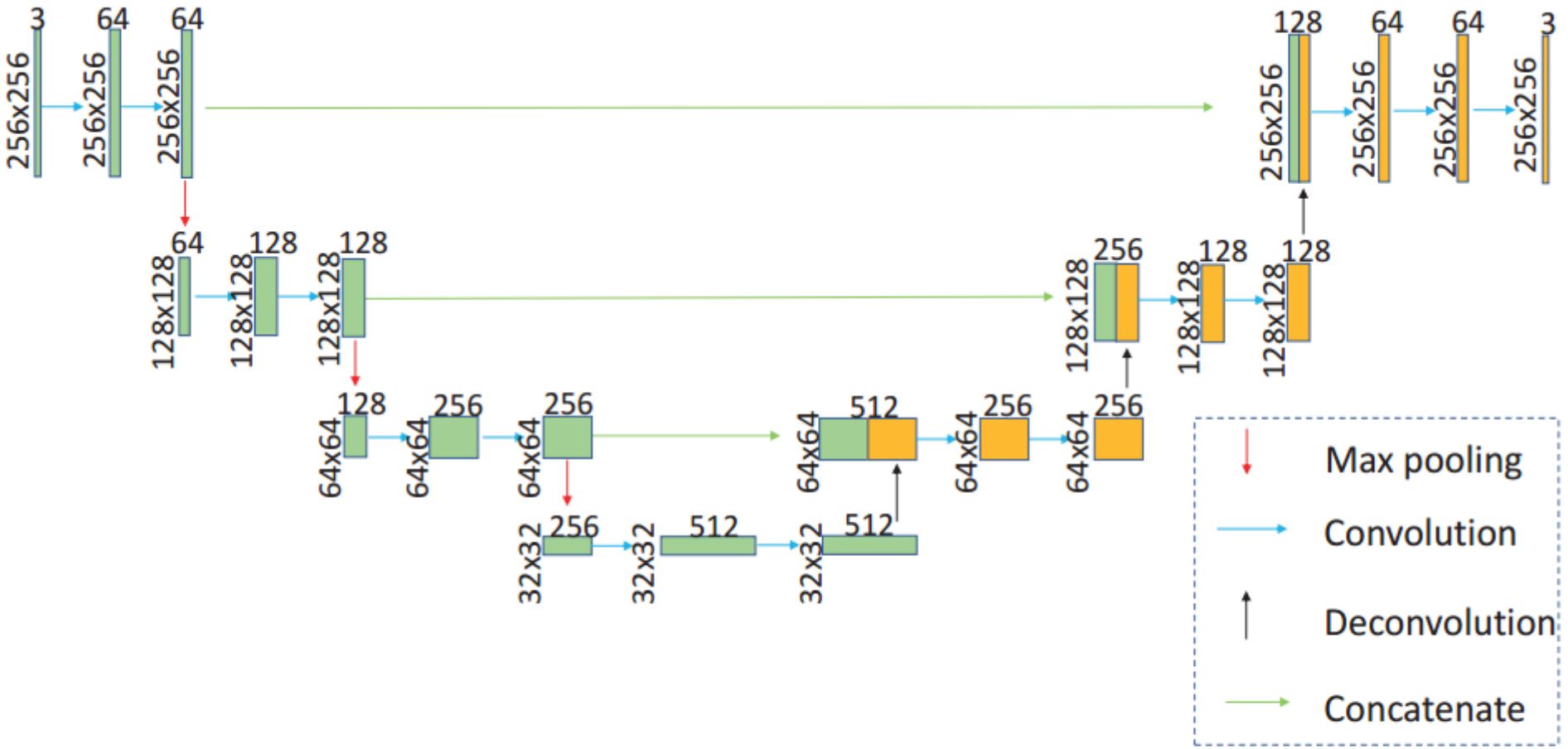

Figure 3. The network architecture of our main prediction network (U-Net). The resolutions of input and output are the same.

> We sequentially stack all these frames and use them to predict a future frame I_{t+1}

至于具体是怎么操作这些前序帧的，还待看代码进一步了解。

3.2 各种约束

数学公式描述，给定一些连续帧$I_1, I_2, \ldots, I_t$,预测出的帧记为$\hat{I}_{t+1}$,ground truth帧为I_{t+1}，我们最小化两种之间的intensity distance和gradient distance，也就是之前提到的appearance constrians，分别按下式计算：

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

$$L_{gd}(\hat{I}, I) = \sum_{i,j} \left\| |\hat{I}_{i,j} - \hat{I}_{i-1,j}| - |I_{i,j} - I_{i-1,j}| \right\|_1 + \left\| |\hat{I}_{i,j} - \hat{I}_{i,j-1}| - |I_{i,j} - I_{i,j-1}| \right\|_1$$

gradient constrain目的就是保证在prediction中的每个像素，让它和它左边和它上面的像素之前的梯度保持和gound truth中的情况一致。

> To preserve the temporal coherence between neighboring frames, we enforce the optical flow between I_{t+1} and I_t and that between $\hat{I}_{t+1}$and I_t to be close.

optical flow constrain数学公式表示如下：

$$L_{op}\left(\hat{I}_{t+1}, I_{t+1}, I_t\right) = \left\| f\left(\hat{I}_{t+1}, I_t\right) - f\left(I_{t+1}, I_t\right) \right\|_1$$

其中，optical flow的求解用到了Flownet【40】，这是一个基于CNN的光流场estimator。

3.4 对抗训练

根据GAN的思路，有一个生成器G和判别器D，当对抗训练收敛后，G可以生成一些图片，让判别器D判别为真的（其实是假的），而判别器又是可以对图片进行判别是真还是假。于是训练分为两大步骤：

- 一、训练判别器D

判别器的目的就是把I_{t+1} 分类为1,把$\hat{I}_{t+1}$ 分类为0,这里的1和0分别代表正常和异常帧。在训练D的时候，fix生成器G的参数，文中用的是MSE损失函数：

$$L_{adv}^{\mathcal{D}}(\hat{I}, I) = \sum_{i,j} \frac{1}{2} L_{MSE}\left(\mathcal{D}(I)_{i,j}, 1\right) + \sum_{i,j} \frac{1}{2} L_{MSE}\left(\mathcal{D}(\hat{I})_{i,j}, 0\right)$$

于是整个判别器D的Object function为:

$$L_{\mathcal{D}} = L_{adv}^{\mathcal{D}}\left(\hat{I}_{t+1}, I_{t+1}\right)$$

这里注意一下，为什么$D(I)$还会有下标i, j呢？是不是觉得有点奇怪？原因是文中follow了【17】的工作，采用了一个patch discriminator，也就是说可以对一张大的图片进行分割（例如3*3的小patches），然后把整个图片送入判别器discriminator，输出每个patch对应的类别。所以就i, j其实是每个小patch的索引啦。

- 二、训练生成器G

生成器G的目标是生成一些帧，让判别器判别为1，也就是正常帧，在训练G的时候，fix判别器D的参数：

$$L_{adv}^{\mathcal{G}}(\hat{I}) = \sum_{i,j} \frac{1}{2} L_{MSE}\left(\mathcal{D}(\hat{I})_{i,j}, 1\right)$$

于是整个生成器G的Object function为:

$$L_{\mathcal{G}} = \lambda_{\text{int}} L_{\text{int}}\left(\hat{I}_{t+1}, I_{t+1}\right) + \lambda_{gd} L_{gd}\left(\hat{I}_{t+1}, I_{t+1}\right) + \lambda_{op} L_{op}\left(\hat{I}_{t+1}, I_{t+1}, I_t\right) + \lambda_{adv} L_{adv}^{\mathcal{G}}\left(\hat{I}_{t+1}\right)$$

$\lambda_{int}, \lambda_{gd}, \lambda_{op}, \lambda_{adv}^{g}$分别是各部分的权重，具体取值参见论文，经验值。

在训练的适合，把像素值normalize到$[-1, 1]$,每帧的大小是256*256，和【27】类似，t取值为4，也就是说每个video clip一共有5帧，mini-batch size 是 4

3.6 测试数据异常检测

同样接上之前人群过马路的例子，我们说，只要当预测出来的帧和真实的帧相差很大，我们就认为是异常了，那么，度量两张图片之间的差异程度有什么方法呢？ 欧氏距离或PSNR，本文就选取了PSNR（Peak Signal to Noise Ratio），定义如下：

$$PSNR(I,\hat{I}) = 10\log_{10}\frac{[\max_j]^2}{\frac{1}{N}\sum_{i=0}^{N}\left(I_i-\hat{I}_i\right)^2}$$

于是我们可以将test video中的每一帧都和对应的预测帧求一个PSNR（当然，最开始前t帧是无法求的）。PSNR越大，就表示两张图片越相似，也就是说更有可能是正常帧，为了后续更好地选取阈值，将PSNR也归一化到$[0,1]$。

$$S(t) = \frac{PSNR\left(I_t,\hat{I}_t\right) - \min_t PSNR\left(I_t,\hat{I}_t\right)}{\max_t PSNR\left(I_t,\hat{I}_t\right) - \min_t PSNR\left(I_t,\hat{I}_t\right)}$$

然后我们就可以给定一个阈值，例如，0.6，然后把低于0.6的全部判断为anomaly。

4. 对比实验

这部分作者做了比较多的对比实验，比如prediction network部分的设计，和【27】中的Beyond-MSE进行了对比；也分别比较了各个constrain和对抗训练的必要性；最后还将本文这种基于future frame prediction策略的模型和基于自编码器策略的模型（Conv-AE 【14】）做了一个综合对比，也证明本文中的方法略好。

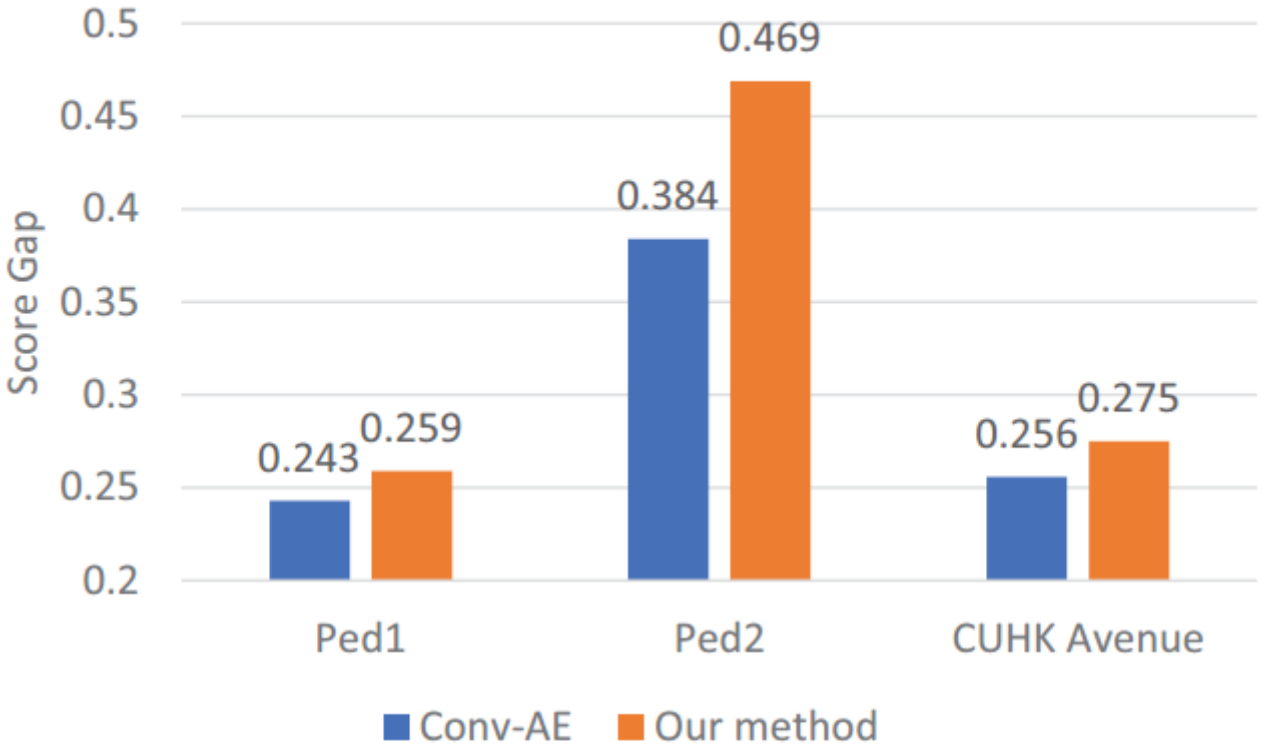

Figure 7. We firstly compute the average score for normal frames and that for abnormal frames in the testing set of the Ped1, Ped2 and Avenue datasets. Then, we calculate the difference of these two scores(Δ_s) to measure the ability of our method and Conv-AE to discriminate normal and abnormal frames. A larger gap(Δ_s) corresponds to small false alarm rate and higher detection rate. The results show that our method consistently outperforms Conv-AE in term of the score gap between normal and abnormal events.

最后是作者在一个toy数据集上进行的验证实验：

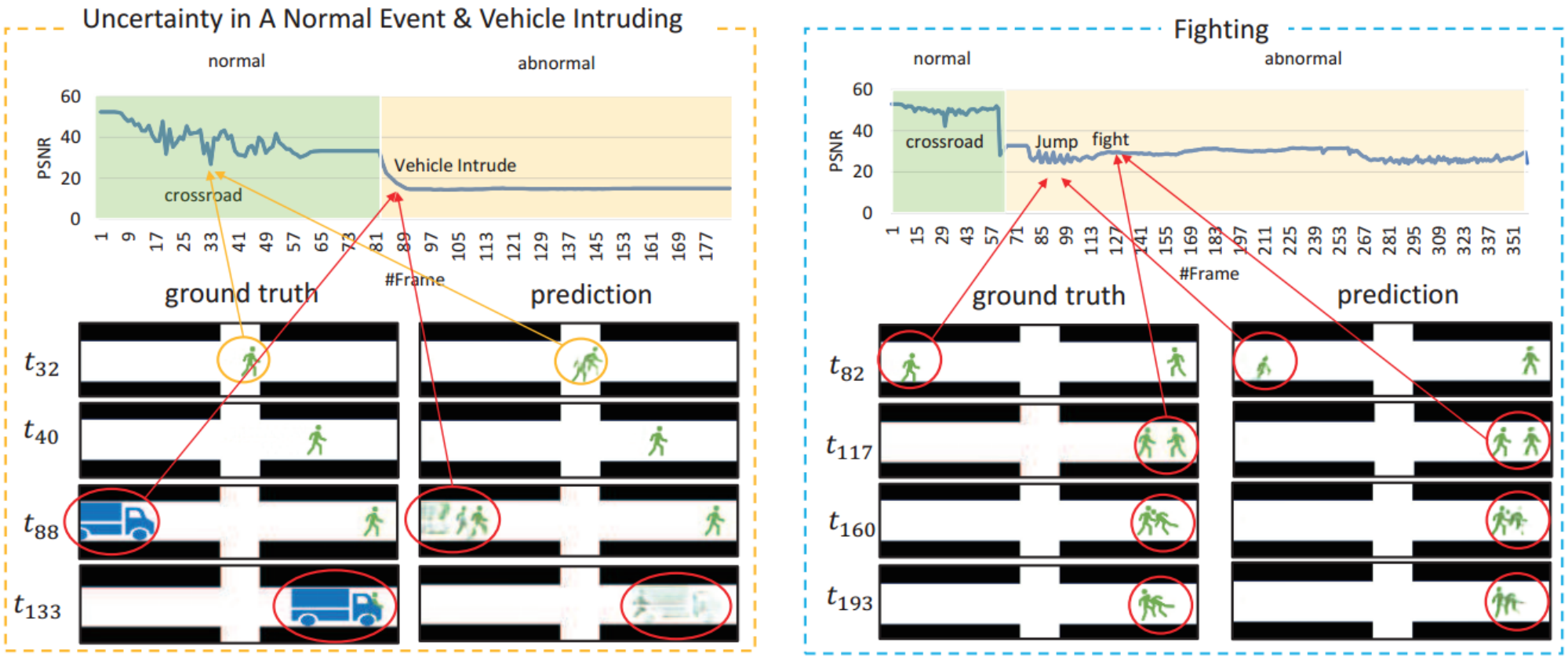

Figure 8. The visualization of predicted testing frames in our toy pedestrian dataset. There are two abnormal cases including vehicle intruding(left column) and humans fighting(right column). The orange circles correspond to normal events with uncertainty in prediction while the red ones correspond to abnormal events. It is noticeable that the predicted truck is blurred, because no vehicles appear in the training set. Further, in the fighting case, two persons cannot be predicted well because fighting motion never appear in the training phase.

Reference

- Wen Liu∗, Weixin Luo∗, Dongze Lian, Shenghua Gaoy: Future Frame Prediction for Anomaly Detection – A New Baseline, CVPR 2018