# Future Frame Prediction for Anomaly Detection – A New Baseline

CVPR 2018
Wen Liu, Weixin Luo, Dongze Lian, Shenghua Gao
ShanghaiTech University

SeulGi Park

February 27, 2020

# Contents

# 0. Abstract

- 이상(Anomaly)?
  - 예상(prediction)하지 못한 행동
- 이상 탐지(Anomaly Detection)?
  - 예상되는 행동이 아닌 이벤트를 식별하는 것

    (to the identification of events that do not conform to expected behavior)
- Proposed
  - U-Net을 변형한 Video prediction framework를 이용한 이상 탐지
  - 정상 이벤트를 학습시킴으로써 미래 프레임을 prediction

    (prediction된 미래 프레임과 ground truth의 차이를 줄이는 방법으로 모델을 만듦)
  - 시공간의 제약을 둠으로써 정상 이벤트를 prediction 할 수 있는 framework를 만듦

# 1. Introduction

- 현실에서의 비정상적인 이벤트는 제한적이지 않기 때문에 모든 종류의 비정상 이벤트를 수집하고 분류하는 방법으로 문제를 해결하는 것은 불가능

- 최근 이상 탐지에서는 training data를 이용한 feature reconstruction을 일반적으로 사용
  1) Hand-crafted features based methods
  2) Deep learning based methods

- Training data를 이용한 Feature reconstruction 방법은 reconstruction의 오류를 줄이면서 이상을 탐지하기 위해 학습하기 때문에 dataset에 국한된 모델을 만들어 일반화할 수 있는 이상 탐지 모델을 만들기에는 한계가 있음

Thus, we can see that almost all training data reconstruction based methods cannot guarantee the finding of abnormal events.

# 1. Introduction

- 이상 탐지를 위한 training data를 reconstruction하는 것이 아니라, 정상적인 사건을 prediction하는 모델을 만듦으로써, 비정상적인 사건이 나타났을 경우 prediction된 프레임과 비교하여 식별하는 framework를 제안

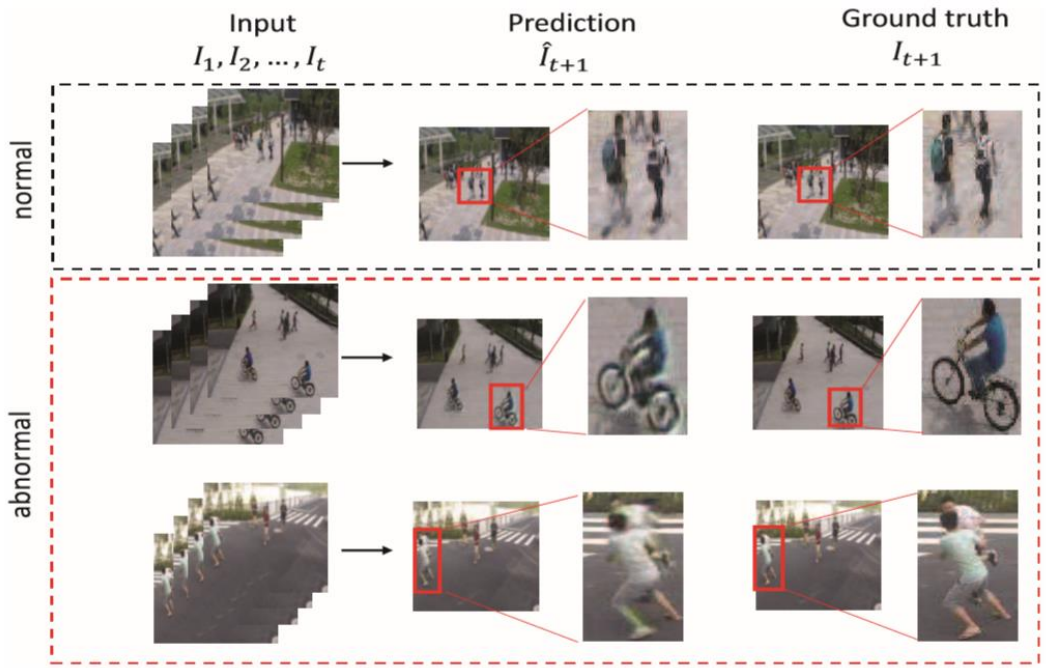



Figure 1. Some predicted frames and their ground truth in normal and abnormal events. Here the region is walking zone. When pedestrians are walking in the area, the frames can be well predicted. While for some abnormal events (a bicycle intrudes/ two men are fighting), the predictions are blurred and with color distortion. Best viewed in color.

# 2. Future Frame Prediction Based Anomaly Detection Method

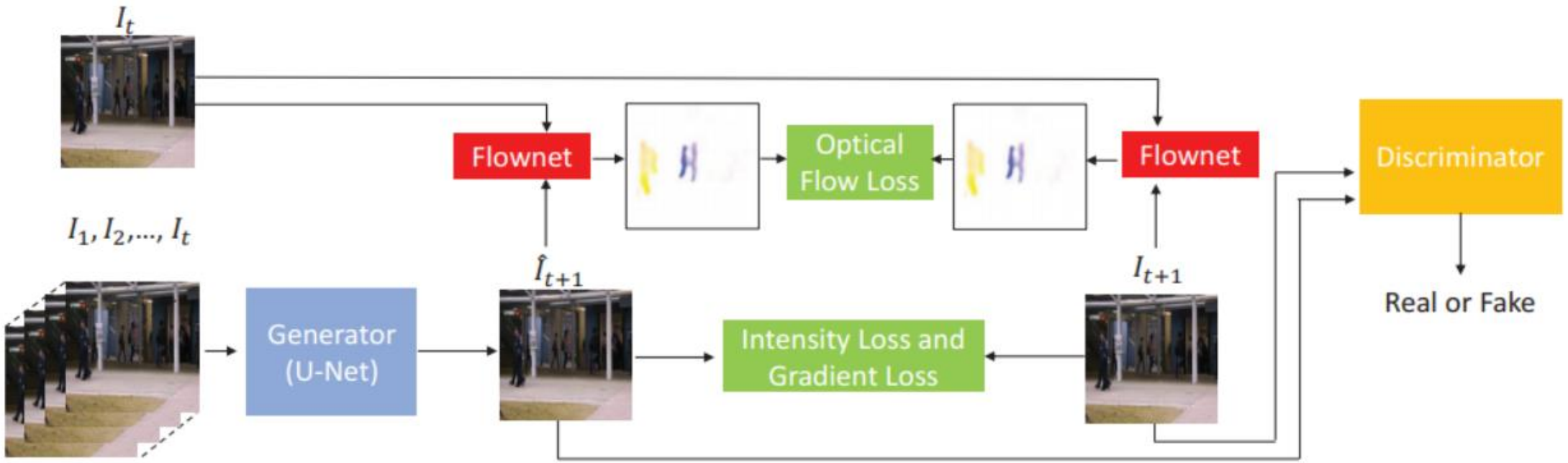
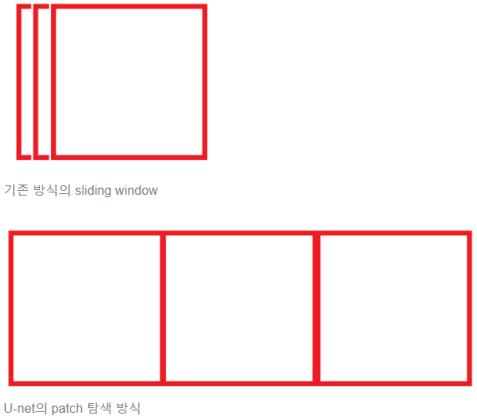



Figure 2. The pipeline of our video frame prediction network. Here we adopt U-Net as generator to predict next frame. To generate high quality image, we adopt the constraints in terms of appearance (intensity loss and gradient loss) and motion (optical flow loss). Here Flownet is a pretrained network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

# 2. Future Frame Prediction Based Anomaly Detection Method

## 2.1 Future Frame Prediction

– U-Net 사용(①기존의 Segmentation network와 비교하여 속도가 빠름
②output layer를 동시에 검증하기 때문에 localization과 context 인식이 가능)




1) Patch: 이미지 인식 단위

2) Contracting path: 공간 해상도를 감소시켜 특징을 추출하는 인코더
Expanding path: 공간 해상도를 증가시켜 프레임을 복구하는 디코더

3) 각 계층에서 gradient loss 및 정보 불균형을 → 로 해결

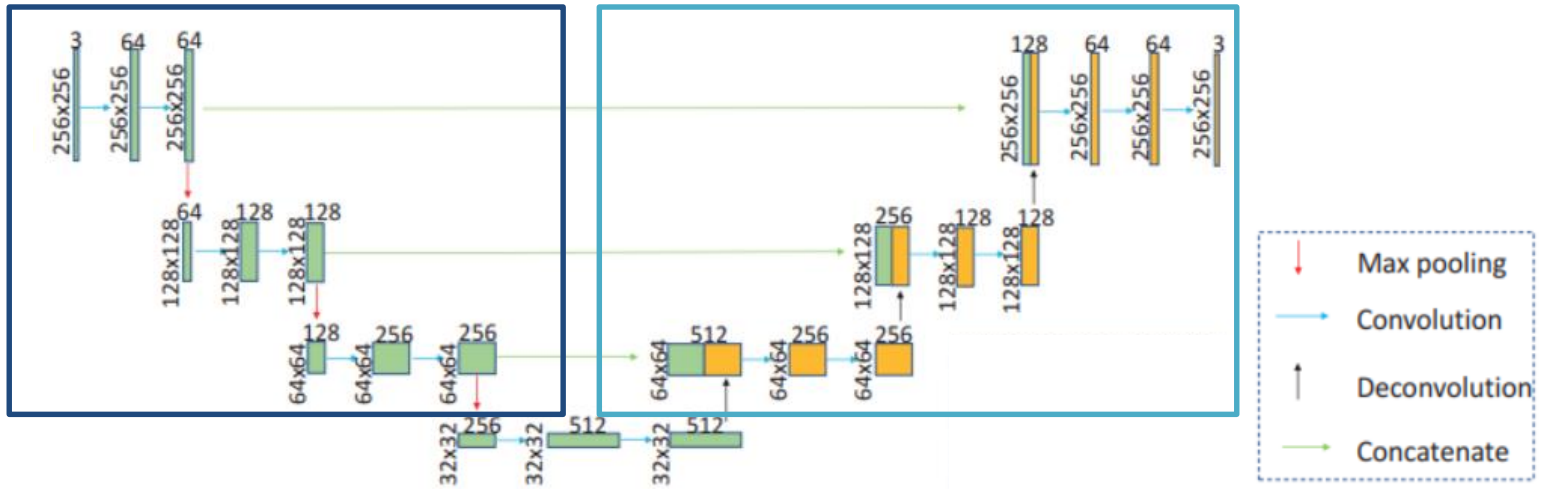



Figure 3. The network architecture of our main prediction network (U-Net). The resolutions of input and output are the same.

# 2. Future Frame Prediction Based Anomaly Detection Method

## 2.2 The Constraints on Intensity and Gradient

- Prediction을 ground truth에 가깝게 학습시키기 위해, intensity와 gradient를 제약
- Intensity: prediction과 ground truth의 RGB 색상을 유사하게 제약(L2 distance를 통해 동일한 위치)

$$L_{int}(\hat{I}, I) = \|\hat{I} - I\|_2^2 \tag{1}$$

- Gradient: 비디오 프레임 공간 인덱스 i, j를 사용하여 prediction frame을 선명하게 만듦

$$\begin{aligned} L_{gd}(\hat{I}, I) = \sum_{i,j} & \left\| |\hat{I}_{i,j} - \hat{I}_{i-1,j}| - |I_{i,j} - I_{i-1,j}| \right\|_1 \\ & + \left\| |\hat{I}_{i,j} - \hat{I}_{i,j-1}| - |I_{i,j} - I_{i,j-1}| \right\|_1 \end{aligned} \tag{2}$$

# 2. Future Frame Prediction Based Anomaly Detection Method

## 2.3 The Constraint on Motion

- Prediction된 frame 내 객체의 motion을 제약해야 할 필요가 있음(Intensity와 Gradient 제약으로는 부족)
- Motion은 시간적 제약이 필요(optical flow를 이용)
- 최근 제안된 Optical Flow estimation을 위한 CNN 기반의 접근법인 Flownet을 사용

$$L_{op}(\hat{I}_{t+1}, I_{t+1}, I_t) = \|f(\hat{I}_{t+1}, I_t) - f(I_{t+1}, I_t)\|_1 \quad (3)$$

# 2. Future Frame Prediction Based Anomaly Detection Method

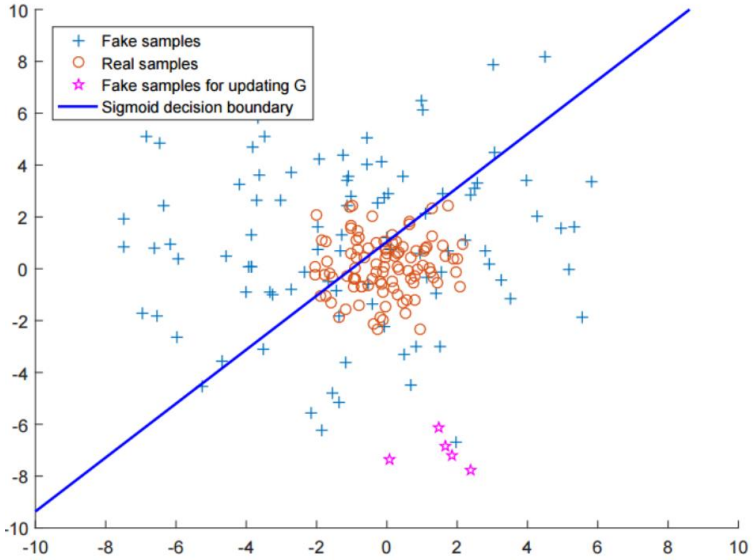
## 2.4 Adversarial Training

- GAN(Generative adversarial network): 이미지 및 비디오 생성에 유용
- 현실적인 미래 프레임을 생성하기 위해 Least Square GAN을 사용

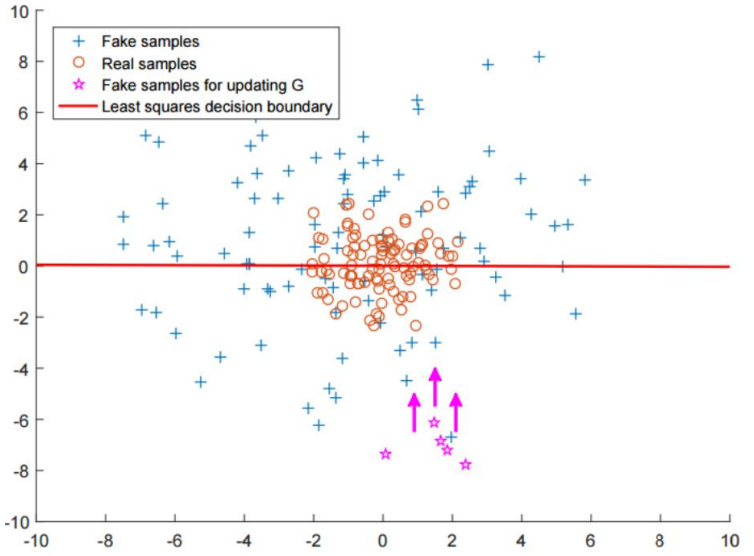
- Regular GAN: discriminator, sigmoid cross entropy loss function
  - discriminator는 자신이 받은 sample이 "진짜"일 확률
- loss 함수 때문에 generator update시, gradient vanishing이 생김
  - generator,이미 discriminator를 매우 잘 속이고 있기 때문에 딱히 더 학습할 의지가 없음(**vanishing gradient**)

- 별표 sample들은 실제 진짜 sample들이 모여 있는 분포에서는 멀리 떨어져 있음=>별표 sample들을 진짜 data 방향으로 끌고 와보자

- **Discriminator에 sigmoid cross entropy loss 대신 least square loss를 사용, decision boundary에서 멀리 있는 sample들에게 penalty를 주자!**



Sigmoid decision boundary



Least square decision boundary


http://jaejunyoo.blogspot.com/2017/03/lsgan-1.html

# 2. Future Frame Prediction Based Anomaly Detection Method

## 2.4 Adversarial Training – **D(discriminator) Training**

- Ground truth의 $I_{t+1}$ 를 True(1) / prediction frame $\mathcal{G}(I_1, I_2, ..., I_t) = \hat{I}_{t+1}$ 를 False(0)로 구분

$$L_{adv}^{\mathcal{D}}(\hat{I}, I) = \sum_{i,j} \frac{1}{2} L_{MSE}(\mathcal{D}(I)_{i,j}, 1) + \sum_{i,j} \frac{1}{2} L_{MSE}(\mathcal{D}(\hat{I})_{i,j}, 0) \quad (4)$$

- D를 Training 시키면서, G의 가중치를 fix하고 Mean Square Error(MSE)를 부여
- Ground truth의 Y는 {0, 1} 이며, prediction frame의 $\hat{Y} \in [0, 1]$

$$L_{MSE}(\hat{Y}, Y) = (\hat{Y} - Y)^2 \quad (5)$$

* Mean Square Error(MSE):
추정 값 또는 모델이 예측한 값과
실제 환경에서 관찰되는 값의 차이를
다룰 때 흔히 사용

## 2.4 Adversarial Training – **G(generator) Training**

- D를 True(1)로 분류하는 프레임을 생성하는 것
- G가 Training 될 때, D의 weight는 fix

$$L_{adv}^{\mathcal{G}}(\hat{I}) = \sum_{i,j} \frac{1}{2} L_{MSE}(\mathcal{D}(\hat{I})_{i,j}, 1) \quad (6)$$

# 2. Future Frame Prediction Based Anomaly Detection Method

## 2.5 Objective Function

– Appearance(intensity and gradient)와 Motion, Adversarial training의 제약 조건을 결합

$$L_{\mathcal{G}} = \lambda_{int} L_{int}(\hat{I}_{t+1}, I_{t+1}) + \lambda_{gd} L_{gd}(\hat{I}_{t+1}, I_{t+1}) + \lambda_{op} L_{op}(\hat{I}_{t+1}, I_{t+1}, I_t) + \lambda_{adv} L^{\mathcal{G}}_{adv}(\hat{I}_{t+1}) \quad (7)$$

– D Training 시, loss function

$$L_{\mathcal{D}} = L^{\mathcal{D}}_{adv}(\hat{I}_{t+1}, I_{t+1}) \quad (8)$$

# 2. Future Frame Prediction Based Anomaly Detection Method

## 2.6 Anomaly Detection on Testing Data

- 가정: 정상적인 이벤트를 기준으로 prediction 할 수 있는 objective function을 만들었기 때문에, 정상적인 이벤트를 더 잘 prediction 할 수 있음
- 비정상 프레임 예측을 위해, prediction frame 과 ground truth를 비교
- MSE(Mean Square Error): RGB 색상 공간의 모든 픽셀에 대한 prediction과 ground truth 간의 L2 distance(Euclidean distance)를 계산하여 prediction된 이미지의 품질을 측정하는 방법
- 최근 연구된 이미지 품질 평가에 더 좋은 PSNR(Peak Signal to Noise Ratio)를 사용

# 3. Experiments

Table 1. AUC of different methods on the Avenue, Ped1, Ped2 and ShanghaiTech datasets. All methods are listed by the published year.

| | CUHK Avenue | UCSD Ped1 | UCSD Ped2 | ShanghaiTech |
|---|---|---|---|---|
| MPPCA [18] | N/A | 59.0% | 69.3% | N/A |
| MPPC+SFA [25] | N/A | 66.8% | 61.3% | N/A |
| MDT [25] | N/A | 81.8% | 82.9% | N/A |
| Del *et al.* [11] | 78.3% | N/A | N/A | N/A |
| Conv-AE [14] | 80.0% | 75.0% | 85.0% | 60.9% |
| ConvLSTM-AE [23] | 77.0% | 75.5% | 88.1% | N/A |
| GrowingGas [34] | N/A | 93.8% | 94.1% | N/A |
| AbnormalGAN [29] | N/A | **97.4%** | 93.5% | N/A |
| DeepAppearance [33] | 84.6% | N/A | N/A | N/A |
| Hinami *et al.* [15] | N/A | N/A | 92.2% | N/A |
| Unmasking [16] | 80.6% | 68.4% | 82.2% | N/A |
| Stacked RNN [24] | 81.7% | N/A | 92.2% | 68.0% |
| **Our proposed method** | **85.1%** | 83.1% | **95.4%** | **72.8%** |

Table 2. The gap ($\Delta_s$) and AUC of different prediction networks in the Ped1 and Ped2 datasets.

| | Ped1 | | Ped2 | |
|---|---|---|---|---|
| | $\Delta_s$ | AUC | $\Delta_s$ | AUC |
| Beyond-MSE | 0.200 | 75.8% | 0.396 | 88.5% |
| U-Net | **0.243** | **81.8%** | **0.435** | **93.5%** |

Table 3. AUC for anomaly detection of networks with/wo the motion constraint in Ped1 and Ped2.

| | Ped1 | Ped2 |
|---|---|---|
| without motion constraint | 81.8% | 93.5% |
| with motion constraint | **83.1%** | **95.4%** |

# 3. Experiments

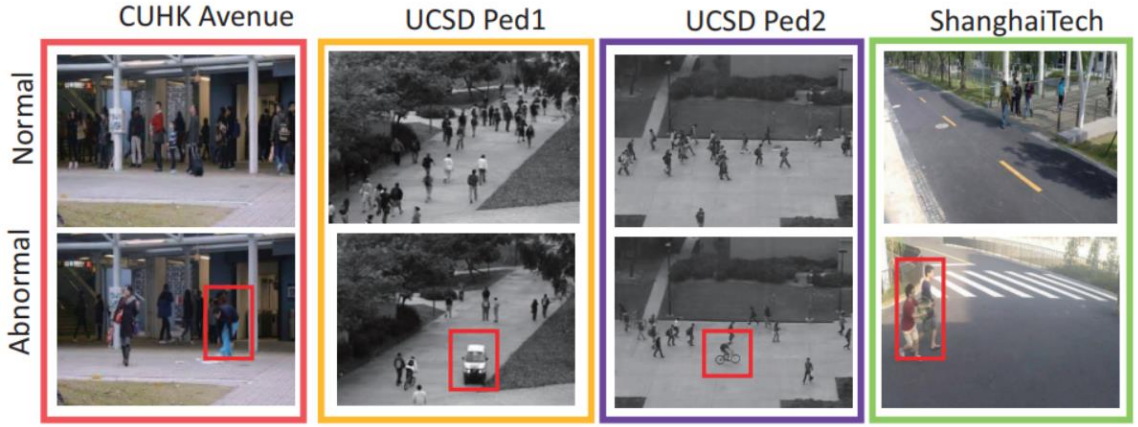



Figure 4. Some samples including normal and abnormal frames in the UCSD, CUHK Avenue and ShanghaiTech datasets are illustrated. Red boxes denote anomalies in abnormal frames.

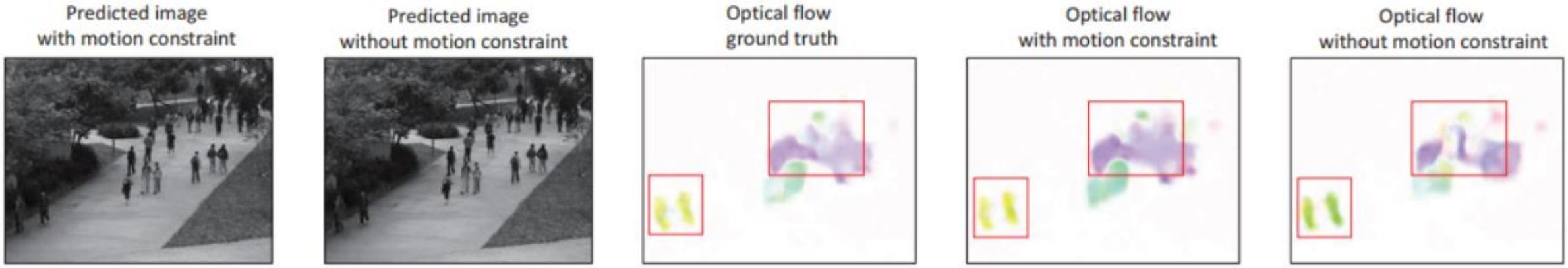
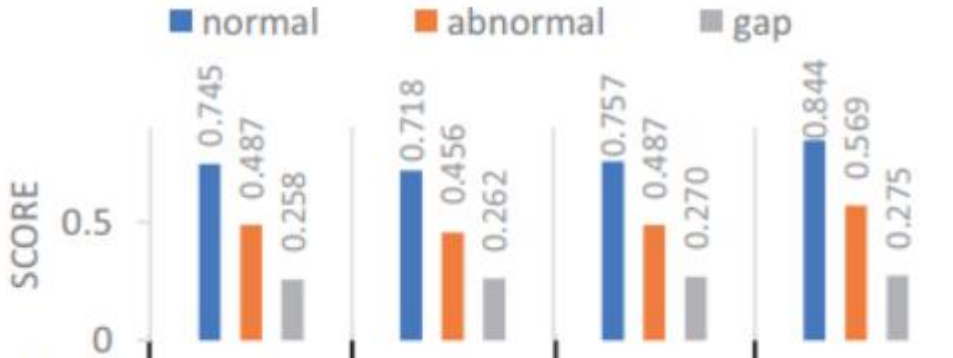



Figure 5. The visualization of optical flow and the predicted images on the Ped1 dataset. The red boxes represent the difference of optical flow predicted by the model with/without motion constraint. We can see that the optical flow predicted by the model with motion constraint is closer to ground truth. Best viewed in color.

# 3. Experiments




| | | | | |
|---|---|---|---|---|
| $L_{int}$ | ✓ | ✓ | ✓ | ✓ |
| $L_{gd}$ | – | ✓ | ✓ | ✓ |
| $L_{adv}$ | – | – | ✓ | ✓ |
| $L_{op}$ | – | – | – | ✓ |
| *AUC* | 82.0% | 82.6% | 83.7% | 85.1% |

Figure 6. The evaluation of different components in our future frame prediction network in the Avenue dataset. Each column in the histogram corresponds to a method with different loss functions. We calculate the average scores of normal and abnormal events in the testing set. The gap is calculated by subtracting the abnormal score from the normal one.

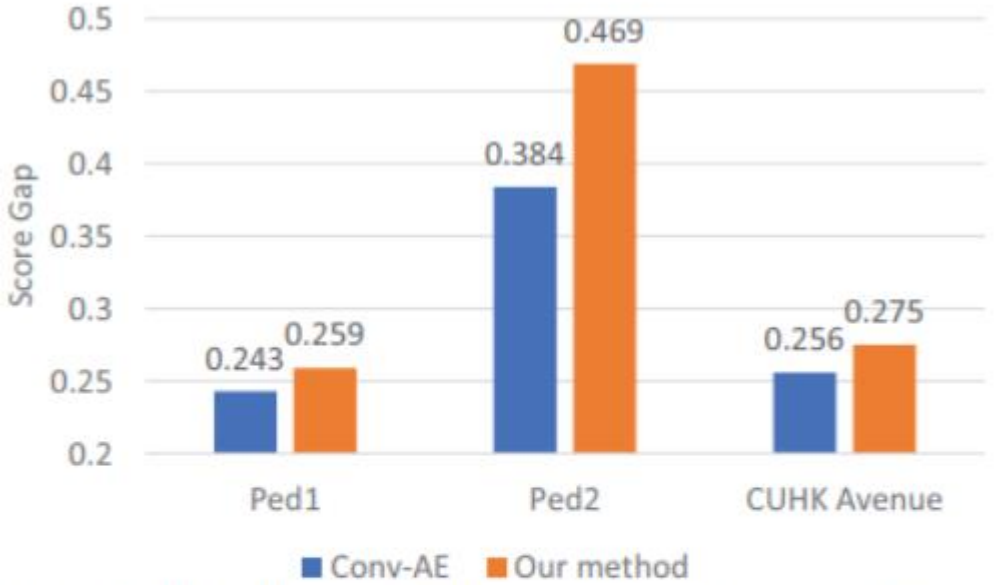



Figure 7. We firstly compute the average score for normal frames and that for abnormal frames in the testing set of the Ped1, Ped2 and Avenue datasets. Then, we calculate the difference of these two scores($\Delta_s$) to measure the ability of our method and Conv-AE to discriminate normal and abnormal frames. A larger gap($\Delta_s$) 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.

# 3. Experiments

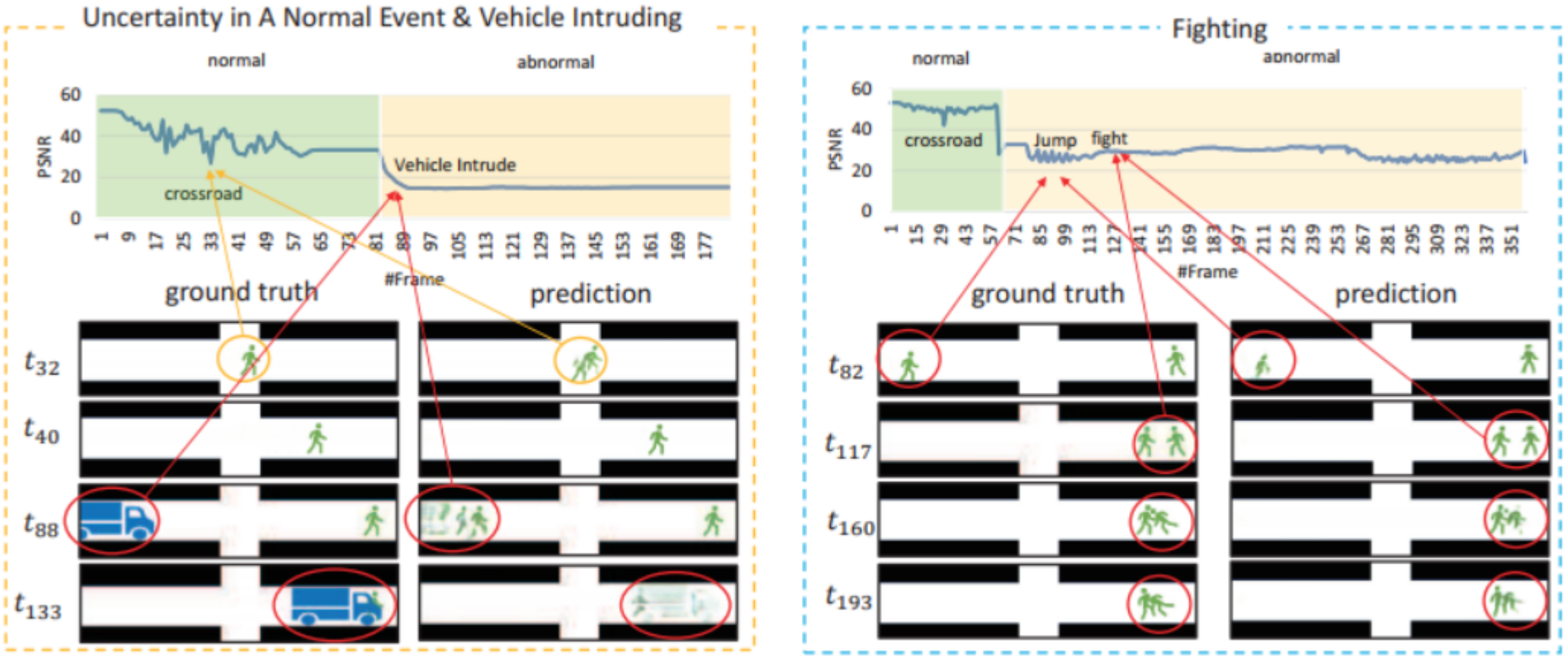



Figure 8. The visualization of predicted testing frames in our toy pedestrian dataset. There are two abnormal cases including vehicle intruding(left column) and humans fighting(right column). The orange circles correspond to normal events with uncertainty in prediction while the red ones correspond to 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.

# 4. Conclusion

- 정상적인 이벤트는 예측 가능하지만 비정상적인 이벤트는 예측과 일치하지 않기 때문에, 이상 감지를 위한 미래 프레임 예측 네트워크를 제안
- U-Net을 기본으로 사용
- 정상 이벤트를 잘 예측할 수 있는 prediction framework를 생성하기 위해, GAN을 사용하고 appearance로 프레임의 공간 정보를 제약하며 추가로 예측된 미래 프레임과 ground truth 사이의 Optical Flow를 사용함으로써 시간 정보를 제약하여 정상 이벤트를 보다 잘 예측 할 수 있음

# Thank you!