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

Zhian Liu¹  Yongwei Nie¹*  Chengjiang Long²  Qing Zhang³  Guiqing Li¹

¹South China University of Technology, ²JD Finance America Corporation, ³Sun Yat-sen University
Video Anomaly Detection

• Motivation
  • Surveillance cameras are widely used.
  • VAD is an essential task to save human labor.
Video Anomaly Detection

• Goal: to identify unexpected behaviours in a video.

Video Anomaly Detection

• Goal: to identify unexpected behaviours in a video.

• Useful but challenging task.
Video Anomaly Detection

• Challenges
  • Anomaly rarely happens.
  • What is anomaly?

• Solution
Related work

- Reconstruction-based method
  - Train AE with L1 or L2 loss.
  - Assume the anomalies lead to larger reconstruction errors.
Related work

• Reconstruction-based method
  • Memory-augmented AE to mitigate the "over-generalization" problem.

\[ \hat{z} = wM = \sum_{i=1}^{N} w_i m_i \]
\[ w_i = \frac{\exp(d(z, m_i))}{\sum_{j=1}^{N} \exp(d(z, m_j))} \]
Related work

- Prediction-based method
  - Take the temporal information into consideration [Liu. et al, 2018].

\[
\mathcal{L} = \left\| \hat{I}_{t+1} - I_{t+1} \right\|_2^2
\]

[Future Frame Pred.] W. Liu et.al, CVPR, 2018
Our approach

• Insight
  • Previous work rarely exploits the consistency between flows and frames.
  • For an abnormal event, what if we manipulate the flows beforehand, and try to produce a poor prediction?
  • Propose to reconstruct the flows first, then using the reconstructed flows as condition to predict future frame.
HF$^2$-VAD pipeline

A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction
ML-MemAE-SC

- Observations
  - Memory only in bottleneck cannot remember all normal patterns.
  - AE with multi-level memories (ML-MemAE) leads to degradation.
  - Skip connection helps.

(a) MemAE  
(b) ML-MemAE  
(c) ML-MemAE-SC
ML-MemAE-SC

- Flow reconstruction objective

\[ \mathcal{L}_{\text{recon}} = \| y_{1:t} - \hat{y}_{1:t} \|_2 \]

\[ \mathcal{L}_{\text{ent}} = \sum_{i=1}^{M} \sum_{k=1}^{N} -\hat{w}_{i,k} \log(\hat{w}_{i,k}) \]

\[ \mathcal{L}_{\text{ML-MemAE-SC}} = \lambda_{\text{recon}} \mathcal{L}_{\text{recon}} + \lambda_{\text{ent}} \mathcal{L}_{\text{ent}} \]
CVE for prediction

- Formulation
  - Let $x_{1:t}$ and $x_{t+1}$ be the previous and future frame, $y_{1:t}$ be the reconstructed flows, $z$ be the hidden variables that control the content information:

$$
\log p(x_{t+1} \mid y_{1:t}) \geq \mathbb{E}_q \log \frac{p(x_{t+1} \mid z, y_{1:t})p(z \mid y_{1:t})}{q(z \mid x_{t+1}, y_{1:t})} \quad \text{(Evidence Lower Bound)}
$$

$$
\approx \mathbb{E}_q \log \frac{p(x_{t+1} \mid z, y_{1:t})p(z \mid y_{1:t})}{q(z \mid x_{1:t}, y_{1:t})} \quad \text{(Short Duration Assumption)}
$$

$$
= -KL[q(z \mid x_{1:t}, y_{1:t}) \| p(z \mid y_{1:t})] + \mathbb{E}_q [\log p(x_{t+1} \mid z, y_{1:t})]
$$

- Resort the conditional Variational Autoencoder (CVAE).
CVE for prediction

- Frame prediction objective

\[ \mathcal{L}_{CVAE} = KL[q(z | x_{1:t} | y_{1:t}) || p(z | y_{1:t})] + \| x_{t+1} - \hat{x}_{t+1} \|_2^2 \]

\[ \mathcal{L}_{gd}(X, \hat{X}) = \sum_{i,j} |X_{i,j} - X_{i-1,j}| - |\hat{X}_{i,j} - \hat{X}_{i-1,j}| \\
|X_{i,j} - X_{i,j-1}| - |\hat{X}_{i,j} - \hat{X}_{i,j-1}| \]

\[ \mathcal{L} = \lambda_{CVAE}\mathcal{L}_{CVAE} + \lambda_{gd}\mathcal{L}_{gd}(\hat{x}_{t+1}, x_{t+1}) \]
Anomaly detecting

- At test time, the anomaly score is composed of two parts:
  - Reconstruction error $S_r = \|\hat{y}_{1:t} - y_{1:t}\|_2^2$
  - Prediction error $S_p = \|\hat{x}_{t+1} - x_{t+1}\|_2^2$

- Frame-level anomaly score

$$S_{O_i} = w_r \cdot S_r + w_p \cdot S_p \quad S = max\{S_{O_1}, S_{O_2}, \ldots, S_{O_N}\}$$
Anomaly detecting

- At test time, the anomaly score is composed of two parts:
  - Reconstruction error $S_r = \| \hat{y}_{1:t} - y_{1:t} \|^2_2$
  - Prediction error $S_p = \| \hat{x}_{t+1} - x_{t+1} \|^2_2$

- Frame-level anomaly score

$$S_{O_i} = w_r \cdot S_r + w_p \cdot S_p \quad S = \max\{S_{O_1}, S_{O_2}, \ldots, S_{O_N}\}$$
Experimental results

• Datasets

a) UCSD Ped2

b) CUHK Avenue

c) ShanghaiTech

• Quantitative results

<table>
<thead>
<tr>
<th>Method</th>
<th>UCSD Ped2</th>
<th>CUHK Avenue</th>
<th>SHTech</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv-AE [11]</td>
<td>90.0</td>
<td>70.2</td>
<td>-</td>
</tr>
<tr>
<td>ConvLSTM-AE [32]</td>
<td>88.1</td>
<td>77.0</td>
<td>-</td>
</tr>
<tr>
<td>GMFC-VAE [7]</td>
<td>92.2</td>
<td>83.4</td>
<td>-</td>
</tr>
<tr>
<td>MemAE [8]</td>
<td>94.1</td>
<td>83.3</td>
<td>71.2</td>
</tr>
<tr>
<td>MNAD-R [39]</td>
<td>90.2</td>
<td>82.8</td>
<td>69.8</td>
</tr>
<tr>
<td>Frame-Pred. [26]</td>
<td>95.4</td>
<td>85.1</td>
<td>72.8</td>
</tr>
<tr>
<td>Conv-VRNN [31]</td>
<td>96.1</td>
<td>85.8</td>
<td>-</td>
</tr>
<tr>
<td>MNAD-P [39]</td>
<td>97.0</td>
<td>88.5</td>
<td>70.5</td>
</tr>
<tr>
<td>VEC [50]</td>
<td>97.3</td>
<td>90.2</td>
<td>74.8</td>
</tr>
<tr>
<td>ST-AE [53]</td>
<td>91.2</td>
<td>80.9</td>
<td>-</td>
</tr>
<tr>
<td>AMC [37]</td>
<td>96.2</td>
<td>86.9</td>
<td>-</td>
</tr>
<tr>
<td>AnoPCN [49]</td>
<td>96.8</td>
<td>86.2</td>
<td>73.6</td>
</tr>
<tr>
<td>HF²-VAD w/o FP</td>
<td>98.8</td>
<td>86.8</td>
<td>73.1</td>
</tr>
<tr>
<td>HF²-VAD w/o FR</td>
<td>94.5</td>
<td>90.2</td>
<td>76.0</td>
</tr>
<tr>
<td>HF²-VAD</td>
<td>99.3</td>
<td>91.1</td>
<td>76.2</td>
</tr>
</tbody>
</table>
Experimental results

• Qualitative results

(a) Skateboarding and riding bicycle of Ped2.

(b) Kid running of Avenue.
A Hybrid Video Anomaly Detection Framework via Memory-Augmented Flow Reconstruction and Flow-Guided Frame Prediction

Table 2. Ablation study results on UCSD Ped2 [35] dataset. The anomaly detection performance is reported in terms of AUROC ↑ (%). Number in bold indicates the best result.

<table>
<thead>
<tr>
<th>Memory-augmented Reconstruction Models</th>
<th>Prediction Models</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>VEC</td>
<td>VAE</td>
<td>96.27</td>
</tr>
<tr>
<td>MNAD-P</td>
<td>CVAE</td>
<td>97.75</td>
</tr>
<tr>
<td>Flow</td>
<td></td>
<td>98.81</td>
</tr>
<tr>
<td>Frame</td>
<td></td>
<td>89.96</td>
</tr>
<tr>
<td>Hybrid</td>
<td></td>
<td>94.48</td>
</tr>
</tbody>
</table>

[VEC] G. Yu et.al, ACM-MM, 2020
[MNAD-P] H Park et.al, CVPR, 2020
Video anomaly detection demo

- On Ped2 dataset

Abnormal events: unusual lorry and bicycle.
Video anomaly detection demo

• On Avenue dataset

Abnormal event: kid running.
Video anomaly detection demo

• On ShanghaiTech dataset

ShanghaiTech Test Video 04_0001

Abnormal events: chasing and jumping.
Conclusion

- Design the Multi-Level Memory Autoencoder with Skip Connections (ML-MemAE-SC) for flow reconstruction.
- Propose to model the consistency between flows and frames by leveraging the conditional Variational Autoencoder (CVAE).
- Design a novel *hybrid* framework in a combination of *flow* reconstruction and flow-guided *frame* prediction, named as $HF^2$-VAD.
Project QR Code

https://github.com/LiUzHiAn/hf2vad

Thank you!