*STAN: Spatio-Temporal Adversarial Networks for Abnormal Event Detection*

Generator: Encoder - ConvLSTM - Combined ConvLSTM – Decoder. Discriminator: 3D CNN.

Generator consists of three network modules: a spatial encoder for encoding spatial features of frames, a bidirectional ConvLSTM for encoding temporal feature of the scene, and a spatial decoder for generating an inter-frame. ConvLSTM consists of forward, backward and combined parts.

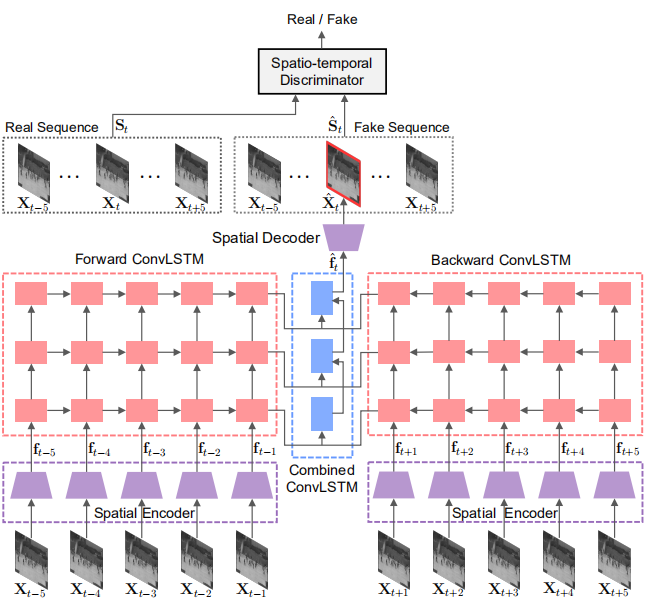
In training steps, the input is , generator tries to generate the inter-frame . Fake sequence: consecutive frames including generated inter-frame . Real sequence (Label): consecutive frames with original inter-frame .

First, previous 5 frames are fed into encoder and forward ConvLSTM, and the later 5 frames are fed into encoder and backward ConvLSTM.

Next, the hidden states and cell states getting from two ConvLSTM modules are concatenated and fed into combined ConvLSTM.

Finally, using spatial decoder to get the generated inter-frame .

At testing time, abnormal events are detected by using a pixel-wise loss for the generator and an adversarial loss for the discriminator.



To quantify detected abnormalities, they devise a novel abnormality score by using the losses of the generator and the discriminator. The proposed abnormality loss can be defined as:

Where is a hyper-parameter to balance the generator detection and the discriminator detection.

Then, by normalizing , abnormality score at *t-th* frame can be written as:

Note that video sequences containing abnormal events have higher abnormality scores.