Identifying anomalies in real-world scenarios is a challenging task that cannot solely rely on action-based knowledge.

To effectively recognize such complex actions, it becomes crucial to consider the objects involved and their interrelationships within the contextual scenes.

We propose VADOR, a method understands complex scenes through the integration of action information and object relations.

VADOR employs a two-stage approach, Video Clip Encoders (VCE) in first stage generates video clip features, TALNet in second stage localizes anomalies in time by using sequentially organized video clip features.

VCE involves the object relation encoder and the action encoder.
- The object relation encoder processes the object features and bounding boxes.
- The action encoder handles the action features.

Cross-attention layers between the encoders enable cross-relations between objects and actions within the same video clip.

TALNet consists of 1D convolution layers, temporal BiFPN blocks and dense prediction heads. The model is similar to dense object detection methods.

Fusion of action and object relation information increases performance of VADOR.

Qualitative and quantitative results show that transformer encoders with cross attention layers provides better temporal anomaly segmentation performance.

The results show that encoders with cross attention is important to get better performance. Furthermore, the results show that object relations are more useful than action to recognize anomalies in UCF Crime dataset.