Video Object Detection

A pilot study

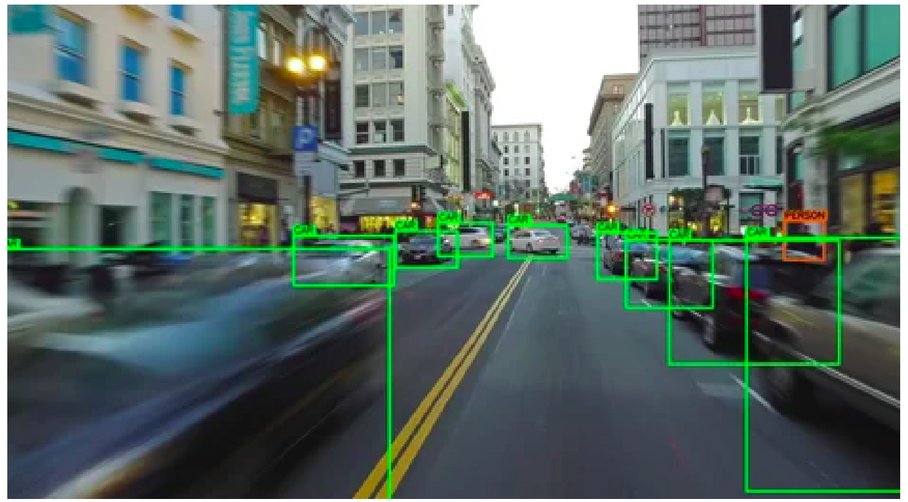
Hello everyone.

Abhishek_Rajora
brillard1

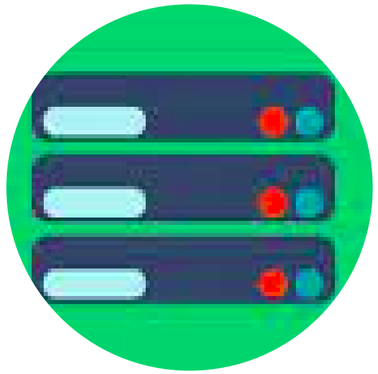
Abu_Shahid
ceyxasm

Introduction

- Video object detection is a vital task in computer vision.
- Requires considering temporal information and tracking objects across frames.
- Challenges include motion blur, occlusions, lighting variations, and complex object interactions.



Meet The Study



DATASET

LITERATURE REVIEW

IMPLEMENTATION

ANALYSIS

Dataset Used

ImageNet-VidVRD Video Visual Relation Dataset

- IMAGENET-VIDVRD CONTAINS 1,000 VIDEOS FROM **ILVSRC2016-VID**
- SPLIT INTO 800 TRAINING SET AND 200 TEST SET
- COVERS 35 CATEGORIES OF COMMON SUBJECTS/OBJECTS
- INCLUDES 132 CATEGORIES OF PREDICATES
- OBJECT TRAJECTORY AND RELATION LABELING
- SUPPLEMENTING ANNOTATIONS FOR 5 CATEGORIES
- LABOR-SAVING APPROACH FOR RELATION LABELING
- TRAINING SET: LABELED TYPICAL SEGMENTS
- TEST SET: ENTIRE VIDEO LABELING
- DATASET STATISTICS PROVIDED

	training set	test set
video	800	200
subject/object category	35	35
predicate category	132	132
relation triplet	2,961	1,011
visual relation instance	–	4,835

Literature Review

- Mask R-CNN
- RetinaNet- Focal Loss for Dense Object Detection
- YOLOV: Making Still Image Object Detectors Great at Video Object Detection
- TransVOD-
- VSTAM- Video Sparse Transformer With Attention-Guided Memory for Video Object Detection
- BoxMask
- Temporal ROI Align- for Video Object Recognition

Mask R-CNN

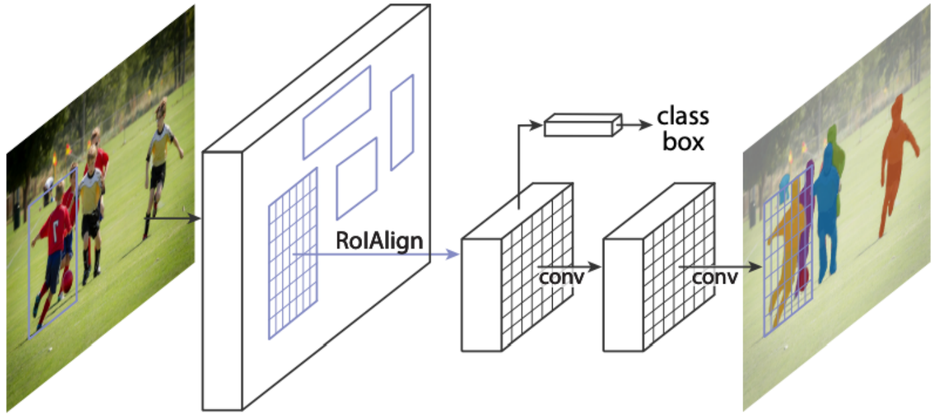


Figure 1. The **Mask R-CNN** framework for instance segmentation.

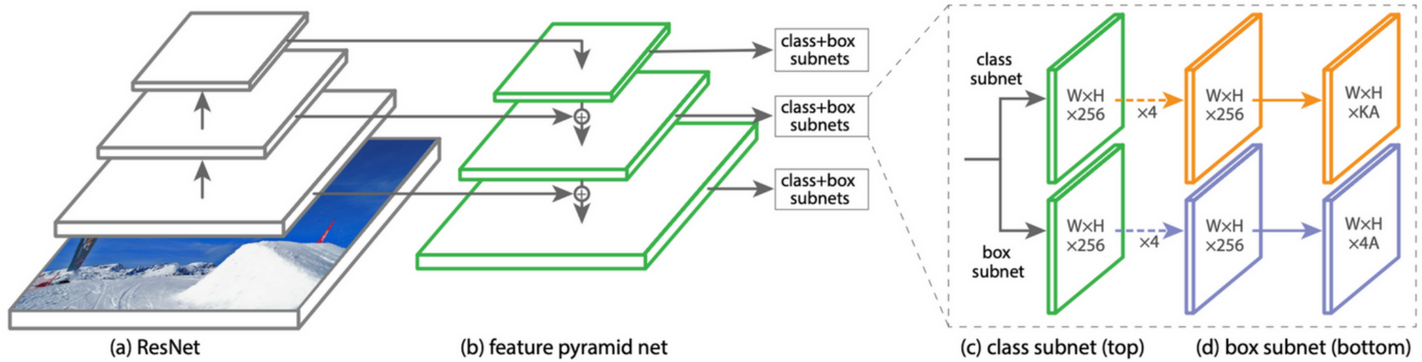
BACKBONE ARCHITECTURE

Mask R-CNN is a meta algorithm which is designed to simple and intuitive.

Utilizes Fast R-CNN and FCN with parallel heads to maintain a mask object at each frame

Uses RoI Align to preserve spatial location

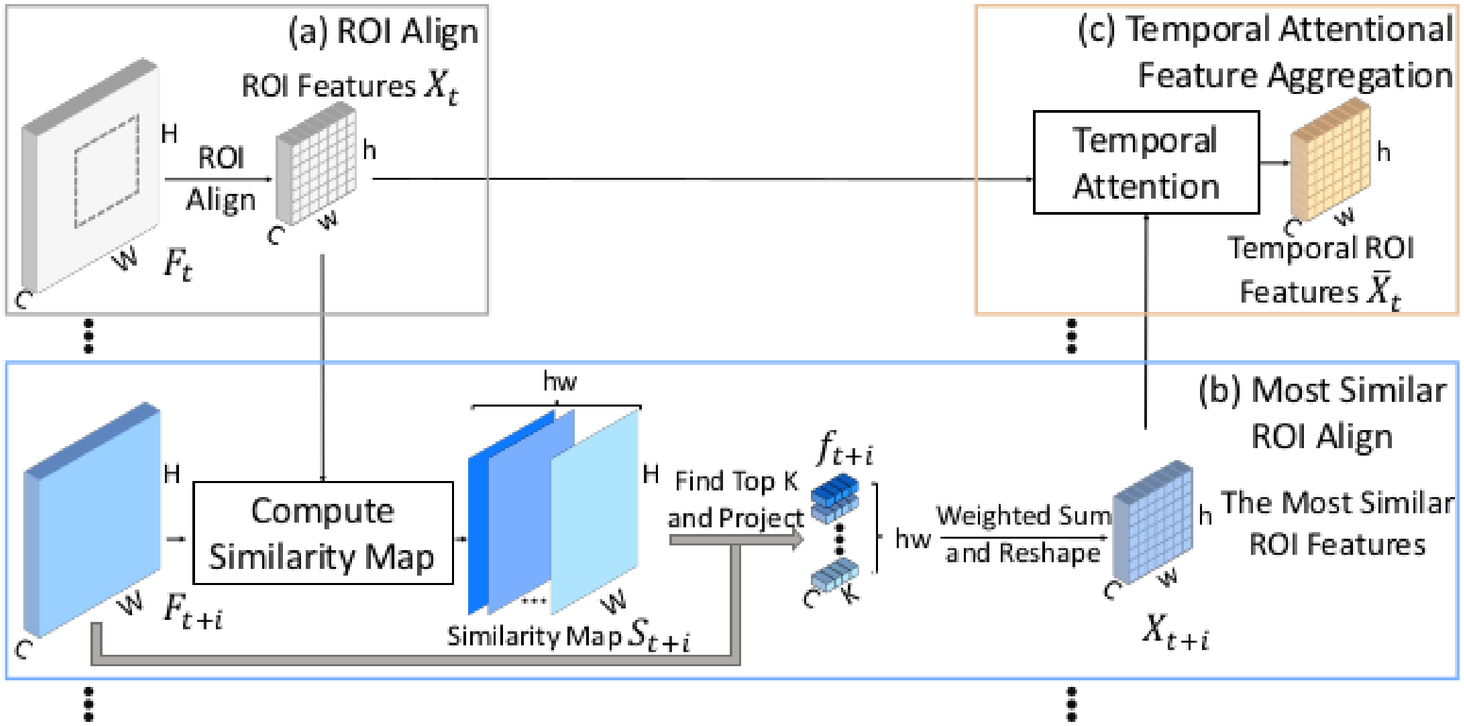
Retina-net



Backbone - Feature Pyramid Network (FPN) on top of a feedforward ResNet architecture

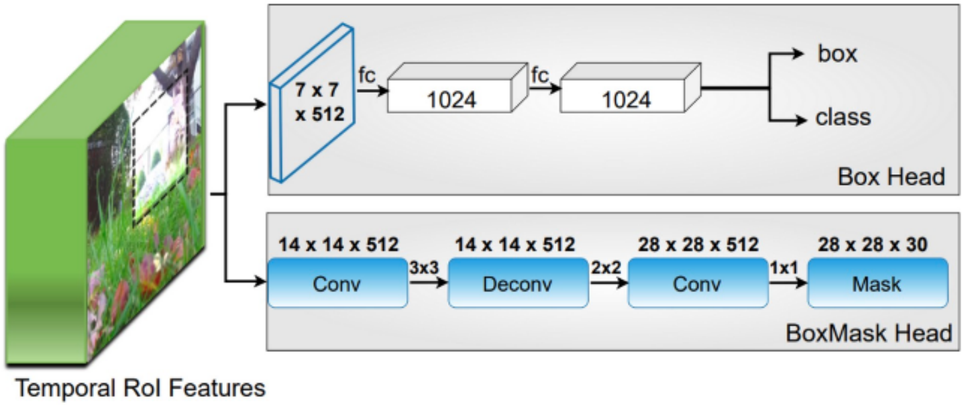
Introduces a novel focal loss function to address the class imbalance problem in dense object detection

Temporal ROI Align



An innovative Temporal ROI Align operator is introduced as a substitute for the traditional ROI Align operator in videos. This operator enables the extraction of temporal information from a complete video of any length for a given proposal.

Box-Mask



GENERAL OVERVIEW

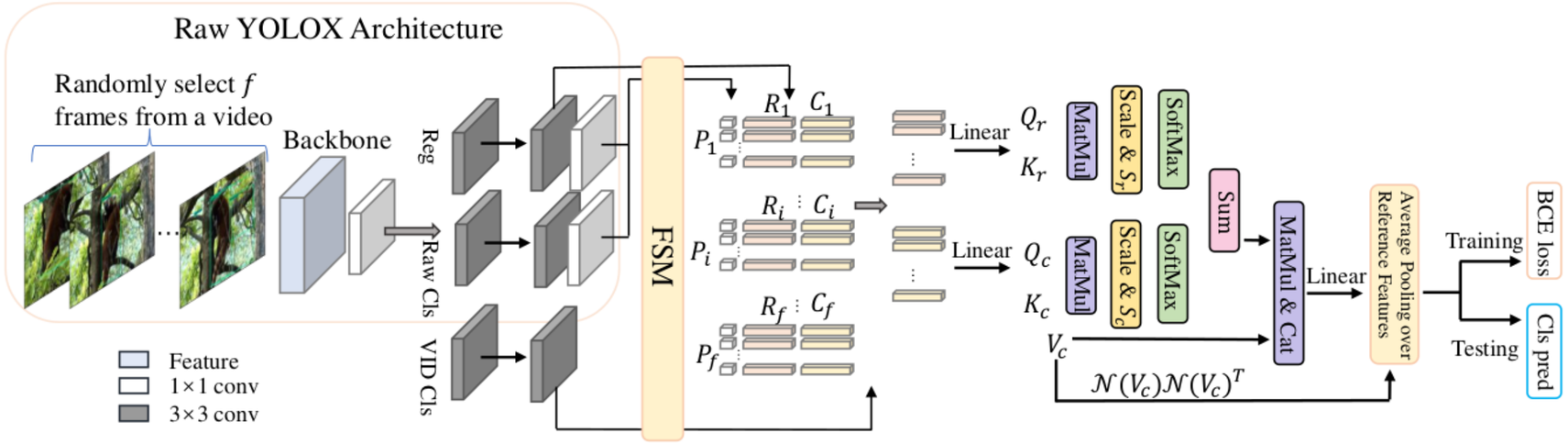
Incorporating key ideas from the two-stage detection model Mask R-CNN

In detection phase equipped with a BoxMask head at the bottom

RPN generates candidate object regions

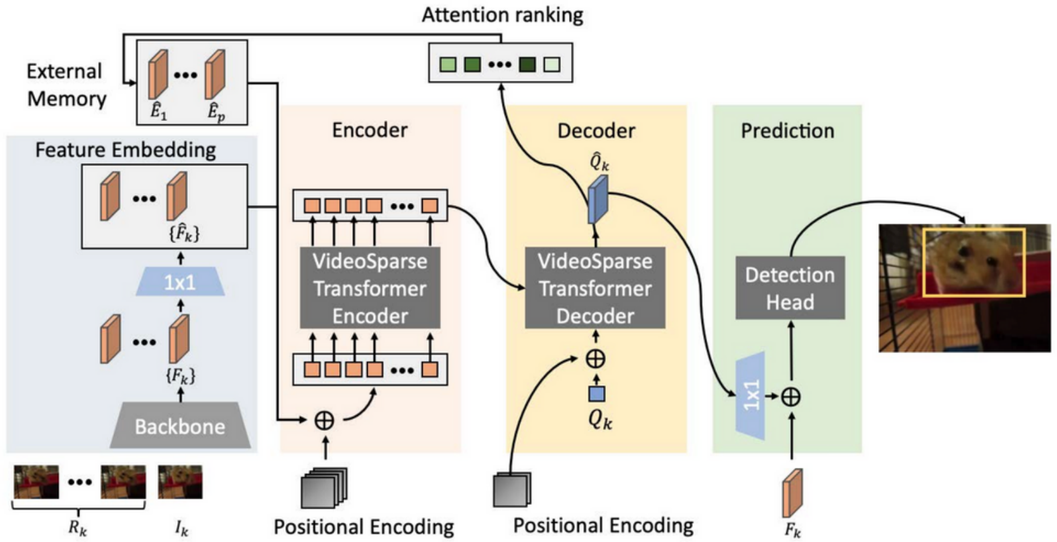
Box and mask head network classifies the object in RoI

YOLOV



In aits different approach, the initial phase involves making predictions, where numerous regions with low confidence are eliminated. The subsequent phase can be seen as refining at the region level, utilizing the aggregation of information from other frames.

VSTAM



OVERVIEW

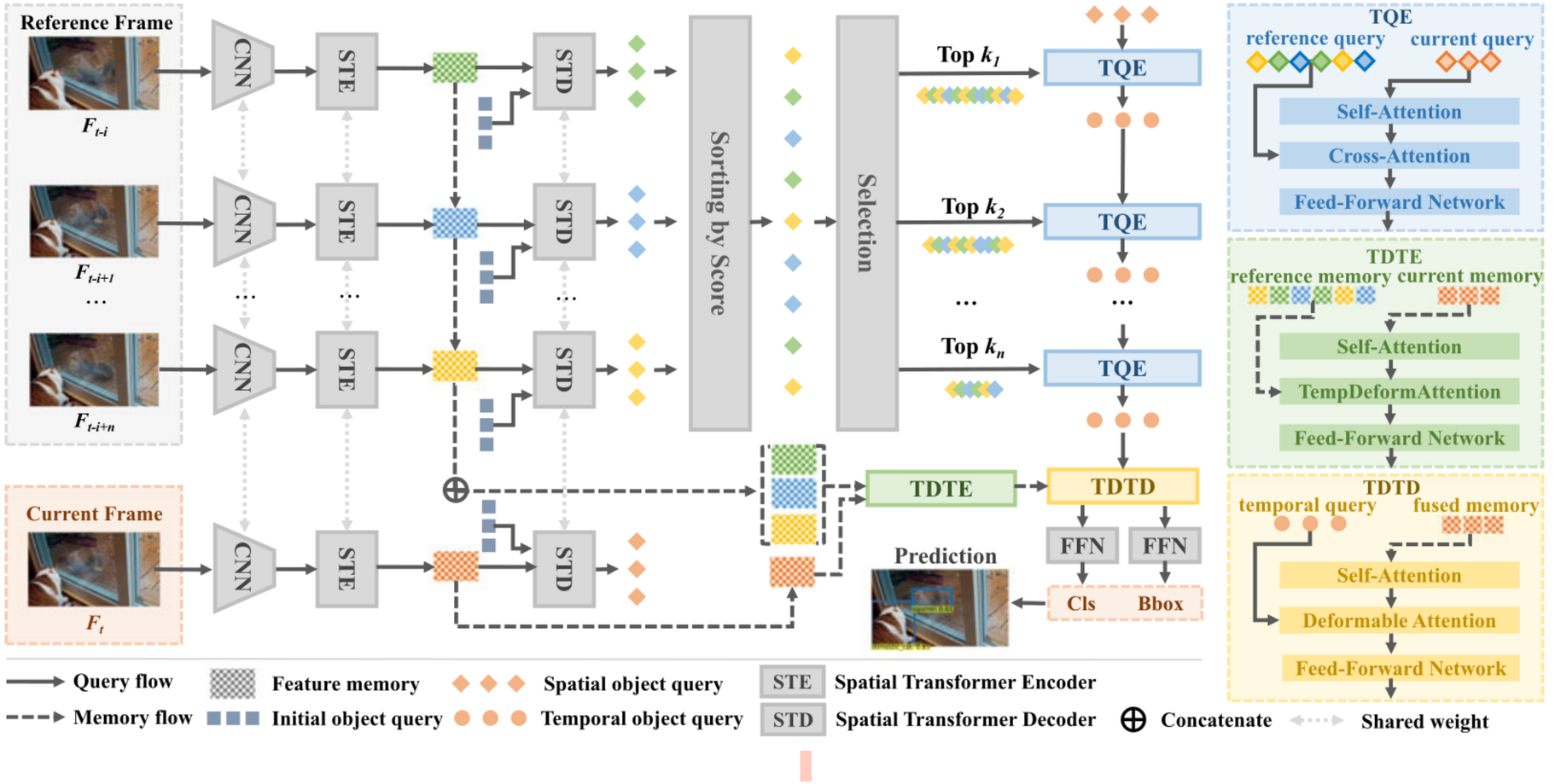
Sparse attention mechanism

Mechanism is guided by the memory module

random and positional attention mechanisms

adaptively store the features of the most vital frames in the external memory

TransVOD


Reference Frame
F_{t-i}
F_{t-i+1}
F_{t-i+n}
Current Frame
F_t
CNN
STE
STD
Sorting by Score
Selection
Top k_1
Top k_2
Top k_n
TQE
TDTE
TDTD
FFN
Cls
Bbox
Prediction
TQE
reference query
current query
Self-Attention
Cross-Attention
Feed-Forward Network
TDTE
reference memory current memory
Self-Attention
TempDeformAttention
Feed-Forward Network
TDTD
temporal query
fused memory
Self-Attention
Deformable Attention
Feed-Forward Network
Query flow
Memory flow
Feature memory
Initial object query
Spatial object query
Temporal object query
STE Spatial Transformer Encoder
STD Spatial Transformer Decoder
Concatenate
Shared weight

Metrics

- INTERSECTION OVER UNION
- MEAN AVERAGE PRECISION

MEAN AVERAGE PRECISION

- mAP (mean Average Precision) evaluates precision-recall trade-off.
- Precision measures accuracy, recall measures completeness.
- mAP calculates average precision across object classes.
- Helps compare models, tune hyperparameters, assess progress.

INTERSECTION OVER UNION

- IOU (Intersection over Union) measures overlap between predicted and ground truth bounding boxes.
- Ratio of intersection area to union area.
- IOU ranges from 0 to 1 (0 = no overlap, 1 = perfect match).

Analysis

Mask R-CNN Improvements

	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}	AP^{kp}_{M}	AP^{kp}_{L}
RoIPool	59.8	86.2	66.7	55.1	67.4
RoIAlign	**64.2**	**86.6**	**69.7**	**58.7**	**73.0**

Table 6. **RoIAlign *vs.* RoIPool** for keypoint detection on `minival`. The backbone is ResNet-50-FPN.

Analysis

Focal Loss with varying alpha and gamma values

TABLE II. mAP analysis by varying α and γ weights for focal loss in ResNet architecture.

α	γ	**AP (R-50)**
0	0.75	49.4
0.1	0.75	49.9
0.2	0.75	50.7
0.5	0.5	51.7
1	0.25	52.0
2	0.25	52.5
5	0.25	49.6

Analysis

BoxMask + RetinaNet Object Detection Results

Let's switch to demo