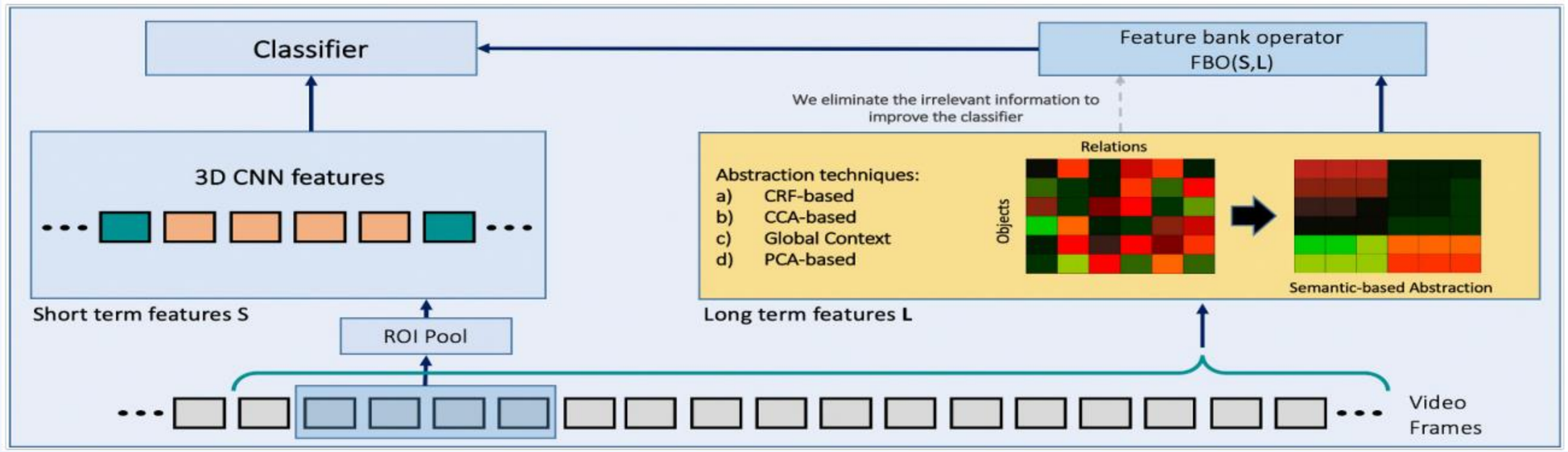
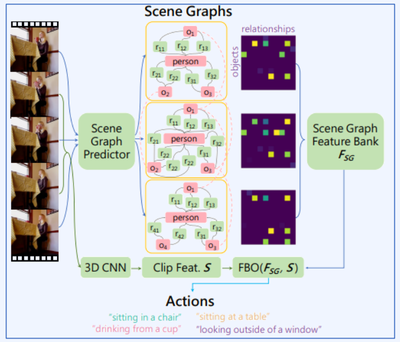
Video Understanding

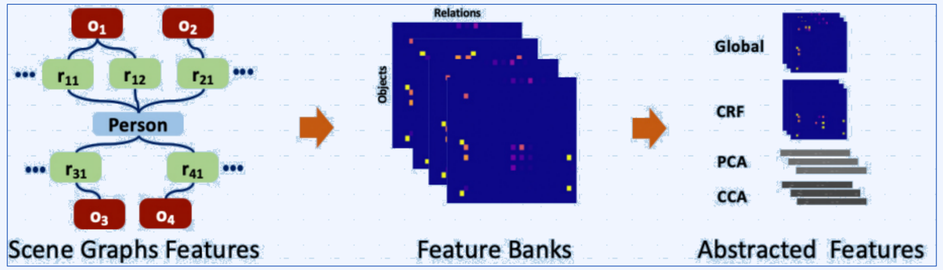
3 High-level Computer Vision for Video Data

Video Understanding
High-level Computer Vision for Video Data

A. Rahimi et al. Toward Improving The Visual Characterization of Sport Activities With Abstracted Scene Graphs. CVPR workshop



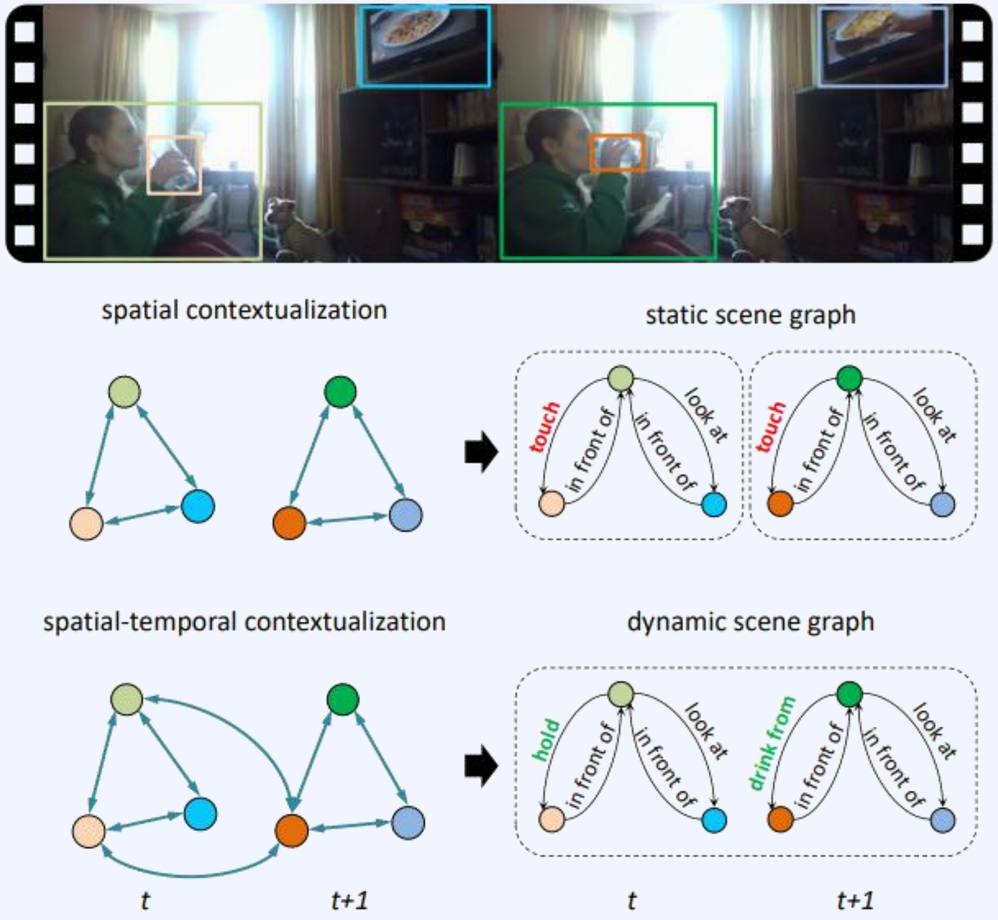




Video Understanding
High-level Computer Vision for Video Data

Y. Cong et al. Spatial-Temporal Transformer for Dynamic Scene Graph Generation. CVPR

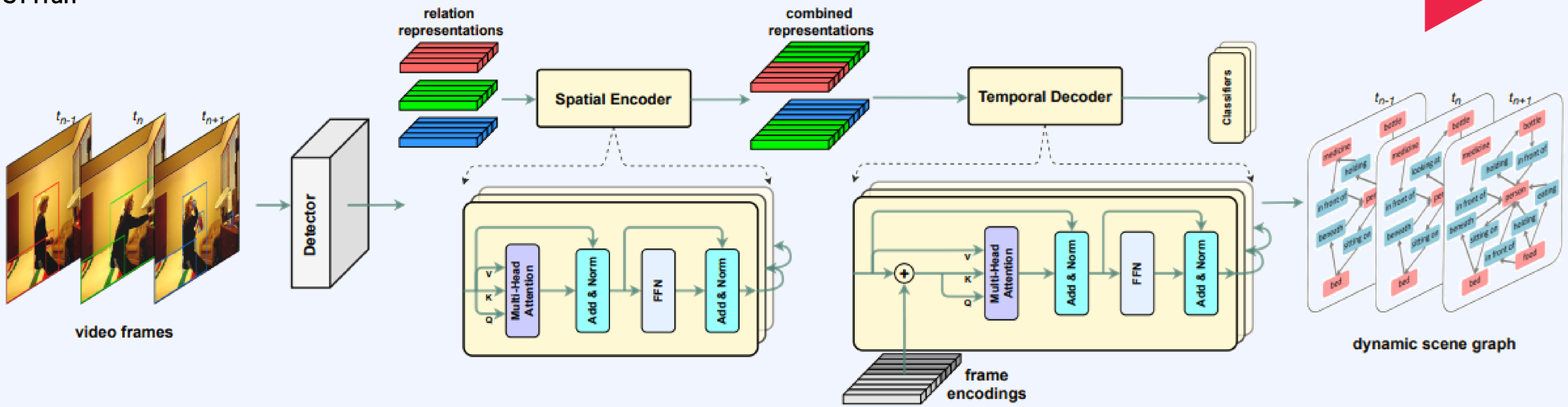
STTran



Video Understanding
High-level Computer Vision for Video Data

Y. Cong et al. Spatial-Temporal Transformer for Dynamic Scene Graph Generation. CVPR

STTran



(a) spatial encoder only (b) complete STTran

Spatial Encoder	Temporal Decoder	Frame Encoding	PredCLS-R@20 With	PredCLS-R@20 Semi	SGDET-R@20 With	SGDET-R@20 Semi
✓	-	-	69.6	78.7	32.9	35.1
-	✓	-	71.0	82.2	33.7	35.5
✓	✓	-	71.3	82.7	33.8	35.6
✓	✓	sinusoidal	71.3	82.8	33.9	35.7
✓	✓	learned	**71.8**	**83.1**	**34.1**	**35.9**

Video Understanding
High-level Computer Vision for Video Data

Y. Cong et al. Spatial-Temporal Transformer for Dynamic Scene Graph Generation. CVPR

STTran

Method	With Constraint									No Constraint								
	PredCLS			SGCLS			SGDET			PredCLS			SGCLS			SGDET		
	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50
VRD[42]	51.7	54.7	54.7	32.4	33.3	33.3	19.2	24.5	26.0	59.6	78.5	99.2	39.2	49.8	52.6	19.1	28.8	40.5
Motif Freq[65]	62.4	65.1	65.1	40.8	41.9	41.9	23.7	31.4	33.3	73.4	92.4	**99.6**	50.4	60.6	64.2	22.8	34.3	46.4
MSDN[35]	65.5	68.5	68.5	43.9	45.1	45.1	24.1	32.4	34.5	74.9	92.7	99.0	51.2	61.8	65.0	23.1	34.7	46.5
VCTREE[51]	66.0	69.3	69.3	44.1	45.3	45.3	24.4	32.6	34.7	75.5	92.9	99.3	52.4	62.0	65.1	23.9	35.3	46.8
RelDN[66]	66.3	69.5	69.5	44.3	45.4	45.4	24.5	32.8	34.9	75.7	93.0	99.0	52.9	62.4	65.1	24.1	35.4	46.8
GPS-Net[40]	66.8	69.9	69.9	45.3	46.5	46.5	24.7	33.1	35.1	76.0	93.6	99.5	53.6	63.3	66.0	24.4	35.7	47.3
STTran	**68.6**	**71.8**	**71.8**	**46.4**	**47.5**	**47.5**	**25.2**	**34.1**	**37.0**	**77.9**	**94.2**	99.1	**54.0**	**63.7**	**66.4**	**24.6**	**36.2**	**48.8**

With Constraint : <subject-object> 에 대해 하나의 predicate만 허용

No Constraint : <subject-object> 에 대해 여러 개의 predicates 허용

Video Understanding
High-level Computer Vision for Video Data

Y. Cong et al. Spatial-Temporal Transformer for Dynamic Scene Graph Generation. CVPR

STTran

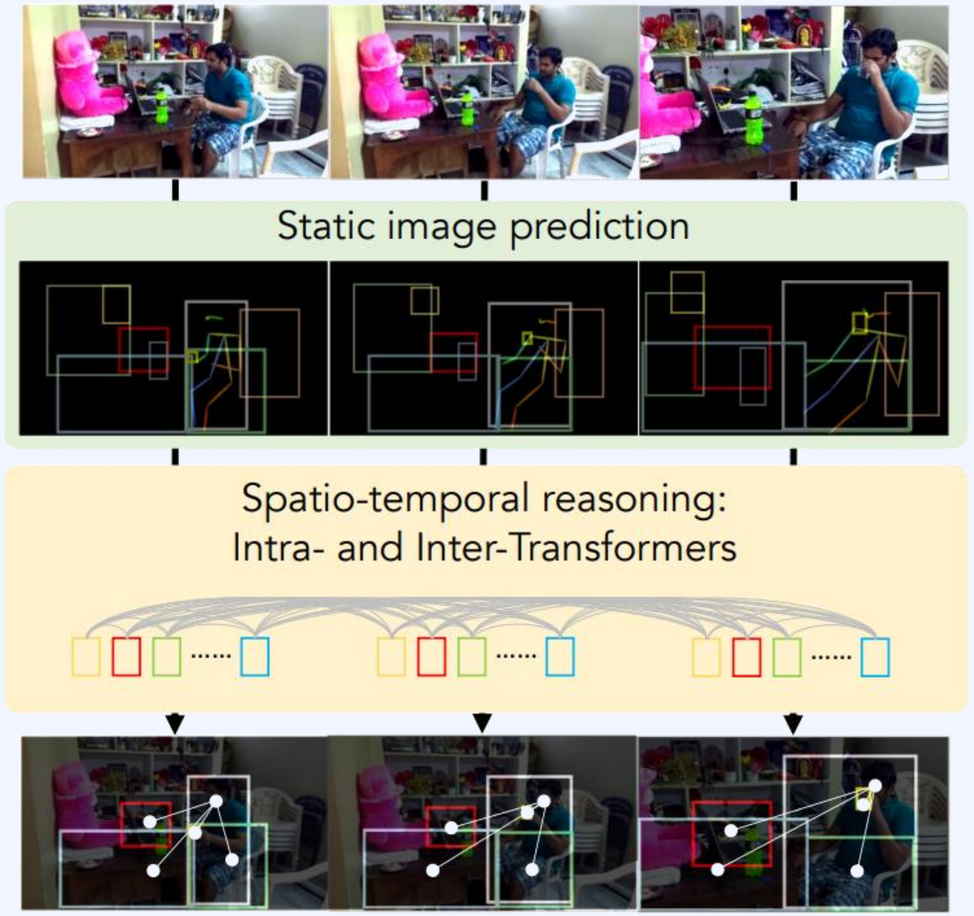
Method	Semi Constraint								
	PredCLS			SGCLS			SGDET		
	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50
VRD[42]	55.5	64.9	65.2	36.2	39.7	40.1	19.0	27.1	32.4
Motif Freq[65]	65.7	74.1	74.5	45.5	49.3	49.5	22.9	33.7	39.0
MSDN[35]	69.6	78.9	79.9	48.3	54.1	54.5	23.2	34.2	41.5
VCTREE[51]	70.1	78.2	79.6	49.0	53.7	54.0	23.7	34.8	40.4
RelDN[66]	70.7	78.8	80.3	49.4	53.9	54.1	24.1	35.0	40.7
GPS-Net[40]	71.3	81.2	82.0	50.2	55.0	55.2	24.5	35.3	41.9
STTran	**73.2**	**83.1**	**84.0**	**51.2**	**56.5**	**56.8**	**24.6**	**35.9**	**44.0**

Semi Constraint : <subject-object> 에 대해 여러 개의 predicates 허용하되 threshold 적용

Video Understanding
High-level Computer Vision for Video Data

J. Ji et al. Detecting Human-Object Relationships in Videos. ICCV

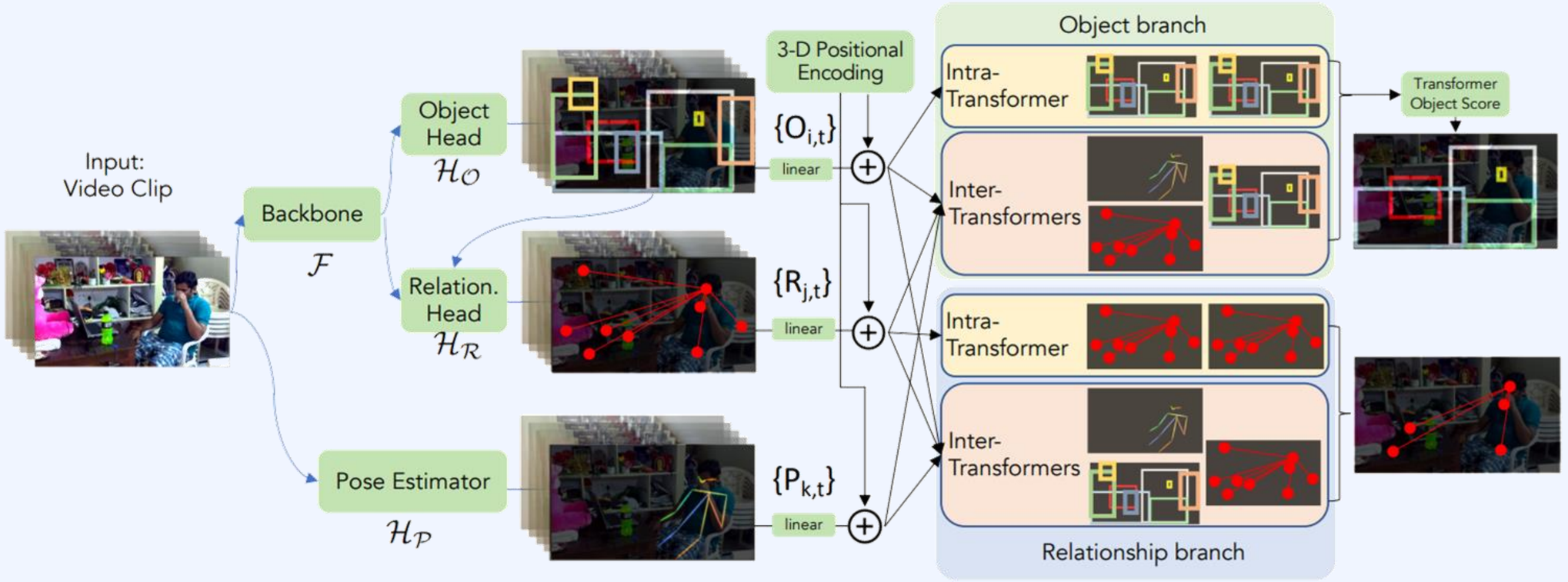
HORT



Video Understanding
High-level Computer Vision for Video Data

J. Ji et al. Detecting Human-Object Relationships in Videos. ICCV

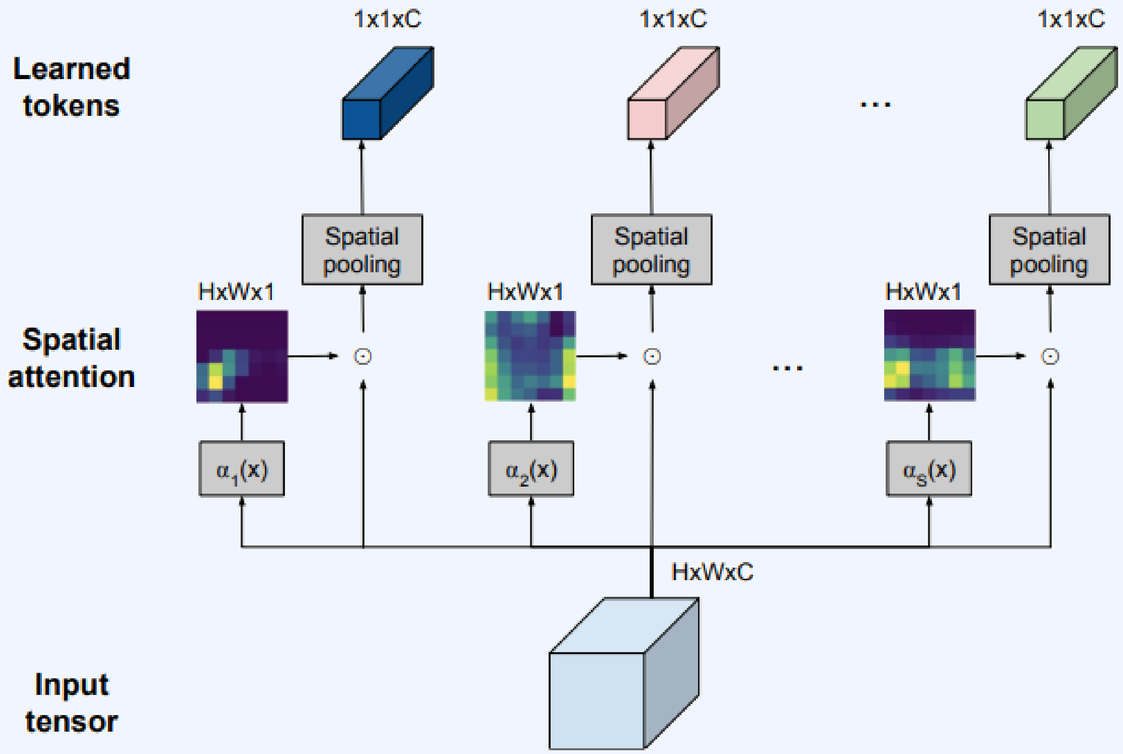
HORT



Video Understanding
High-level Computer Vision for Video Data

M. Ryoo et al. TokenLearner: Adaptive Space-Time Tokenization for Videos. NeurIPS

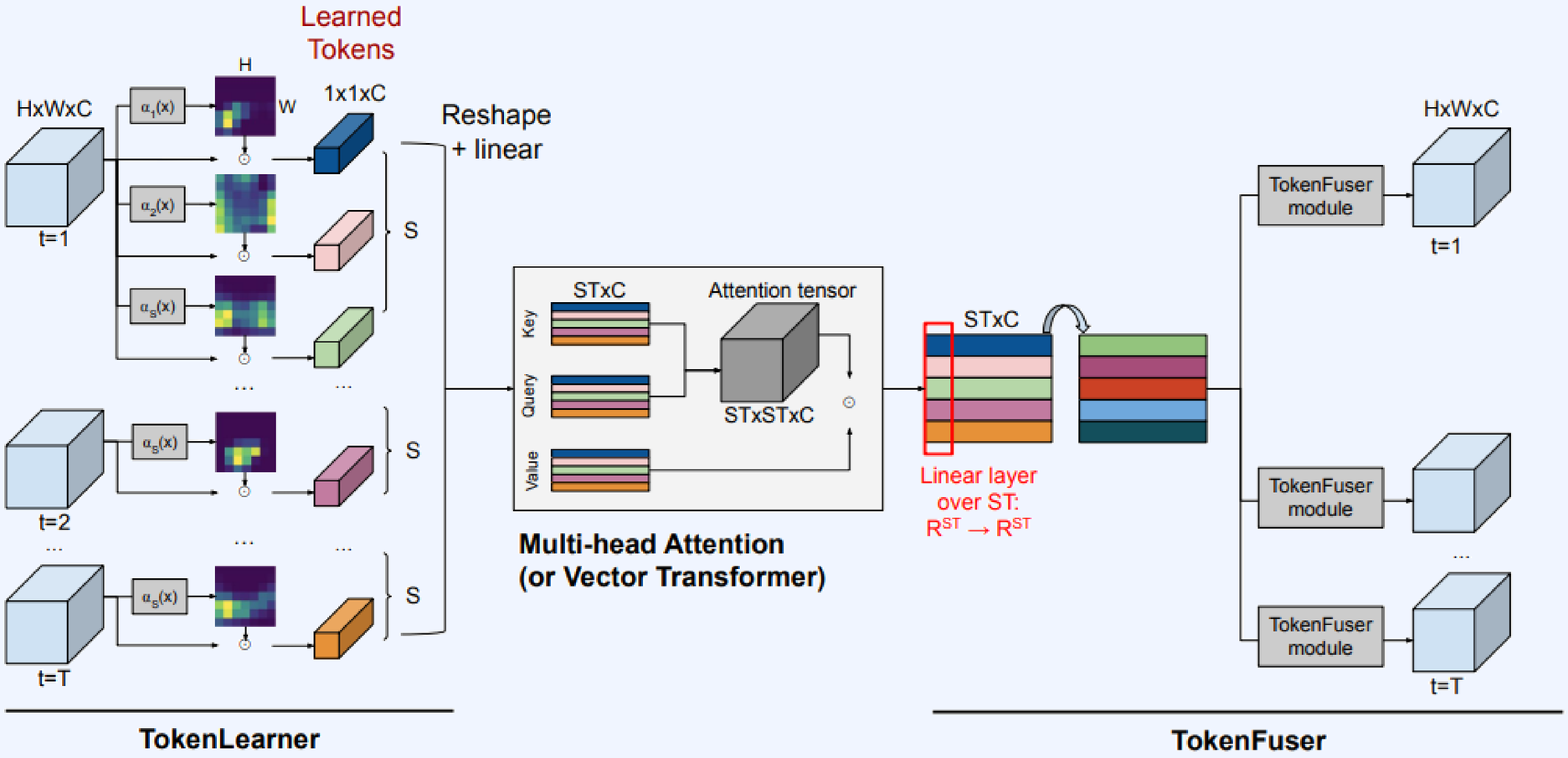
TokenLearner



Video Understanding
High-level Computer Vision for Video Data

M. Ryoo et al. TokenLearner: Adaptive Space-Time Tokenization for Videos. NeurIPS

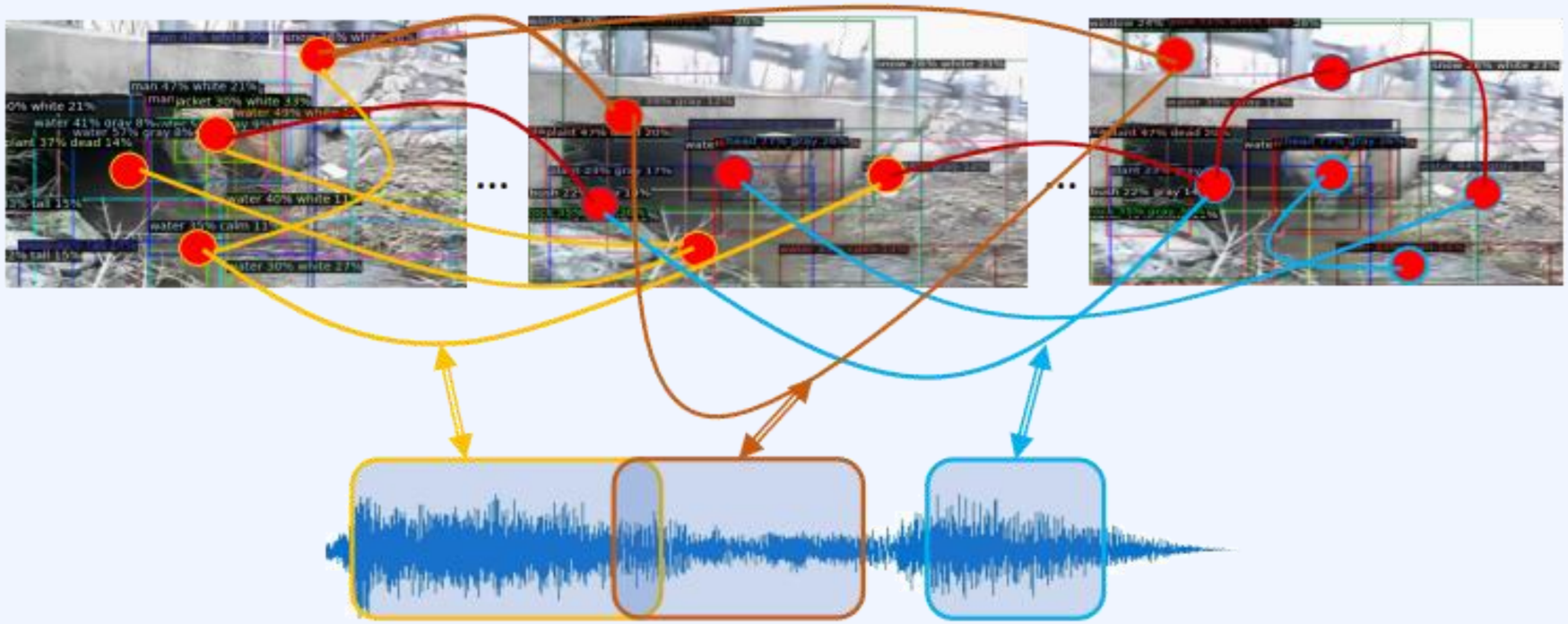
TokenLearner



Video Understanding
High-level Computer Vision for Video Data

M. Chatterjee et al. Visual Scene Graphs for Audio Source Separation. ICCV

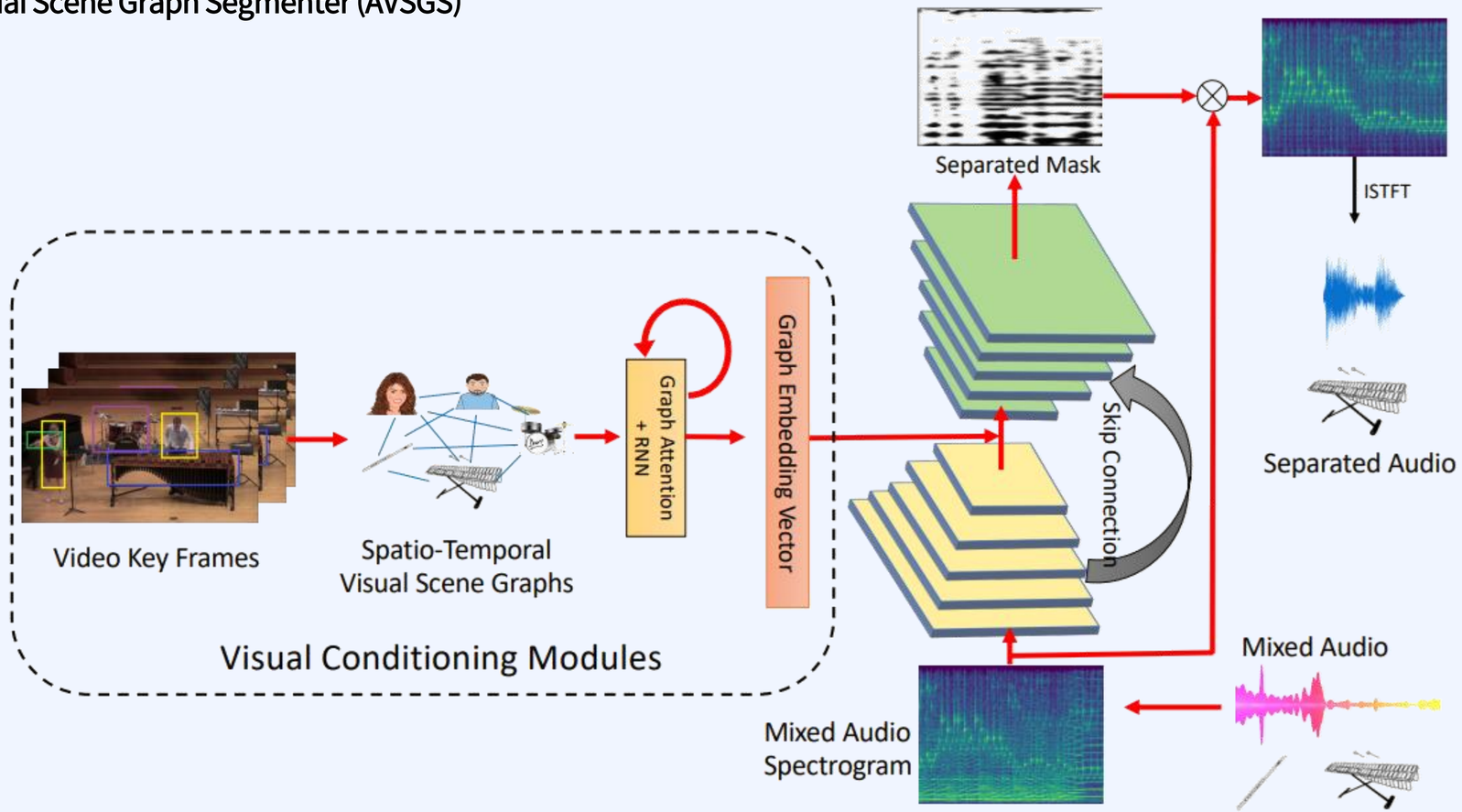
Audio Visual Scene Graph Segmenter (AVSGS)



Video Understanding
High-level Computer Vision for Video Data

M. Chatterjee et al. Visual Scene Graphs for Audio Source Separation. ICCV

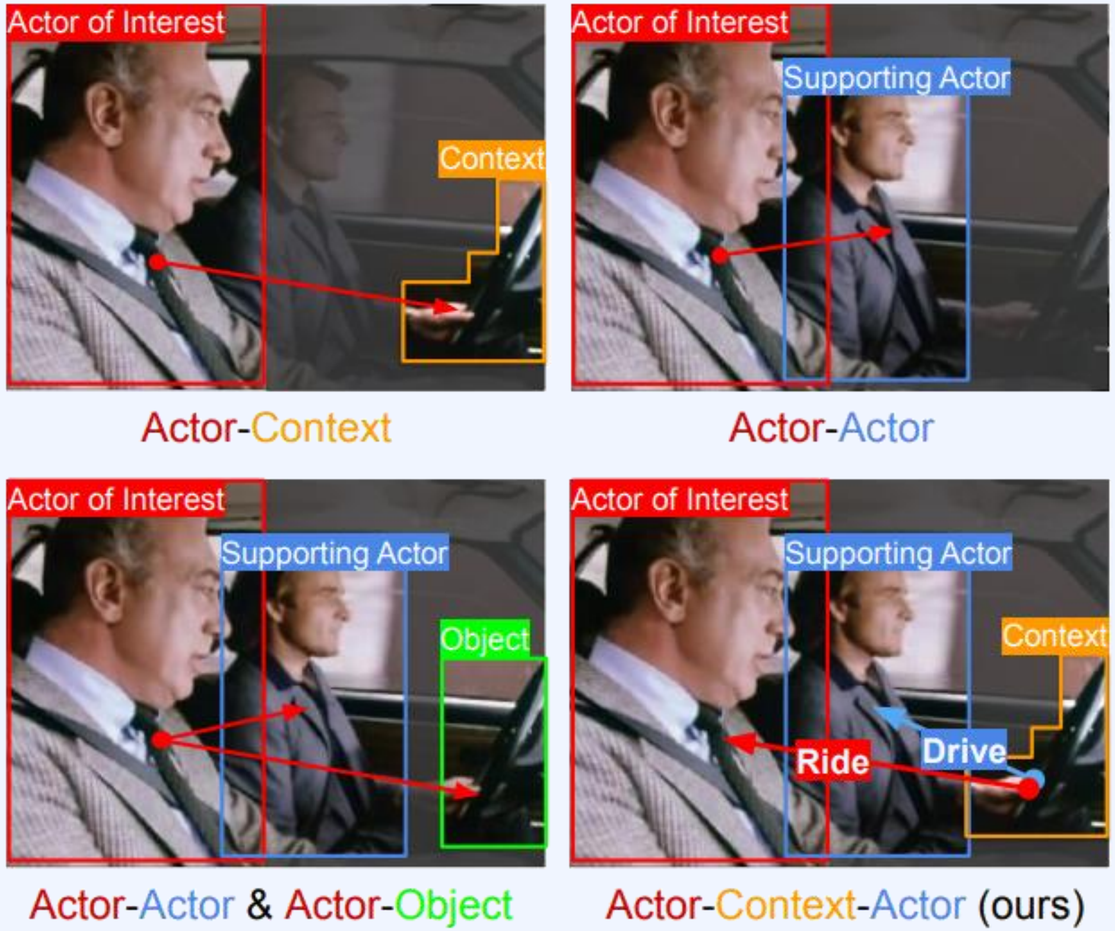
Audio Visual Scene Graph Segmenter (AVSGS)



Video Understanding
High-level Computer Vision for Video Data

J. Pan et al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. CVPR

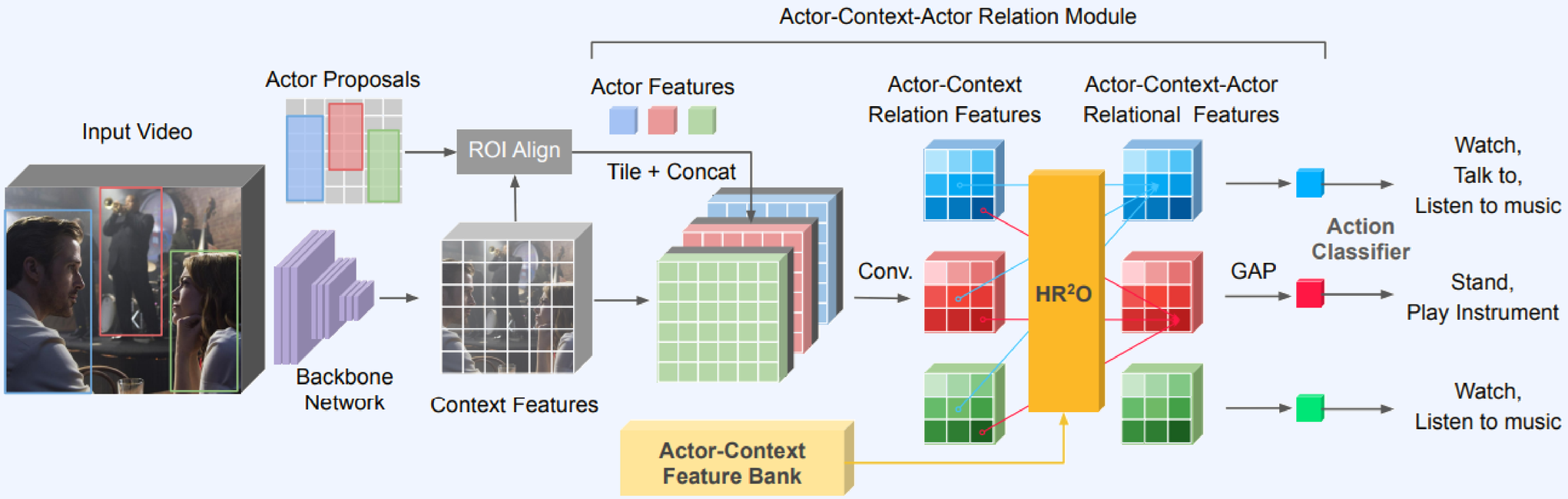
ACAR-Net



Video Understanding
High-level Computer Vision for Video Data

J. Pan et al. Actor-Context-Actor Relation Network for Spatio-Temporal Action Localization. CVPR

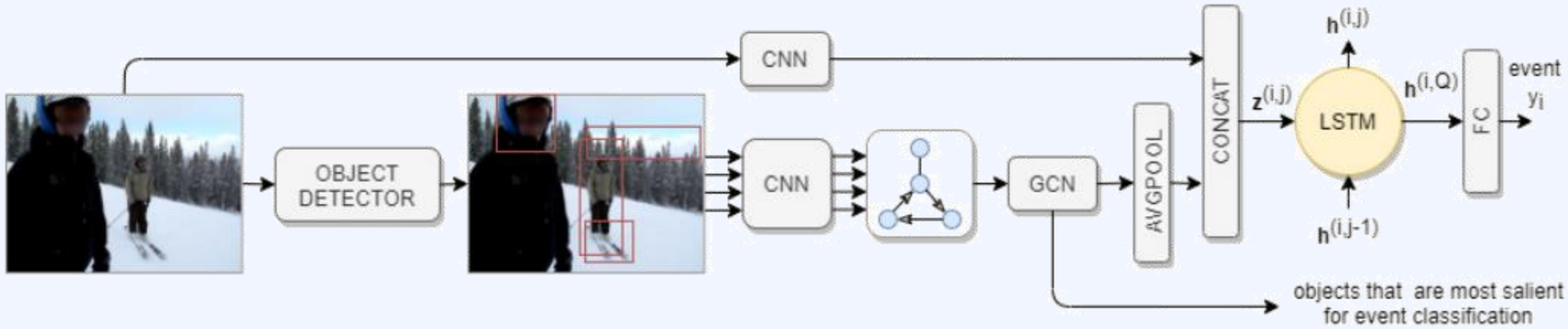
ACAR-Net



Video Understanding
High-level Computer Vision for Video Data

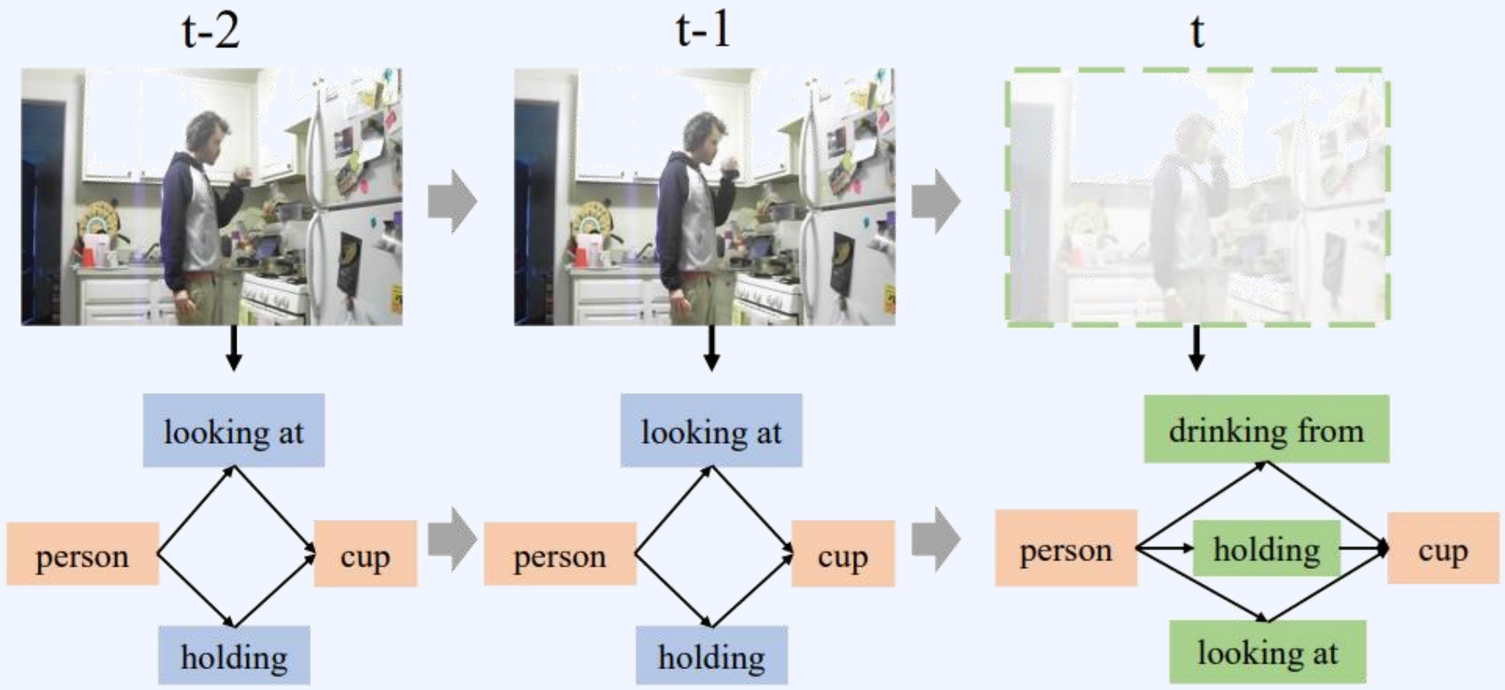
N. Gkalelis et al. ObjectGraphs: Using Objects and a Graph Convolutional Network for the Bottom-up Recognition and Explanation of Events in Video. CVPR Workshop

ObjectGraphs



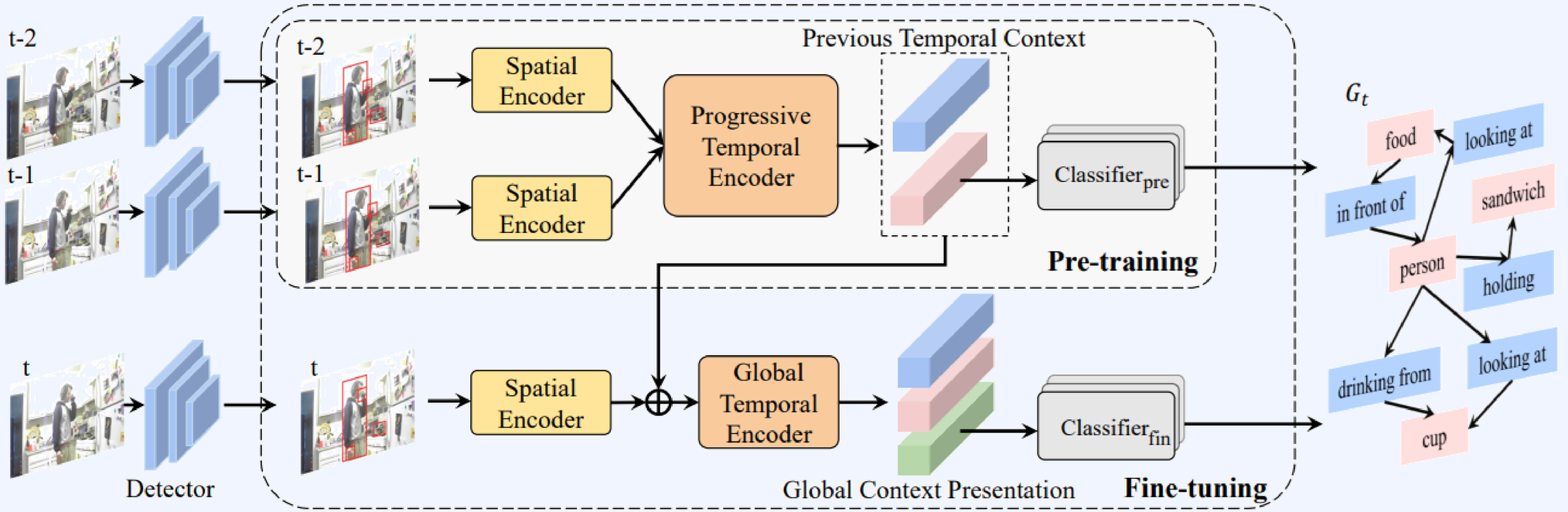
Video Understanding
High-level Computer Vision for Video Data

Y. Li et al. Dynamic Scene Graph Generation via Anticipatory Pre-training. CVPR



Video Understanding
High-level Computer Vision for Video Data

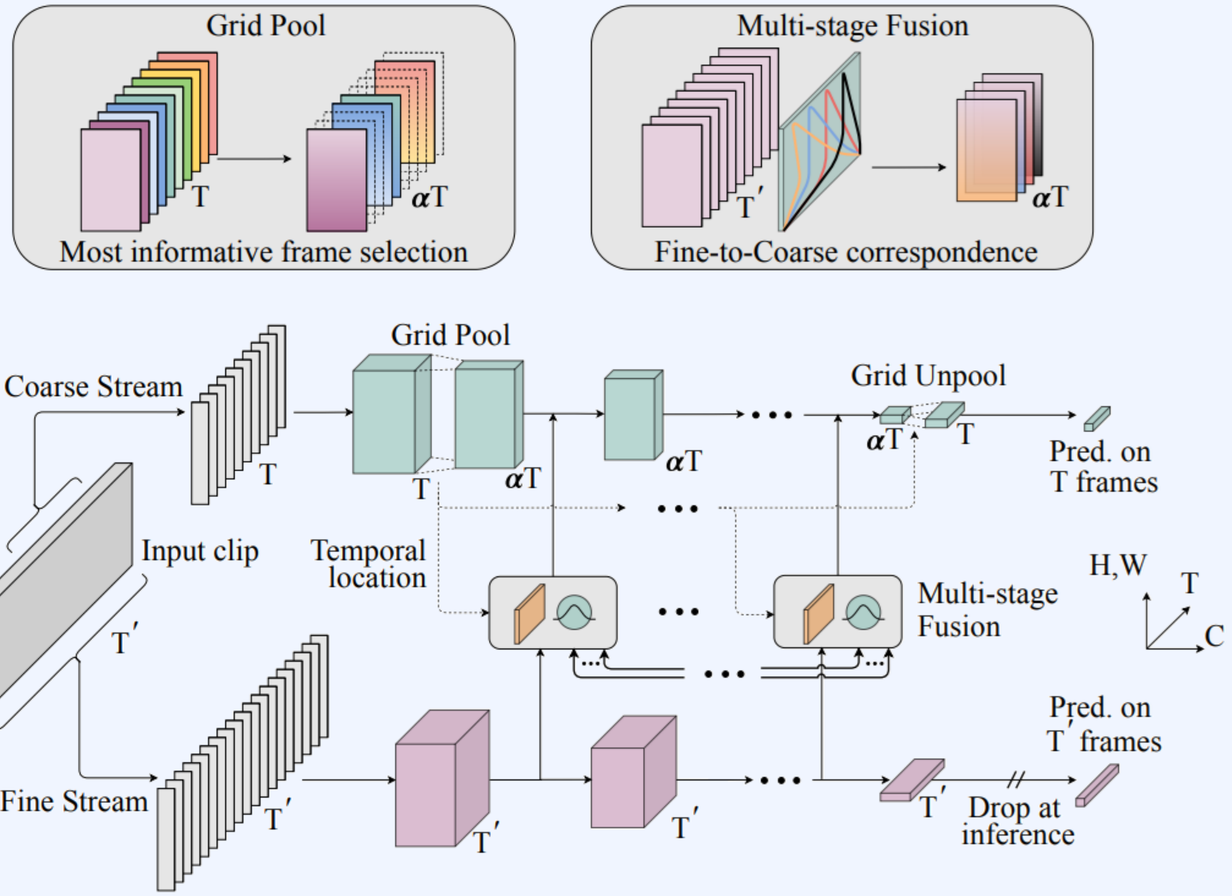
Y. Li et al. Dynamic Scene Graph Generation via Anticipatory Pre-training. CVPR



Video Understanding
High-level Computer Vision for Video Data

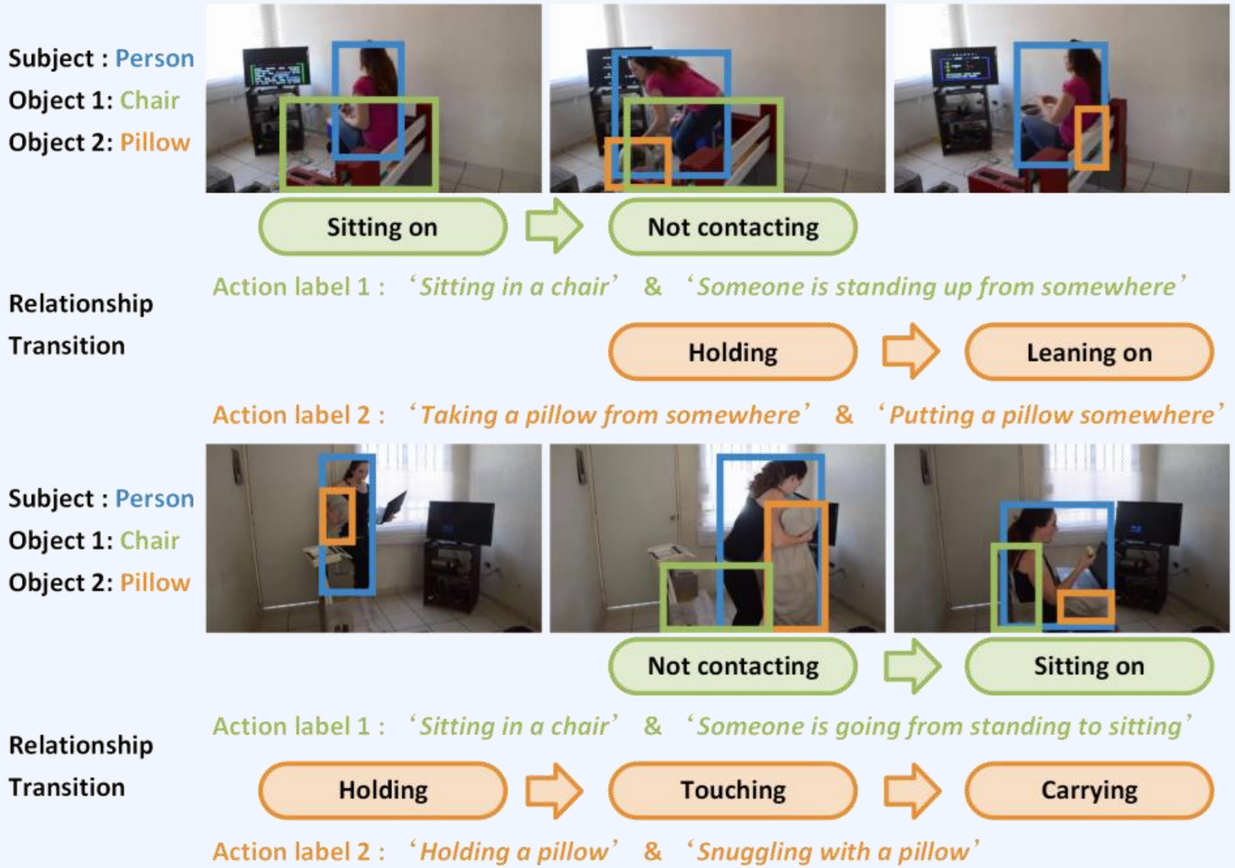
Kahatapitiya and Ryoo. Coarse-Fine Networks for Temporal Activity Detection in Videos. CVPR

Coarse-Fine Network



Video Understanding
High-level Computer Vision for Video Data

Y. Ou et al. Object-Relation Reasoning Graph for Action Recognition. CVPR



Contact relationships describe the different ways the person is contacting an object. A change in contact often indicates the occurrence of an actions: for example, changing from <person - not contacting – book> to <person - holding – book> may show an action of "picking up a book".

– Action Genome

Video Understanding
High-level Computer Vision for Video Data

Y. Ou et al. Object-Relation Reasoning Graph for Action Recognition. CVPR