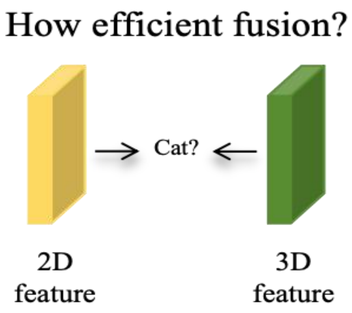
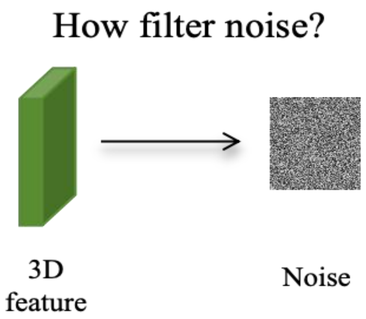
SPATIO-TEMPORAL ACTION DETECTION WITH A MOTION SENSE AND SEMANTIC CORRECTION FRAMEWORK

Yong Zhang[1,2], Chunan Yu[1], Chenglong Fu[1], Yuanqi Hu[1], Ying Zang[1,†]
[1]Huzhou University, [2]Liaoning Normal University

ICASSP 2024 KOREA

Motivation





Action Detection:

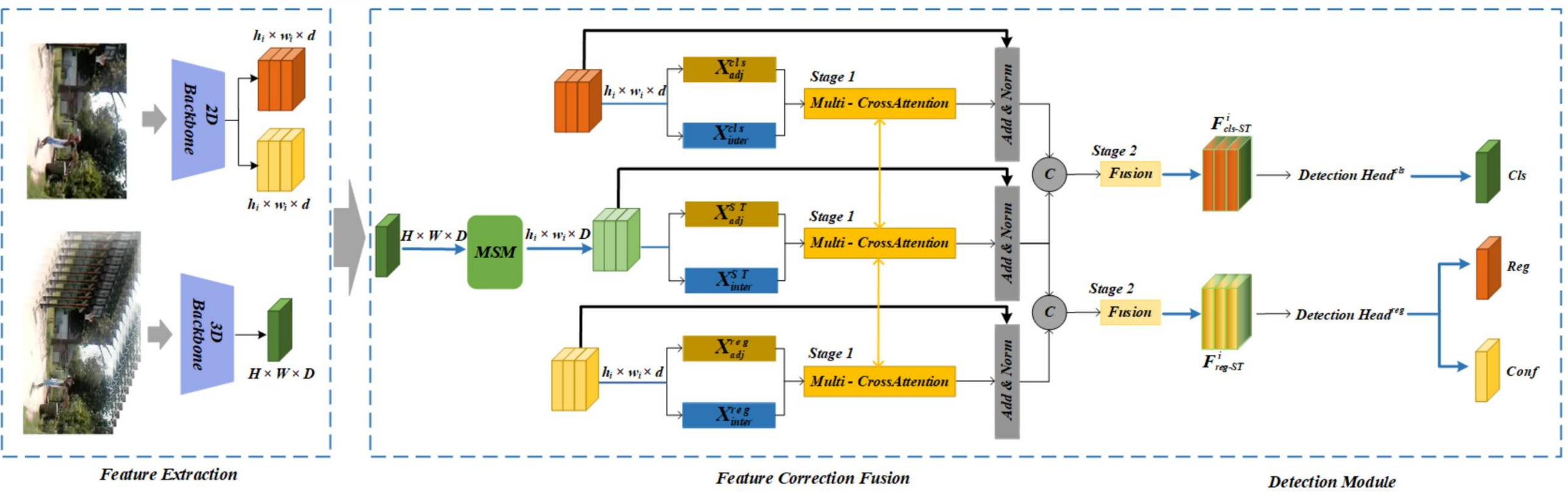
Aimed at utilizing model frameworks to recognize actions in videos or sequences of images.

Introduction

Q1: How to address high feature similarity, making it difficult to differentiate between action and background features?
Contribution 1: Motion Sense Module (MSM) to significantly increase the feature distance between action and non-action features in the semantic space, enhancing feature discriminability.

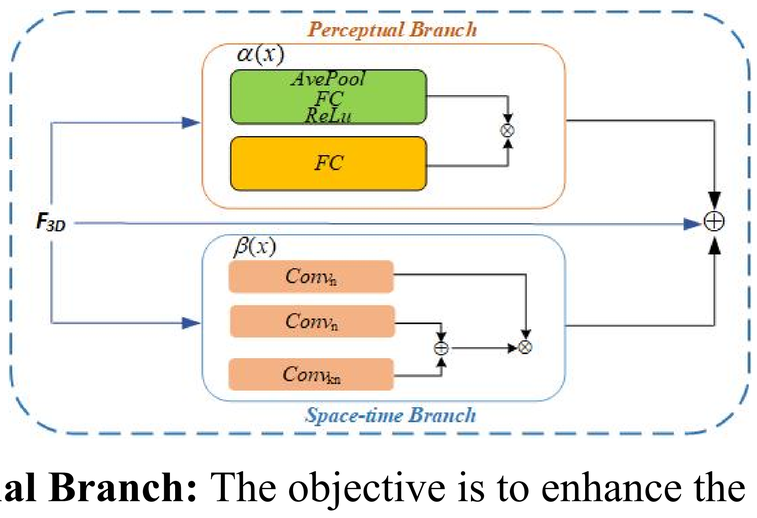
Q2: How to achieve multimodal feature fusion that 1+1>2, fulfilling fusion between different modalities?
Contribution 2: Considering the complementary nature of information across different modalities, an efficient Semantic Correction Fusion Module (SFM) is introduced to facilitate interaction between features of distinct modalities and maximize their complementary information integration.

Method Overview



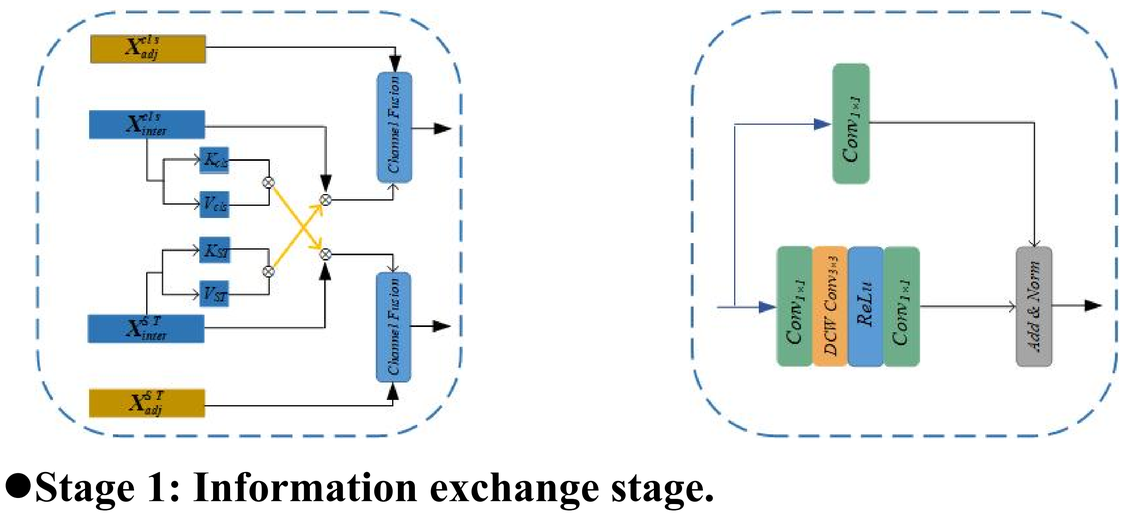
- This paper proposes a novel Motion Sense and Semantic Correction framework (MS-SC).
- The MS-SC framework achieves accurate detection by fusing features from images (spatial dimension) and videos (spatio-temporal dimension).

Motion Sense Module



- **Perceptual Branch:** The objective is to enhance the distinguishability between action features and background features by increasing their feature distance from the average features at the video level.
- **Spacetime Branch:** The objective is to incorporate semantic content from a wider receptive field, which helps in dynamically focusing on features of a specific scale.

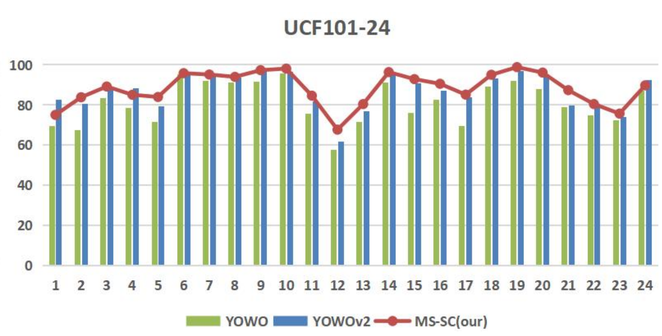
Semantic Correction Fusion Module



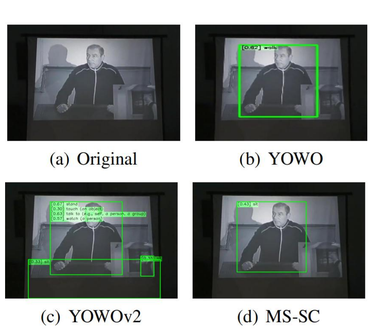
- **Stage 1: Information exchange stage.**
  In order to correct each other between spatiotemporal features and spatial features, cross-attention method is used for feature guidance.
- **Stage 2: Fusion stage.**
  We use a simple channel embedding to merge the features of the two paths, which is achieved by a 1×1 convolutional layer.

Experiments

Modules		UCF101-24		AVA
MSM	SFM	fAP(0.5)	vAP(0.5)	fAP(0.5)
		87.00	52.80	21.70
✓		87.54	53.51	21.94
	✓	88.07	53.68	22.14
✓	✓	**88.28**	**53.74**	**22.30**



Method	Backbone	Input	fAP(0.5)	GFLOPs
I3D[14]	I3D-VGG	32 × 2	14.5	—
SlowFast[1]	SlowFast-R50	16 × 4	24.2	308
X3D-XL[6]	X3D-XL	16 × 5	26.1	290
WOO[23]	SlowFast-R101	8 × 4	28.3	252
Tuber*[8]	CSN-152	32 × 2	31.7	120
YOWO[10]	3D-ResNeXt-101	32 × 4	19.1	82
YOWOv2-L[11]	3D-ResNeXt-101	32 × 4	21.7	92
MS-SC(our)	3D-ResNeXt-101	32 × 4	**22.3**	**94.5**



(a) Original (b) YOWO (c) YOWOv2 (d) MS-SC