Countering Modal Redundancy and Heterogeneity: A Self-Correcting Multimodal Fusion

Pengkun Wang [1], *Xu Wang* [1], *Binwu Wang* [1], *Yudong Zhang* [1], *Lei Bai* [2,*], *and Yang Wang* [1,*]

[1] University of Science and Technology of China (USTC), China
[2] Shanghai AI Laboratory, China

Reporter: Pengkun Wang

ICDM-2022

Outline

- Background
- Our Method
- Experiment
- Conclusion

Background

What is a *Modality*?

- A certain type of information or the representation format in which information is stored

  - ✓ Tactile, auditory, visual and olfactory data
  - ✓ Audio, image, video, text
  - ✓ Radar, infrared, accelerometer
  - ✓ Different languages
  - ✓ Data sets collected under different conditions

Fig. 1 Various Modalities.

Background

What is *Multimodal Learning*?

- Process and understand multi-source modal information by means of machine learning
- Five Challenges

Representati on

Translation

Alignment

Fusion

Co-learning

Multimodal Fusion

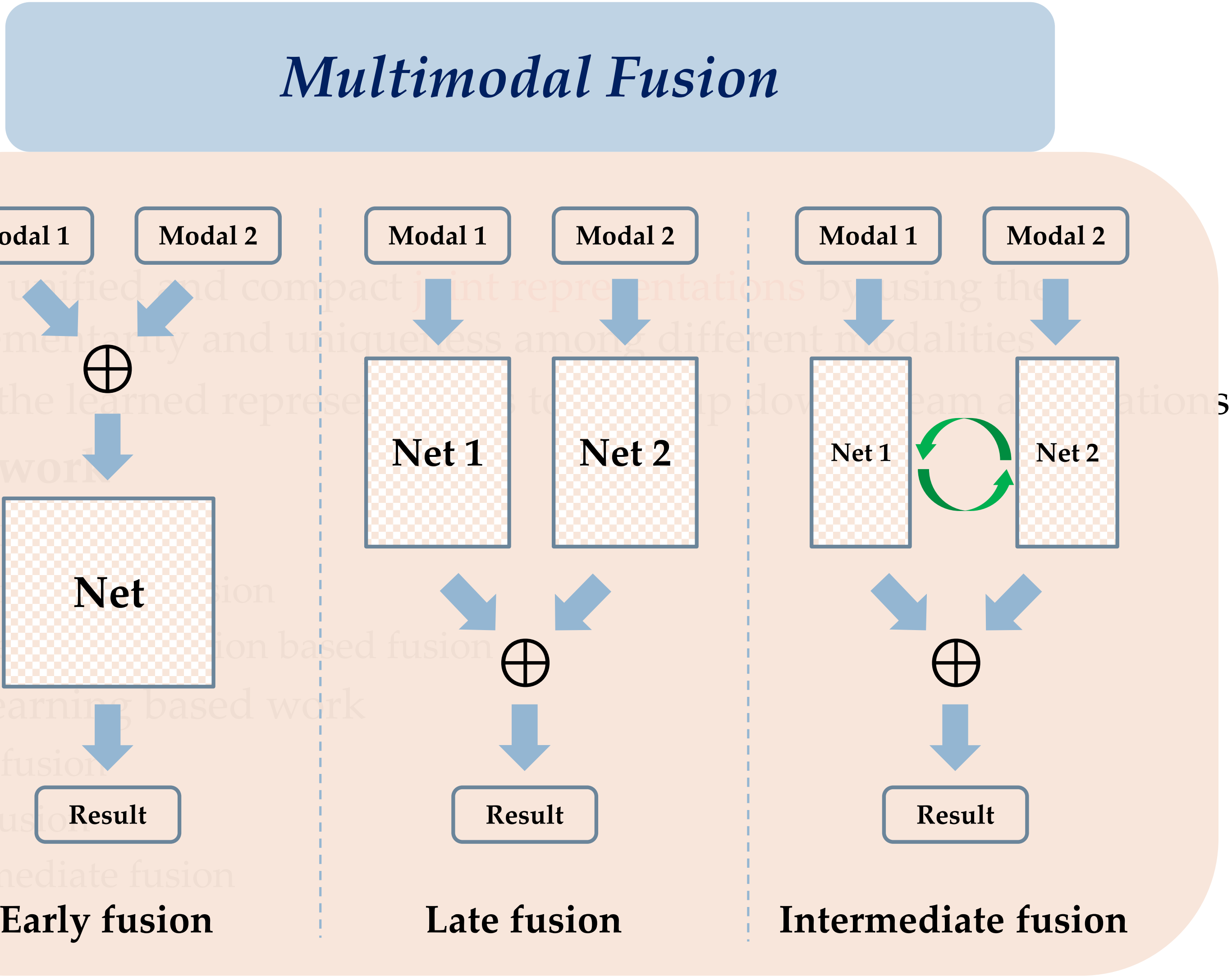
Joining information from two or more modalities to perform a prediction

Background

Multimodal Fusion

- **Purpose**
  - Extract unified and compact joint representations by using the complementarity and uniqueness among different modalities
  - Apply the learned representations to prop up downstream applications
- **Related work**
  - Traditional methods
    - Bayesian based fusion
    - Sparse representation based fusion
  - Deep learning based methods
    - Early fusion
    - Late fusion
    - Intermediate fusion

Background



Multimodal Fusion
Modal 1
Modal 2
Net
Result
Early fusion
Modal 1
Modal 2
Net 1
Net 2
Result
Late fusion
Modal 1
Modal 2
Net 1
Net 2
Result
Intermediate fusion

Our Method

- Challenges I
  - Feature redundancies
    - Irrelevant information
      - Caused by a general task-irrelevant feature extractor
    - Repetitive information
      - Similar information in the modal.
- Motivation I
  - Irrelevant information + Repetitive information (I+R)

    → Accumulation of redundancies → Serious semantic bias of fusion representations
  - Existing methods cannot be directly used to simultaneously deal with I+R

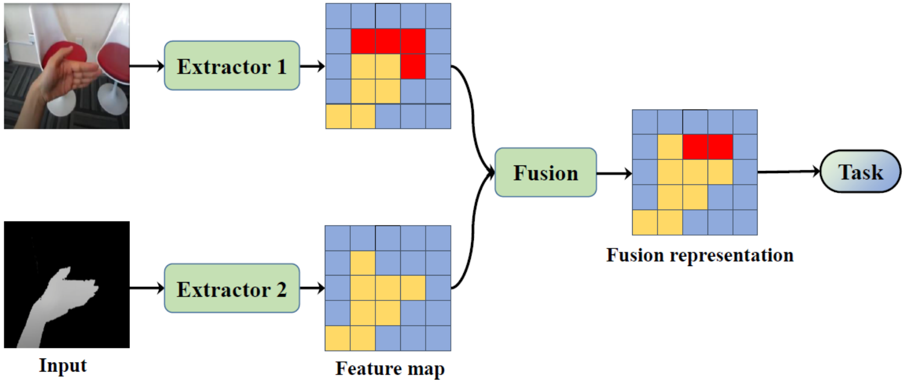
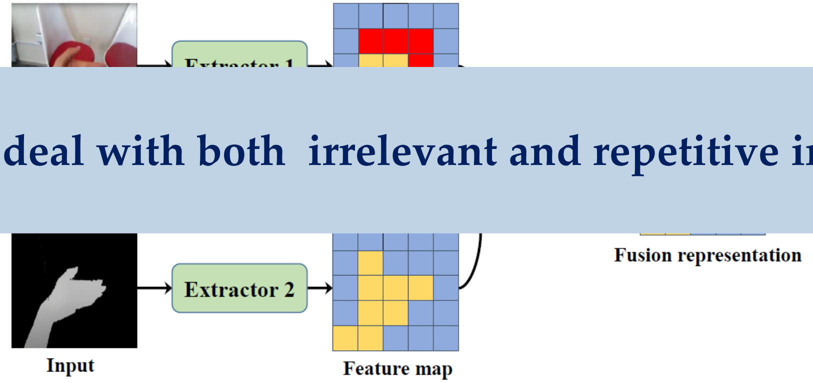


Fig. 2 Illustration of the redundancies in multimodal fusion.

Our Method

- **Challenges I**
  - Feature redundancies
    - Irrelevant information
      - Caused by a general task-irrelevant feature extractor
    - Repetitive information
      - Similar information in the modal.
- **Motivation I**
  - Irrelevant information + Repetitive information (I+R)

    → Accumulation of redundancies → Serious semantic bias of fusion representations
  - Existing methods cannot be directly used to simultaneously deal with I+R



Simultaneously deal with both irrelevant and repetitive information

Fig. 2 Illustration of the redundancies in multimodal fusion.

Our Method

Challenges II

- Feature homogeneity
  - Unified data structure
    - Easy to achieve feature interaction
- Feature heterogeneity
  - Diverse data structure
    - Difficult to achieve feature interaction

Motivation II

- Existing methods fall short in processing data with diverse structures

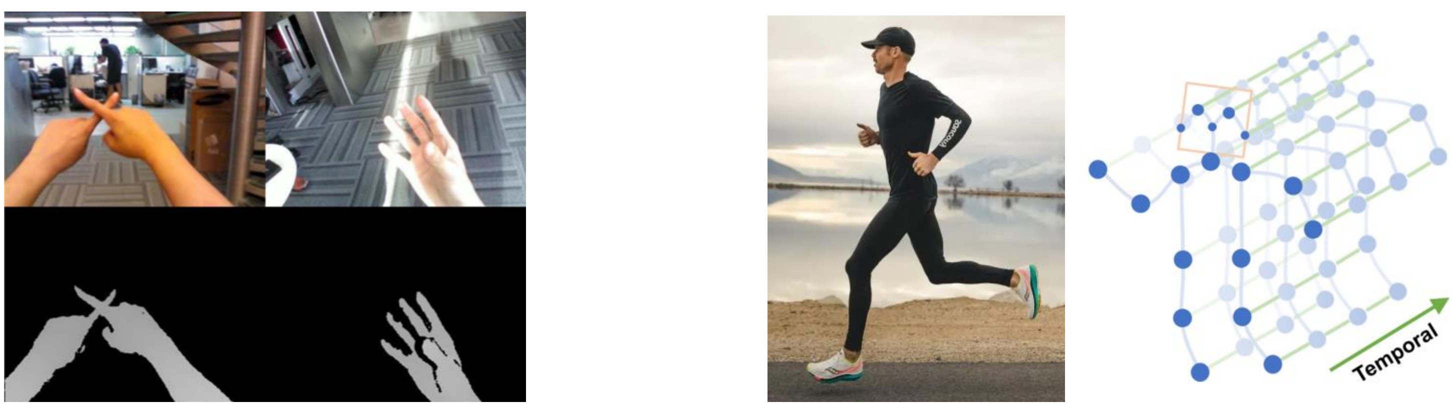


Fig. 3 Feature homogeneity and feature heterogeneity.

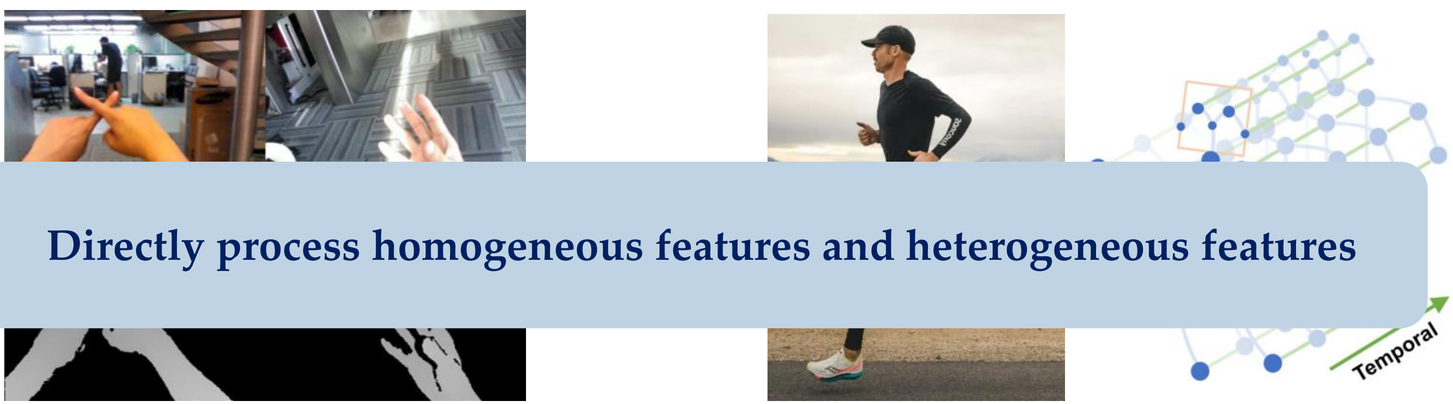
Our Method

Challenges II

- Feature homogeneity
  - Unified data structure
    - Easy to achieve feature interaction
- Feature heterogeneity
  - Diverse data structure
    - Difficult to achieve feature interaction

Motivation II

- Existing methods fall short in processing data with diverse structures



Directly process homogeneous features and heterogeneous features

Fig. 3 Feature homogeneity and feature heterogeneity.

Our Method

- **Countering Modal Redundancy and Heterogeneity (CMRH)**
  - Unified Feature Interaction Module (UFIM)
    - Orthogonal attention component
    - Interactive feedback mechanism

    Countering heterogeneity
  - Self-Correcting Transformer Module (SCTM)
    - Modified transformer
    - Fusion representation correction

    Countering redundancy

- **Contributions**
  - First work that comprehensively understands the modal redundancy problem
  - A unified multimodal fusion strategy to counter modal redundancy and heterogeneity
  - Experiments on four cross-domain datasets show the effectiveness of CMRH

Our Method

Unified Feature Interaction Module (UFIM)

- Orthogonal attention component

Step 1. Obtain the fine-grained attention map $\mathcal{M}_{X^iY^j} = Softmax(\frac{{E_X^i}^{\top} \otimes E_Y^j}{\sqrt{C}}),$

Step2. Obtain the fine-grained attention-based representations

$$\begin{cases} {E_X^i}' = \mathcal{M}_{X^iY^j} \otimes E_X^i \\ {E_Y^j}' = \mathcal{M}_{X^iY^j}^{\top} \otimes E_Y^j \end{cases}$$

- Interactive feedback mechanism

Step 3. Fine-grained attention-based representati
are fed back to the original feature

$$\begin{cases} \widehat{E_X^i} = E_X^i + \alpha * {E_Y^j}' \\ \widehat{E_Y^j} = E_Y^j + \alpha * {E_X^i}' \end{cases}$$

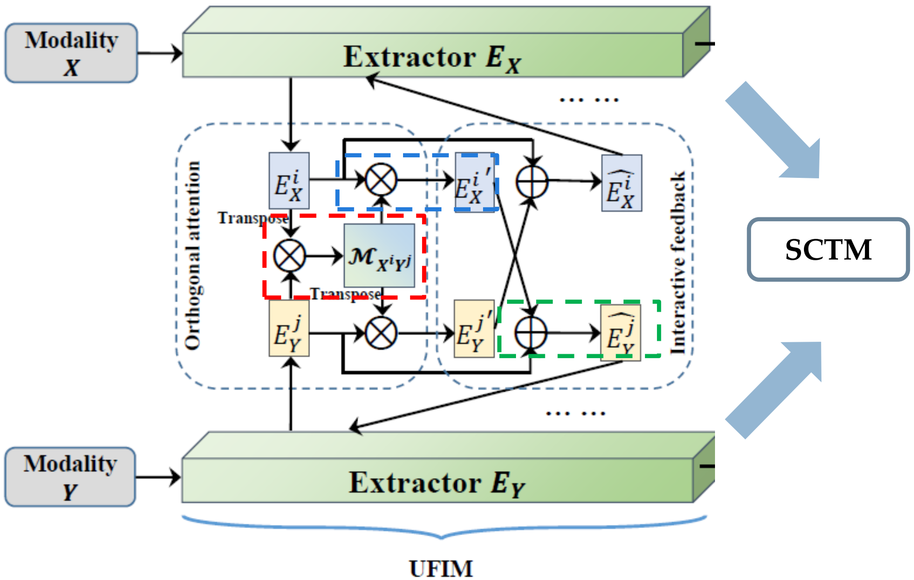


Fig. 4 An illustration of our proposed UFIM.

Our Method

- **S**elf-**C**orrecting **T**ransformer **M**odule (SCTM)
  - Modified transformer

Step 1. Transfer features to a consistent dimension and obtain attention-based feature maps

$$\begin{cases} \mathcal{M}'_X = \mathcal{T}(Func_{X'}(X'), Func_{Y'}(Y'), Func_{Y'}(Y')) \\ \mathcal{M}'_Y = \mathcal{T}(Func_{Y'}(Y'), Func_{X'}(X'), Func_{X'}(X')) \end{cases}$$

Substep 1. Features are equally divided into blocks

Substep 2. Each block is concatenated with the position embedding of this block and the modal-identity embedding of the current modality.

Substep 3. Obtain the attention-based feature map

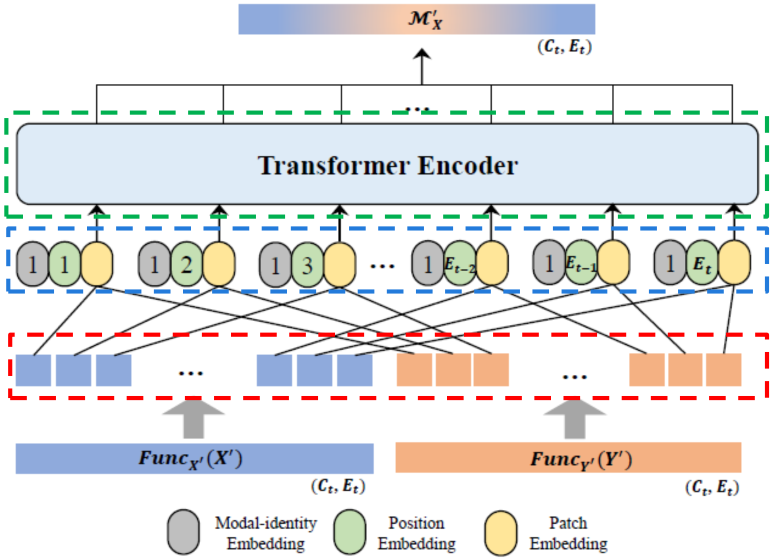


Fig. 5 Illustrated of the modified transformer.

Our Method

- **Self-Correcting Transformer Module (SCTM)**
  - Modified transformer

Step 2. Feed the obtained attention-based feature maps back to their original modal features

$$\begin{cases} \widehat{X'} = \mathcal{M}'_X + X' \\ \widehat{Y'} = \mathcal{M}'_Y + Y' \end{cases}$$

  - Fusion representation correction

Step 3. Obtain the fusion representation

$$\mathcal{P} = \mathcal{F}(\widehat{X'}, \widehat{Y'})$$

Step 4. Calculate the element-wise weighted average feature map as the weights of fusion representation

$$\widehat{\mathcal{P}} = \mathcal{P} \odot Norm(\frac{\mathcal{M}'_X + \mathcal{M}'_Y}{2})$$

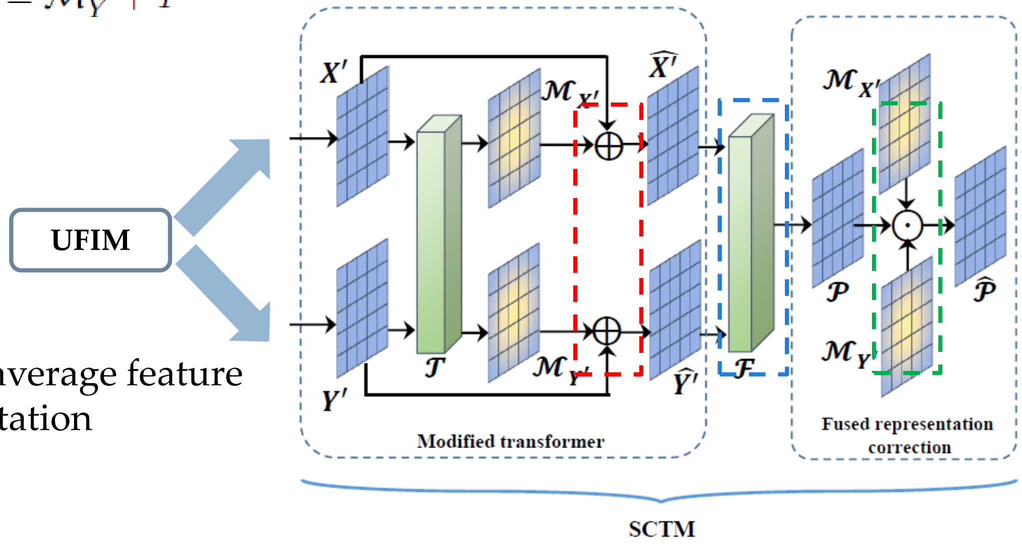


Fig. 6 An illustration of our proposed SCTM.

Experiment

Task 1: Hand Gesture Recognition

- Dataset
  - EgoGesture
- Implementation
  - I3D for RGBs and depth maps
  - Apply UFIMs to the last three inception modules
  - Improve the fusion module with SCTM
- Analysis
  - With UFIM:
    - Perform other non-interactive methods and MMTM
    - Properly feature interactions help improve the expressiveness of representations
  - With SCTM
    - Perform more effectively (compared with I3D late fusion)
    - Correcting the redundancy makes the final representation more can represent the fused modalities
  - With UFIM and SCTM
    - Outperform the top performer MMTM by 1.09%
    - Mutually compatible

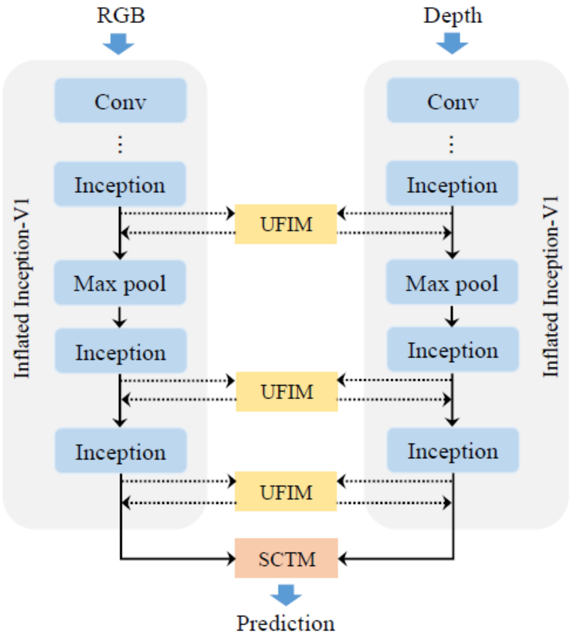


Fig. 6 An overview of the improved hand gesture recognition framework.

TABLE I
PERFORMANCE ON EGOGESTURE DATASET.

Method	Modalities	Accuracy
VGG16+LSTM [25]	RGB+Depth	81.40%
C3D+LSTM+RSTTM [23]	RGB+Depth	92.20%
I3D [20]	Depth	89.47%
I3D [20]	RGB	90.33%
I3D late fusion [20]	RGB+Depth	92.78%
MMTM [1]	RGB+Depth	93.51%
UFIM	RGB+Depth	93.92%
SCTM	RGB+Depth	94.15%
UFIM+SCTM	RGB+Depth	**94.60%**

Experiment

- Task 2: House Price Prediction
  - Dataset: NYC and BEIJING dataset
- Task 3: Action Recognition
  - Dataset: NTU-RGBD
- Task 4: Traffic Accident Forecast
  - Dataset: NYC and SIP dataset

TABLE II
PERFORMANCE ON NYC AND BEIJING DATASET.

Method	NYC		Beijing	
	RMSE	MAPE	RMSE	MAPE
Deep-ST+C+F [27]	27.81	12.69	74.62	10.78
ST-InceptionV4+C+F [28]	27.03	11.11	73.79	10.48
ST-ResNet+C+F [29]	26.04	11.74	73.11	10.63
FTD_DenseNet [26]	22.81	9.98	64.83	9.42
JGC_MMN [2]	21.43	9.16	60.19	9.04
MMTM+JGC [1]	20.37	9.01	58.92	8.96
UFIM+JGC	19.07	8.46	53.29	8.75
SCTM	19.69	8.55	54.14	8.79
UFIM+SCTM	**18.23**	**8.12**	**52.83**	**8.01**

TABLE IV
PERFORMANCE ON NYC AND SIP DATASET.

Method	NYC (Acc@20)	SIP (Acc@6)
STDN [36]	37.48%	42.18%
DFN [37]	40.26%	36.98%
STSGCN [38]	26.46%	33.59%
RiskSeq [4]	56.42%	71.27%
MMTM+RiskSeq [1]	57.69%	73.05%
UFIM+RiskSeq	59.81%	75.33%
SCTM	58.65%	73.60%
UFIM+SCTM	**60.42%**	**76.08%**

TABLE III
PERFORMANCE ON NTU-RGBD DATASET.

Method	Skeleton Model	Modalities	Accuracy (CS)
I3D [20]	-	RGB	85.63%
HCN [31]	-	Pose	85.24%
ST-GCN [32]	-	Pose	81.50%
PoseMap [33]	-	RGB+Pose	91.71%
I3D+HCN late fusion [1]	HCN	RGB+Pose	91.56%
SGM-Net [34]	ST-GCN	RGB+Pose	89.10%
MSAF [35]	HCN	RGB+Pose	92.24%
MMTM [1]	HCN	RGB+Pose	91.99%
MMTM [1]	ST-GCN	RGB+Pose	88.79%
UFIM	HCN	RGB+Pose	92.20%
SCTM	HCN	RGB+Pose	92.37%
UFIM+SCTM	HCN	RGB+Pose	**92.69%**
UFIM	ST-GCN	RGB+Pose	89.14%
SCTM	ST-GCN	RGB+Pose	89.50%
UFIM+SCTM	ST-GCN	RGB+Pose	90.27%

- Discussion of the UFIM and SCTM
  - Location of the interaction
    - Optimal location for module insertion is the tail layer of the model
  - Number of the interaction
    - NOT More than one-third of the entire model layers
  - Selection of weight parameters
    - setting α at about 0.2 and fine-tuning it according to the actual task.

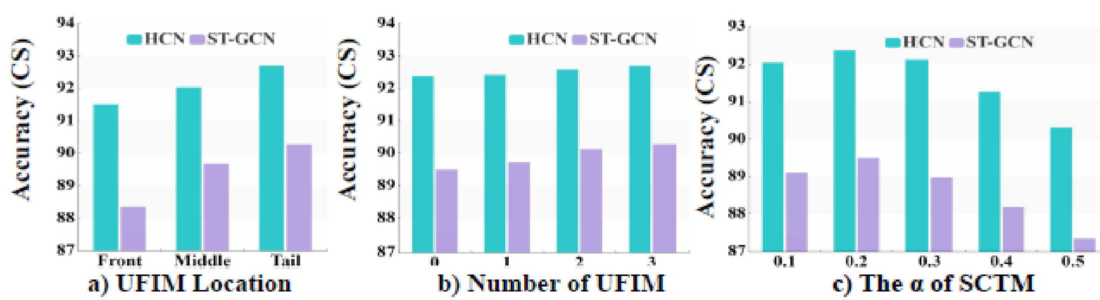


Fig. 7 Comparison of varieties on NTU-RGBD dataset.

Conclusion

- We propose a unified multimodal fusion strategy, including two well-designed modules, UFIM and SCTM, for addressing both modal heterogeneity and redundancy by exploiting the inter-modal complementarity.

- UFIM and SCTM can be flexibly applied to existing multimodal fusion networks at a relatively low cost.

- Extensive experiments on four different cross-domain datasets from the fields of hand gesture recognition, house price prediction, action recognition, and traffic accident forecast show the effectiveness of the proposed modules.

Thanks for your listening!

For more details, please refer to our paper!

Reporter: Pengkun Wang
pengkun@mail.ustc.edu.cn