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

Pengkun Wang†, Xu Wang†, Binwu Wang†, Yudong Zhang†, Lei Bai‡, and Yang Wang†
† *University of Science and Technology of China, China*
‡ *Shanghai AI Laboratory, China*
† {pengkun, wx309, wbw1995, zyd2020}@mail.ustc.edu.cn, angyan@ustc.edu.cn
‡ baisanshi@gmail.com

***Abstract*—Fusing multimodal heterogeneous data plays a vital role in recognition and prediction tasks in various fields, e.g., action recognition and traffic accident forecast. Yet, there remain some key challenges, such as heterogeneous feature interaction and feature redundancies, that significantly affect the performance of multimodal fusion. To tackle these challenges, we first devise a Unified Feature Interaction Module (UFIM) in which a novel orthogonal attention component is designed to obtain fine-grained inter-modal interaction information among heterogeneous features. Then, we propose a novel Self-Correcting Transformer Module (SCTM) which employs a modified transformer to obtain the one-to-many correlation information between the current modal feature and the merged features of other modalities to alleviate the redundancy problem. Extensive experiments on four cross-domain tasks demonstrate the effectiveness and generalization ability of our proposed method.**

***Index Terms*—multimodal fusion, redundancy, heterogeneity, feature interaction, transformer**

I. Introduction

With the developments of hardware devices, especially sensing devices, a variety of once-rare data has been collected abundantly, making the field of multimodal learning attract widespread attention from academia and industry. How to efficiently explore the characteristics in different modalities and use them jointly for recognition or prediction becomes increasingly crucial. Therefore, multimodal fusion, especially multimodal representation learning, has become a core research problem in the field of multimodal learning. Theoretically, multimodal representation learning aims to extract unified and compact joint representations by using the complementarity and uniqueness among different modalities, and apply the learned representations to prop up downstream applications such as hand gesture recognition [1], house price prediction [2], action recognition [3], and traffic accident forecast [4].

To obtain multimodal representations, researchers usually employ specialized modal-corresponding models to respectively pre-extract different modal features. However, due to the different designs and purposes of these models, the extracted features are normally interlaced, complementary, and not in the same semantic space, which brings severe challenges to multimodal fusion. The current popular solution is intermediate fusion [3], [5]–[9], which typically establishes feature interaction channels between extractors and compromises inter-modal commonalities and intra-modal uniqueness.

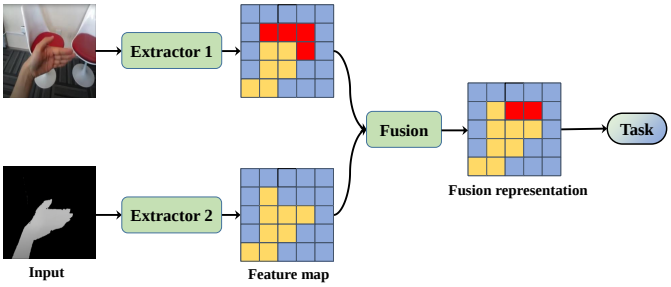

Fig. 1. Illustration of the redundancies in multimodal fusion. The red block indicates irrelevant information, which is caused by a general task-irrelevant feature extractor. The yellow block marks repetitive information. Improperly handling these redundancies will make the fused representation inaccurate and influence the performance.

Challenges: Even though multimodal fusion has been extensively researched in recent years, there remain some key challenges that may significantly affect the performance of multimodal fusion. On the one hand, these traditional multimodal fusion methods, which indiscriminately utilize unimodal information, will lead to semantic bias of fusion representations due to ignoring feature redundancies. In particular, feature redundancies consist of two parts, irrelevant information and repetitive information, as illustrated in Figure 1. The main reason for these redundancies is that general baseline models for unimodal feature extraction will introduce irrelevant information into the fusion process, and this information will make the fused representation inaccurate if not handled properly. Besides, the excessive fusion of modal repetitive information will lead to the suppression of modal-specific information, resulting in redundancy. *The simultaneous appearance of these two kinds of redundancies consequently leads to the accumulation of redundancies, then causes serious semantic bias of fusion representations, and finally results in wrong executions of downstream tasks.* Considering this, many methods [10], [11] have been proposed to alleviate this bias.

* Yang Wang is the corresponding author. Lei Bai is the joint corresponding author.
* Code available at: https://github.com/pongkun/CMRH

Specifically, [12] directly applies the attention mechanism on high-level features to accomplish video-related tasks. [5] proposes a depth-sensitive attention module to explicitly eliminate background distraction in significance detection. Even though some early works have attempted to filter redundancies within unimodal features by using the attention mechanism [13], nevertheless, these existing methods cannot be directly used to simultaneously deal with both irrelevant and repetitive information since they only employ attention unilaterally.

On the other hand, worth noting that existing feature interaction methods are homogeneity-based. For example, [10] proposes an integrated multimodal fusion network, which converts the intermediate modal features into two-dimensional vectors for direct concatenation and transmission. Based on previous work, [1] squeezes and excites features to overcome multimodal feature scale inconsistency. [11] proposes a dynamic channel exchanging strategy to realize parameter-free interaction by directly exchanging some channels. However, these methods fall short in processing data with diverse structures, and intuitively, *it is crucial to simultaneously utilize diversified networks to respectively process different modalities and to extensively extract structure-specific features.* Nevertheless, considering the limitations of existing feature interaction methods, they cannot directly process heterogeneous features.

To tackle these two challenges, we first devise a **U**nified **F**eature **I**nteraction **M**odule (UFIM) in which an orthogonal attention component is designed to obtain inter-modal fine-grained attention information among features which can maximally guarantee the integrality of features, and transfer fine-grained information to the entire network by using an interactive feedback mechanism. Opposite to traditional global average pooling and concatenation-based methods which can only obtain coarse-grained global contexts, fine-grained information can benefit structured feature extraction. To alleviate modal redundancy, we propose a novel **S**elf-**C**orrecting **T**ransformer **M**odule (SCTM) which employs a modified transformer to obtain the one-to-many correlation information between the current modal feature and the merged features of other modalities which can be viewed as the collective concerns of all other modalities to the current modality. And we alleviate both irrelevant and repetitive information by correcting the current modality by using the collective concerns and correcting the fusion representation by utilizing the correlation information of all modalities as weights. Worth noting that the proposed UFIM is inclusive since it can be integrated into existing fusion networks at a relatively low cost.

The main contributions of this paper are as follows:

- To the best of our knowledge, this is the first work that comprehensively understands the modal redundancy problem during multimodal fusion and focuses on addressing both irrelevant and repetitive information.
- We propose a unified multimodal fusion strategy to counter modal redundancy and heterogeneity. For heterogeneity, we design a groundbreaking UFIM approach, which can effectively extract and transfer inter-modal fine-grained correlations among features with its inbuilt orthogonal attention and interactive feedback, achieving unified modal interactions. For redundancy, we utilize a novel SCTM to obtain the one-to-many modal correlation information to alleviate two kinds of redundancies in different ways.
- To evaluate the effectiveness of our method, we conduct extensive experiments on four different cross-domain datasets. Comparisons with alternative state-of-the-art methods demonstrate the superiority of our method.

II. Related Work

Great efforts have been made in addressing the issue of **deep multimodal fusion**, hence achieving joint recognition or prediction on multifarious data. As above discussed, deep multimodal fusion methods can be mainly categorized into three categories: early, late, and intermediate fusions. Regarding early fusion [7], [8] which aims to achieve multimodal recognition or prediction with a single network, it pre-processes multiple input-level features in the form of feature concatenation and highly depends on whether multiple modalities can be directly concatenated or not, hence resulting in modal context loss. On the contrary, late fusion [3], [9], which aims to establish separate models for each modality and fuse them at model-level, ensures the full extraction of modal information and barely relies on whether multiple modalities can be directly concatenated or not. However, none of the existing late fusion methods has considered the inter-modal interaction issue during fusion, and this is antithetical to the original intention of multimodal fusion. Considering the inherent defects of these two fusions, some recent works, which are so-called the intermediate fusion since they attempt to achieve multimodal feature interactions during the intermediate phase, focus on modeling both intra- and inter-modal dynamics [14]. In particular, [15] introduces a dual-attention mechanism where two attention-based feature maps are eventually fused to implement collective land-cover classification. [6] links the attention module after the fusion module to address the camera localization problem. Considering the complexity issue of multimodal fusion, some tensor-based methods, which try to use the low-rank weight [16], tucker decomposition [14], and high-order polynomial pooling [17], have been proposed to obtain compact representations at tensor-level. However, all these existing approaches are generally based on an idealized assumption that all modalities are in a similar semantic space, i.e., the extracted multimodal features can be directly concatenated. However, inter-modal semantic gaps cannot be avoided in real scenarios due to various data collection means. Recently, attention mechanism [13] is applied to achieve accurate fusion representations in multimodal fusion. As pioneers, [12], [18] directly apply the attention mechanism to high-level features to accomplish video-related tasks. [5] proposes a depth-sensitive attention module to explicitly eliminate background distraction in significance detection. As can be easily analyzed, all pre-processing methods including early fusion and intermediate fusion, which assume that all modalities are in a similar semantic space, will eventually result in the

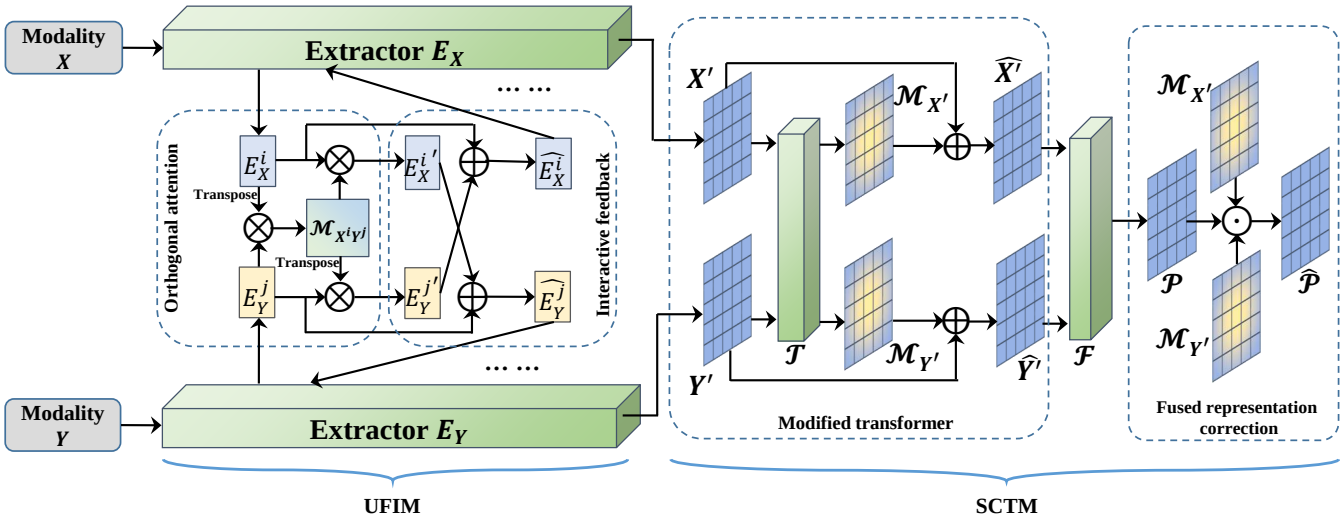

Fig. 2. An illustration of our proposed UFIM and SCTM. Note that the interaction module UFIM between extractors may be one or more.

inevitable pre-redundancy problem. And the issue of post-redundancy is inescapable for all post-processing methods including late fusion and intermediate fusion, which employ an attention mechanism before fusion. In summary, these two kinds of redundancies seem to be antagonistic and none of the existing methods can simultaneously address these two kinds of redundancies.

Furthermore, considering the data structural diversity in multimodal fusion, another indispensable issue is **Heterogeneous Feature Interaction**, and some preliminary works have been studied in recent years. Specifically, to achieve heterogeneous feature interaction, [10] first proposes an integrated multimodal fusion network, which converts the intermediate modal features into two-dimensional vectors for direct concatenation and transmission. Based on previous work, [1] squeezes and excites features to overcome multimodal feature scale inconsistency. [11] proposes a dynamic channel exchanging strategy to realize parameter-free interaction by directly exchanging some channels. Even though these above-mentioned methods can interact with multimodal features, nevertheless, they are all homogeneous feature-based methods. Given these facts, none of the existing works can effectively deal with structurally inconsistent modalities, the issue of heterogeneous feature interaction remains a challenge for existing multimodal fusion methods.

III. Methodology

In this section, we first give an overview of unified multimodal fusion. Then, we introduce the detailed technical implementation of our proposed method, which contains two fundamental parts: Unified Feature Interaction Module (UFIM) and Self-Correcting Transformer Module (SCTM), for respectively addressing the heterogeneity and redundancy issues in multimodal fusion.

A. Model Overview

As shown in Figure 2, to clearly understand the whole fusion process, we take a fusion of two modalities as an example. Given two modalities X and Y, we assume that they may be homogeneous or heterogeneous. For each modality, we select the modal-corresponding mainstream feature extractor to extract the unimodal representation for fusion. Specifically, we define the extractors corresponding to modalities X and Y as E_X and E_Y, which will realize feature interaction through the proposed UFIM in the early stage of fusion. Next, we fuse the high-level features X' and Y' extracted by E_X and E_Y through SCTM, which consists of two parts, modified transformer and fused representation correction. The modified transformer obtains the correlation information between the two modal features which can be viewed as the concerns of one modality to the other, and corrects the two modalities by using these concerns. The corrected features $\widehat{X}$ and $\widehat{Y}$ are be used as the inputs to a task-specific fusion module $\mathcal{T}$, which obtains the final fusion representation $\widehat{P}$ through the fused representation correction strategy.

B. UFIM for heterogeneity

As has been discussed in previous sections, to extract high-level features, it is crucial to simultaneously utilize different extractors to respectively process different modalities extensively and fully extract modality-specific features. Generally, in the intermediate fusion, the multimodal features received by the interaction module do not necessarily correspond to the same layer due to the structural differences of extractors. Here, we assume that the network needs to interact with the output features of i_{th} layer of E_X and j_{th} layer of E_Y. As shown in Figure 2, $E_X^i \in \mathbb{R}^{C_X^i \times N_X^i}$ and $E_Y^j \in \mathbb{R}^{C_Y^j \times N_Y^j}$ represents the output features where C_X^i and C_Y^j indicate the feature dimension, and N_X^i and N_Y^j denote the structured dimension such as the scale of the flattened feature map or the number of nodes in the graph. To achieve orthogonality, no matter for

homogeneous features or heterogeneous features, we utilize feature transformation to unify their feature dimension. Note that for heterogeneous features, we take the dimensions of the structured feature as the target dimension. Here, C_X^i and C_Y^j are converted to the unified dimension C. By receiving these unimodal features, UFIM aims at learning an interactive intermediate embedding by taking advantage of the complementarity between multimodal features and making more sufficient feature-level fusion. In particular, UFIM mainly contains two parts: orthogonal attention and interactive feedback.

1) Orthogonal Attention: Different from homogenous features, heterogeneous features possess modality-specific structured information. Existing direct feature concatenation methods [1], [10] cannot handle the semantic gap between heterogeneous features, and will cause the loss of structured information. To this end, we interact E_X^i with E_Y^j to obtain the fine-grained attention map $\mathcal{M}_{X^iY^j}$ and then utilize a softmax function to normalize $\mathcal{M}_{X^iY^j}$ into a probability distribution, i.e.,

$$\mathcal{M}_{X^iY^j} = Softmax(\frac{{E_X^i}^\top \bigotimes E_Y^j}{\sqrt{C}}), \tag{1}$$

where $\bigotimes$ denotes the dot product operation and ${E_X^i}^\top$ corresponds to the transposition of E_X^i. Note that to prevent the vanishing gradient problem in $Softmax$ function, we here also employ a scale factor $\sqrt{C}$. After calculating $\mathcal{M}_{X^iY^j}$, we then feed back $\mathcal{M}_{X^iY^j}$ to E_X^i and E_Y^j to respectively obtain the fine-grained attention-based representations ${E_X^i}' \in \mathbb{R}^{C\times N_Y^j}$ and ${E_Y^j}' \in \mathbb{R}^{C\times N_X^i}$, i.e.,

$$\begin{cases} {E_X^i}' = \mathcal{M}_{X^iY^j} \bigotimes E_X^i \\ {E_Y^j}' = \mathcal{M}_{X^iY^j}^\top \bigotimes E_Y^j \end{cases} \tag{2}$$

The processes of obtaining ${E_X^i}'$ and ${E_Y^j}'$ can be regarded as orthogonal since they share one common attention map $\mathcal{M}_{X^iY^j}$.

2) Interactive feedback: The fine-grained attention-based representations are finally fed back to the original feature E_X^i and E_Y^j via interactive feedback so that original modalities can learn complementary information from other modalities. This mechanism can be achieved via weighted summation, i.e.,

$$\begin{cases} \widehat{E_X^i} = E_X^i + \alpha * {E_Y^j}' \\ \widehat{E_Y^j} = E_Y^j + \alpha * {E_X^i}' \end{cases} \tag{3}$$

where α represents the weight of the interaction process, and $\widehat{E_X^i}$ and $\widehat{E_Y^j}$ are the output features of UFIM which will be passed to higher layers of the network.

3) Generalization and discussion of UFIM: For fusing more than two modalities, we promote this method by first unifying the feature dimension of features and then utilizing the UFIM module to fuse unified features in a pairwise manner and finally learn an independent attention-based representation for each modality. UFIM, which can effectively learn the complementarity of homogeneous or heterogeneous features, make the integration of heterogeneous features with semantic gaps possible. By exploiting the orthogonal attention, the correlations between different modalities can be learned, so that the useful information of other heterogeneous modalities can be dynamically complemented into each original modality.

C. SCTM for redundancy

After sufficient feature interaction, the feature extractors finally output their high-level features as the output of a task-specific fusion module. Based on the analysis in the introduction, inevitable redundancy can significantly impact the semantic accuracy of multimodal representation. Therefore, we aim at improving multimodal fusion by simultaneously eliminating both irrelevant and repetitive information without substantially modifying the original network structure. Thus, we here propose SCTM to solve these problems by effectively utilizing both modality-common and modality-specific information. As illustrated in Figure 2, we assume that the high-level features extracted from E_X and E_Y are X' and Y'.

1) Modified transformer: As discussed in [15], the attention map can be used to guide complementarity among multiple modalities, hence alleviating irrelevant information. To this end, we here employ and modify the transformer mechanism [19] to achieve modal feature transfer, and denote this modified transformer as $\mathcal{T}$. Regarding $\mathcal{T}$, instead of using a self-attention, its embedded multi-head attention receives two modalities as the inputs. Specifically, to achieve modal feature transfer, we first define X' as Q (Query), Y' as K (Key) and V (Value), and utilize $\mathcal{T}$ to extract the information $\mathcal{M}'_X$. Notice here $\mathcal{M}'_X$ indicates the information of Y' that X' is interested in. Similarly, $\mathcal{M}'_Y$ denotes the information of X' that Y' is interested in. And these two pieces of information can be extracted by,

$$\begin{cases} \mathcal{M}'_X = \mathcal{T}(X' \to Y') = \mathcal{T}(X', Y', Y') \\ \mathcal{M}'_Y = \mathcal{T}(Y' \to X') = \mathcal{T}(Y', X', X') \end{cases} \tag{4}$$

In addition, to empower $\mathcal{T}$ the ability to distinguish different modalities, we additionally embed modal-identity information into traditional position embedding. Note that modal-identity information, which indicates the current in-processing main modality, is only determined by Q. Namely, the modal-identity information of Q, K, and V which are correspondingly inputted into $\mathcal{T}$ are the same. Since the dimension of different high-level modal features may be inconsistent, we transfer them to a consistent dimension before inputting them into the modified transformer by using the corresponding transformation layer $Func$. For example, regarding CNN features, we employ a convolutional layer. Based on this, Equation 4 can be revised as,

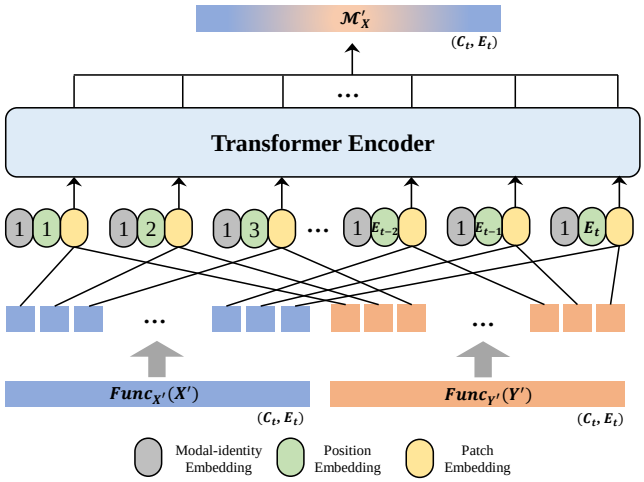

Fig. 3. Illustrated of the calculation of $\mathcal{M}'_X$ with $\mathcal{T}$. Feature X' and Y' are equally divided into E_t blocks, and each block is concatenated with the position embedding of this block and the modal-identity embedding of the current modality. Then, the two concatenated representations are sent into a transformer encoder to obtain the attention-based feature map $\mathcal{M}'_X$. Notice here we assume that X' is the first modality, so the modal identity is 1.

$$\begin{cases} \mathcal{M}'_X = \mathcal{T}(Func_{X'}(X'), Func_{Y'}(Y'), Func_{Y'}(Y')) \\ \mathcal{M}'_Y = \mathcal{T}(Func_{Y'}(Y'), Func_{X'}(X'), Func_{X'}(X')) \end{cases} \tag{5}$$

Notice that the detailed calculation of $\mathcal{M}'_X$ is shown in Figure 3 where the inputs are correspondingly revised. Here, we assume that the unified dimension of the features processed by $Func$ is $\mathbb{R}^{C_t \times E_t}$. We respectively feed the obtained attention-based feature maps, $\mathcal{M}'_X$ and $\mathcal{M}'_Y$, back to their original modal features by employing an elemnet-wise addition directly on an obtained feature map and its corresponding original feature, i.e.,

$$\begin{cases} \widehat{X'} = \mathcal{M}'_X + X' \\ \widehat{Y'} = \mathcal{M}'_Y + Y' \end{cases} \tag{6}$$

Here $\widehat{X'}$ and $\widehat{Y'}$ can be viewed as the corrected unimodal features. In this way, a current modality can perceive the areas of other modalities that it is interested in, and these areas can be regarded as common information of modalities. Once common information is perceived and fed back to a modality, the beneficial information in the original modal feature is then emphasized and the corresponding irrelevant information is then suppressed.

2) Fusion representation correction: Obviously, the corrected modal features, $\widehat{X'}$ and $\widehat{Y'}$, can be used as the input of most existing multimodal fusion networks. Without loss of generality, we here assume the fusion module $\mathcal{F}$ in Figure 2 is an existing task-specific network, and the fusion representation $\mathcal{P}$ can be obtained by,

$$\mathcal{P} = \mathcal{F}(\widehat{X'}, \widehat{Y'}) \tag{7}$$

As has been discussed in [6], even though irrelevant information has been extensively considered before fusion, existing fusion methods will still involve some repetitive information in the final fusion representation. To this end, we reuse the attention-based feature maps obtained by $\mathcal{T}$ to correct the final fusion representation. In particular, we calculate the elemnet-wise weighted average feature map of the attention-based feature maps of all modalities and use this feature map as the weights of fusion representation, hence balancing the modal-common and modal-specific information in the final representation, and suppressing the semantic bias caused by repetitive information. Thus, the calculation can be formulated by

$$\widehat{\mathcal{P}} = \mathcal{P} \odot Norm(\frac{\mathcal{M}'_X + \mathcal{M}'_Y}{2}), \tag{8}$$

where $\odot$ is the Hadamard product, and $Norm$ corresponds to the min-max normalization.

3) Generalization of SCTM: Considering that some multimodal fusion tasks may involve more than two modalities, we generalize SCTM to those common cases which include more than two modalities. Assuming that there are J modalities that need to be fused, and their transformed features as $\{\mathcal{Z}_i \in \mathbb{R}^{C_t \times E_t}\}_{i=1}^{J}$, and the calculation of the attention-based feature map of the ith modality can be written as,

$$\begin{aligned} \mathcal{M}_{\mathcal{Z}_i} &= \mathcal{T}(\mathcal{Z}_i \rightarrow \overline{Z_i}) \\ &= \mathcal{T}(\mathcal{Z}_i, Conc(\overline{Z_i}), Conc(\overline{Z_i})), \end{aligned} \tag{9}$$

where $Conc$ represents feature-wise concatenation, and $\overline{Z_i}$ indicates $\{\mathcal{Z}_1, \cdots, \mathcal{Z}_{i-1}, \mathcal{Z}_{i+1}, \cdots, \mathcal{Z}_J\}$. And the calculation of the final fusion representation can be calculated by,

$$\widehat{\mathcal{P}} = \mathcal{P} \odot Norm(\frac{\sum_{i=1}^{J} \mathcal{M}_{\mathcal{Z}_i}}{J}). \tag{10}$$

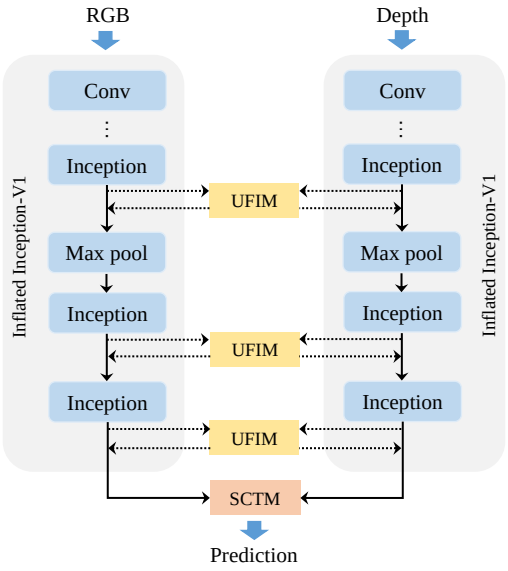

Fig. 4. An overview of the improved hand gesture recognition framework. It is improved from the two-stream model I3D [20], and the two branches are utilize to process RGB and depth modalities respectively. Each branch uses Inflated inception-V1 as the backbone. We insert UFIM after inception modules to interact the two modal information, and correct modal representations with SCTM.

IV. Experiments Results

In this section, we select four representative cross-domain tasks to demonstrate the effectiveness and generality of the proposed model, including gesture recognition, house price prediction, action recognition, and traffic accident forecast. Specifically, we insert UFIM and SCTM as plug-and-play components into state-of-the-art methods for unified feature interaction and redundancy mitigation. For each task, we first compare the proposed method with state-of-the-art methods on popular datasets. Then, we design corresponding ablation experiments to demonstrate the effectiveness and compatibility of the modules. Finally, we analyze the flexibility of modules and how to use them efficiently. Our model is implemented using PyTorch and trained on 8 NVIDIA Tesla V100 GPUs.

A. Task 1: Hand Gesture Recognition

Hand gesture recognition [1], [21] aims to recognize gesture categories from static images or videos, which is a typical visual classification task. Recently, many methods introduce additional modalities such as depth map [1], optical flow [21] and stereo-IR [22] into this task for joint decision-making. Here, we integrate SCTM and UFIM into an existing network and carefully evaluate the integrated network on the EgoGesture dataset [23], [24].

1) Dataset: EgoGesture dataset is a large-scale multimodal dataset for egocentric hand gesture recognition. It contains 2,081 RGB-D videos, 24,161 gesture samples and 2,953,224 frames from 50 distinct subjects. In this dataset, the videos are collected from 6 diverse indoor and outdoor scenes, including both RGB and depth modalities.

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**%

2) Baseline: VGG16+LSTM [25] is a combined method that uses VGG16 to process each frame individually and LSTM to process all frames to extract temporal information. C3D+LSTM+RSTTM [23] utilizes C3D and LSTM to process consecutive frames and introduces a novel recurrent spatiotemporal transformer. I3D [20] is a one-stream model that processes RGB or depth modalities independently. I3D late fusion [20] designs two I3D branches to process RGB and depth modalities respectively and fuse their results. MMTM [1] is an intermediate fusion method that proposes an efficient feature interaction module to fuse different modalities.

3) Implementation: For a fair comparison, we follow the setting of [1] and utilize I3D [20] as the fundamental network architecture when processing RGB maps and depth maps. As shown in Figure 4, I3D takes the inflated inception-V1 as the backbone, which can process time series and obtain spatiotemporal feature maps with four dimensions. In our implementation, we apply UFIMs to the last three inception modules by converting the channel dimensions and flattening the converted features. After the last inception of inflated inception-V1, SCTM is designed to alleviate the redundancy of input features instead of using average pooling.

4) Analysis: As shown in Table I, the improved interactive two-stream method outperforms other non-interactive methods when only UFIM is used to achieve modal interaction, which proves that properly feature interactions help improve the expressiveness of representations. Furthermore, the proposed model is comparable to the specialized homogeneous feature interaction-based method MMTM, which demonstrates the superior performance of UFIM in handling homogeneous feature interactions. When only SCTM is applied, compared with I3D late fusion, SCTM performs more effectively because the former directly concatenates the extracted unimodal features, ignoring the redundancy in the process, while the latter corrects the redundancy before and after feature concatenation, making the final representation more can represent the fused modalities. Although the performance of the model is improved when the proposed modules are used individually, the effect of using multiple modules jointly is still unknown, so we jointly apply UFIM and SCTM in the two-stream network, and the recognition accuracy of the integrated model outperforms the top performer MMTM by 1.09%, suggesting the proposed modules are mutually compatible.

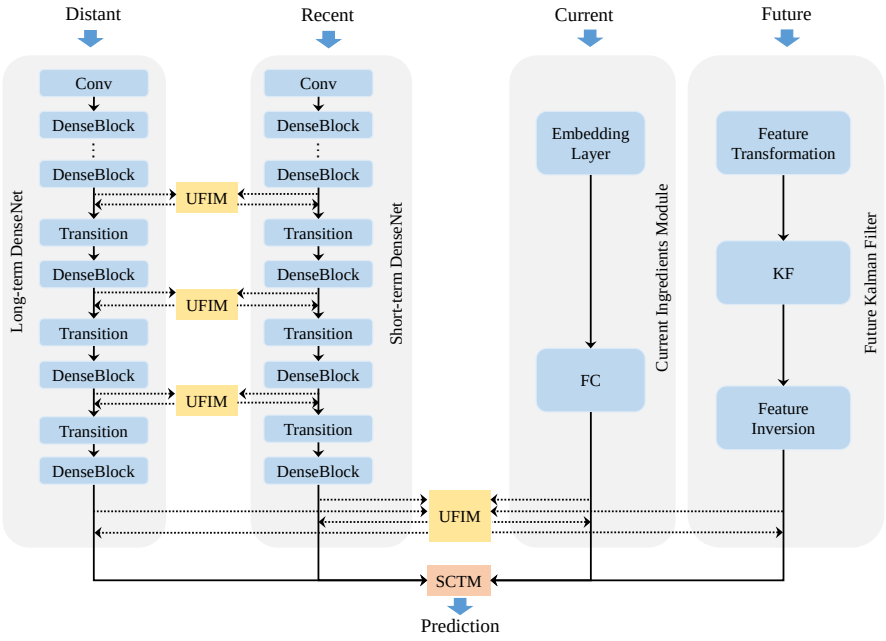

Fig. 5. Architecture of the improved JGC_MMN. The architecture consists of four branches, including long-term DenseNet, short-term DenseNet, current ingredients module, and future kalman filter. We apply UFIM hierarchically across branches.

B. Task 2: House Price Prediction

House price prediction methods generally utilize multi-source structured information to comprehensively predict sub-region house prices. Specifically, some methods take distance, recent, current, and future ingredients as multiple modalities to realize the fine-grained prediction. We improve an existing method JGC_MMN and evaluate them on the NYC and Beijing datasets [26].

1) Dataset: The house transaction price dataset of NYC is provided by NYC Open Data with a 13-year timespan. Current ingredients of NYC are obtained from the Federal Reserve Economic Data. The house transaction dataset of Beijing, which spans 7 years, is taken from Kaggle, and its current ingredients are provided on the website of the State Statistics Bureau. Both datasets contain statistical indicators from various fields.

2) Baseline: Deep-ST+C+F [27], ST-InceptionV4+C+F [28], and ST-ResNet+C+F [29] use Deep-ST, ST-InceptionV4, and ST-ResNet respectively to model spatiotemporal information, and other factors are also considered in the process of training. FTD_DenseNet [26] is a DenseNet-based framework that fully considers four time periods for depicting spatiotemporal dependencies. JGC_MMN [2] is a joint gated co-attention based multimodal networks for subregion house price prediction. MMTM+JGC [1] inserts the MMTM module into the JGC network to realize feature interaction.

3) Implementation: As shown in Figure 5, JGC_MMN believes that the influence factors subregion housing price can be generalized into four parts: long-term spatiotemporal correlations, short-term spatiotemporal correlations, current economic and social ingredients, and future price-growth expectations. It treats each factor as a modality and uses a four-branch network to extract each information separately, which contains two deep branches (DenseNet) and two shallow branches. Considering this, we plug UFIM into the last layer of the four branches and add three additional UFIMs to two DenseNet branches to ensure the interaction of the deep features. Finally, we improve the joint gated co-attention based fusion module with our SCTM.

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**

4) Analysis: From the quantitative comparison results shown in Table II, we find that the integrated variant respectively decreases the RMSE by 14.9% and 11.4% and the MAPE by 12.2% and 11.3% compared with JGC_MMN, which indicates that modal redundancy will significantly affect the performance of the fusion model. Further, UFIM-based JGC is better than MMTM-based JGC, mainly because MMTM tends to deal with semantically similar features, while deep features and shallow features have a significant semantic gap. Qualitatively, considering the trend correlation between two deep branches in JGC_MMN, feature interaction

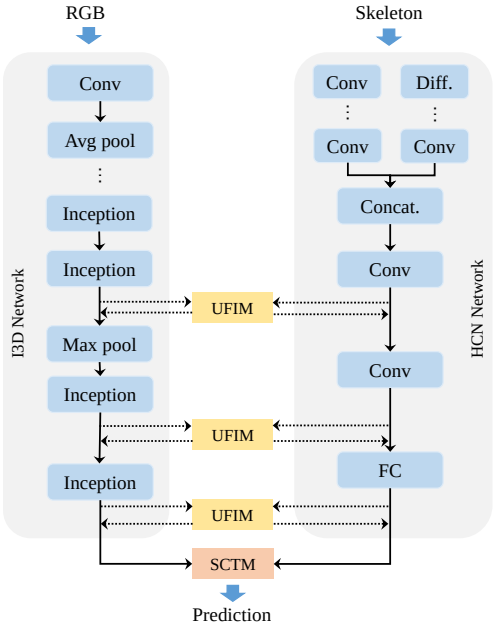

Fig. 6. Improved action recognition architecture based on I3D and HCN. We utilize I3D and HCN to process RGB and skeleton modalities respectively and introduce UFIMs to fuse the output feature maps of the Inception and Conv layers. Likewise, SCTM is used to improve a simple late fusion module.

in the intermediate layer is obviously beneficial to feature extraction. In addition, since the other two branches do not utilize deep extractors, some irrelevant information is extracted indiscriminately, but SCTM alleviates this problem by using the correlation information between multiple branches.

C. Task 3: Action Recognition

Action recognition is a popular computer vision task. State-of-the-art methods integrate action video with skeletons and depth maps to improve recognition accuracy. Here, we introduce two skeleton backbones and evaluate improved networks on the well-known NTU-RGBD dataset [30].

1) Dataset: NTU-RGBD is a large-scale multimodal action recognition dataset that contains 56,880 action samples captured from 40 subjects. The data form of each sample includes depth, 3D skeletons, RGB, and IR sequences. For a fair comparison, we follow the Cross-Subject (CS) evaluation and only use skeleton and RGB modalities.

2) Baseline: I3D [20] is a one-stream model for processing RGB modalities individually. HCN [31] and ST-GCN [32] are unimodal-based methods that utilize hierarchical aggregation and GCN to handle the skeleton modality. PoseMap [33] recognizes human actions as the evolution of pose estimation maps. I3D+HCN late fusion processes RGB and skeleton modalities with I3D and HCN respectively and fuses their outputs. SGM-Net [34] proposes the guided block to take full use of the complementarity of RGB and skeleton modalities at the feature level. MSAF [35] utilizes a sequential architecture exploration method to find the optimal multimodal fusion architecture. MMTM [1] is an intermediate fusion method

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%

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%**

that proposes an efficient feature interaction module to fuse different modalities.

3) Implementation: Similar to hand gesture recognition, we use I3D for the RGB video stream. Then, we employ two models for the skeletal stream, HCN and ST-GCN [32], which are based on CNN and GCN, respectively. Specifically, for HCN, according to [1], we add 3 UFIMs that receive inputs from the last three inception modules of I3D and conv5, conv6, and fc7 of HCN. For ST-GCN, we add 3 UFIMs between the last three G-Convs of ST-GCN and the last three inception modules of I3D. In both cases, SCTM is utilized to fuse features and alleviate redundancy.

4) Analysis.: As shown in Table III, whether HCN or ST-GCN is used as the skeleton model, our UFIM consistently outperforms the intermediate fusion model MMTM, which proves that the proposed unified feature interaction is suitable for homogeneous and heterogeneous features. Qualitatively, the interaction module of MMTM is very suitable for processing homogeneous features, but when ST-GCN is utilized as the baseline, its result is not ideal because the structural information of heterogeneous features is ignored. Similar to gesture recognition, the RGB branch brings irrelevant information into the fusion process, resulting in two kinds of redundancy in fusion representation. Our proposed modules retain structured information and explicitly use an improved attention mechanism to suppress unnecessary redundant information, which is more suitable for unified multimodal fusion.

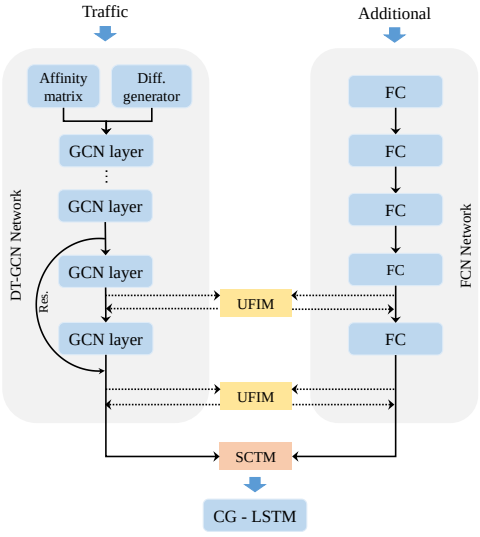

Fig. 7. Architecture of interaction-based spatiotemporal feature map extraction network. The architecture consists of two branches, the DT-GCN branch is used to extract spatiotemporal features from traffic data, and the FCN branch is used to model additional information. We embed UFIMs at the tail of the network to enable heterogeneous feature interactions. For the outputs of the two branches, we utilize SCTM to correct them and feed them into the CG-LSTM module.

D. Task 4: Traffic Accident Forecast

Traffic accident forecast typically uses multimodal heterogeneous information collected from multiple sensors, such as road structure, traffic, and weather, to forecast whether an accident will occur within a period of time [39]. Based on a state-of-the-art model RiskSeq [4], we propose several variants and validate them on NYC and SIP datasets.

1) Dataset: NYC Opendata contains modalities such as accidents, speed values, weather, demographics, and road network, which utilizes the taxi trip volumes in each subregion as the indicator of human mobilities. SIP dataset contains traffic flows and speeds, which is integrated with another traffic accident dataset collected from Microblog.

2) Baseline: STDN [36] proposes the flow gating and shifted attention to jointly model volume and flow interactions. DFN [37] combines a hierarchical recurrent structure with a context-aware embedding module to perform daily accident prediction. STSGCN [38] captures localized spatiotemporal correlations and heterogeneities with a synchronous network. RiskSeq [4] foresees sparse urban accidents with finer granularities and multiple steps in a spatiotemporal perspective. MMTM+RiskSeq [1] uses MMTM modules to exchange information between the features of the two modalities.

3) Implementation: Our improved models are based on RiskSeq. After necessary data preprocessing, we employ DT-GCN to process traffic information to obtain fine-grained risk feature maps and coarse-grained sequences, respectively. Then we concatenate all additional data such as weather and time, and send them to a five-layer fully connected network (FCN) to obtain a context embedding. Then, We apply UFIM after the last two layers of DT-GCN and FCN. Finally, for the outputs of the two branches, we utilize SCTM to correct them and feed them into the CG-LSTM module.

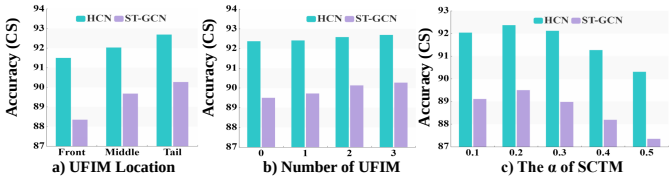

Fig. 8. Comparison of varieties on NTU-RGBD dataset. a) Explore the optimal insertion Location of UFIM. b) Explore the effect of the number of UFIM on the model. c) Explore the hyperparameter settings of SCTM.

4) Analysis: Table IV illustrates the quantitative comparisons on NYC and SIP datasets. Encouragingly, our integrated method UFIM+SCTM achieves the highest accuracy and outperforms the top performer MMTM+RiskSeq, which means that the proposed module can effectively improve the performance of the prediction model. From the perspective of fusion, Riskseq utilizes additional factors as an independent modality. However, these factors are partially related to spatiotemporal features, so it will inevitably lead to redundancy. SCTM enables spatiotemporal features and additional factors to understand each other to suppress irrelevant information and make the fusion representation more accurate.

E. Discussion of the UFIM and SCTM

In this subsection, we explore the flexibility of modules and how to use them efficiently on NTU-RGBD.

1) Location of the interaction: Like other intermediate fusion methods, our UFIM can theoretically be applied to different feature extraction layers of the model, which usually depends on the backbone used for the specific task. To this end, based on the implementation details of the action recognition task, we plug UFIMs into different locations of the learning pipeline to observe fluctuations in model performance. As shown in Figure 8 a), we find that the optimal location for module insertion is the tail layer of the model, which is consistent with the conclusions of previous methods [1], [40]. The reason is that the semantic information contained in the feature maps extracted by different feature layers of the model is different, and the high-level features often have richer semantic features and stronger inter-modality correlation than the low-level features.

2) Number of the interaction: The number of interaction modules also affects the performance of the model. Based on the optimal insertion location, we gradually insert interaction modules from the tail of the model. The experimental results in Figure 8 b) show that sufficient interaction between high-level features is beneficial for the model to learn semantically richer multimodal representations. Therefore, we recommend inserting as many UFIMs as possible at the tail of the model, while avoiding inserting more than one-third of the entire model layers.

3) Selection of weight parameters: In the interactive feedback stage of SCTM, the weight α is used to control the influence of the fine-grained attention-based representation on

the original representation. Figure 8 c) shows the effect of weight changes on model performance. We find that a too large weight will damage the original semantic information, while a too small weight will not feed back the inter-modality correlation to each modality. To better stimulate the potential of SCTM, we recommend setting α at about 0.2 and fine-tuning it according to the actual task.

V. Conclusion

In this paper, 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.

VI. Acknowledgments

This paper is partially supported by the Anhui Science Foundation for Distinguished Young Scholars (No.1908085J24), Natural Science Foundation of China (No.62072427), and the Project of Stable Support for Youth Team in Basic Research Field, CAS (No.YSBR005).

References

[1] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "Mmtm: Multimodal transfer module for cnn fusion," in *CVPR*, 2020, pp. 13 289–13 299.

[2] P. Wang, C. Ge, Z. Zhou, X. Wang, Y. Li, and Y. Wang, "Joint gated co-attention based multi-modal networks for subregion house price prediction," *TKDE*, 2021.

[3] X. Bruce, Y. Liu, and K. C. Chan, "Multimodal fusion via teacher-student network for indoor action recognition," in *AAAI*, vol. 35, no. 4, 2021, pp. 3199–3207.

[4] Z. Zhou, Y. Wang, X. Xie, L. Chen, and C. Zhu, "Foresee urban sparse traffic accidents: A spatiotemporal multi-granularity perspective," *TKDE*, 2020.

[5] P. Sun, W. Zhang, H. Wang, S. Li, and X. Li, "Deep rgb-d saliency detection with depth-sensitive attention and automatic multi-modal fusion," in *CVPR*, 2021, pp. 1407–1417.

[6] K. Zhou, C. Chen, B. Wang, M. R. U. Saputra, N. Trigoni, and A. Markham, "Vmloc: Variational fusion for learning-based multimodal camera localization," in *AAAI*, vol. 35, no. 7, 2021, pp. 6165–6173.

[7] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *AAAI*, vol. 33, no. 01, 2019, pp. 3558–3565.

[8] J. Chen and A. Zhang, "Hgmf: heterogeneous graph-based fusion for multimodal data with incompleteness," in *KDD*, 2020, pp. 1295–1305.

[9] F. R. Valverde, J. V. Hurtado, and A. Valada, "There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge," in *CVPR*, 2021, pp. 11 612–11 621.

[10] J. Nie, J. Yan, H. Yin, L. Ren, and Q. Meng, "A multimodality fusion deep neural network and safety test strategy for intelligent vehicles," *TIV*, vol. 6, no. 2, pp. 310–322, 2020.

[11] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," in *NeurIPS*, 2020.

[12] X. Long, C. Gan, G. De Melo, X. Liu, Y. Li, F. Li, and S. Wen, "Multimodal keyless attention fusion for video classification," in *AAAI*, 2018.

[13] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *NeurIPS*, 2014, pp. 2204–2212.

[14] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *ACL*, 2018.

[15] S. Mohla, S. Pande, B. Banerjee, and S. Chaudhuri, "Fusatnet: Dual attention based spectrospatial multimodal fusion network for hyperspectral and lidar classification," in *CVPR*, 2020, pp. 92–93.

[16] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *ICCV*, 2017, pp. 2612–2620.

[17] M. Hou, J. Tang, J. Zhang, W. Kong, and Q. Zhao, "Deep multimodal multilinear fusion with high-order polynomial pooling," *NeurIPS*, vol. 32, pp. 12 136–12 145, 2019.

[18] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *ICCV*, 2017, pp. 4193–4202.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[20] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017, pp. 6299–6308.

[21] M. Abavisani, H. R. V. Joze, and V. M. Patel, "Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training," in *CVPR*, 2019, pp. 1165–1174.

[22] P. Gupta, K. Kautz *et al.*, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks," in *CVPR*, vol. 1, no. 2, 2016, p. 3.

[23] C. Cao, Y. Zhang, Y. Wu, H. Lu, and J. Cheng, "Egocentric gesture recognition using recurrent 3d convolutional neural networks with spatiotemporal transformer modules," in *ICCV*, 2017, pp. 3763–3771.

[24] Y. Zhang, C. Cao, J. Cheng, and H. Lu, "Egogesture: a new dataset and benchmark for egocentric hand gesture recognition," *TMM*, vol. 20, no. 5, pp. 1038–1050, 2018.

[25] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *CVPR*, 2015, pp. 2625–2634.

[26] C. Ge, Y. Wang, X. Xie, H. Liu, and Z. Zhou, "An integrated model for urban subregion house price forecasting: A multi-source data perspective," in *ICDM*. IEEE, 2019, pp. 1054–1059.

[27] J. Zhang, Y. Zheng, D. Qi, R. Li, and X. Yi, "Dnn-based prediction model for spatio-temporal data," in *SIGSPATIAL/GIS*, 2016, pp. 1–4.

[28] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017.

[29] J. Zhang, Y. Zheng, and D. Qi, "Deep spatio-temporal residual networks for citywide crowd flows prediction," in *AAAI*, 2017.

[30] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *CVPR*, 2016, pp. 1010–1019.

[31] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," in *IJCAI*, 2018.

[32] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[33] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *CVPR*, 2018, pp. 1159–1168.

[34] J. Li, X. Xie, Q. Pan, Y. Cao, Z. Zhao, and G. Shi, "Sgm-net: Skeleton-guided multimodal network for action recognition," *PR*, vol. 104, p. 107356, 2020.

[35] L. Su, C. Hu, G. Li, and D. Cao, "Msaf: Multimodal split attention fusion," *arXiv preprint arXiv:2012.07175*, 2020.

[36] H. Yao, X. Tang, H. Wei, G. Zheng, and Z. Li, "Revisiting spatial-temporal similarity: A deep learning framework for traffic prediction," in *AAAI*, vol. 33, no. 01, 2019, pp. 5668–5675.

[37] C. Huang, C. Zhang, P. Dai, and L. Bo, "Deep dynamic fusion network for traffic accident forecasting," in *CIKM*, 2019, pp. 2673–2681.

[38] C. Song, Y. Lin, S. Guo, and H. Wan, "Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting," in *AAAI*, vol. 34, no. 01, 2020, pp. 914–921.

[39] P. Wang, C. Zhu, X. Wang, Z. Zhou, G. Wang, and Y. Wang, "Inferring intersection traffic patterns with sparse video surveillance information: An st-gan method," *TVT*, 2022.

[40] F. Li, N. Neverova, C. Wolf, and G. Taylor, "Modout: Learning multi-modal architectures by stochastic regularization," in *FG*. IEEE, 2017, pp. 422–429.