DIN: Dual Impulse Network for Multi-View Representation Learning

Yilin Wu[1], **Weihong Lin** [1], **Renjie Lin** [2], **Zihan Fang**[1], **Shide Du**[1], **Shiping Wang**[1]*

[1] College of Computer and Data Science, Fuzhou University, Fuzhou, China
[2] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China
yinnwu510@gmail.com, whlin192@163.com, renjielin1993@gmail.com
fzihan11@163.com, dushidems@gmail.com, shipingwangphd@163.com

Abstract

Multi-view representation learning, which utilizes multiple channels to improve perceptual accuracy, is recognized for its effectiveness in the analysis of multi-view data. However, deploying these methods in real-world scenarios presents two primary challenges. 1) Lack of Variegation: Multi-view representation techniques commonly observe along a singular axis, *i.e.*, the attribute axis; 2) Insufficient Relationship: Most multi-view models lack mechanisms for exploring potential relationships between attribute axis and channel axis. To mitigate these obstacles, we design a Dual Impulse Network framework for multi-view representation learning (DIN) to train a feature representation. In this framework, a strategy observed along the channel axis and attribute axis simultaneously is introduced, and two different representations are generated by two analogous impulse networks, which are capable of extracting information corresponding to different axes. Furthermore, we incorporate an integration network that analyzes the potential relationship between attribute axis and channel axis to generate two attention matrices. The final two feature representations derived from these attention matrices are aggregated to amplify the expression of internal information. Comprehensive experimental results support the efficacy and superiority of the proposed framework, demonstrating improvements in classification performance compared to state-of-the-art methods.

Code — https://github.com/wyl510fz/DIN_AAAI_2026

Introduction

Recent developments in multimedia technologies have fundamentally transformed human perceptual processes and established a basis for sophisticated analytical paradigms. Multi-view learning has emerged as a critical domain of artificial intelligence for addressing multi-view data, since it emphasizes the integration and exploitation of diverse data to improve accuracy and robustness. This learning paradigm acquires data pertaining to real-world objects from various channels, leveraging commonalities cross-channels to improve efficacy in applications such as computer vision (Hwang et al. 2021; Ning et al. 2024; Bao and Lu 2024), text processing (Chen et al. 2022; Song et al. 2024; Fang et al. 2024a), large-scale language models (Guo et al. 2023; Liu et al. 2025b; Zheng 2025), and other fields (Ye and Li 2024; Hu et al. 2024; Du et al. 2025).

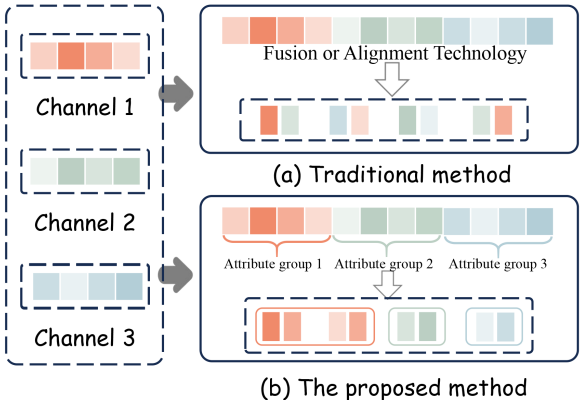


Figure 1: Difference between traditional method and proposed framework in observing datasets.

Among them, multi-view representation learning synthesizes diverse feature spaces and extracts common insights from different channels to obtain a final representation. Many methods leverage diverse technologies to explore the information between attributes and channels, involving fusion (Wang et al. 2025a; Chen et al. 2025a; Mao et al. 2025; Chen et al. 2025b) and alignment technology (Wang et al. 2024b; Liu et al. 2024b; Wang et al. 2024a; Luo et al. 2025). These methods demonstrate superior capability in extracting representative features and exploring the correlations among different channels. However, these approaches encounter two principal challenges. **Challenge I: Lack of Variegation.** As illustrated in Fig. 1, traditional multi-view learning methods observe from the attribute axis, *i.e.*, consider attributes as separate entities, which limits the exploration of information among attribute groups. However, if observing along the channel axis, the attributes within each channel can be treated as multiple attribute groups. This strategy will improve the identification of relationships among attribute groups. **Challenge II: Insufficient Relationship.** Current frameworks employ a feature representation with its specific information. Their capacity to model and elucidate the relationship between different feature representations is

*Corresponding author.

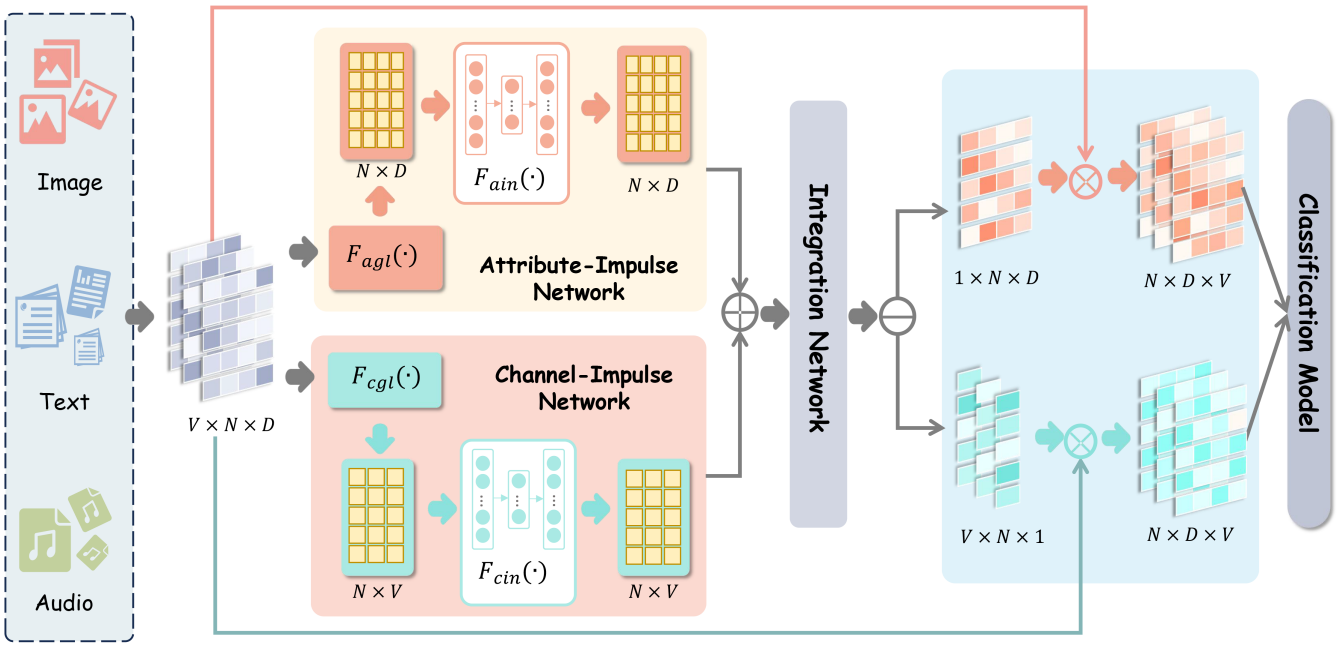


Figure 2: An overview of the proposed framework, which consists of two impulse networks and integration network. These two impulse networks generate two different feature representations from attribute axis and channel axis, respectively. While the integration network is responsible for fusion between these two representations to construct attention weight matrices.

limited. Consequently, it is difficult for them to generalize a representation that summarizes and preserves the majority of informative information across views, thereby restricting the expressive power of the learned features. This issue results in the inadequate modeling and weak generalization ability of multi-view feature representations from these methods.

In light of these challenges, we propose an effective multi-view framework called Dual Impulse Network for multi-view representation learning (DIN). Specifically, DIN begins with feeding multi-view data into two impulse networks from attribute axis and channel axis, respectively. These two networks perform dimensionality reduction and impulse to enhance their expression with two novel feature representations. Besides, DIN devises an integration network to interact with these two feature representations. The interaction between these feature representations facilitates the generation of more precise attention weight matrices. For better prediction results, DIN executes the integration of the attention matrix and input feature space to create the final representation. Finally, the learned representation is integrated into a classification model for training. The main contributions of this work are summarized as below:

- We introduce a perspective of observation from the channel axis and attribute axis simultaneously, and design two analogous impulse networks for the attribute axis and channel axis, respectively.
- We propose an interaction network to model the potential relationships between attribute representation and channel representation, resulting in the aggregation of comprehensive information.
- The proposed framework is applied to multi-view semi-supervised classification tasks on eight widely used datasets, demonstrating superior performance relative to existing state-of-the-art algorithms.

Related Work

In this section, we present an introduction of the pertinent work on this paper, including multi-view representation learning and multi-view attention learning.

Multi-View Representation Learning

Multi-view representation learning is concerned with the problem of learning representations that facilitate extracting valuable information (Li, Yang, and Zhang 2018; Hassani and Khasahmadi 2020; Lin et al. 2022; Zheng et al. 2023). In brief, this paradigm leverages additional information to achieve better performance. Since the performance is heavily related to the expressive power of representation, it has emerged as a highly promising topic with wide applicability (Huang, Liang, and Jia 2024; Lin et al. 2025b). For example, (Xie et al. 2021) preserved the local geometric structure and incorporated hyper-Laplacian regularization into feature representation space. (Chen et al. 2023a) implemented a differentiable node selection mechanism to augment the graph fusion module's capacity, thereby improving the expressive power of the resulting representation. Recently, (Liu et al. 2025a) implemented a quality-aware sub-network to dynamically assign quality scores. (Fu et al. 2025) obtained a consistent representation with a global coordinate form anchors. (Fang et al. 2025a) incorporated uncertainty calibration and

view-wise debiasing mechanism to learn generalizable features in multi-view open-set learning. However, these approaches only focus on the attribute axis in each view and ignore the premise that additional information is present in other axes, thereby limiting their ability to derive comprehensive representations. In our framework, both attribute axis and channel axis are considered to facilitate comprehensive information extraction.

Multi-View Attention Learning

Attention technology can be considered as a mechanism that selectively modulates the distribution of computational resources, enhancing processing emphasis on the most salient features. Owing to its advantageous properties, numerous research integrates it with multi-view learning to develop multi-view attention learning (Yao et al. 2022; Liu et al. 2024a; Lee et al. 2025). The benefits of multi-view attention learning have been shown across a range of multi-view fields, from disease treatment (Sun et al. 2024; Ouyang et al. 2024) to civil engineering (Li et al. 2023; Ma et al. 2025). Some notable works, for example, (Ma et al. 2023) introduced a correspondence attention block designed to selectively amplify pertinent feature subsets. (Yang et al. 2024) designed a multi-view consistency block that facilitated the exchange of information across multiple single-view diffusion processes. (Lu et al. 2024) established a graph diffusion to derive the attention matrix with multi-view consistent information. (Wang et al. 2025b) considered the multi-view fusion as a decision making problem and constructed a pairwise integrator to merge the feature with the selected view. However, these attention methods primarily analyze the relationship between channels or attributes and ignore the potential relationship between channels and attributes. Our framework significantly designs an integration network to facilitate integration between the two different representations and produce attention matrices.

The Proposed Method

In this section, we present a general framework which introduces two different impulse networks to capture information along attribute axis and channel axis, respectively. Besides, we aggregate them together to obtain the relationship between channel and attribute for generating a feature representation with strong ability.

Overview

Since an instance is generated through multiple sources, the information corresponding to different sources sharing the same label may exhibit differences. These differences induce intra-class inconsistency, thereby impairing recognition accuracy. We leverage squeeze-excitation mechanism from attribute axis and channel axis, respectively. Then two novel feature representations with corresponding information are intermixed with each other for modeling attention weight matrices. Third, an element-wise addition is executed between the above attention weight matrices and original features to obtain the final representation reflecting attribute and channel information. Finally, the learned feature representation and an adjacency matrix derived from the original feature space serve as dual inputs to the classification model.

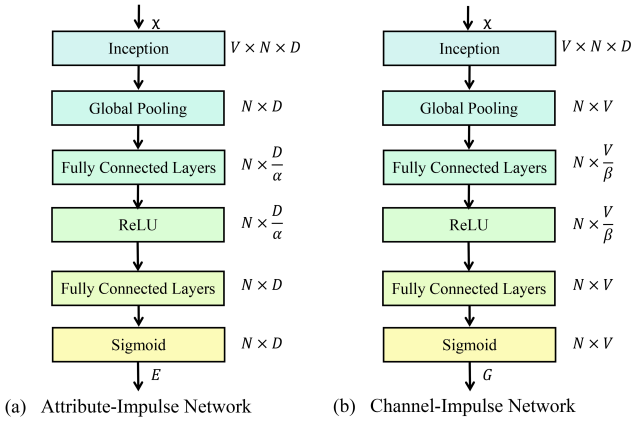


Figure 3: Explanation of the proposed impulse networks.

Specifically, divergent feature space dimensions pose challenges in subsequent processing modules. To address this issues, we apply a feature transformation to convert them into the same dimension D. Then we concatenate these feature spaces to obtain a novel $V \times N \times D$ feature space for processing, where N, V, and D represent the number of instances, views, and features, respectively. As illustrated in Fig. 3, two different impulse networks are designed to draw information from attribute axis and channel axis generated by squeezing appropriate direction, thus obtaining better feature representations for label prediction.

Attribute-Impulse Network (AIN)

Discriminant feature representations are fundamental for multi-view learning and can be derived through the aggregation of attribute information. However, the observation from many works (Liu et al. 2013; Lin et al. 2025a) suggests that the internal information can preserve information without the nonlinear mapping, thereby contributing to the classification of instances. To effectively capture the internal information in feature space, an attribute-excitation mechanism is proposed. AIN encodes a wider range of information into feature representation, thus augmenting their descriptive capacity. Next, we elaborate on the process of aggregating internal information.

As illustrated in Fig. 3(a), given a multi-view feature space $\mathcal{X} \in \mathbb{R}^{V \times N \times D}$, we squeeze channel information into an attribute descriptor. This is accomplished with the application of global average pooling to compute attribute descriptors. Formally, a feature space $\mathbf{B} \in \mathbb{R}^{N \times D}$ is generated by shrinking channel dimension V, where the $(n, d)-$th element of $\mathbf{B}$ is calculated by:

$$\mathbf{B}_{n,d} = F_{agl}(\mathcal{X}) = \frac{1}{V} \sum_{v=1}^{V} \mathcal{X}(v, n, d). \tag{1}$$

The output $\mathbf{B}$ can be interpreted as an attribute collection

of each channel whose statistics are expressive for the information of whole attributes.

A squeeze-excitation operation is performed to comprehensively capture the attribute information, effectively utilizing the information aggregated during the squeeze operation. For this objective, the function should satisfy two criteria: it must exhibit flexibility, specifically requiring the capacity to model nonlinear interactions among input channels; at the same time, it must emphasize the high-quality features while eliminating the low-quality ones, thereby improving the quality and discriminative capacity of the novel feature representation. To satisfy these requirements, a fundamental function is implemented for obtaining the output $\mathbf{E} \in \mathbb{R}^{N \times D}$ as follows:

$$\mathbf{E} = F_{ain}(\mathbf{B}, \mathbf{W}_b) = f_{aex}(f_{asq}(\mathbf{B}, \mathbf{W}_{b1}), \mathbf{W}_{b2}), \quad (2)$$

where f_{asq} denotes a squeeze network, and f_{aex} represents an impulse network with sigmoid function. Here, $\mathbf{W}_{b1} \in \mathbb{R}^{D \times \frac{D}{\alpha}}$, $\mathbf{W}_{b2} \in \mathbb{R}^{\frac{D}{\alpha} \times D}$, and α is the hyperparameter for attribute dimension reduction ratios. Specifically, the global pooling executes an initial dimensionality reduction of channel dimension and fails to achieve comprehensive extraction of the internal information among attributes. Consequently, additional compression techniques are necessary to achieve feature extraction. To achieve condensing internal information, the formula of f_{asq} is defined as:

$$f_{asq}(\mathbf{B}, \mathbf{W}_{b1}) = ReLU(\varphi_a(\mathbf{B}, \mathbf{W}_{b1})), \quad (3)$$

where φ_a denotes the fully connected network. With Eq. (3), the internal information is condensed to construct an appropriate foundational basis for the following activation process. Then $ReLU(\cdot)$ function is executed on it, whereby each attribute is activated only if it exceeds threshold parameter. After the squeeze process, it is necessary to flexibly restore the original dimension and activate the information. For this purpose, the formula of f_{aex} is defined as follows:

$$f_{aex}(\mathbf{B}', \mathbf{W}_{b2}) = Sigmoid(\zeta_a(\mathbf{B}', \mathbf{W}_{b2})), \quad (4)$$

where $\mathbf{B}' \in \mathbb{R}^{N \times \frac{D}{\alpha}}$ is the feature space after attribute compression, and ζ_a also represents another fully connected network. $Sigmoid(\cdot)$ ensures the non-negativity of the feature representation. From above processes, it can be inferred that the resultant attributes feature representation matrix $\mathbf{E}$ is computed across all channels. Besides, this network employs a comprehensive attribute aggregation mechanism, selectively integrating attribute information. In this network, similar features mutually reinforce each other, thereby enhancing semantic consistency.

Channel-Impulse Network (CIN)

In multi-view learning, each view is regarded as a channel. Existing methods analyze multi-view data along the attribute axis, *i.e.*, considering each attribute as an individual entity, which can effectively leverage the information internal to each attribute. However, this strategy introduces a new challenge, failing to take into account the problem that attributes are from different channels. While the attributes within a channel are naturally a group. Within an attribute group, certain attributes function as synergistic components, exhibiting limited contribution alone but manifesting a multiplicative enhancement in a group. Investigating this potential relationship, higher-quality attribute groups can be identified, thereby allowing more comprehensive retrieval of essential information. Based on the above rationale, we aim to conduct the impulsing mechanism along channel axis, considering each channel as an attribute group to amplify the contributions of high-quality channels and augment the representational capacity of their attributes. Therefore, a channel-impulse network is constructed to explicitly capture the information among attribute groups. The structure of CIN is illustrated in Fig. 3(b). Different from AIN, CIN squeezes attribute information into a channel descriptor. The channel representation matrix $\mathbf{C} \in \mathbb{R}^{N \times V}$ is calculated from original feature $\mathcal{X}$. Specifically, this network reshapes the original feature $\mathcal{X}$ to $\mathbf{C}$ by squeezing the attribute dimension D. In this process, the attributes in each channel are aggregated as an attribute group into channel vectors. The $(n, v)-$th element of $\mathbf{C}$ is computed by:

$$\mathbf{C}_{n,v} = F_{cgl}(\mathcal{X}) = \frac{1}{D} \sum_{d=1}^{D} \mathcal{X}(v, n, d). \quad (5)$$

To prioritize the salient attribute group through comparison of differentials across multiple channels, a function similar to AIN is defined as follows:

$$\mathbf{G} = F_{cin}(\mathbf{C}, \mathbf{W}_c) = f_{cex}(f_{csq}(\mathbf{C}, \mathbf{W}_{c1}), \mathbf{W}_{c2}), \quad (6)$$

where $\mathbf{G} \in \mathbb{R}^{N \times V}$ is the channel feature representation after squeeze-excitation, $\mathbf{W}_{c1} \in \mathbb{R}^{V \times \frac{V}{\beta}}$, $\mathbf{W}_{c2} \in \mathbb{R}^{\frac{V}{\beta} \times V}$, and β denotes the hyperparameter of channel dimension reduction ratios.

The detailed formulas of f_{csq} and f_{cex} are introduced below. First, to make the relationship among attribute groups be reflected in the process of squeezing, f_{csq} is defined as:

$$f_{csq}(\mathbf{C}, \mathbf{W}_{c1}) = ReLU(\varphi_c(\mathbf{C}, \mathbf{W}_{c1})), \quad (7)$$

where φ_c denotes the fully connected network. This function amplifies the significant attribute group and limits the inconsiderable one. Subsequently, $ReLU(\cdot)$ is employed to preserve strong attribute groups while eliminating the low ones.

To achieve a similar purpose to f_{aex}, f_{cex} is defined as follows:

$$f_{cex}(\mathbf{C}', \mathbf{W}_{b2}) = Sigmoid(\zeta_c(\mathbf{C}', \mathbf{W}_{b2})), \quad (8)$$

where $\mathbf{C}' \in \mathbb{R}^{N \times \frac{V}{\beta}}$ is the feature space after channel compression, and ζ_c denotes another fully connected network. Consequently, the channel feature representation $\mathbf{G}$ is computed from CIN. Essential attribute groups contribute additional descriptive information, thereby augmenting the discriminative capacity of channel information.

Integration Network

Since existing methods have not considered observing from a channel axis before, the research about the interaction between attribute axis and channel axis remains limited. In

order to explore the information between attribute axis and channel axis, we aggregate the feature representations from these two networks. Specifically, we concatenate the outputs of two networks and perform an integration network to accomplish feature fusion. Both feature representations are jointly input into an integration network M_i to produce attention weight matrix. The integration network is composed of multi-layer perceptrons (MLPs) with one hidden layer. Concisely, the general attention is computed as follows:

$$\begin{aligned} M_i(\mathbf{G}, \mathbf{E}) &= \sigma(MLP(\phi(\mathbf{E}, \mathbf{G}))) \\ &= \sigma(\mathbf{W}_1(\mathbf{W}_0(\phi(\mathbf{E}, \mathbf{G})))), \end{aligned} \tag{9}$$

where σ denotes the sigmoid function, and $\phi(\cdot)$ represents the concatenation function. Note that the weight matrices $\mathbf{W}_0$ and $\mathbf{W}_1$ of the MLP are shared across both feature representation spaces, with the $ReLU$ activation function applied subsequent to $\mathbf{W}_0$. With the above specified operation, the channel information can interact with the attribute information. After the attribute feature representation and channel feature representation have fully interacted with each other, attribute attention $\mathbf{E}' \in \mathbb{R}^{N \times D \times 1}$ and channel attention $\mathbf{G}' \in \mathbb{R}^{N \times 1 \times V}$ are derived through the partitioning of the output feature space along their respective dimension.

Following the application of the shared network to each feature representation, the resulted feature vectors are combined with element-wise addition. The processes of attention matrices to born novel feature representation are summarized respectively:

$$\begin{aligned} \mathbf{T}_a &= \psi_e(\mathbf{E}') \otimes \mathcal{X}, \\ \mathbf{T}_c &= \psi_g(\mathbf{G}') \otimes \mathcal{X}, \end{aligned} \tag{10}$$

where ψ_e, ψ_g are the functions which contain a broadcasting and reshape the corresponding attention matrices into $V \times N \times D$, $\mathbf{T}_a, \mathbf{T}_c \in \mathbb{R}^{V \times N \times D}$, and $\otimes$ denotes element-wise multiplication. Finally, in order to obtain the final prediction, these feature representations are processed with classification model as follows:

$$\widehat{\mathbf{Y}} = g(\theta(AvgPool(\mathbf{T}_a), AvgPool(\mathbf{T}_c)), \mathbf{A}, \mathbf{W}_g), \tag{11}$$

where $\widehat{\mathbf{Y}} \in \mathbb{R}^{N \times L}$ is the predicted label of classification model, $g(\cdot)$ indicates the learnable classification model, $\mathbf{A}$ is the adjacency matrix by using a traditional method in the processed feature space, $\theta(\cdot)$ denotes a function of stitching process, $AvgPool(\cdot)$ conducts the function of compressing feature representation into $N \times D$, and $\mathbf{W}_g$ denotes the learnable matrix of the classification model. We summarize the complete procedure of DIN in Algorithm 1.

Algorithm 1: DIN

Require: Multi-view data $\mathcal{X} \in \mathbb{R}^{V \times N \times D}$, α, and β.
Ensure: $\widehat{\mathbf{Y}}$: Predictive labels.

```
Initialize adjacency matrix A form X;
while not convergent do
    Generalize the feature space B and feature representation E by Eq. (1) and Eq. (2);
    Produce the feature space C_{n,v} and feature representation G by Eq. (5) and Eq. (6);
    Integrate feature representations E and G by Eq. (9);
    Separate the attention matrix derived through M_share into two attention weight matrices E' and G';
    Calculate the according feature representations T_a and T_c by Eq. (10);
    Input the feature representations T_a, T_c, and A to train the W_g of classification model by Eq. (11);
    Optimize the trainable parameters;
end while
Acquire the predictive label Ŷ from the trained model;
return Ŷ as finally result.
```

Experiments and Studies

Datasets. Eight multi-view datasets, extensively utilized across application domains, are employed to quantitatively assess the performance of DIN. Table 1 briefly illustrates the pivotal information of all eight datasets.

Compared Methods. We compare the proposed framework with classical and state-of-the-art baseline models, including, Co-GCN (Li, Li, and Wang 2020), LGCNFF (Chen et al. 2023b), PDMF (Xu et al. 2023), ECMGD (Lu et al. 2024), ORLNet (Fang et al. 2024b), RCML (Xu et al. 2024), and SMMRL (Fang et al. 2025b).

Datasets	# Samples	# Views	#Feature	# Classes
ALOI	1,079	4	64/64/77/13	10
BBCSports	544	2	3,183/3,203	5
Flickr	12,154	2	100/100	7
IAPR	7,855	2	100/100	6
MNIST	10,000	3	30/9/9	10
NUS-WIDE	1,600	6	64/144/73/128/225/500	8
UCI	2,000	3	240/76/6	10
WebKB-texas	187	2	187/1,703	5

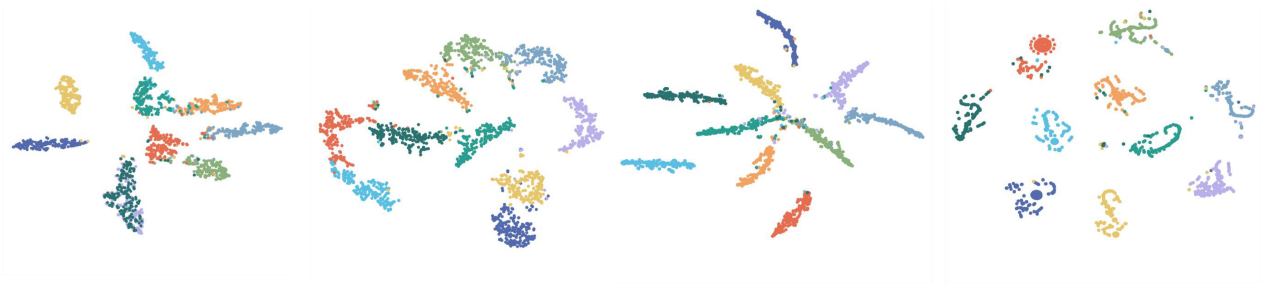
Table 1: A brief description of experimental datasets.

Experimental Setting. DIN is implemented using the PyTorch on an NVIDIA GeForce RTX 4060ti GPU with 16GB of memory. We train DIN for 500 epochs with α and β selected from $\{1, 2, \cdots, 5\}$. In this experiment, graph convolution network (GCN) serves as the classification model. The number of GCN layers is set to 2. Experimental evaluations of the proposed framework and seven competitive methods are under 10% randomly labeled instances. Each experiment is repeated five times, and the average performance metrics are reported. Two widely used classification evaluation metrics are employed, *i.e.*, Accuracy (ACC), and F1 Score (F1). ACC is defined as the ratio of correctly classified samples to the total number of samples. F1 is calculated as the arithmetic mean of the F1 values across each class. Both metrics are normalized to a range of $[0, 1]$, with higher values indicating superior model performance.

Comparison with State-of-the-art Methods. We conduct an extensive comparison of DIN against several state-of-the-art methods. The experimental results, as detailed in Table 2, provide the following principal observations. Intuitively, DIN exhibits a substantial performance improvement over the traditional method. Although LGCNFF occasion-

Metric	Method\Dataset	ALOI	BBCSports	Flickr	IAPR	MNIST	NUS-WIDE	UCI	WebKB-texas
ACC	Co-GCN	87.26 (1.92)	85.39 (3.77)	61.24 (2.59)	61.78 (2.46)	92.00 (0.51)	41.74 (1.43)	94.24 (2.60)	60.53 (8.25)
	LGCNFF	90.61 (8.06)	93.65 (1.19)	37.34 (3.87)	36.44 (3.35)	**93.84 (0.09)**	33.83 (3.51)	92.61 (2.22)	61.85 (0.98)
	PDMF	88.61 (0.70)	81.79 (3.91)	70.10 (0.33)	56.96 (0.88)	90.77 (0.14)	45.17 (1.05)	94.41 (0.73)	56.55 (1.51)
	ECMGD	94.36 (1.20)	59.90 (0.27)	63.78 (0.18)	64.70 (0.10)	88.30 (0.11)	47.51 (0.39)	94.62 (0.44)	66.07 (1.13)
	ORLNet	83.10 (1.45)	78.00 (4.79)	70.06 (0.30)	58.27 (0.85)	88.49 (0.60)	41.95 (1.60)	88.73 (0.52)	56.43 (7.22)
	RCML	83.44 (1.41)	80.55 (0.25)	72.22 (0.32)	60.10 (0.18)	85.10 (0.04)	40.65 (0.17)	93.87 (0.16)	62.14 (0.48)
	SMMRL	56.97 (2.12)	86.35 (2.05)	72.28 (0.15)	56.34 (0.43)	90.20 (8.12)	41.10 (0.99)	81.54 (1.47)	57.26 (5.91)
	DIN	**97.59 (0.27)**	**97.56 (0.00)**	**72.29 (0.03)**	**69.84 (0.25)**	93.81 (0.06)	**52.01 (0.00)**	**95.58 (0.48)**	**73.21 (0.00)**
F1	Co-GCN	87.15 (1.97)	82.82 (5.54)	60.95 (2.43)	62.41 (2.45)	91.93 (0.53)	39.73 (1.21)	94.22 (2.65)	31.37 (10.6)
	LGCNFF	90.67 (7.86)	93.31 (1.71)	35.14 (3.91)	34.41 (5.22)	**93.77 (0.09)**	31.71 (4.54)	92.59 (2.43)	37.79 (0.93)
	PDMF	88.81 (0.64)	79.20 (4.55)	70.05 (0.35)	59.44 (0.82)	90.60 (0.15)	44.15 (1.18)	94.44 (0.73)	19.17 (2.71)
	ECMGD	94.39 (1.20)	73.20 (1.04)	63.68 (0.19)	65.66 (0.11)	88.09 (0.11)	46.54 (0.37)	94.64 (0.44)	32.93 (1.35)
	ORLNet	82.74 (1.60)	76.04 (6.56)	69.96 (0.28)	60.03 (0.72)	88.28 (0.56)	40.75 (1.64)	88.72 (0.54)	39.94 (7.60)
	RCML	83.63 (1.57)	79.38 (0.38)	71.84 (0.37)	62.81 (0.15)	84.66 (0.05)	39.40 (0.18)	93.87 (0.16)	28.19 (0.50)
	SMMRL	51.93 (2.10)	83.82 (2.49)	72.05 (0.15)	58.32 (0.39)	90.20 (0.13)	41.28 (1.05)	81.53 (1.46)	29.70 (2.81)
	DIN	**97.58 (0.27)**	**97.76 (0.00)**	**72.07 (0.03)**	**70.78 (0.28)**	93.71 (0.06)	**50.00 (0.00)**	**95.58 (0.47)**	**52.04 (0.00)**

Table 2: Classification results (mean% and standard deviation%) of compared algorithms. The best performance is highlighted in bold and the second-best is underlined.



(a) LGCNFF (b) ECMGD (c) RCML (d) DIN

Figure 4: Visualization results of the learned graph representations of compared algorithms on the UCI dataset.

ally outperforms DIN in specific cases, it demonstrates more stable performance across both metrics. More specifically, DIN improves over the sub-optimal method by 12.43% on the WebKB-texas dataset in terms of F1. On one hand, DIN demonstrates a higher performance metric compared to other methods when applied to high-channel datasets. For example, on NUS-WIDE, it surpasses the compared method by a margin of 4.5% (ACC), and 3.5% (F1), respectively. This remarkable improvement highlights the efficacy of the proposed channel-impulse network in multi-view learning. On the other hand, for high-dimensional feature datasets, like BBCSports, DIN still achieves superior classification performance across both evaluation metrics. This confirms that the attribute-impulse network in DIN effectively facilitates the extraction of information across the feature space. In conclusion, DIN not only demonstrates the capacity to process multi-channel input information but also has the ability to extract the information located within the high-dimensional feature space.

Ablation Analysis. We conduct ablation experiment to verify the effectiveness of each network contributes to addressing multi-view tasks. Table 3 shows the corresponding results of the ablation experiments, with different ablation models explained as follows:

Dataset\Method	Base-1	Base-2	Base-3	DIN
ALOI	78.22	76.88	87.10	**97.59**
BBCSports	93.89	94.30	97.35	**97.56**
Flickr	70.92	56.70	70.09	**72.29**
IAPR	65.88	61.24	62.33	**69.84**
MNIST	91.30	91.89	93.81	**93.88**
NUS-WIDE	50.62	50.00	50.00	**52.01**
UCI	85.06	81.61	86.50	**95.58**
WebKB-texas	64.88	66.67	72.02	**73.21**

Table 3: An ablation study of the proposed framework on various datasets, where different networks of the framework are turned on or off and their accuracy is measured. Best accuracy values are highlighted in bold.

- **Base-1:** It only employs AIN to generate the attribute representation.
- **Base-2:** Compared with Base-1, it only utilizes CIN for the channel representation.
- **Base-3:** It employs two impulse networks, however, the feature representations derived from these networks are directly input into the classification model.
- **DIN:** It incorporates integration network into Base-3.

The results shown in Table 3 demonstrate that the

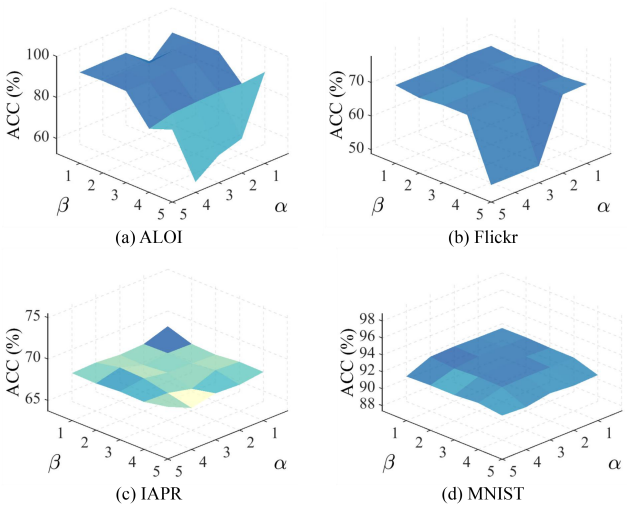


Figure 5: Parameter sensitivity of α and β in DIN.

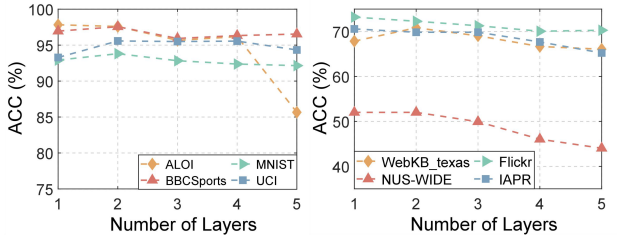


Figure 6: Parameter sensitivity of DIN w.r.t. layer numbers.

proposed framework consistently outperforms the variants methods across all datasets. We can discover that each component contributes to the performance. Comparing the performance of Base-1 and Base-2, it can be seen that channel and attribute networks exhibit distinct advantages in different datasets. For example, Base-1 outperformes Base-2 on the Flickr dataset, while Base-2 outperformes Base-1 on the WebKB-texas dataset. This demonstrates that conducting along the channel axis can achieve performance comparable to or even better than conducting along the attribute axis. Compared to the two impulse networks, adding the integration network can also promote performance improvement. For example, after adding the integration network, the performance of DIN improves from 87.10% to 97.59% on ALOI. This enhancement can be attributed to the integration between attribute representation and channel representation, which enhances the ability of final feature representation.

Visualization. To quantitatively evaluate the quality of the multi-view representations learned by DIN, as well as by all outstanding competitive methods on the test sets of the UCI dataset. The visualizations are presented in Fig. 4 (Fig. 2 in Appendix), where the multi-view representations are projected into a two-dimensional space using t-SNE. Observation demonstrates that the multi-view representations generated by DIN manifest a clearer and more discrete classification structure compared to other methods. In contrast, other methods struggle to capture attribute-channel information, leading to learned representations with a more ambiguous classification structure. The results indicate that the proposed framework can improve the ability to capture internal information, which is crucial for learning robust and class-distinctive representations.

Parameter Sensitivity Analysis. We present a detailed analysis of the sensitivity of the hyper-parameters α and β in Eq. (2) and Eq. (6). The parameter sensitivity of DIN is illustrated in Fig. 5 on four representative datasets in terms of α and β of ACC (detalis in Fig. 3 in Appendix). The hyper-parameters α and β are systematically varied within $[1, 5]$. From the experimental results, it can be clearly observed that as these two values increase, the performance of DIN will be influenced. Specifically speaking, when both α and β parameters are below the value of 4, DIN maintains stable performance on MNIST. Once the threshold of 4 is met, the performance of DIN slightly declines. This degradation results from information loss is induced by overzealous compression of the feature space, thereby impairing the performance. This suggests that while DIN is robust to hyper-parameters variations, excessive dimension reduction may suppress the expression of discriminative information, particularly in more challenging datasets.

Furthermore, we also analyze the impact of layer number on DIN performance to evaluate its structural robustness and scalability. Fig. 6 depicts the performance trend of DIN with increasing layer depth, showing how the number of layers influences its capability. We can observe that although DIN with different layers occasionally exhibits superior performance, the overall performance of DIN when integrated with two-layer GCN is more comprehensive and balanced. It is worth noting that subsequent layer additions result in a gradual performance degradation, this decline remains modest and within a small range, indicating a robust layer-scalability of the model.

Conclusion and Future Work

In this paper, we proposed a dual impulse network framework for multi-view learning, namely DIN, to improve the classification performance across multiple views. To extract the internal information, DIN introduced two different impulse networks for channels and attributes, respectively. These networks squeezed the specific axis to facilitate the activation of the internal information. To model attention matrices with implied relationship between attribute and channel, DIN employed an integration network that amalgamates channel representation with attribute representation. Finally, DIN derived the feature representation by applying these attention matrices, and fed this along with the adjacency matrix into a classification model. Experimental evaluations exhibited superior performance in multi-view learning tasks compared to existing state-of-the-art methods. In future work, we will explore more advanced and sophisticated impulse methods and implement novel interaction paradigms to handle more complex, large-scale multi-view data, such as incorporating dynamic graph structures.

Acknowledgments

This work is in part supported by the National Natural Science Foundation of China (Grant Nos. U25A20527 and 62276065), and the Fujian Provincial Natural Science Foundation of China (Grant No. 2024J01510026).

References

Bao, Y.; and Lu, F. 2024. Unsupervised gaze representation learning from multi-view face images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1419–1428.

Chen, L.; Xue, Z.; Li, Y.; Liang, M.; Wang, Y.; van den Hengel, A.; and Qi, Y. 2025a. Medusa: A multi-scale high-order contrastive dual-diffusion approach for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10295–10304.

Chen, M.-S.; Lai, P.-Y.; Liao, D.-Z.; Wang, C.-D.; and Lai, J.-H. 2025b. Graph Prompt Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7): 5794–5805.

Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; and Chen, H. 2022. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *The forty-fifth International ACM SIGIR Conference on Research and Development in Information Retrieval*, 904–915.

Chen, Z.; Fu, L.; Xiao, S.; Wang, S.; Plant, C.; and Guo, W. 2023a. Multi-view graph convolutional networks with differentiable node selection. *ACM Transactions on Knowledge Discovery from Data*, 18(1): 1–21.

Chen, Z.; Fu, L.; Yao, J.; Guo, W.; Plant, C.; and Wang, S. 2023b. Learnable graph convolutional network and feature fusion for multi-view learning. *Information Fusion*, 95: 109–119.

Du, S.; Fang, Z.; Tan, Y.; Wang, C.; Wang, S.; and Guo, W. 2025. OpenViewer: Openness-aware multi-view learning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 16389–16397.

Fang, Y.; Rao, X.; Gao, X.; Li, W.; and Min, Z. 2024a. MTSNet: Joint feature adaptation and enhancement for text-guided multi-view martian terrain segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5092–5101.

Fang, Z.; Du, S.; Cai, Z.; Lan, S.; Wu, C.; Tan, Y.; and Wang, S. 2024b. Representation learning meets optimization-derived networks: from single-view to multi-view. *IEEE Transactions on Multimedia*, 26: 8889–8901.

Fang, Z.; Xu, Z.; Du, L.; Du, S.; Cai, Z.; and Wang, S. 2025a. Enhancing Multi-view Open-set Learning via Ambiguity Uncertainty Calibration and View-wise Debiasing. In *Proceedings of the Thirty-three ACM International Conference on Multimedia*, 1220–1228.

Fang, Z.; Zou, Y.; Lan, S.; Du, S.; Tan, Y.; and Wang, S. 2025b. Scalable multi-modal representation learning networks. *Artificial Intelligence Review*, 58(7): 209.

Fu, L.; Deng, B.; Huang, S.; Liao, T.; Zhang, C.; and Chen, C. 2025. Learn from Global Rather Than Local: Consistent Context-Aware Representation Learning for Multi-View Graph Clustering. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 5145–5153.

Guo, Z.; Tang, Y.; Zhang, R.; Wang, D.; Wang, Z.; Zhao, B.; and Li, X. 2023. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, 15372–15383.

Hassani, K.; and Khasahmadi, A. H. 2020. Contrastive multi-view representation learning on graphs. In *International Conference on Machine Learning*, 4116–4126.

Hu, D.; Liu, S.; Wang, J.; Zhang, J.; Wang, S.; Hu, X.; Zhu, X.; Tang, C.; and Liu, X. 2024. Reliable Attribute-missing Multi-view Clustering with Instance-level and feature-level Cooperative Imputation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1456–1466.

Huang, Z.; Liang, Z.; and Jia, K. 2024. Sur2f: A hybrid representation for high-quality and efficient surface reconstruction from multi-view images. In *European Conference on Computer Vision*, 1–18.

Hwang, H.; Kim, G.; Hong, S.; and Kim, K. 2021. Multi-view representation learning via total correlation objective. In *Advances in Neural Information Processing Systems*, 12194–12207.

Lee, D. I.; Park, H.; Seo, J.; Park, E.; Park, H.; Baek, H.; Shin, S.; Kim, S.; and Kim, S. 2025. EditSplat: Multi-view fusion and attention-guided optimization for view-consistent 3D scene editing with 3D gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11135–11145.

Li, S.; Li, W.-T.; and Wang, W. 2020. Co-GCN for multi-view semi-supervised learning. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, 4691–4698.

Li, Y.; Wang, H.; Dang, L. M.; Song, H.; and Moon, H. 2023. Attention-guided multiscale neural network for defect detection in sewer pipelines. *Computer-Aided Civil and Infrastructure Engineering*, 38(15): 2163–2179.

Li, Y.; Yang, M.; and Zhang, Z. 2018. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10): 1863–1883.

Lin, J.; Huang, T.-Z.; Zhao, X.-L.; Ji, T.-Y.; and Zhao, Q. 2025a. Tensor robust kernel PCA for multidimensional data. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2): 2662–2674.

Lin, R.; Li, J.; Du, S.; Wang, S.; and Zhang, L. 2025b. OIMGC-Net: Optimization-inspired Interpretable Multi-view Graph Clustering Network. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 1957–1966.

Lin, Y.; Gou, Y.; Liu, X.; Bai, J.; Lv, J.; and Peng, X. 2022. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4447–4461.

Liu, C.; Jia, J.; Wen, J.; Liu, Y.; Luo, X.; Huang, C.; and Xu, Y. 2024a. Attention-induced embedding imputation for incomplete multi-view partial multi-label classification. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence*, 13864–13872.

Liu, C.; Wen, J.; Xu, Y.; Zhang, B.; Nie, L.; and Zhang, M. 2025a. Reliable representation learning for incomplete multi-view missing multi-label classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6): 4940–4956.

Liu, J.; Musialski, P.; Wonka, P.; and Ye, J. 2013. Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1): 208–220.

Liu, S.; Liang, K.; Dong, Z.; Wang, S.; Yang, X.; Zhou, S.; Zhu, E.; and Liu, X. 2024b. Learn from view correlation: An anchor enhancement strategy for multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26151–26161.

Liu, Z.; Wu, L.; He, M.; Guan, Z.; Zhao, H.; and Feng, N. 2025b. Multi-view empowered structural graph wordification for language models. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 24714–24722.

Lu, J.; Wu, Z.; Chen, Z.; Cai, Z.; and Wang, S. 2024. Towards multi-view consistent graph diffusion. In *Proceedings of the Thirty-Second ACM International Conference on Multimedia*, 186–195.

Luo, R.; Huang, H.; Lee, I.; Xu, C.; Qi, J.; and Feng, X. 2025. FairGP: A Scalable and Fair Graph Transformer Using Graph Partitioning. In *Proceedings of The 39th Annual AAAI Conference on Artificial Intelligence*, 12319–12327.

Ma, D.; Wang, N.; Fang, H.; Chen, W.; Li, B.; and Zhai, K. 2025. Attention-optimized 3D segmentation and reconstruction system for sewer pipelines employing multi-view images. *Computer-Aided Civil and Infrastructure Engineering*, 40(5): 594–613.

Ma, J.; Wang, Y.; Fan, A.; Xiao, G.; and Chen, R. 2023. Correspondence attention transformer: A context-sensitive network for two-view correspondence learning. *IEEE Transactions on Multimedia*, 25: 3509–3524.

Mao, K.; Lian, Y.; Wang, Y.; Liu, M.; Zheng, N.; and Wei, P. 2025. Unveiling multi-view anomaly detection: Intra-view decoupling and inter-view fusion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 12, 12381–12389.

Ning, X.; Yu, Z.; Li, L.; Li, W.; and Tiwari, P. 2024. DILF: Differentiable rendering-based multi-view image–language fusion for zero-shot 3D shape understanding. *Information Fusion*, 102: 102033.

Ouyang, D.; Liang, Y.; Wang, J.; Li, L.; Ai, N.; Feng, J.; Lu, S.; Liao, S.; Liu, X.; and Xie, S. 2024. HGCLAMIR: Hypergraph contrastive learning with attention mechanism and integrated multi-view representation for predicting miRNA-disease associations. *Public Library of Science Computational Biology*, 20(4): 1011927.

Song, S.; Zhao, S.; Wang, C.; Yan, T.; Li, S.; Mao, X.; and Wang, M. 2024. A dual-way enhanced framework from text matching point of view for multimodal entity linking. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 19008–19016.

Sun, J.; Hu, M.; Wu, X.; Tang, C.; Lahza, H.; Wang, S.; and Zhang, Y. 2024. MVSI-Net: Multi-view attention and multi-scale feature interaction for brain tumor segmentation. *Biomedical Signal Processing and Control*, 95: 106484.

Wang, R.; Sun, H.; Lin, Y.; Zuo, C.; Gong, Y.; Yin, Y.; and Meng, W. 2025a. SeqMvRL: A sequential fusion framework for multi-view representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25822–25831.

Wang, R.; Sun, H.; Lin, Y.; Zuo, C.; Gong, Y.; Yin, Y.; and Meng, W. 2025b. SeqMvRL: A sequential fusion framework for multi-view representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25822–25831.

Wang, X.; Gao, H.; Wei, X.; Peng, L.; Li, R.; Liu, C.; Wu, S.; and Wong, H.-S. 2024a. Contrastive graph distribution alignment for partially view-aligned clustering. In *Proceedings of the Thirty-Second ACM International Conference on Multimedia*, 5240–5249.

Wang, Y.; Chang, D.; Fu, Z.; Wen, J.; and Zhao, Y. 2024b. Partially view-aligned representation learning via cross-view graph contrastive network. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8): 7272–7283.

Xie, Y.; Liu, J.; Qu, Y.; Tao, D.; Zhang, W.; Dai, L.; and Ma, L. 2021. Robust kernelized multiview self-representation for subspace clustering. *IEEE Transactions on Neural Networks Learning Systems*, 32(2): 868–881.

Xu, C.; Si, J.; Guan, Z.; Zhao, W.; Wu, Y.; and Gao, X. 2024. Reliable conflictive multi-view learning. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 16129–16137.

Xu, C.; Zhao, W.; Zhao, J.; Guan, Z.; Yang, Y.; Chen, L.; and Song, X. 2023. Progressive deep multi-view comprehensive representation learning. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 10557–10565.

Yang, J.; Cheng, Z.; Duan, Y.; Ji, P.; and Li, H. 2024. Consistnet: Enforcing 3d consistency for multi-view images diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7079–7088.

Yao, K.; Liang, J.; Liang, J.; Li, M.; and Cao, F. 2022. Multi-view graph convolutional networks with attention mechanism. *Artificial Intelligence*, 307: 103708.

Ye, F.; and Li, S. 2024. MileCut: A multi-view truncation framework for legal case retrieval. In *Proceedings of the ACM Web Conference 2024*, 1341–1349.

Zheng, Q.; Zhu, J.; Li, Z.; Tian, Z.; and Li, C. 2023. Comprehensive multi-view representation learning. *Information Fusion*, 89: 198–209.

Zheng, Z. 2025. Zero-Shot multi-view australian sign language recognition. In *Companion Proceedings of the ACM on Web Conference*, 2463–2467.