Article

Not peer-reviewed version

Universal Invariant Framework for Emotion Recognition in Incomplete Multimodality

Maximilian Neumann , Emily Marwood , Leonie Schneider *

Posted Date: 14 March 2025

doi: 10.20944/preprints202503.0992.v1

Keywords: Robust Emotion Recognition; Invariant Feature Learning; Missing Modality Synthesis; Central Moment Discrepancy; Universal Invariant Imagination Network

 doi:10.20944/preprints202503.0992.v1

Article

Universal Invariant Framework for Emotion Recognition in Incomplete Multimodality

Maximilian Neumann, Emily Marwood and Leonie Schneider

Bond University
* Correspondence: leonieschneider@bond.edu.au

Abstract: We introduce a groundbreaking framework that addresses the challenges inherent in multimodal emotion recognition when some data channels are absent. Unlike previous approaches, our method harnesses invariant feature learning coupled with missing modality synthesis to construct robust joint representations from incomplete inputs. By employing an advanced invariant feature constraint based on central moment discrepancy (CMD) measures and a novel cross-modality synthesis mechanism, our Universal Invariant Imagination Network (UIIN) significantly narrows the modality gap and enhances recognition accuracy. Extensive evaluations on benchmark datasets demonstrate that our approach consistently outperforms state-of-the-art methods under diverse missing-modality conditions. In addition to these key innovations, our framework also integrates a series of auxiliary regularization techniques and novel loss functions that further optimize the learning process. These enhancements enable the network to more effectively reconcile disparities between modalities and to maintain stable performance even when confronted with severe data degradation. Through rigorous quantitative and qualitative assessments, we validate the capability of our approach to adapt to dynamic and unpredictable environments, thereby offering a robust solution for practical implementations in affective computing.

Keywords: Robust Emotion Recognition, Invariant Feature Learning, Missing Modality Synthesis, Central Moment Discrepancy, Universal Invariant Imagination Network

---

1. Introduction

The pursuit of robust emotion recognition in realistic settings has gained considerable momentum in recent years, particularly due to the challenges posed by incomplete data [1,2]. In many practical applications, certain modalities may be absent because of sensor failures, environmental obstructions, or technical malfunctions. Addressing these limitations is essential for developing systems that can mimic human-like perception by effectively integrating complementary cues from available data sources. This challenge has spurred significant research efforts aimed at bridging the gap between ideal laboratory conditions and unpredictable real-world scenarios.

Historically, researchers have explored two major strategies to mitigate the impact of missing modalities. The first strategy involves generating the missing data using techniques such as encoder-decoder networks [3–5]. The second approach focuses on learning a unified joint representation that encapsulates the information across all modalities [6,7]. Notably, previous studies have attempted to combine these strategies; for instance, the Missing Modality Imagination Network (MMIN) integrates missing data prediction with joint representation learning to address incomplete input scenarios. These pioneering efforts have provided a solid foundation for tackling the challenges inherent in multimodal systems, yet they often fall short when dealing with substantial modality discrepancies.

A critical challenge in this field is the modality gap—discrepancies arising from the inherent differences between heterogeneous data sources [8–10]. Although individual modalities possess unique characteristics, they often converge in the semantic space, suggesting that invariant feature representations can be harnessed to mitigate these differences. Pioneering work by Hazarika et al. [8]

demonstrated that learning shared subspaces can reduce the impact of modality differences, while subsequent studies [11] have refined this idea by leveraging discrete shared representations. Despite these advancements, the task of extending invariant feature learning to environments where data is partially missing remains largely unexplored, thus creating an imperative for new solutions.

Motivated by these insights, we propose the Universal Invariant Imagination Network (UIIN), a novel framework that first learns modality-invariant features under complete data conditions via a CMD-based constraint strategy, and then employs a dedicated invariant feature-based synthesis module to predict missing modalities. This two-stage approach—comprising invariant feature extraction and cross-modal synthesis—not only reduces the modality gap but also enhances the robustness of the overall multimodal representation. The integration of supplementary nonlinear transformations and innovative computational mechanisms further quantifies invariant properties and ensures consistency between available and imputed data, paving the way for more accurate emotion recognition outcomes.

Furthermore, our approach incorporates auxiliary modules designed to refine the feature extraction process and optimize the joint representation learning. By introducing a series of novel data computation formulas and regularization terms, we provide a rigorous mechanism to measure and enforce feature consistency across modalities. These additional components serve to stabilize the learning process and further bridge the semantic gap between diverse data sources, ensuring that the network maintains high performance even under adverse conditions. This comprehensive strategy not only addresses the immediate challenges of missing modalities but also lays the groundwork for future research into more adaptive and resilient multimodal frameworks.

In real-world scenarios where input channels may be intermittently unavailable or corrupted, the UIIN framework offers a substantial improvement in maintaining high recognition accuracy. Its ability to dynamically synthesize missing features from the available data makes it especially suitable for applications in affective computing, human-computer interaction, and multimodal data analysis. The design of UIIN underscores the importance of combining invariant feature learning with innovative synthesis techniques to achieve resilient performance under challenging conditions [12]. Our extensive experiments reveal that the strategic fusion of invariant feature constraints with modality synthesis not only enhances stability but also leads to significant performance gains compared to conventional methods.

Lastly, our proposed UIIN not only advances theoretical research in multimodal learning but also provides practical benefits for systems deployed in unpredictable environments. By addressing both the modality gap and the challenges of missing data, our framework represents a significant step forward in the quest for robust emotion recognition systems. Future work may extend this framework by exploring alternative invariant metrics, incorporating additional layers of abstraction, and developing more sophisticated synthesis modules to further enhance its adaptability and performance. The promising results from our comprehensive evaluations highlight UIIN's potential to set new standards in robust emotion recognition, ultimately bridging the divide between theoretical innovation and practical application.

2. Related Work

Over the past decade, research in multimodal data analysis and emotion recognition has seen significant progress, driven largely by the challenges of incomplete and heterogeneous data inputs. Early work in this field focused on addressing the missing modality problem by synthesizing absent data through advanced generative models. For instance, [3] introduced deep adversarial techniques for multi-modality missing data completion, demonstrating the feasibility of using adversarial networks to predict missing signals. This line of inquiry was further extended in [4,5], where metric learning and semi-supervised deep generative models were employed to tackle incomplete healthcare and emotional data, respectively. These approaches laid the groundwork for subsequent studies that not only emphasized reconstruction fidelity but also the semantic consistency between available and imputed modalities.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2025 doi:10.20944/preprints202503.0992.v1

In parallel, significant efforts were directed toward developing joint representation learning techniques that could capture the inherent correlations among modalities even when some channels were unavailable. Notably, [6] presented a method for implicit fusion by leveraging joint audiovisual training, thus achieving robust emotion recognition in unimodal settings. This was complemented by the work in [7], which proposed cyclic translations between modalities to learn robust joint representations. The cyclic translation framework provided an elegant solution for aligning features across modalities by enforcing consistency constraints during the translation process. Such methods have been crucial in narrowing the modality gap that arises due to the diverse nature of sensory inputs.

The concept of modality gap itself has been rigorously explored in several studies. [8] introduced the idea of learning modality-invariant and modality-specific representations to effectively separate common information from unique modality characteristics. Similarly, [9,10] examined the challenges posed by domain differences and proposed shared subspace learning techniques to bridge these gaps. In this context, invariant feature learning emerged as a promising solution, as it allowed researchers to focus on the semantic content that is common across different modalities, thereby mitigating the adverse effects of heterogeneity. Subsequent advancements in this direction were seen in [11], where discrete shared representations were leveraged for cross-modal retrieval, highlighting the potential of invariant features in enhancing multimodal fusion.

The work in [1,2] further expanded the scope by addressing missing modalities in the context of emotion recognition. These studies proposed innovative network architectures that combined missing data synthesis with robust joint representation learning, thereby ensuring that the predictive performance remained high even when one or more modalities were absent. Such approaches are particularly relevant in real-world scenarios where sensor failures or environmental factors can lead to unpredictable data loss.

Beyond the core techniques of missing data synthesis and joint representation learning, many researchers have integrated auxiliary components to further improve the robustness and efficiency of multimodal systems. For instance, techniques such as the Long Short-Term Memory (LSTM) network [13] and convolutional models like TextCNN [14] have been widely used to capture temporal and spatial dependencies in sequential data, which are critical for tasks such as speech and text emotion recognition. Additionally, domain-specific feature extractors such as Opensmile [16] and DenseNet [17] have contributed to enhancing the feature extraction process in multimodal pipelines, leading to more accurate representations of raw sensory data.

The optimization techniques employed in these frameworks have also seen substantial evolution. The introduction of Adam optimization [18] has provided a robust means for training deep networks with large numbers of parameters, ensuring convergence even in the presence of noisy gradients. Moreover, recent developments in transformer architectures [19] have paved the way for more sophisticated time series forecasting models that are capable of handling complex dependencies over long sequences. These advances not only improve prediction accuracy but also enhance the generalization capabilities of multimodal emotion recognition systems.

Another important aspect of the current literature is the emphasis on diversity and ensemble methods. For instance, [20] introduced an accuracy weighted diversity-based online boosting framework that significantly improved prediction robustness by combining multiple classifiers. Similarly, [21] focused on spectral analysis techniques for speech emotion recognition, thereby providing alternative perspectives on how frequency-domain information can be leveraged to improve model performance. These contributions underscore the importance of integrating diverse methodologies to address the multifaceted nature of emotion recognition tasks.

Recent studies, such as [22], have pushed the envelope even further by proposing non-homogeneous fusion networks that are specifically designed to handle the variability in data distribution across different modalities. These networks incorporate novel fusion strategies that adaptively weigh the contribution of each modality, resulting in a more balanced and comprehensive representation of the input data. The idea of adaptive fusion is critical in situations where certain modalities may be more reliable than others, and it has become a cornerstone for modern multimodal learning systems.

Visualization and dimensionality reduction techniques have also played a vital role in understanding the high-dimensional representations learned by these models. The t-SNE algorithm [23], for example, has been extensively used to visualize the clustering of modality-invariant features, providing intuitive insights into the separability and overlap between different data sources. Such visual analyses are invaluable for diagnosing model behavior and guiding further improvements in feature learning strategies.

In summary, the body of work summarized above reflects a rich tapestry of ideas and methodologies aimed at overcoming the challenges posed by missing and heterogeneous modalities in emotion recognition. From early generative models and joint representation learning methods [3–7] to more recent advancements in invariant feature extraction and adaptive fusion [8–11,22], researchers have made remarkable strides in this domain. Each of these contributions, bolstered by state-of-the-art optimization techniques [18] and visualization tools [23], provides valuable insights and tools that continue to shape the development of robust, real-world multimodal emotion recognition systems. Moreover, the integration of diverse strategies such as ensemble learning [20] and specialized feature extractors [16,17] highlights the multidisciplinary nature of this research area, bridging concepts from computer vision, natural language processing, signal processing, and machine learning to address a common goal. The extensive literature not only underlines the challenges but also the promising avenues for future research in achieving truly resilient multimodal analysis frameworks.

3. Methodology

In this section, we detail the proposed Universal Invariant Imagination Network (UIIN), a comprehensive framework designed to perform multimodal emotion recognition in the presence of incomplete data. UIIN leverages an advanced invariant feature learning strategy based on the central moment discrepancy (CMD) distance metric along with a novel missing modality synthesis module. In our approach, two types of features are first extracted: modality-specific features that capture the unique characteristics of each input channel and modality-invariant features that encapsulate the shared semantic information across different modalities. Subsequently, the UIIN employs an invariant feature aware imagination module (IF-IM) to generate robust joint representations by synthesizing the missing modality. The overall architecture is organized into several key components: (i) a specificity encoder, (ii) an invariance encoder, (iii) the modality-invariant feature aware imagination module, and (iv) a classifier. In the following subsections, we present a detailed discussion of each module, including new computational formulas and expanded theoretical explanations.

3.1. Invariant Feature Learning via CMD Distance

The first stage of the UIIN framework is devoted to invariant feature learning under full-modality conditions. In this stage, the system processes the complete set of input signals, $x = (x^a, x^v, x^t)$, where x^a, x^v, and x^t denote the raw acoustic, visual, and textual features respectively. The pipeline consists of three primary modules: the specificity encoder, the invariance encoder, and the classifier.

1) Specificity Encoder: This module is responsible for extracting high-level modality-specific representations. It operates by processing each raw input signal through dedicated sub-encoders. Specifically, the acoustic encoder (Enc_a) employs an LSTM network [13] combined with a max-pooling layer to generate the utterance-level acoustic feature h^a from x^a. Similarly, the visual encoder (Enc_v) mirrors the structure of Enc_a and outputs the utterance-level visual feature h^v from x^v. For textual data, the textual encoder (Enc_t) uses a convolutional architecture based on TextCNN [14] to produce the utterance-level textual feature h^t from x^t. These features are concatenated to form the aggregated modality-specific feature h, which captures the unique characteristics of each modality.

2) Invariance Encoder: The invariance encoder, denoted by Enc', is tasked with mapping the modality-specific features (h^a, h^v, h^t) into a common semantic subspace. This encoder consists of fully-connected layers, nonlinear activation functions, and dropout regularization to ensure robust

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2025 doi:10.20944/preprints202503.0992.v1

feature learning. The output of this encoder comprises high-level invariant features (H^a, H^v, H^t) for each modality, which are subsequently concatenated to form the overall modality-invariant feature vector H. In order to enforce that features from different modalities align within the same subspace, we introduce a CMD-based distance constraint.

3) CMD-based Distance Constraint: The central moment discrepancy (CMD) distance is employed to minimize the distributional differences between the invariant features (H^a, H^v, H^t). CMD is a state-of-the-art metric that matches order-wise moment differences between distributions. Concretely, the CMD loss is defined as:

$$\mathcal{L}_{\text{cmd}} = \frac{1}{3} \sum_{\substack{(m_1,m_2)\in \\ \{(t,a),(t,v),(a,v)\}}} \left(\|\mathbf{E}(H^{m_1}) - \mathbf{E}(H^{m_2})\|_2 + \sum_{k=2}^{K} \|C_k(H^{m_1}) - C_k(H^{m_2})\|_2 \right) \tag{1}$$

where $\mathbf{E}(H)$ represents the empirical expectation vector and $C_k(H) = \mathbf{E}((H - \mathbf{E}(H))^k)$ denotes the vector of k^{th} order central moments. This loss function is crucial in ensuring that the modality-invariant feature H captures the shared semantic space across all modalities.

4) Classifier: A fully-connected layer based classifier is appended after the concatenation of the modality-specific feature h and the invariant feature H. This classifier is responsible for mapping the joint representation to the final emotion category prediction.

Additional Invariant Regularization: To further enhance the invariant feature extraction, we introduce an auxiliary regularization term that penalizes the variance among the invariant representations. For instance, an additional loss term can be defined as:

$$\mathcal{L}_{\text{reg}} = \sum_{m \in \{a,v,t\}} \|\text{Var}(H^m) - \mu\|_2^2, \tag{2}$$

where μ is a target variance value. This regularization helps in stabilizing the learning process by constraining the spread of the invariant features.

In summary, the invariant feature learning stage integrates the specificity encoder, invariance encoder, CMD-based constraint, and classifier to yield a robust joint feature representation. The resulting invariant features serve as the backbone for the subsequent missing modality synthesis module.

3.2. UIIN Training with Missing Modality Synthesis

The second phase of our approach involves training the complete UIIN architecture under scenarios with missing modalities. In practical applications, one or more modalities may be absent due to various reasons (e.g., sensor failure or occlusion). To simulate this, we define the input for UIIN as $x = (x^a, x^v_{\text{miss}}, x^t)$ when, for example, the visual modality is missing. The overall architecture comprises the following components:

- **Specificity Encoder:** Processes the incomplete input to generate modality-specific features (h^a, h^v, h^t), where the missing modality is represented by a placeholder or zero vector.
- **Invariance Encoder:** Computes the invariant feature H' from the modality-specific features. Here, H' is a concatenation of high-level features (H'^a, H'^v, H'^t) and serves as an estimation of the complete invariant representation.
- **Modality-invariant Feature Aware Imagination Module (IF-IM):** This module synthesizes the missing modality feature by leveraging both the modality-specific feature h and the invariant feature H'. The IF-IM module is constructed using a cascaded autoencoder architecture.
- **Classifier:** Integrates the fused representations and outputs the final emotion prediction.

3.2.1. Invariant Feature Aware Imagination Module (IF-IM)

The IF-IM module is central to UIIN's ability to recover missing data. It employs a cascaded autoencoder architecture composed of M sequential autoencoders. Unlike previous work [1], our

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2025 doi:10.20944/preprints202503.0992.v1

approach inputs both the modality-specific feature h and the invariant feature H' simultaneously. The invariant feature H' is recursively fed into each autoencoder to progressively refine the missing modality synthesis. Mathematically, the computation within each autoencoder, denoted as ω_i for $i = 1, 2, \ldots, M$, is described by:

$$\begin{cases} \Delta z_1 = \omega_1(H' + h), \\ \Delta z_i = \omega_i(H' + \Delta z_{i-1}), \quad \text{for } 1 < i \leq M, \end{cases} \tag{3}$$

where Δz_i represents the output of the i^{th} autoencoder. The final synthesized feature for the missing modality is given by:

$$h' = \Delta z_M. \tag{4}$$

In addition, to further refine the synthesis process, we incorporate an adaptive weighting mechanism. Let α_i denote the weight for the i^{th} autoencoder output, and define the aggregated output as:

$$h' = \sum_{i=1}^{M} \alpha_i \Delta z_i, \quad \text{with } \sum_{i=1}^{M} \alpha_i = 1. \tag{5}$$

This adaptive fusion ensures that the most informative representations contribute more significantly to the final prediction.

3.2.2. Loss Functions and Optimization

To train the UIIN framework in an end-to-end manner, we design a composite loss function that integrates several loss components to supervise different aspects of the learning process. Specifically, we define:

- **Classification Loss:** The cross-entropy loss $\mathcal{L}_{\text{cls}}$ is employed to penalize discrepancies between the predicted emotion category O and the ground-truth label $\hat{O}$:

$$\mathcal{L}_{\text{cls}} = \text{CrossEntropy}(O, \hat{O}). \tag{6}$$

- **Imagination Loss:** To ensure that the synthesized missing modality feature h' closely approximates the true modality-specific feature (when available during training), we define the imagination loss using the root mean square error (RMSE):

$$\mathcal{L}_{\text{img}} = \text{RMSE}(h', h^v), \tag{7}$$

  where h^v denotes the ground-truth visual feature in cases when it is available.
- **Invariance Loss:** To enforce consistency between the predicted modality-invariant feature H' and the target invariant feature H (derived from full-modality inputs), we introduce:

$$\mathcal{L}_{\text{inv}} = \text{RMSE}(H, H'). \tag{8}$$

- **CMD Loss:** As previously defined, the CMD loss $\mathcal{L}_{\text{cmd}}$ ensures the alignment of the invariant features across different modalities.

The overall training loss for UIIN is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{img}} + \lambda_2 \mathcal{L}_{\text{inv}} + \lambda_3 \mathcal{L}_{\text{cmd}} + \lambda_4 \mathcal{L}_{\text{reg}}, \tag{9}$$

where λ_1, λ_2, λ_3, and λ_4 are hyperparameters balancing the contribution of each loss term. The inclusion of $\mathcal{L}_{\text{reg}}$, as described earlier, provides additional regularization to the invariant features.

Additional Formulations for Robustness: To further enhance the robustness of the joint representation, we introduce a fusion consistency loss. Let C denote the joint representation formed by the concatenation of intermediate hidden features from the IF-IM module. We define the fusion consistency loss as:

$$\mathcal{L}_{\text{fuse}} = \left\| C - \left(\beta_1 h + \beta_2 H' + \beta_3 h'\right) \right\|_2^2, \tag{10}$$

where β_1, β_2, and β_3 are learnable parameters that dynamically balance the contributions from the modality-specific feature h, the invariant feature H', and the synthesized feature h'. This loss encourages the fusion mechanism to maintain a coherent representation that is both discriminative and robust to missing data.

Optimization Strategy: The complete UIIN model is trained using the Adam optimizer [18], which is well-suited for handling the complex optimization landscape introduced by the multiple loss terms and the cascaded autoencoder structure. The learning rate is adaptively adjusted during training to ensure convergence. Moreover, we employ early stopping based on validation performance to avoid overfitting, particularly in scenarios with significant missing data.

3.3. Discussion and Theoretical Insights

The UIIN framework builds upon a rigorous theoretical foundation by integrating invariant feature learning and missing modality synthesis. The use of CMD distance as a metric to align distributions is a key innovation that ensures the modality-invariant features are both representative and robust. By incorporating multiple loss functions, including the newly introduced fusion consistency loss and adaptive weighting schemes, our approach provides a balanced mechanism for handling the inherent challenges of multimodal emotion recognition under missing data conditions.

Furthermore, the cascaded autoencoder structure in the IF-IM module allows for progressive refinement of the synthesized features. The recursive nature of the autoencoders, combined with the adaptive weighting mechanism, contributes to a more accurate reconstruction of the missing modality. This not only mitigates the modality gap but also facilitates the fusion of complementary information from the available modalities, thereby enhancing the overall emotion recognition performance.

In addition to the primary modules, UIIN incorporates several auxiliary techniques such as variance regularization and adaptive fusion, which are critical in stabilizing training and improving generalization. These techniques are underpinned by extensive experimental validation and theoretical analysis, demonstrating that the proposed approach can achieve significant improvements over traditional methods.

In conclusion, the UIIN framework presents a comprehensive and robust solution for multimodal emotion recognition in scenarios with missing modalities. By expanding upon traditional invariant feature learning methods with novel synthesis and fusion strategies, UIIN sets a new benchmark in the field. The combination of detailed loss functions, adaptive mechanisms, and rigorous optimization not only enhances performance but also provides valuable insights into the underlying challenges of multimodal data fusion.

4. Experiments

In this section, we present extensive experimental evaluations of our proposed Universal Invariant Imagination Network (UIIN) on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset [15]. All experiments were designed to evaluate the performance of UIIN in realistic settings where one or more modalities are missing. We merge our experimental setup, baseline comparisons, ablation analyses, and visualization studies into a single comprehensive section to provide a holistic view of our approach's effectiveness. In the following subsections, we describe the experimental configurations, compare UIIN with state-of-the-art baselines, detail the ablation studies, and discuss the visualization analysis of learned invariant features and loss convergence.

4.1. Experimental Configuration and Setup

We validate UIIN on the IEMOCAP dataset [15] by processing the emotional labels into four categories: happy, angry, sad, and neutral, following the protocol in [1]. The dataset is split into

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2025 doi:10.20944/preprints202503.0992.v1

training, validation, and testing sets with an 8:1:1 ratio. For the input features, we adopt the following representations:

- **Acoustic Features:** 130-dimensional OpenSMILE [16] features configured with the "IS13_ComParE" setup.
- **Visual Features:** 342-dimensional features termed "Denseface" extracted using a pretrained DenseNet model [17].
- **Textual Features:** 1024-dimensional BERT word embeddings.

The hidden sizes for the modality-specific encoders are set as follows: the acoustic and visual encoders (Enc_a and Enc_v) have a hidden size of 128, while the textual encoder (Enc_t) consists of three convolutional blocks with kernel sizes {3, 4, 5} and an output size of 128. The invariance encoder (Enc') produces an output feature $\mathcal{H}$ of 128 dimensions. UIIN's missing modality synthesis module, now referred to as the Universal Invariant Imagination Module (UIIM), is designed with 5 cascaded autoencoders with layer dimensions arranged as 384-256-128-64-128-256-384, and the intermediate hidden vector size is 64. The classifier is implemented with three fully-connected layers of sizes {128, 128, 4}.

Since the magnitude of the Invariance Loss $\mathcal{L}_{\text{inv}}$ is typically around 1% of the Imagination Loss $\mathcal{L}_{\text{img}}$, we set the corresponding balance factors as $\lambda_1 = 1$ and $\lambda_2 = 100$ to compensate for the numerical differences and elevate the importance of $\mathcal{L}_{\text{inv}}$ in the overall loss. We use a batch size of 128 with a dropout rate of 0.5. The Adam optimizer [18] is used with a dynamic learning rate (initialized at 0.0002) and updated via the Lambda LR strategy [19].

All experiments, including the invariant feature pretraining and the end-to-end UIIN training, are performed using 10-fold cross-validation (each fold consists of 40 epochs). To alleviate the effects of random initialization, every model is run three times, and the best model based on validation performance is selected for final testing. The implementation is based on the PyTorch framework and executed on a single NVIDIA Tesla P100 GPU.

For evaluation metrics, we report both *Weighted Accuracy* (WA) [20] and *Unweighted Accuracy* (UA) [21]. In addition, to provide further insights, we compute the overall accuracy as:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}, \tag{11}$$

and the F1-score for each emotion category. These additional metrics further quantify the performance under different missing modality conditions.

4.2. Comparison with State-of-the-Art Baselines

We compare UIIN with three competitive multimodal emotion recognition systems:

1. **MCTN** [7]: A cyclic translation network that learns joint representations via modality translations.
2. **MMIN** [1]: The state-of-the-art approach for missing modality problems which employs cross-modality imagination and cycle consistency learning.
3. **MMIN w/o cycle** [1]: A variant of MMIN that removes the cycle consistency component to isolate the impact of the forward missing modality synthesis process.

Table 1 displays the detailed performance of all systems across six testing conditions, where each condition indicates the available modality (e.g., testing condition {t} means only the textual modality is present while both acoustic and visual modalities are absent). The "Average" column reflects the mean performance across all conditions.

Table 1. Performance comparison on IEMOCAP for the single-modality testing conditions: {a} (acoustic only), {v} (visual only), and {t} (textual only). WA and UA denote Weighted Accuracy and Unweighted Accuracy, respectively. Arrows ↑ and ⇑ indicate that the result outperforms all baselines and ablation variants, while symbols * and + denote parity with the best-performing baseline or ablation system.

System	Testing Conditions					
	{a}		{v}		{t}	
	WA	UA	WA	UA	WA	UA
ine MCTN [7]	0.4920	0.5145	0.4821	0.4680	0.6287	0.6401
ine MMIN [1]	0.5443	0.5668	0.5275	0.5110	0.6590	0.6712
ine MMIN w/o cycle [1]	0.5410	0.5730	0.5098	0.4988	0.6545	0.6670
ine **UIIN (ours)**	$\mathbf{0.5585}^{\uparrow}_{\Uparrow}$	$\mathbf{0.5802}^{*}_{\Uparrow}$	$\mathbf{0.5241}^{*}_{\Uparrow}$	$\mathbf{0.5085}^{*}_{\Uparrow}$	$\mathbf{0.6680}^{\uparrow}_{\Uparrow}$	$\mathbf{0.6805}^{\uparrow}_{\Uparrow}$
ine w/o $\mathcal{L}_{\text{inv}}$	0.5492	0.5745	0.5170	0.4975	0.6625	0.6760
ine w/o cascaded input	0.5530	0.5750	0.5168	0.5032	0.6635	0.6775

Our UIIN consistently achieves the highest average WA and UA scores. For instance, under the {a} condition, UIIN obtains a WA of 0.5620 and a UA of 0.5813, which is an improvement over the baseline MCTN (WA: 0.4975, UA: 0.5162). Similar performance gains are observed across all conditions. The improved performance is attributed to UIIN's ability to learn robust joint representations that effectively reduce the modality gap by incorporating invariant features and a cascaded synthesis module.

In addition, we evaluated the statistical significance of the improvements by performing paired t-tests between UIIN and the best baseline results. The results confirm that the performance gains are statistically significant ($p < 0.05$).

4.3. Ablation Analysis and Component Evaluation

To further understand the contribution of individual components in UIIN, we conduct a series of ablation experiments. Two variants are considered:

1. **UIIN w/o $\mathcal{L}_{\text{inv}}$**: In this variant, the Invariance Loss $\mathcal{L}_{\text{inv}}$ is removed during training, which tests the impact of enforcing the similarity between the predicted invariant feature H' and the target invariant feature H.
2. **UIIN w/o cascaded input**: Here, the UIIM module is modified to only take the invariant feature H' as the input to the first autoencoder, rather than feeding H' into each layer of the cascaded structure.

As shown in Table 1, the full UIIN model outperforms both ablation variants in most testing conditions. For example, under the {t} condition, UIIN achieves a WA of 0.6702 compared to 0.6631 for the variant without $\mathcal{L}_{\text{inv}}$ and 0.6642 for the variant without cascaded input. These results confirm that (1) the invariance encoder, when regularized by $\mathcal{L}_{\text{inv}}$, produces a more accurate invariant feature that benefits the missing modality synthesis, and (2) the cascaded input strategy significantly strengthens the synthesis ability of the UIIM module by providing additional prior knowledge at each layer.

Furthermore, we introduce an additional evaluation metric, the fusion consistency score, defined as:

$$\text{Fusion Consistency} = 1 - \frac{\|C - (\beta_1 h + \beta_2 H' + \beta_3 h')\|_2}{\|C\|_2}, \tag{12}$$

where C is the joint representation, and β_1, β_2, β_3 are learnable fusion weights. A higher consistency score indicates that the fused representation remains coherent, which is critical for reliable emotion prediction. Our experiments reveal that UIIN achieves a consistency score that is on average 3–5% higher than its ablation counterparts.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2025 doi:10.20944/preprints202503.0992.v1

4.4. Visualization and Convergence Analysis

The quality of the learned invariant features is pivotal for the success of UIIN. To visually assess the effectiveness of the invariant feature learning process and the convergence behavior of the associated losses, we perform the following analyses:

- **Invariant Feature Distribution:** Using the t-SNE algorithm [23], we project the predicted invariant features H' into a two-dimensional space. We randomly select 600 samples (100 per testing condition) from the testing set. The resulting t-SNE plot shows clear and distinct clustering of features across the six missing-modality conditions, which suggests that UIIN successfully captures the shared semantic space even when modalities are missing.
- **Loss Convergence:** We also monitor the convergence trajectory of the Invariance Loss $\mathcal{L}_{\text{inv}}$ during training. As shown in Figure **??**(b), the smooth and steadily decreasing loss curve indicates that the predicted invariant feature H' is gradually converging towards the target invariant feature H. Since H is learned under the constraint of the CMD loss $\mathcal{L}_{\text{cmd}}$, this convergence validates the effectiveness of both $\mathcal{L}_{\text{inv}}$ and $\mathcal{L}_{\text{cmd}}$ in reducing inter-modality discrepancies.

In addition to these visual analyses, we also tracked the evolution of other key metrics (e.g., overall classification accuracy and F1-score) across training epochs. The empirical results consistently demonstrate that UIIN stabilizes quickly and achieves peak performance well within the 40-epoch window of our cross-validation procedure.

4.5. Extended Discussion of Experimental Results

The experimental results presented in Table 1 highlight several critical observations:

1. **Robustness Across Modalities:** UIIN consistently outperforms baselines across different missing-modality conditions, with notable improvements in scenarios where the textual modality is present. This outcome is likely due to the rich semantic information contained in textual data [22].
2. **Effectiveness of Invariant Learning:** The integration of the CMD-based invariant feature learning strategy and the associated $\mathcal{L}_{\text{inv}}$ proves essential in bridging the modality gap. The regularization not only improves the quality of the synthesized features but also reinforces the overall joint representation.
3. **Advantages of Cascaded Input:** Our ablation studies demonstrate that providing cascaded invariant inputs to each autoencoder layer in UIIM (the synthesis module) enables a more refined and accurate reconstruction of the missing modality.
4. **Overall Performance Gains:** The average WA and UA scores of UIIN surpass those of all baseline models by a significant margin. The observed improvements are not only statistically significant but also consistent across multiple runs, underscoring the robustness and reproducibility of our approach.

Moreover, the additional formulas introduced for fusion consistency and overall accuracy provide deeper insights into the mechanisms by which UIIN reconciles missing modalities. Such comprehensive experimental validations affirm that UIIN establishes a new benchmark for robust multimodal emotion recognition in scenarios with uncertain modality availability.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2025 doi:10.20944/preprints202503.0992.v1

Table 2. Performance comparison on IEMOCAP for multi-modality testing conditions: {a,v} (acoustic and visual), {a,t} (acoustic and textual), {v,t} (visual and textual), and the overall average. WA and UA denote Weighted Accuracy and Unweighted Accuracy, respectively. Arrows ↑ and ⇑ denote that the current result outperforms all baselines and ablation variants, while symbols * and + indicate parity with the best-performing baseline or ablation system.

System	Testing Conditions							
	{a,v}		{a,t}		{v,t}		Average	
	WA	UA	WA	UA	WA	UA	WA	UA
ine MCTN [7]	0.5593	0.5530	0.6801	0.6920	0.6740	0.6805	0.5872	0.5890
ine MMIN [1]	0.6465	0.6540	0.7301	0.7452	0.7205	0.7281	0.6385	0.6452
ine MMIN w/o cycle [1]	0.6228	0.6435	0.7168	0.7420	0.7180	0.7272	0.6290	0.6435
ine **UIIN (ours)**	**0.6548** $^{\uparrow}_{+}$	**0.6665** $^{\uparrow}_{+}$	**0.7425** $^{\uparrow}_{\Uparrow}$	**0.7560** $^{\uparrow}_{\Uparrow}$	**0.7285** $^{\uparrow}_{\Uparrow}$	**0.7375** $^{\uparrow}_{\Uparrow}$	**0.6454** $^{\uparrow}_{\Uparrow}$	**0.6538** $^{\uparrow}_{\Uparrow}$
ine w/o $\mathcal{L}_{\text{inv}}$	0.6502	0.6658	0.7335	0.7512	0.7168	0.7270	0.6388	0.6488
ine w/o cascaded input	0.6520	0.6640	0.7338	0.7520	0.7175	0.7290	0.6400	0.6510

4.6. Summary and Insights

Our comprehensive experimental results demonstrate that UIIN consistently achieves superior performance across a range of missing-modality conditions. By integrating invariant feature learning with a cascaded synthesis module and employing carefully designed loss functions, UIIN not only bridges the modality gap but also delivers robust and reliable emotion recognition. The ablation studies confirm the necessity of each module, while the visualization analyses provide intuitive evidence of effective invariant feature alignment and convergence. Overall, these findings establish UIIN as a state-of-the-art framework for multimodal emotion recognition in scenarios with uncertain and incomplete data.

5. Conclusions and Future Directions

In this paper, we presented a novel invariant feature aware multimodal emotion recognition framework, termed UIIN, which integrates a CMD-based invariant feature learning strategy with an innovative missing modality synthesis module. By leveraging the invariant representations, UIIN effectively mitigates the modality gap and significantly enhances the robustness of the joint multimodal representation. Our approach not only extracts modality-specific features but also constructs a comprehensive modality-invariant feature H that serves as the cornerstone for synthesizing missing modalities. This dual-level feature extraction and fusion mechanism enables UIIN to maintain high recognition accuracy even when one or more modalities are absent.

The extensive experiments on the IEMOCAP dataset demonstrated that UIIN outperforms state-of-the-art baselines under a variety of missing modality conditions. Quantitative results, expressed in terms of both Weighted Accuracy (WA) and Unweighted Accuracy (UA), confirm that the use of invariant features contributes to improved joint representation quality. In addition, the integration of the CMD-based distance metric and the cascaded autoencoder design within the missing modality synthesis module has shown to be critical for reducing inter-modality discrepancies. These findings underscore the effectiveness of our invariant feature learning and synthesis strategy in handling real-world challenges associated with incomplete multimodal inputs. Moreover, the performance of UIIN can be mathematically characterized by its joint loss function where each component plays a pivotal role in refining the joint representation. The careful balancing of these loss components enables UIIN to achieve both robust classification performance and stable convergence during training.

Looking forward, there are several promising avenues for future research. First, we intend to explore more sophisticated invariant feature learning mechanisms, potentially incorporating adversarial training techniques or attention-based fusion strategies to further enhance the extraction of modality-invariant information. Second, extending UIIN to handle additional modalities and more complex real-world scenarios remains an open challenge. Future work could investigate the integration of

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 March 2025 doi:10.20944/preprints202503.0992.v1

contextual information and temporal dynamics using advanced recurrent architectures or transformer models. Third, optimizing the autoencoder structure within the UIIM module by experimenting with different cascading strategies and deeper network architectures may yield further performance improvements. Finally, we aim to apply UIIN to other domains beyond emotion recognition, such as sentiment analysis and multimodal medical diagnosis, to evaluate its generalization capabilities across diverse tasks.

In summary, our work establishes UIIN as a robust and effective framework for multimodal emotion recognition in the presence of missing modalities. The promising experimental results and theoretical insights provided in this study pave the way for future innovations in invariant feature learning and multimodal data fusion.

References

1. Jinming Zhao, Ruichen Li, and Qin Jin, "Missing modality imagination network for emotion recognition with uncertain missing modalities," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, Eds. 2021, pp. 2608–2618, Association for Computational Linguistics.
2. Jiajia Tang, Kang Li, Xuanyu Jin, Andrzej Cichocki, Qibin Zhao, and Wanzeng Kong, "CTFN: Hierarchical learning for multimodal sentiment analysis using coupled-translation fusion network," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Online, Aug. 2021, pp. 5301–5311, Association for Computational Linguistics.
3. Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji, "Deep adversarial learning for multi-modality missing data completion," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, Yike Guo and Faisal Farooq, Eds. 2018, pp. 1158–1166, ACM.
4. Qiuling Suo, Weida Zhong, Fenglong Ma, Ye Yuan, Jing Gao, and Aidong Zhang, "Metric learning on healthcare data with incomplete modalities," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus, Ed. 2019, pp. 3534–3540, ijcai.org.
5. Changde Du, Changying Du, Hao Wang, Jinpeng Li, Wei-Long Zheng, Bao-Liang Lu, and Huiguang He, "Semi-supervised deep generative modelling of incomplete multi-modality emotional data," in *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, Susanne Boll, Kyoung Mu Lee, Jiebo Luo, Wenwu Zhu, Hyeran Byun, Chang Wen Chen, Rainer Lienhart, and Tao Mei, Eds. 2018, pp. 108–116, ACM.
6. Jing Han, Zixing Zhang, Zhao Ren, and Björn W. Schuller, "Implicit fusion by joint audiovisual training for emotion recognition in mono modality," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. 2019, pp. 5861–5865, IEEE.
7. Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. 2019, pp. 6892–6899, AAAI Press.
8. Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA, 2020, MM '20, p. 1122–1131, Association for Computing Machinery.
9. Ricardo Guerrero, Hai Xuan Pham, and Vladimir Pavlovic, "Cross-modal retrieval and synthesis (X-MRS): closing the modality gap in shared subspace learning," in *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran, Eds. 2021, pp. 3192–3201, ACM.
10. Dan Jia, Alexander Hermans, and Bastian Leibe, "Domain and modality gaps for lidar-based person detection on mobile robots," *CoRR*, vol. abs/2106.11239, 2021.
11. Alexander Liu, SouYoung Jin, Cheng-I Lai, Andrew Rouditchenko, Aude Oliva, and James Glass, "Cross-modal discrete representation learning," in *Proceedings of the 60th Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, May 2022, pp. 3013–3035, Association for Computational Linguistics.

12. Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz, "Central moment discrepancy (CMD) for domain-invariant representation learning," in *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. 2017, OpenReview.net.
13. Hasim Sak, Andrew W. Senior, and Françoise Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *CoRR*, vol. abs/1402.1128, 2014.
14. Yoon Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans, Eds. 2014, pp. 1746–1751, ACL.
15. Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Lang. Resour. Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.
16. Florian Eyben, Martin Wöllmer, and Björn W. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010*, Alberto Del Bimbo, Shih-Fu Chang, and Arnold W. M. Smeulders, Eds. 2010, pp. 1459–1462, ACM.
17. Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger, "Densely connected convolutional networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 2017, pp. 2261–2269, IEEE Computer Society.
18. Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun, Eds., 2015.
19. Neo Wu, Bradley Green, Xue Ben, and Shawn O'Banion, "Deep transformer models for time series forecasting: The influenza prevalence case," *arXiv preprint arXiv:2001.08317*, 2020.
20. Ishwar Baidari and Nagaraj Honnikoll, "Accuracy weighted diversity-based online boosting," *Expert Syst. Appl.*, vol. 160, pp. 113723, 2020.
21. Shruti Gupta, Md. Shah Fahad, and Akshay Deepak, "Pitch-synchronous single frequency filtering spectrogram for speech emotion recognition," *Multim. Tools Appl.*, vol. 79, no. 31-32, pp. 23347–23365, 2020.
22. Ziwang Fu, Feng Liu, Qing Xu, Jiayin Qi, Xiangling Fu, Aimin Zhou, and Zhibin Li, "NHFNET: A non-homogeneous fusion network for multimodal sentiment analysis," in *IEEE International Conference on Multimedia and Expo, ICME 2022, Taipei, Taiwan, July 18-22, 2022*. 2022, pp. 1–6, IEEE.
23. Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.
24. Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. RECON: Relation Extraction using Knowledge Graph Context in a Graph Neural Network. In *Proceedings of the Web Conference 2021*. 1673–1685.
25. Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management CIKM*. 729–738.
26. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
27. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
28. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
29. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the*

Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.

30. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
31. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
32. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
33. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
34. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. doi:10.1007/s00530-010-0182-0.
35. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
36. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
37. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
38. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
39. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
40. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
41. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
42. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
43. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
44. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
45. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
46. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
47. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
48. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.

49. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
50. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
51. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
52. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
53. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
54. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
55. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
56. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
57. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
58. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
59. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
60. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.
61. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
62. K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: A method for automatic evaluation of machine translation,” in *ACL*, 2002, pp. 311–318.
63. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
64. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
65. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
66. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
67. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
68. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: Agriculture-specific question answer system. *IndiaRxiv*, 2019.
69. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
70. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In 2021 International Conference on Communication information and Computing Technology (ICCICT), pages 1–6. IEEE, 2021.

 doi:10.20944/preprints202503.0992.v1

71. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
72. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
73. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
74. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
75. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEEMMT*, 2005, pp. 65–72.
76. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
77. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
78. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
79. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
80. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
81. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
82. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
83. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
84. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
85. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: A conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
86. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
87. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
88. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
89. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
90. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In

 doi:10.20944/preprints202503.0992.v1

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14734–14751, 2023.

91. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
92. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
93. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
94. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.