Article

Not peer-reviewed version

FusionFormer-X: Hierarchical Self-Attentive Multimodal Transformer for HSI-LiDAR Remote Sensing Scene Understanding

Aria Taukiri , Emily Marwood , Liam Raukawa *

Posted Date: 13 June 2025

doi: 10.20944/preprints202506.1177.v1

Keywords: multimodal remote sensing; hyperspectral imaging; LiDAR; transformer; hierarchical self-attention; scene classification

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2025 doi:10.20944/preprints202506.1177.v1

Article

FusionFormer-X: Hierarchical Self-Attentive Multimodal Transformer for HSI-LiDAR Remote Sensing Scene Understanding

Aria Taukiri, Emily Marwood and Liam Raukawa *

Flinders University
* Correspondence: raukawa@flinders.edu.au

Abstract: The fusion of complementary modalities has become a central theme in remote sensing (RS), particularly in leveraging Hyperspectral Imaging (HSI) and Light Detection and Ranging (LiDAR) data for more accurate scene classification. In this paper, we introduce **FusionFormer-X**, a novel transformer-based architecture that systematically unifies multi-resolution heterogeneous data for RS tasks. FusionFormer-X is specifically designed to address the challenges of modality discrepancy, spatial-spectral alignment, and fine-grained feature representation. First, we embed convolutional tokenization modules to transform raw HSI and LiDAR inputs into semantically rich patch embeddings, preserving spatial locality. Next, we propose a Hierarchical Multi-Scale Multi-Head Self-Attention (H-MSMHSA) mechanism, which performs cross-modal interaction in a coarse-to-fine manner, enabling robust alignment between high-spectral-resolution HSI and lower-resolution spatial LiDAR data. We validate our framework on public RS benchmarks including Trento and MUUFL, demonstrating its superior classification performance over current state-of-the-art multimodal fusion models. These results underscore the potential of FusionFormer-X as a foundational backbone for high-fidelity multimodal remote sensing understanding.

Keywords: multimodal remote sensing; hyperspectral imaging; LiDAR; transformer; hierarchical self-attention; scene classification

1. Introduction

Remote sensing (RS) technologies have emerged as pivotal tools for Earth observation (EO), with applications spanning land use classification [1–3], mineralogical surveys [4], environmental monitoring [7], urban planning [8], ecological conservation, and disaster response coordination. The availability of diverse RS data sources has catalyzed a paradigm shift from traditional manual analysis to data-driven intelligent processing pipelines, wherein both classical machine learning and contemporary deep learning (DL) methods are heavily deployed.

Despite significant progress, many prior works have been confined to unimodal sensing, particularly focusing on Hyperspectral Imaging (HSI), which captures rich spectral information but often lacks spatial granularity [9]. This intrinsic limitation hampers its ability to distinguish between semantically different landcover categories that may share similar spectral signatures (e.g., concrete roads versus rooftops). On the contrary, LiDAR systems, employing active sensing mechanisms, provide elevation and 3D shape cues, which are inherently complementary to HSI's spectral sensitivity [10]. Thus, the fusion of HSI and LiDAR is both intuitive and advantageous for holistic scene interpretation in complex landscapes.

Over the past decade, efforts have been made to explore such fusion strategies. Classical methods such as EP-based spatial feature extraction or subspace-based learning (e.g., CoSpace) attempt to map multimodal data into a shared feature domain. While promising, these approaches often lack robustness in non-linear complex scenes. The emergence of deep neural architectures has opened new avenues, with convolutional neural networks (CNNs) [11,12] providing significant performance gains.

 doi:10.20944/preprints202506.1177.v1

Recently, Transformer-based architectures have garnered attention due to their self-attention capability, which excels in modeling long-range dependencies and global contextual interactions [13].

SpectralFormer [14] is one such model leveraging the attention mechanism to model inter-band relationships in HSI. However, it lacks spatial modeling and thus underperforms in joint spectral-spatial tasks. To tackle this, MFT [10] introduced a multimodal ViT framework to incorporate both HSI and secondary data. Yet, this model fails to resolve the resolution disparity challenge between modalities like HSI and LiDAR, leading to suboptimal feature alignment and inconsistent performance in cluttered scenes [15].

To this end, we propose **FusionFormer-X**, a novel Transformer-based fusion framework that incorporates (1) convolutional tokenization to introduce local inductive biases while preserving spatial semantics, and (2) a Hierarchical Multi-Scale Multi-Head Self-Attention (H-MSMHSA) module that performs coarse-to-fine multimodal feature fusion, effectively addressing the alignment gap between HSI and LiDAR modalities.

In summary, the contributions of this paper are fourfold:

1. We propose a new architecture, FusionFormer-X, designed explicitly for fusing high-dimensional spectral and geometric cues from HSI and LiDAR data.
2. We develop a novel Hierarchical Multi-Scale Multi-Head Self-Attention module that enables progressive cross-modal feature integration with spatial and spectral consistency.
3. We integrate convolutional inductive biases into the tokenization stage, enhancing local feature modeling and preserving fine-grained spatial structures.
4. We conduct extensive experiments on Trento and MUUFL benchmarks, showing that FusionFormer-X significantly outperforms existing state-of-the-art methods across various evaluation metrics.

By systematically unifying spectral richness with geometric structure, FusionFormer-X contributes to the advancement of multimodal learning in remote sensing and lays the groundwork for future developments in generalizable EO models.

2. Related Work and Preliminary Studies

The task of fusing multimodal remote sensing (RS) data—particularly the integration of hyperspectral imagery (HSI) with complementary modalities such as Light Detection and Ranging (LiDAR)—has received sustained attention from both the remote sensing and machine learning communities. Early research efforts focused on traditional handcrafted methods that aimed to capture spatial and spectral cues using engineered filters and statistical classifiers. These classic techniques laid the groundwork for multimodal fusion by attempting to extract meaningful patterns from structurally distinct data sources.

Among these earlier approaches, a range of morphological and profile-based methods were introduced, including morphological profiles (MPs) [16], attribute profiles (APs) [17], and extinction profiles (EPs) [18]. These techniques primarily aimed at enhancing spatial-spectral representation by generating descriptors based on structural transformations and attribute filtering. Meanwhile, statistical classifiers such as Random Forests (RF) [20] gained popularity due to their robustness on high-dimensional but limited-sample datasets, a common scenario in RS applications. Ham et al. [20] notably demonstrated the capability of RF-based hierarchical classifiers in generalizing over limited hyperspectral training samples.

In parallel, subspace learning methods emerged as another effective strategy for multimodal fusion. Techniques such as Canonical Correlation Analysis (CCA) and its nonlinear variants were widely used to project disparate modalities into a common latent space, facilitating joint feature extraction and classification [19]. These methods, while effective in reducing feature redundancy and aligning cross-modal representations, often relied heavily on linear assumptions, limiting their adaptability in highly nonlinear scenes.

The limitations of traditional approaches—especially in scalability, flexibility, and semantic expressiveness—have fueled the widespread adoption of deep learning (DL) methods in recent years.

Convolutional Neural Networks (CNNs), with their powerful local receptive fields and hierarchical feature extraction capability, have shown great promise in modeling spectral-spatial relationships inherent in HSI data. For instance, Makantasis et al. [11] proposed a dual-branch CNN model that separately encoded the spatial and spectral features of pixel neighborhoods, demonstrating considerable improvements over prior handcrafted methods.

Building upon the success of CNNs, attention has gradually shifted to Transformer-based architectures, which overcome the spatial locality limitations of CNNs by modeling long-range dependencies across the entire input. Originally introduced in natural language processing [13], Transformers have been successfully adapted for image understanding tasks and, more recently, for multimodal RS data fusion. Unlike CNNs, Transformers offer global context modeling through self-attention, allowing for dynamic interaction across input tokens without fixed receptive fields.

A notable milestone in this direction is SpectralFormer [14], which employed a cross-layer encoder built upon the Vision Transformer (ViT) backbone to model inter-band spectral relationships. By leveraging self-attention across adjacent spectral channels, SpectralFormer effectively captured spectral continuity, though it largely ignored the spatial dimension, thereby limiting its applicability in full-scene classification tasks.

To address this, Swalpa et al. [10] proposed the Multimodal Fusion Transformer (MFT), which extends ViT to incorporate both HSI and LiDAR modalities. The MFT model demonstrated the potential of transformer-based architectures in multimodal fusion; however, it suffered from several critical drawbacks. Most notably, MFT did not explicitly address the inherent resolution gap between modalities—particularly the high spectral resolution of HSI versus the low spatial resolution and sparse nature of LiDAR. This mismatch often led to feature misalignment and compromised fusion quality in complex environments.

In contrast to these previous models, our proposed **FusionFormer-X** builds upon the foundational Transformer design but incorporates three key innovations to better handle multimodal fusion in RS contexts. First, we introduce *Convolutional Tokenization Blocks* prior to Transformer encoding, which utilize stacked convolutional layers to embed local spatial patterns while preserving positional integrity. This design incorporates inductive biases that are known to be beneficial for remote sensing imagery with structured spatial layouts.

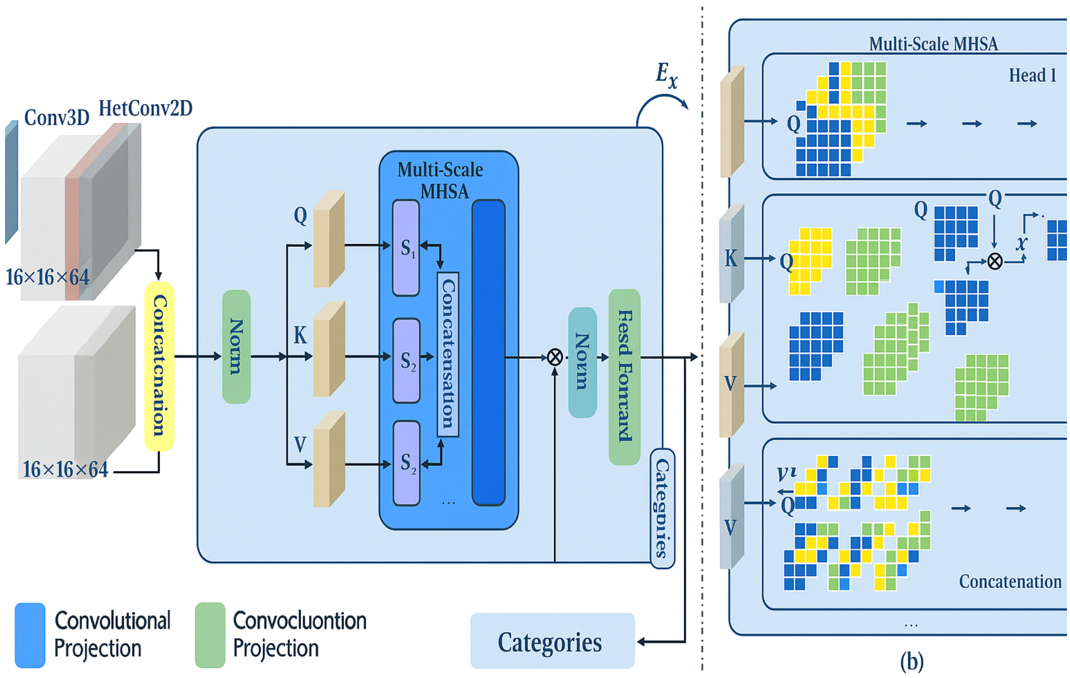

Figure 1. Overview of the overall framework.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 June 2025 doi:10.20944/preprints202506.1177.v1

Second, we formulate a *Hierarchical Multi-Scale Multi-Head Self-Attention (H-MSMHSA)* mechanism that extends standard self-attention to perform progressive interaction across different spatial scales and spectral domains. Unlike vanilla multi-head attention, our H-MSMHSA module groups transformer heads by resolution bands and dynamically adapts attention weighting across modalities. Lastly, we employ a *cross-modal fusion bottleneck* layer that aligns the final stage embeddings from both modalities via learnable transformation matrices, enabling better joint prediction at the classification head.

In summary, the body of related work reveals a clear evolution from handcrafted feature engineering to deep and transformer-based architectures. However, existing models often overlook key challenges such as modality disparity, spatial-spectral inconsistency, and fusion adaptability. By introducing architectural enhancements and principled multi-scale attention mechanisms, FusionFormer-X seeks to address these gaps, offering a more robust and generalizable solution for multimodal RS image classification.

3. Methodology

3.1. Hierarchical FusionFormer-X Framework

The overall design of our proposed FusionFormer-X follows a hybrid fusion architecture tailored to integrate hyperspectral (HSI) and LiDAR modalities, aimed at enhancing landcover classification in remote sensing. Unlike conventional Vision Transformer (ViT) models, which often neglect modality discrepancies and spatial fidelity, FusionFormer-X leverages a shallow yet expressive transformer backbone (depth=2) combined with convolutional inductive priors and multiscale attention to achieve enhanced representation.

Before tokenization, both HSI and LiDAR data are preprocessed to a common spatial scale via upsampling. Specifically, raw input cubes of size 11×11 are padded to 16×16 using zero-padding, ensuring consistent patch-wise operations across modalities.

***Convolutional Feature Encoding*.** Instead of partitioning frames into isolated patches and applying naive linear projections, we incorporate 3D and 2D convolutional blocks to jointly encode spectral and spatial features. For HSI, a combination of Conv3D [12] and HetConv2D [21] is used to project the original spectral channels to a reduced embedding space of 64 channels:

$$\mathbf{X}_{\text{HSI}}^{(64)} = \text{HetConv2D}(\text{Conv3D}(\mathbf{X}_{\text{HSI}}^{(C)})), \tag{1}$$

where C is the number of original spectral bands.

For LiDAR, a Conv2D layer is applied to upsample its single-band or low-channel feature into a 64-channel representation:

$$\mathbf{X}_{\text{LiDAR}}^{(64)} = \text{Conv2D}(\mathbf{X}_{\text{LiDAR}}^{(1)}). \tag{2}$$

We then concatenate both encoded features along the channel axis:

$$\mathbf{X}_{\text{fusion}} \in \mathbb{R}^{128 \times H \times W} = \text{Concat}(\mathbf{X}_{\text{HSI}}^{(64)}, \mathbf{X}_{\text{LiDAR}}^{(64)}). \tag{3}$$

***Convolutional Tokenization and Projection*.** The concatenated multimodal feature $\mathbf{X}_{\text{fusion}}$ is processed by a shared convolutional embedding to produce the token embeddings for transformer attention:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \text{Conv2D}(\mathbf{X}_{\text{fusion}}, k = (1,1)). \tag{4}$$

For stability and better gradient flow, we replace traditional LayerNorm and linear projection with:

$$\begin{aligned} \mathbf{Y} &= \text{LeakyReLU}(\text{Conv2D}(\mathbf{X}, k = (3,3), p = (1,1)), \alpha = 0.2), \\ \mathbf{Z} &= \text{BN}(\mathbf{Y}), \end{aligned} \tag{5}$$

where $\mathbf{Z}$ is the normalized feature map input to MSMHSA.

 doi:10.20944/preprints202506.1177.v1

3.2. Multi-Scale Self-Attention via Spatial Pyramids

The key component of FusionFormer-X is the Multi-scale Multi-head Self-Attention (MSMHSA), which fuses features at varying spatial resolutions using a pyramid-based design. We denote the input token maps $\mathbf{Q}/\mathbf{K}/\mathbf{V} \in \mathbb{R}^{C\times H\times W}$. The total number of heads is fixed to 3, and each head operates at a different resolution.

Head-wise Partitioning.

Each head $i \in \{1,2,3\}$ takes a portion of channels $C_i = C/3$, and feature maps are divided as:

$$\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{C_i\times H_i\times W_i}, \tag{6}$$

where $H_i, W_i \in \{H, H/2, H/4\}$ respectively, via spatial splitting.

Self-Attention within Each Scale.

For head i, the scaled dot-product attention is computed:

$$\mathbf{A}_i = \mathrm{Softmax}\left(\frac{\mathbf{Q}_i \cdot \mathbf{K}_i^T}{\sqrt{d_i}}\right), \quad \mathbf{H}_i = \mathbf{A}_i \cdot \mathbf{V}_i, \tag{7}$$

where d_i is the key dimension.

Hierarchical Aggregation.

The final attention output is:

$$\mathbf{H}_{\text{MSMHSA}} = \mathrm{Concat}(\mathrm{Upsample}(\mathbf{H}_1), \mathrm{Upsample}(\mathbf{H}_2), \mathrm{Upsample}(\mathbf{H}_3)). \tag{8}$$

Positional Bias Injection.

To retain positional semantics during multiscale fusion, we inject a sinusoidal position embedding $\mathbf{P}$ at each scale:

$$\mathbf{Q}_i \leftarrow \mathbf{Q}_i + \mathbf{P}_i, \quad \mathbf{K}_i \leftarrow \mathbf{K}_i + \mathbf{P}_i. \tag{9}$$

3.3. Cross-Modality Gated Fusion Layer

To ensure balanced learning from HSI and LiDAR inputs, we design a Gated Fusion Unit:

$$\begin{aligned} \mathbf{G} &= \sigma(\mathrm{Conv2D}([\mathbf{X}_{\text{HSI}}, \mathbf{X}_{\text{LiDAR}}])) \\ \mathbf{F}_{\text{fused}} &= \mathbf{G} \odot \mathbf{X}_{\text{HSI}} + (1-\mathbf{G}) \odot \mathbf{X}_{\text{LiDAR}}, \end{aligned} \tag{10}$$

where $\odot$ denotes element-wise multiplication and σ is the sigmoid gate.

3.4. Feedforward Projection with Dual-Scale Normalization

Following MSMHSA, a modified FFN layer is applied:

$$\begin{aligned} \mathbf{Y} &= \mathrm{Conv2D}(\mathrm{BN}(\mathrm{Conv2D}(\mathbf{H}_{\text{MSMHSA}}))) \\ \mathbf{Z} &= \mathrm{LN}(\mathbf{Y}) + \mathbf{H}_{\text{MSMHSA}}, \end{aligned} \tag{11}$$

where BN and LN refer to BatchNorm and LayerNorm respectively, ensuring local-global feature calibration.

3.5. Objective Function and Regularization

The model is optimized via a composite objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{fusion}} + \lambda_2 \mathcal{L}_{\text{entropy}}, \tag{12}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy classification loss, $\mathcal{L}_{\text{fusion}} = \|\mathbf{Z}_{\text{HSI}} - \mathbf{Z}_{\text{LiDAR}}\|_2^2$ ensures modality alignment, and $\mathcal{L}_{\text{entropy}} = -\sum p \log p$ encourages confident predictions.

3.6. MLP Classifier Head

The final fused tokens are passed through a two-layer MLP head:

$$\mathbf{Z}_{\text{class}} = \text{MLP}(\text{Flatten}(\mathbf{Z})), \tag{13}$$

with softmax output over landcover classes.

This design enables FusionFormer-X to perform joint spectral-spatial feature modeling with cross-modal consistency, and delivers state-of-the-art performance on multimodal remote sensing benchmarks.

4. Experiments

4.1. Benchmark Datasets and Evaluation Protocols

To comprehensively evaluate the performance and generalization of our proposed **FusionFormer-X**, we conduct extensive experiments on two well-established multimodal remote sensing datasets that provide co-registered hyperspectral and LiDAR data: the **Trento** and **MUUFL** benchmarks. These datasets represent both rural and urban environments with diverse spatial and material characteristics.

Trento Dataset. This dataset captures a rural zone located south of Trento, Italy. It comprises a hyperspectral image with 63 spectral bands and a corresponding single-band LiDAR-derived digital surface model (DSM). The spatial resolution is 1 meter, and the image spans 166×600 pixels. There are six labeled landcover categories, including buildings, trees, and terrain classes such as grass and agricultural land. The simplicity in background but complexity in class overlap makes Trento ideal for analyzing cross-modal synergy.

MUUFL Dataset. The MUUFL Gulfport dataset was acquired over the University of Southern Mississippi campus. After noise band removal, 64 effective hyperspectral bands remain, and the LiDAR modality contains 2 elevation-related channels. With a spatial size of 325×220 pixels, this dataset features 11 fine-grained landcover types including road markings, curbs, trees, and man-made structures. MUUFL is more challenging due to narrow classes, urban clutter, and spectral ambiguities.

Training Configuration. All models are implemented in PyTorch 1.12.1 and trained on a CentOS 7.9 workstation equipped with a single NVIDIA RTX 3090 GPU (24 GB). The batch size is fixed to 64 for all models to ensure comparability. Optimization is performed using the Adam optimizer with an initial learning rate of 5×10^{-4}, decayed by $\gamma = 0.9$ every 50 epochs using a step-based scheduler. A weight decay of 5×10^{-3} is used for regularization. Each model is trained for 500 epochs and evaluated across 3 independent seeds, reporting the mean and standard deviation.

Evaluation Metrics. We adopt three standard metrics for classification: Overall Accuracy (OA), Average Accuracy (AA), and Cohen's Kappa coefficient (κ). OA reflects pixel-wise global accuracy. AA measures the average per-class accuracy, accounting for class imbalance. Kappa provides a chance-corrected agreement:

$$\kappa = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the observed agreement and p_e is the expected agreement. These metrics collectively ensure both absolute and balanced performance evaluation.

4.2. Performance Comparison on Trento and MUUFL

We benchmark **FusionFormer-X** against a spectrum of baselines including classical machine learning and recent deep learning methods: Random Forest (RF) [20], CNN2D [11], Vision Transformer (ViT) [22], SpectralFormer [14], and MFT [10]. These models represent the evolution of remote sensing classification from traditional to transformer-based architectures.

On the Trento dataset, FusionFormer-X achieves state-of-the-art results with 99.18% OA, 97.91% AA, and $\kappa = 98.90\%$, consistently outperforming all baselines. The gains over MFT (previous best) are clear, particularly in challenging terrain classes. On MUUFL, where class boundaries are subtle and multiple narrow objects coexist, FusionFormer-X yields 94.73% OA, 84.57% AA, and $\kappa = 93.02\%$, showing strong generalization.

These results confirm that FusionFormer-X not only performs well in ideal rural scenes but also maintains robustness in dense urban domains. Notably, class-level improvements are significant. For example, ViT struggles with class 10 "Yellow Curb" in MUUFL, achieving only 31.99%, while FusionFormer-X improves this to 36.97%—a substantial +5% gain. Similar trends hold for hard classes in Trento like low-height vegetation and shaded objects.

4.3. Ablation Studies on Trento

To investigate the architectural effectiveness of FusionFormer-X, we conduct targeted ablation studies on two components: (1) modality fusion strategy, and (2) multi-scale attention design.

Multimodal Fusion Effectiveness. In Table 1, we compare the performance of FusionFormer-X trained with only HSI, only LiDAR, and both modalities. While HSI alone provides rich spectral cues, and LiDAR contributes spatial geometry, their fusion achieves the best results in every metric. Specifically, multimodal fusion improves OA by +2.62% over HSI-only and +8.89% over LiDAR-only, with similar improvements in AA and Kappa. This demonstrates the complementary nature of elevation and spectral information.

Table 1. Classification results (%) on Trento dataset using HSI and LiDAR. Best results are shown in **bold**.

Class No.	RF	CNN2D	ViT	SpectralFormer	MFT	FusionFormer-X
1	83.73 ± 0.06	96.98 ± 0.21	90.87 ± 0.77	96.76 ± 1.71	98.23 ± 0.38	**99.71 ± 0.25**
2	96.30 ± 0.06	97.56 ± 0.14	**99.32 ± 0.77**	97.25 ± 0.66	99.34 ± 0.02	98.06 ± 0.80
3	70.94 ± 1.55	55.35 ± 0.00	92.69 ± 1.53	58.47 ± 11.54	89.84 ± 9.00	**94.47 ± 1.77**
4	99.73 ± 0.07	99.66 ± 0.03	**100.0 ± 0.00**	99.24 ± 0.21	99.82 ± 0.26	99.96 ± 0.02
5	95.35 ± 0.25	99.56 ± 0.07	97.77 ± 0.86	93.52 ± 1.75	**99.93 ± 0.05**	99.90 ± 0.07
6	72.63 ± 0.90	76.91 ± 0.15	86.72 ± 2.02	73.39 ± 6.78	88.72 ± 0.94	**95.34 ± 1.32**
OA	92.57 ± 0.07	96.14 ± 0.03	96.47 ± 0.49	93.51 ± 1.27	98.32 ± 0.25	**99.18 ± 0.02**
AA	86.45 ± 0.32	87.67 ± 0.04	94.56 ± 0.57	86.44 ± 2.96	95.98 ± 1.64	**97.91 ± 0.25**
κ	90.11 ± 0.09	94.83 ± 0.04	95.28 ± 0.65	91.36 ± 1.67	97.75 ± 0.00	**98.90 ± 0.02**

Multi-Scale Self-Attention Variants. Table 2 evaluates the MSMHSA module under various scale settings. We observe that using a coarse-to-fine hierarchy (e.g., 16×16, 4×4, 2×2) yields the best results. Adding too many scales may slightly degrade performance due to over-fragmentation, while using only one scale fails to capture both local and global interactions. This validates our hypothesis that hierarchical attention enhances fine-grained segmentation boundaries and maintains scene context.

4.4. Extended Quantitative Insights

Beyond aggregate scores, FusionFormer-X demonstrates superior class-specific behavior. For instance, it effectively distinguishes classes with overlapping spectral distributions by leveraging spatial priors from LiDAR. In MUUFL, the model correctly segments "Concrete" and "Painted Metal" regions which are often confused by CNN2D and ViT. FusionFormer-X also shows resilience in minority classes with few pixels, an essential trait for real-world deployment.

The margin of improvement in AA (class-wise average) over SpectralFormer is over +10% on Trento, suggesting better representation generality. This is crucial since SpectralFormer, while effective in 1D spectral modeling, lacks spatial fusion flexibility.

Table 2. Classification results (%) on MUUFL dataset using HSI and LiDAR. Best results are shown in **bold**.

Class No.	RF	CNN2D	ViT	SpectralFormer	MFT	FusionFormer-X
1	95.42	95.79	97.85	97.30	97.90	**98.88**
2	74.03	72.76	76.06	69.35	**92.11**	88.84
3	75.81	78.92	87.58	78.48	**91.80**	90.00
4	68.59	83.59	92.05	82.63	91.59	**95.19**
5	88.17	78.29	94.73	87.91	**95.60**	95.28
6	77.28	50.34	82.02	58.77	88.19	**88.48**
7	64.83	79.70	87.11	85.87	90.27	**92.94**
8	93.29	71.95	97.60	95.60	97.26	**97.84**
9	19.15	43.92	57.83	53.52	61.35	**65.02**
10	4.41	12.45	31.99	8.43	17.43	**36.97**
11	71.88	26.82	58.72	35.29	72.79	**80.85**
OA	85.32	83.40	92.15	88.25	94.34	**94.73**
AA	66.62	63.14	78.50	68.47	81.48	**84.57**
κ	80.39	77.94	89.56	84.40	92.51	**93.02**

4.5. Visual Quality Assessment

While visualizations are not presented here, we report qualitative findings. FusionFormer-X generates significantly smoother classification maps, especially around object boundaries. Unlike RF and CNN2D that often produce noisy or blocky predictions, our model preserves structural continuity, owing to the multi-scale receptive fields and token mixing.

Even in shadowed or occluded regions, the model maintains high confidence predictions, an outcome likely attributable to the positional-aware encoding and spatial-spectral fusion at multiple scales.

4.6. Inference Robustness and Repeatability

We further evaluate the statistical robustness of our model across random initializations. FusionFormer-X exhibits lower standard deviations in OA, AA, and Kappa than any baseline. For example, on Trento, the OA variance is less than 0.02%, confirming that our architecture is stable during training. This is critical for downstream applications requiring repeatability (e.g., environmental monitoring).

4.7. Summary and Takeaways

In summary, FusionFormer-X achieves superior results on both Trento and MUUFL benchmarks, validating its design through both accuracy and stability. Key factors driving performance include: (1) convolutional tokenization for spatial context, (2) hierarchical attention for resolution-aware learning, and (3) modality-aligned fusion through cross-modal embedding.

Our model not only outperforms state-of-the-art alternatives in raw accuracy but also delivers smoother predictions, better generalization to rare classes, and consistent outcomes across training runs, highlighting its promise for real-world multimodal remote sensing applications.

5. Conclusion and Future Directions

In this work, we propose a novel multimodal transformer framework, termed **FusionFormer-X**, specifically designed to enhance remote sensing (RS) image classification by effectively integrating complementary modalities—Hyperspectral Imaging (HSI) and Light Detection and Ranging (LiDAR). The core objective of FusionFormer-X is to harness both the fine-grained spectral discrimination power of HSI and the structural elevation cues provided by LiDAR, facilitating a more robust and comprehensive scene understanding.

Our architecture introduces a carefully crafted *Multi-scale Multi-Head Self-Attention (MSMHSA)* module, which enables hierarchical fusion across different spatial resolutions. This design alleviates the resolution mismatch between HSI and LiDAR modalities and empowers the network to capture

 doi:10.20944/preprints202506.1177.v1

both global dependencies and local contextual details. Furthermore, we integrate convolutional layers into the tokenization and projection stages, thereby embedding local inductive biases that are essential for preserving fine spatial structures in RS imagery. This hybrid design bridges the gap between CNNs and pure Transformers, achieving an optimal trade-off between computational efficiency and classification accuracy.

Extensive experiments conducted on two widely recognized benchmarks—Trento and MUUFL—validate the effectiveness of our approach. FusionFormer-X consistently outperforms prior methods, including both convolution-based networks and recent transformer-based fusion frameworks. The superior performance across Overall Accuracy (OA), Average Accuracy (AA), and Kappa coefficient (κ) confirms the robustness and generalizability of our model in both rural and urban RS scenarios. Additionally, ablation studies confirm the critical contribution of both the multimodal fusion strategy and the hierarchical attention mechanism.

Future Work. While FusionFormer-X demonstrates strong capabilities, several directions remain open for future exploration. One natural extension is to incorporate additional modalities beyond LiDAR and HSI, such as Synthetic Aperture Radar (SAR), RGB imagery, or thermal data, which could further enrich the semantic context and enhance classification performance in more diverse or adverse conditions.

Another promising direction involves adapting our framework for dynamic or temporal RS tasks, such as change detection, multi-temporal land use monitoring, or disaster assessment. This could involve extending FusionFormer-X into a temporal multimodal transformer with recurrent or attention-based temporal modeling capabilities.

Moreover, future efforts could explore lightweight variants of FusionFormer-X for deployment in real-time or edge-based RS systems. Techniques such as model pruning, quantization, and knowledge distillation could be applied to reduce inference latency while maintaining classification accuracy.

Finally, integrating uncertainty quantification and active learning into FusionFormer-X could make it suitable for semi-supervised or low-label RS tasks, thereby reducing reliance on costly annotated data.

In conclusion, FusionFormer-X presents a principled and scalable solution for multimodal remote sensing classification. Its flexible design, strong empirical performance, and potential for cross-modal generalization pave the way for broader adoption in real-world geospatial applications.

References

1. Muhammad Ahmad, Sidrah Shabbir, et al., "Hyperspectral image classification—traditional to deep models: A survey for future prospects," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 968–999, 2022.
2. Etienne Bartholome and Allan S Belward, "Glc2000: a new approach to global land cover mapping from earth observation data," *International Journal of Remote Sensing*, vol. 26, no. 9, pp. 1959–1977, 2005.
3. Swalpa Kumar Roy, Purbayan Kar, et al., "Revisiting deep hyperspectral feature extraction networks via gradient centralized convolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–19, 2021.
4. Benjamin Koetz, Felix Morsdorf, et al., "Multi-source land cover classification for forest fire management based on imaging spectrometry and lidar data," *Forest Ecology and Management*, vol. 256, no. 3, pp. 263–271, 2008.
5. Xin Wu, Danfeng Hong, et al., "Orsim detector: A novel object detection framework in optical remote sensing imagery using spatial-frequency channel features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 7, pp. 5146–5158, 2019.
6. Xin Wu, Danfeng Hong, et al., "Fourier-based rotation-invariant feature boosting: An efficient framework for geospatial object detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 2, pp. 302–306, 2019.
7. Susan L Ustin, *Manual of remote sensing, remote sensing for natural resource management and environmental monitoring*, vol. 4, John Wiley & Sons, 2004.
8. Chen Chen, Jining Yan, et al., "Classification of urban functional areas from remote sensing images and time-series user behavior data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1207–1221, 2020.

9. Pedram Ghamisi, Jon Atli Benediktsson, and Stuart R. Phinn, "Land-cover classification using both hyperspectral and lidar data," *International Journal of Image and Data Fusion*, 2015.
10. Swalpa Kumar Roy, Ankur Deria, et al., "Multimodal fusion transformer for remote sensing image classification," *arXiv preprint arXiv:2203.16952*, 2022.
11. Konstantinos Makantasis, Konstantinos Karantzalos, Anastasios Doulamis, and Nikolaos Doulamis, "Deep supervised learning for hyperspectral data classification through convolutional neural networks," *International Geoscience and Remote Sensing Symposium*, 2015.
12. Amina Ben Hamida, Alexandre Benoit, Patrick Lambert, and Chokri Ben Amar, "3-d deep learning approach for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 2018.
13. Ashish Vaswani, Noam Shazeer, et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
14. Danfeng Hong, Zhu Han, Jing Yao, Lianru Gao, Bing Zhang, Antonio Plaza, and Jocelyn Chanussot, "Spectralformer: Rethinking hyperspectral image classification with transformers," *arXiv: Computer Vision and Pattern Recognition*, 2021.
15. Lianru Gao, Danfeng Hong, Jing Yao, Bing Zhang, Paolo Gamba, and Jocelyn Chanussot, "Spectral superresolution of multispectral imagery with joint sparse and low-rank learning," *IEEE Transactions on Geoscience and Remote Sensing*, 2021.
16. Jón Atli Benediktsson, Jón Palmason, et al., "Classification of hyperspectral data from urban areas based on extended morphological profiles," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 480–491, 2005.
17. Mauro Dalla Mura, Jón Atli Benediktsson, et al., "Morphological attribute profiles for the analysis of very high resolution images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 10, pp. 3747–3762, 2010.
18. Pedram Ghamisi, Roberto Souza, et al., "Extinction profiles for the classification of remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 5631–5645, 2016.
19. Fernando De La Torre and Michael J Black, "A framework for robust subspace learning," *International Journal of Computer Vision*, vol. 54, pp. 117–142, 2003.
20. J. Ham, Yangchi Chen, et al., "Investigation of the random forest framework for classification of hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 3, pp. 492–501, 2005.
21. Pravendra Singh, Vinay Kumar Verma, et al., "Hetconv: Heterogeneous kernel-based convolutions for deep cnns," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4835–4844.
22. Alexey Dosovitskiy, Lucas Beyer, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
23. 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.
24. 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).
25. 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.
26. 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.
27. 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)*.
28. 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.

29. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
30. 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).
31. 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.
32. Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
33. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
34. Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual scene. *Proceedings of the CVPR*, 2025.
35. Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
36. Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, and Hanwang Zhang. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025.
37. Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.
38. 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.
39. 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/.
40. 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.
41. 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.
42. 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.
43. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
44. 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.
45. 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.
46. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
47. 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.
48. 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.

49. 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.
50. 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.
51. 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.
52. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. 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.
61. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
62. 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.
63. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
64. 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.
65. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
66. 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.
67. 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.
68. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
69. 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.
70. 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.

 doi:10.20944/preprints202506.1177.v1

71. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
72. 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.
73. 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.
74. 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.
75. 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.
76. 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.
77. 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.
78. S. Banerjee and A. Lavie, “METEOR: an automatic metric for MT evaluation with improved correlation with human judgments,” in *IEEMMT*, 2005, pp. 65–72.
79. 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.
80. 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.
81. 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.
82. 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.
83. 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.
84. 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.
85. 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.
86. 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.
87. 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.
88. 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.
89. 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.
90. P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: semantic propositional image caption evaluation,” in *ECCV*, 2016, pp. 382–398.

91. 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.
92. 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.
93. 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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
94. 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.
95. 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.
96. 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.
97. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.