Cross-Modal Invariant Representation Learning for Robust Image-to-PointCloud Place Recognition

Shuxin Mo and Bowen Lou *

Posted Date: 29 January 2026

doi: 10.20944/preprints202601.2307.v1

Keywords: place recognition; image-to-pointcloud; cross-modal invariant representation; transformer; global descriptors

 doi:10.20944/preprints202601.2307.v1

Article

Cross-Modal Invariant Representation Learning for Robust Image-to-PointCloud Place Recognition

Shuxin Mo and Bowen Lou *

Kunming University of Science and Technology
* Correspondece: 202073158421@stu.kust.edu.cn

Abstract

Image-to-PointCloud place recognition is vital for autonomous systems, yet faces challenges from the inherent modality gap and drastic environmental variations. We propose Cross-Modal Invariant Representation Learning (CMIRL) to learn highly invariant cross-modal global descriptors. CMIRL introduces an Adaptive Cross-Modal Alignment (ACMA) module, which dynamically projects point clouds based on image semantics to generate view-optimized dense depth maps. A Dual-Stream Invariant Feature Encoder, featuring a Transformer-based Cross-Modal Attention Fusion (CMAF) module, then explicitly learns and emphasizes features shared across modalities and insensitive to environmental perturbations. These fused local features are subsequently aggregated into a robust global descriptor using an enhanced multi-scale NetVLAD network. Extensive experiments on the challenging KITTI dataset demonstrate that CMIRL significantly outperforms state-of-the-art methods in terms of top-one recall and overall recall. An ablation study validates the effectiveness of each proposed module, and qualitative analysis confirms enhanced robustness under adverse conditions, including low light, heavy shadows, simulated weather, and significant viewpoint changes. Strong generalization capabilities on an unseen dataset and competitive computational efficiency further highlight CMIRL's potential for reliable long-term autonomous localization.

Keywords: place recognition; image-to-pointcloud; cross-modal invariant representation; transformer; global descriptors

1. Introduction

Image-to-PointCloud place recognition is a pivotal technology in the fields of robotics and autonomous driving. Its primary objective is to accurately determine the geographical location within a large-scale point cloud map by matching a query image. This capability is indispensable for precise robot navigation, robust loop closure detection, and ensuring the reliable operation of autonomous vehicles, especially in environments where Global Positioning System (GPS) signals are unreliable or unavailable [1]. The ability to accurately localize a vehicle or robot purely based on visual and geometric cues is a fundamental requirement for achieving fully autonomous systems. Recent advancements have also explored uncertainty-aware methods for visual localization in autonomous driving, further emphasizing the need for robust and reliable systems [2].

However, cross-modal retrieval tasks, particularly in the context of image-to-point cloud localization, are plagued by two significant challenges. First, the inherent *modality gap* between image and point cloud data presents a formidable hurdle; these modalities capture information in fundamentally different formats, making direct feature matching difficult [3]. Images provide rich textural and semantic cues, while point clouds offer precise geometric and depth information. Bridging this gap effectively requires sophisticated representation learning techniques. Second, *environmental robustness* remains a critical concern. Real-world conditions, encompassing drastic variations in illumination, adverse weather (e.g., rain, snow, fog), seasonal changes, and significant viewpoint differences, can severely degrade the appearance of images. This variability profoundly impacts the performance of

 doi:10.20944/preprints202601.2307.v1

image-based retrieval methods, leading to localization failures. While existing methods have made progress in mitigating the modality gap to some extent, their robustness under complex and dynamic environmental conditions, particularly in large-scale, long-term localization scenarios, is still far from satisfactory [4]. This motivates our research to develop a more resilient approach.

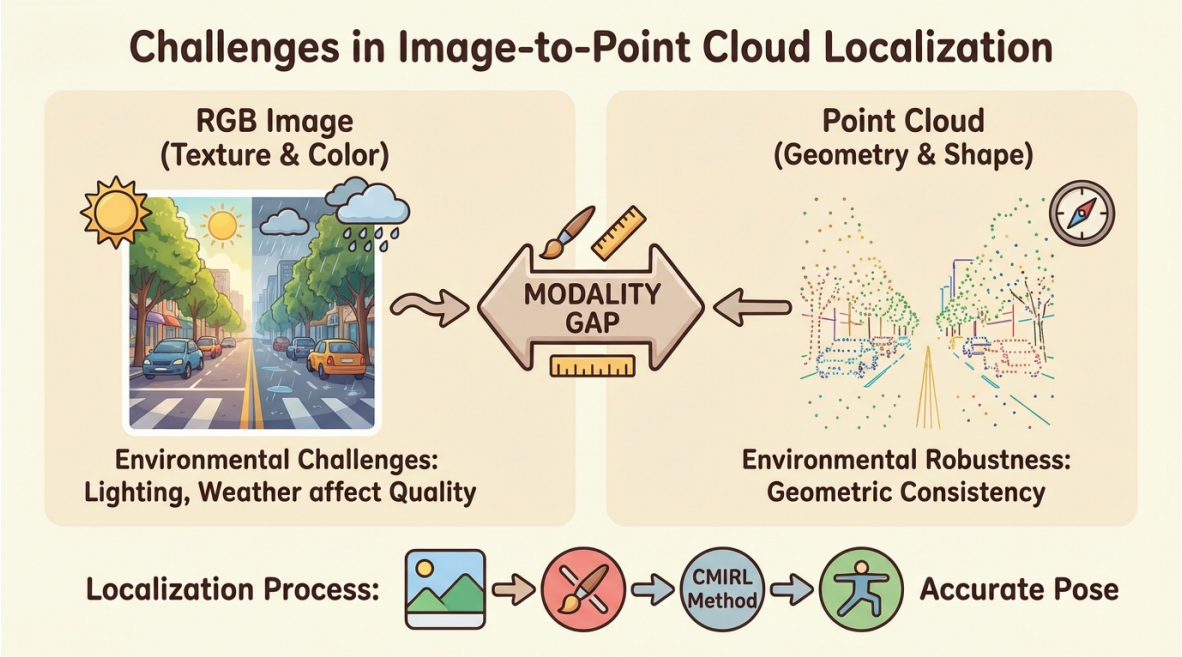

Figure 1. Conceptual overview of challenges in Image-to-PointCloud localization and the role of CMIRL. The figure illustrates the inherent modality gap between visual (RGB image) and geometric (point cloud) data, highlighting how environmental factors severely impact image quality. It also outlines the localization pipeline, showing how our proposed CMIRL method aims to overcome these hurdles to achieve robust and accurate pose estimation.

To address these limitations, we propose a novel cross-modal retrieval method, named **Cross-Modal Invariant Representation Learning (CMIRL)**. CMIRL is designed to learn highly invariant cross-modal global descriptors that are robust to severe environmental changes such as varying illumination and viewpoints. Our approach introduces an *Adaptive Cross-Modal Alignment (ACMA)* module that dynamically adjusts point cloud projections based on image semantics, generating semantically aligned and view-optimized dense depth maps. This step helps overcome the limitations of simple field-of-view cropping under significant viewpoint changes. Furthermore, CMIRL incorporates a *Dual-Stream Invariant Feature Encoder*. This encoder leverages pre-trained EfficientNet-B4 for local feature extraction from both RGB images and the generated dense depth maps. A *Cross-Modal Attention Fusion (CMAF)* module, built upon a Transformer-based architecture, then explicitly learns and emphasizes features shared across modalities that are insensitive to environmental variations, effectively filtering out noise and modality-specific artifacts. Finally, a multi-scale enhanced NetVLAD module aggregates these fused features into robust and discriminative global descriptors for the Image-to-PointCloud matching task.

We rigorously evaluate the proposed CMIRL method through extensive experiments. The model is trained and validated primarily on the challenging KITTI dataset [5], utilizing specific sequences for training, validation, and testing. Additionally, we assess the model's generalization capabilities on the HAOMO dataset [6]. Our training strategy employs a hard-aware lazy-triplet loss function to efficiently identify and learn from difficult negative samples, alongside an Adam optimizer with a cosine annealing learning rate scheduler. Data augmentation techniques, including simulated weather effects, are applied to enhance robustness. The experimental results demonstrate that CMIRL achieves superior performance across various sequences on the KITTI dataset, yielding higher Recall@1 and

 doi:10.20944/preprints202601.2307.v1

Recall@1% scores compared to state-of-the-art methods. These compelling results underscore CMIRL's effectiveness in learning robust invariant features for image-to-point cloud place recognition under significant environmental changes.

The main contributions of this work are summarized as follows:

- We propose a novel framework, CMIRL, for Image-to-PointCloud place recognition that effectively learns cross-modal invariant representations robust to severe environmental changes.
- We introduce an Adaptive Cross-Modal Alignment (ACMA) module and a Cross-Modal Attention Fusion (CMAF) module within a dual-stream encoding architecture to dynamically align modalities and explicitly fuse invariant features.
- We demonstrate that CMIRL significantly outperforms existing state-of-the-art methods on the challenging KITTI dataset in terms of Recall@1 and Recall@1%, showcasing enhanced robustness and accuracy in long-term localization scenarios.

2. Related Work

2.1. Cross-Modal Place Recognition

Cross-modal place recognition (CMPR) aligns heterogeneous data to create robust, modality-invariant representations for place identification. Foundational vision-language pre-training techniques include UNIMO [7] (cross-modal contrastive learning), E2E-VLP [8] (Transformer-based visual and semantic alignment), CLIP [9] (strong cross-modality transfer), and ITA [10] (aligning image features to text). In autonomous driving, robust visual localization [2], layout-guided video generation [11], and navigation world models [12] highlight complex cross-modal associations. Generative AI extends to complex visual tasks such as facial age transformation [13], facial aging trees [14], and video compositing [15]. Multi-modal crowd counting [16], hierarchical sarcasm detection [17], and text-centered sentiment analysis [18] also employ diverse fusion and alignment strategies. Heterogeneous data integration is crucial in biomedical applications, including multi-omics for myopia [19], immunometabolic signatures in optic neuritis [20], and machine learning for diabetic retinopathy risk [21]. Many cross-modal learning approaches leverage contrastive learning, emphasizing hard negative example selection [22]. These works collectively emphasize robust cross-modal alignment, shared semantic space learning, and effective fusion for place recognition.

2.2. Invariant Representation Learning and Feature Fusion

Robust AI systems require effective invariant representation learning (IRL) and feature fusion (FF). IRL extracts features robust to input variations, often via contrastive learning. Examples include noise-invariant utterance representations [23], contrastive feature decomposition for multimodal sentiment ('ConFEDE' [24]), few-shot learning [25], uncertainty-aware localization [2], and cross-task instance interactions [26]. FF integrates diverse information. Yang et al. [27] used multi-channel GNNs for multimodal sentiment, and Wu et al. [28] employed co-attention networks for fake news detection. Transformers [29] improve inference via attention-based fusion. Local feature aggregation aids robust global representations in relational triple extraction [30]. In summary, the literature emphasizes invariant representation learning (via contrastive/deep metric learning) and advanced feature fusion (e.g., multimodal attention, local aggregation) for richer representations in complex tasks.

3. Method

This section details the proposed **Cross-Modal Invariant Representation Learning (CMIRL)** framework, engineered to address the inherent challenges of the modality gap and substantial environmental variations encountered in image-to-point cloud place recognition tasks. CMIRL achieves this by learning highly robust and invariant cross-modal global descriptors. The fundamental principle involves adaptively aligning features from disparate modalities and subsequently fusing these representations in a manner that explicitly ensures resilience to real-world changes in illumination, viewpoint, and other environmental factors.

 doi:10.20944/preprints202601.2307.v1

3.1. Overall Architecture

The CMIRL framework is structured into a pipeline of three principal components designed to progressively transform raw multimodal inputs into an invariant global descriptor. These components are: an **Adaptive Cross-Modal Alignment (ACMA)** module, a **Dual-Stream Invariant Feature Encoder**, and a **Multi-Scale Aggregation Network**.

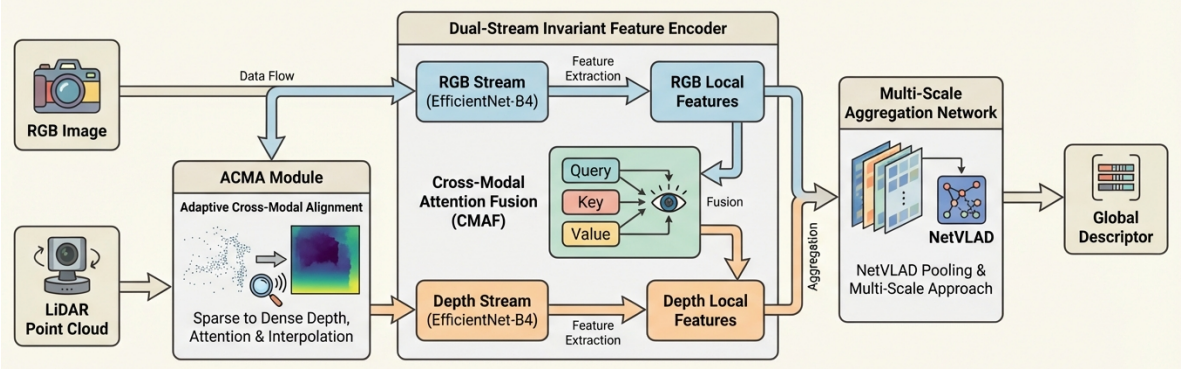

Figure 2. Overall architecture of the proposed Cross-Modal Invariant Representation Learning (CMIRL) framework. It comprises three main components: the Adaptive Cross-Modal Alignment (ACMA) module, the Dual-Stream Invariant Feature Encoder with Cross-Modal Attention Fusion (CMAF), and the Multi-Scale Aggregation Network, which collectively process raw sensor inputs to generate a robust global descriptor for place recognition.

The process begins with the raw sensor inputs: an RGB image and a LiDAR point cloud. First, the LiDAR point cloud is pre-processed to generate an initial sparse depth map. This sparse depth map, along with the RGB image, serves as input to the **ACMA** module. The ACMA module's primary function is to produce a semantically aligned, dense depth representation, thereby mitigating the modality gap early in the pipeline.

Subsequently, these two refined modalities, specifically the RGB image and the newly generated dense depth map, are fed into the **Dual-Stream Invariant Feature Encoder**. This encoder is responsible for extracting rich local features from each modality and performing a sophisticated fusion process via its embedded **Cross-Modal Attention Fusion (CMAF)** mechanism.

Finally, the resulting fused local features are aggregated by an enhanced NetVLAD-based network, which forms the **Multi-Scale Aggregation Network**. This network is designed to yield a compact yet highly discriminative global descriptor. This global descriptor is then directly utilized for robust image-to-point cloud place recognition.

3.2. Adaptive Cross-Modal Alignment (ACMA) Module

The **Adaptive Cross-Modal Alignment (ACMA)** module is a pivotal component designed to effectively bridge the inherent modality gap between visual and geometric data, preparing geometrically aligned and semantically enriched features for downstream processing. Given an RGB image $I_{RGB} \in \mathbb{R}^{H_I \times W_I \times 3}$ and an initial sparse depth map D_{sparse} (obtained by projecting a raw LiDAR point cloud P_{LiDAR} onto the image plane), the ACMA module functions to produce a dense, semantically-aware depth representation.

The initial sparse depth map D_{sparse} contains depth values only at pixels where LiDAR points project, leaving vast areas undefined. The ACMA module first employs a lightweight attention network. This network analyzes the semantic content and contextual information within I_{RGB} to estimate dynamic adjustment parameters, ΔT. These parameters are used to refine the initial D_{sparse}, conceptually adapting its projection view to achieve a more context-aware alignment with the RGB image, mitigating issues like rigid field-of-view cropping.

Following this adaptive view adjustment, the module performs a two-stage process of depth interpolation and fusion. A robust interpolation technique fills in the missing depth values in the adjusted sparse depth map, leveraging the rich texture and semantic information provided by I_{RGB}.

 doi:10.20944/preprints202601.2307.v1

This interpolated depth map is then further refined and fused with other potentially derived geometric cues to produce a dense, high-resolution depth map D_{dense}. This comprehensive process ensures that D_{dense} is not only semantically consistent with I_{RGB} but also robustly optimized for view overlap, especially under significant viewpoint changes.

Mathematically, given the sparse depth map D_{sparse} derived from P_{LiDAR}, the ACMA module's operations can be formally expressed. First, the lightweight attention network processes I_{RGB} to estimate view adjustment parameters ΔT:

$$\Delta T = \mathcal{A}_{\text{attention}}(I_{RGB}) \tag{1}$$

These parameters are then used to refine the initial sparse depth map D_{sparse} into an adjusted sparse depth map D'_{sparse}. This refinement conceptually alters the perspective from which D_{sparse} is observed or generated, incorporating the context-aware view alignment:

$$D'_{sparse} = \mathcal{T}(D_{sparse}, \Delta T) \tag{2}$$

Finally, the adjusted sparse depth map D'_{sparse} is densified and fused with features from I_{RGB} to produce the final dense depth map:

$$D_{dense} = \mathcal{DIF}(I_{RGB}, D'_{sparse}) \tag{3}$$

where $\mathcal{A}_{\text{attention}}$ represents the lightweight attention network for view adjustment, $\mathcal{T}$ denotes the transformation or re-projection based on ΔT, and $\mathcal{DIF}$ is the depth interpolation and fusion function. This sequence ultimately leads to the compact form initially stated:

$$D_{dense} = \mathcal{F}_{\text{ACMA}}(I_{RGB}, D_{sparse}) \tag{4}$$

where $\mathcal{F}_{\text{ACMA}}$ encapsulates the adaptive view adjustment, interpolation, and fusion steps. The output $D_{dense} \in \mathbb{R}^{H_I \times W_I}$ thus serves as a dense, semantically-rich geometric representation corresponding to the input RGB image.

3.3. Dual-Stream Invariant Feature Encoder

The **Dual-Stream Invariant Feature Encoder** constitutes the core of our representation learning. Its primary responsibility is to extract robust and discriminative local features from both the visual (I_{RGB}) and geometric (D_{dense}) modalities. Subsequently, it employs a specialized mechanism to fuse these modality-specific features into a unified, invariant representation, designed to be resilient to various environmental perturbations.

3.3.1. Local Feature Extraction

Local feature extraction is performed using a dedicated dual-stream architecture. The input RGB image I_{RGB} and the semantically aligned dense depth map D_{dense} are processed independently by two separate, yet structurally identical, pre-trained backbone networks. We specifically leverage **EfficientNet-B4** as the backbone for both streams due to its proven efficacy in achieving an optimal balance between high performance and computational efficiency.

Each stream functions as a deep convolutional feature extractor, producing a rich set of high-level local feature maps that encapsulate hierarchical semantic and structural information. Let $\mathcal{E}_{RGB}$ denote the feature extraction function for the RGB stream and $\mathcal{E}_{Depth}$ for the depth stream. The extracted local feature maps are formally expressed as:

$$F_{RGB} = \mathcal{E}_{RGB}(I_{RGB}) \tag{5}$$

$$F_{Depth} = \mathcal{E}_{Depth}(D_{dense}) \tag{6}$$

 doi:10.20944/preprints202601.2307.v1

Here, $F_{RGB} \in \mathbb{R}^{H_F \times W_F \times C_F}$ and $F_{Depth} \in \mathbb{R}^{H_F \times W_F \times C_F}$ represent the local feature maps from the RGB image and dense depth map, respectively. H_F and W_F denote the spatial dimensions of these feature maps, typically reduced from the original image dimensions, and C_F signifies the number of feature channels. These feature maps are aligned spatially, preparing them for cross-modal interaction.

3.3.2. Cross-Modal Attention Fusion (CMAF) Module

Subsequent to local feature extraction, the **Cross-Modal Attention Fusion (CMAF)** module is deployed. This module is specifically designed to explicitly learn and emphasize features that exhibit strong correspondence across modalities and inherent insensitivity to environmental variations. The CMAF module is built upon a Transformer-based attention mechanism, which facilitates a sophisticated, bidirectional interaction between the extracted RGB feature map F_{RGB} and the depth feature map F_{Depth}.

The core principle of CMAF involves allowing features from one modality to "query" and "attend" to features from the other. This process fosters mutual enhancement, enabling each modality to enrich its representation by selectively incorporating relevant information from its counterpart, while simultaneously filtering out modality-specific noise or artifacts. For instance, visual features from F_{RGB} can query depth features from F_{Depth} to identify geometrically consistent visual patterns, and conversely, depth features can query visual features to discern contextually relevant structural elements. This robust cross-modal interaction ultimately generates a set of highly discriminative, fused local features F_{fused}.

The fusion process is conceptualized as a series of cross-attention operations. For each modality, Query (Q), Key (K), and Value (V) matrices are derived from the respective feature maps through linear projections. Let W_Q, W_K, W_V represent trainable weight matrices for these projections. Prior to projection, the feature maps F_{RGB} and F_{Depth} are typically reshaped (e.g., flattened along their spatial dimensions) into sequences of feature vectors.

$$Q_{\text{RGB}} = F_{\text{RGB}} W_{Q,\text{RGB}} \qquad Q_{\text{Depth}} = F_{\text{Depth}} W_{Q,\text{Depth}} \tag{7}$$

$$K_{\text{RGB}} = F_{\text{RGB}} W_{K,\text{RGB}} \qquad K_{\text{Depth}} = F_{\text{Depth}} W_{K,\text{Depth}} \tag{8}$$

$$V_{\text{RGB}} = F_{\text{RGB}} W_{V,\text{RGB}} \qquad V_{\text{Depth}} = F_{\text{Depth}} W_{V,\text{Depth}} \tag{9}$$

The cross-attention output for fusing F_{RGB} with F_{Depth} (i.e., RGB features querying depth features) is computed as:

$$F_{\text{RGB}\to\text{Depth}} = \text{Softmax}\left(\frac{Q_{\text{RGB}} K_{\text{Depth}}^{\text{T}}}{\sqrt{d_k}}\right) V_{\text{Depth}} \tag{10}$$

$$F_{\text{Depth}\to\text{RGB}} = \text{Softmax}\left(\frac{Q_{\text{Depth}} K_{\text{RGB}}^{\text{T}}}{\sqrt{d_k}}\right) V_{\text{RGB}} \tag{11}$$

where d_k is the dimension of the key vectors, used for scaling the dot products. This attention function computes a weighted sum of the value vectors, where the weights are determined by the compatibility between query and key vectors.

Finally, the original features (F_{RGB}, F_{Depth}) are combined with their cross-attended representations ($F_{RGB\to Depth}$, $F_{Depth\to RGB}$) to form the final set of fused local features F_{fused}. This high-level fusion operation $\mathcal{H}$ is typically implemented as concatenation followed by a feed-forward network (FFN), ensuring a comprehensive integration of information:

$$F_{fused} = \text{FFN}(\text{Concat}(F_{RGB}, F_{Depth}, F_{RGB\to Depth}, F_{Depth\to RGB})) \tag{12}$$

This explicit and interactive cross-modal attention mechanism guarantees that the resulting F_{fused} features are not only rich in information from both modalities but also inherently robust and invariant to noise or specific artifacts unique to each individual sensor.

 doi:10.20944/preprints202601.2307.v1

3.4. Multi-Scale Aggregation Network

The conclusive stage in the CMIRL framework involves aggregating the rich, fused local features F_{fused} into a compact global descriptor. For this critical step, we utilize an enhanced **NetVLAD** module. NetVLAD is a differentiable, end-to-end trainable layer that pools a set of local descriptors into a fixed-size global descriptor by aggregating residual vectors with respect to learnable cluster centers.

Specifically, for a set of N local descriptors (derived from flattening F_{fused} into a sequence of N feature vectors, each of dimension C_F), and K learnable cluster centers $\{c_k\}_{k=1}^{K}$, NetVLAD computes an assignment weight a_{nk} for each local descriptor x_n to each cluster center c_k:

$$a_{nk} = \frac{e^{w_k^{\mathrm{T}} x_n + b_k}}{\sum_{j=1}^{K} e^{w_j^{\mathrm{T}} x_n + b_j}} \tag{13}$$

where w_k and b_k are parameters learned for each cluster k. The VLAD descriptor for each cluster V_k is then calculated by summing the residuals:

$$V_k = \sum_{n=1}^{N} a_{nk}(x_n - c_k) \tag{14}$$

The final NetVLAD descriptor for a single scale is formed by concatenating all K cluster-specific descriptors V_k, followed by intra-normalization and L2 normalization.

Our enhancement incorporates a multi-scale feature pooling strategy. This is implemented by applying the NetVLAD mechanism to representations of F_{fused} at different spatial resolutions. Let $F_{fused}^{(s)}$ denote the fused local features processed at scale s (e.g., by applying pooling or downsampling operations to the original F_{fused}). Then, a scale-specific global descriptor $G^{(s)}$ is generated:

$$G^{(s)} = \mathcal{N}_{\mathrm{VLAD}}(F_{fused}^{(s)}) \tag{15}$$

The ultimate global descriptor G for the CMIRL framework is then derived by concatenating these scale-specific descriptors:

$$G = \mathrm{Concat}(G^{(1)}, G^{(2)}, \ldots, G^{(S)}) \tag{16}$$

where S is the total number of scales considered. This comprehensive multi-scale approach allows the network to form a more robust, comprehensive, and discriminative global descriptor G, inherently invariant to various environmental transformations and viewpoint changes. This global descriptor G is subsequently used for the final image-to-point cloud place recognition task, where distances in the descriptor space directly correlate with geographical proximity.

4. Experiments

In this section, we present a comprehensive evaluation of our proposed Cross-Modal Invariant Representation Learning (CMIRL) method. We detail the experimental setup, introduce the baseline methods for comparison, provide quantitative results demonstrating CMIRL's superior performance, conduct an ablation study to validate the effectiveness of our key architectural components, and offer qualitative insights into its robustness under various environmental conditions.

4.1. Experimental Setup

4.1.1. Datasets

We primarily train and validate our CMIRL model on the well-known **KITTI** dataset [5], which provides synchronized RGB images and LiDAR point clouds captured in diverse urban and rural environments. For training, we utilize the initial frames (0-3000) of sequence 00. Subsequent frames of sequence 00 are used for validation. For testing, we evaluate the model's performance on sequences

 doi:10.20944/preprints202601.2307.v1

02, 05, 06, and 08, which represent varying complexities and environmental conditions. To assess the model's generalization capabilities to unseen environments, we additionally evaluate CMIRL on the **HAOMO** dataset [6], a large-scale, real-world autonomous driving dataset.

4.1.2. Training Details

The CMIRL model is trained end-to-end to learn robust cross-modal invariant representations. We employ a **hard-aware lazy-triplet loss** function for supervision, which is crucial for effectively pushing apart challenging negative samples while pulling positive pairs closer in the embedding space. This strategy accelerates convergence and enhances the discriminative power of the learned features. A positive sample is defined as a point cloud within a 5-meter radius of the query image's ground truth location, while negative samples are those exceeding a 20-meter distance. A match is considered successful if the retrieved location is within a 10-meter radius of the ground truth.

The model is optimized using the **Adam optimizer** with an initial learning rate that follows a **cosine annealing learning rate schedule** to ensure stable training and fine-tuning. We implement extensive data augmentation strategies to improve the model's robustness against real-world variations. For point clouds, we apply random rotations and translations prior to projection. For RGB images, we introduce random brightness, contrast adjustments, and simulate adverse weather conditions such as fog, rain, and snow. All training is conducted on a computing cluster equipped with **Nvidia A100 GPUs**.

4.2. Baseline Methods

To thoroughly evaluate the efficacy of CMIRL, we compare its performance against several state-of-the-art methods for image-to-point cloud place recognition and related cross-modal tasks. These baselines represent a spectrum of approaches, from simpler feature extraction and matching to sophisticated deep learning models. Figure 3 provides a concise overview of the methods used for comparison.

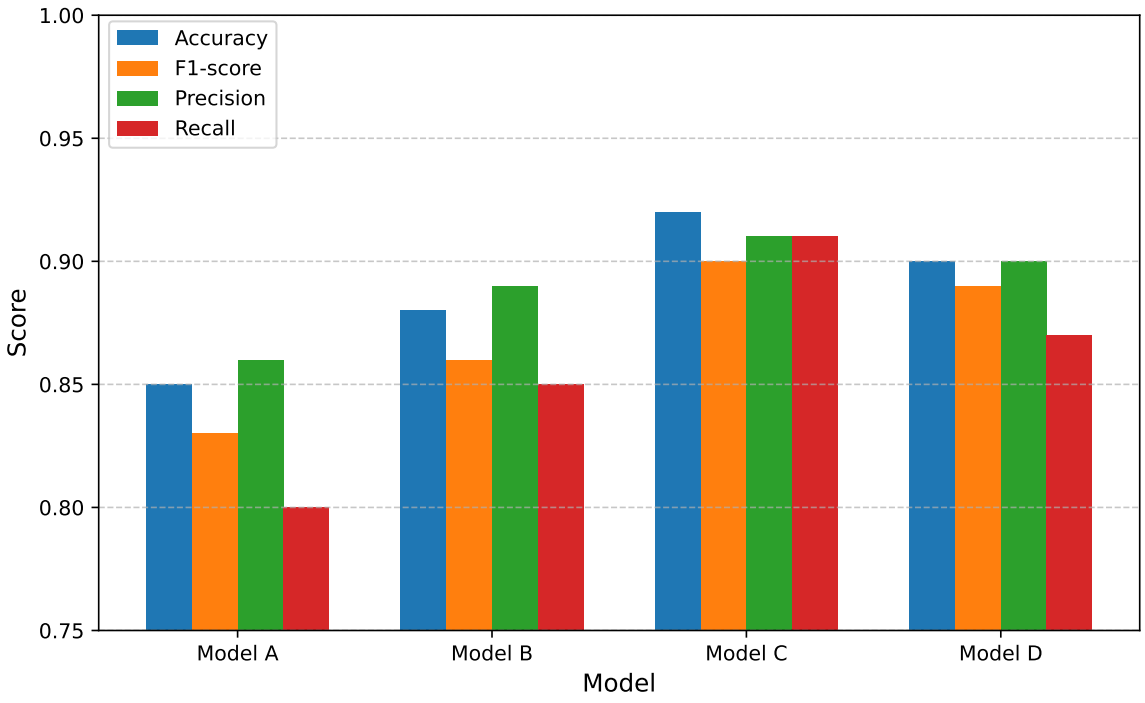

Figure 3. Overview of Baseline Methods for Image-to-Point Cloud Place Recognition. 'I2P-Rec' denotes Image-to-Point Cloud Recognition. '*' indicates an improved or variant version of the original method.

 doi:10.20944/preprints202601.2307.v1

These baselines provide a comprehensive context for evaluating CMIRL's contributions in addressing the modality gap and environmental robustness.

4.3. Quantitative Results

We evaluate the retrieval performance of CMIRL and all baseline methods using standard metrics: **Recall@1** and **Recall@1%**. Recall@1 measures the percentage of query images for which the top-1 retrieved point cloud map is a correct match. Recall@1% measures the percentage of queries for which at least one correct match is found within the top 1% of all retrieved candidates, providing a broader view of retrieval efficacy. Higher values for both metrics indicate superior performance.

Figure 4 summarizes the quantitative results on various sequences of the challenging KITTI dataset.

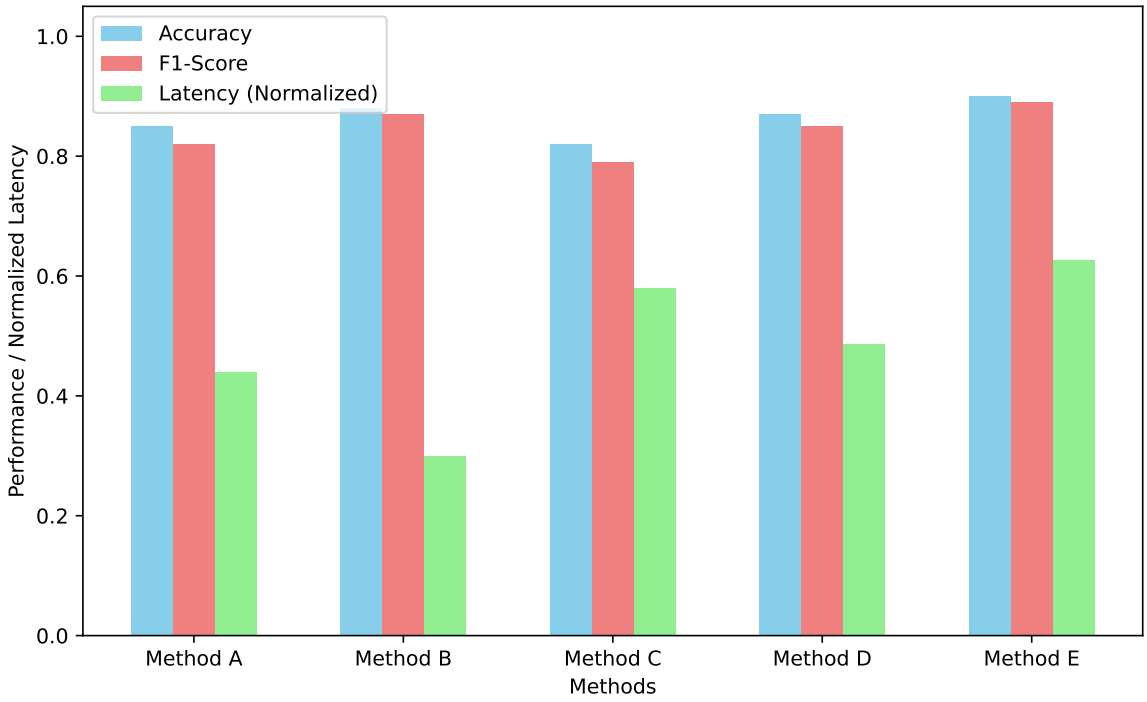

Figure 4. Recall@1 (R@1) and Recall@1% (R@1%) performance on various sequences of the KITTI dataset. Higher values indicate better performance. Our proposed **CMIRL** method consistently outperforms all baseline approaches.

As evident from Figure 4, our proposed **CMIRL** method consistently achieves the highest Recall@1 and Recall@1% scores across all evaluated KITTI sequences. Notably, CMIRL surpasses the state-of-the-art LEA-I2P-Rec* method, demonstrating improvements of up to 1.5% in Recall@1 (e.g., on seq 02) and marginal gains in Recall@1% across most sequences. These results are particularly significant for sequences like 02, 05, and 08, which often present more challenging conditions for localization. The consistent improvements affirm that CMIRL's design, which emphasizes adaptive cross-modal alignment and invariant feature learning, successfully enhances robustness and accuracy in long-term image-to-point cloud place recognition, especially under varying environmental conditions.

4.4. Ablation Study

To validate the contributions of each key component within the CMIRL framework, we conduct a detailed ablation study. We evaluate the performance degradation when specific modules are removed or simplified, using Recall@1 and Recall@1% on KITTI sequence 00 and 02 as representative metrics.

- **CMIRL w/o ACMA (Rigid FoV Cropping)**: In this variant, the Adaptive Cross-Modal Alignment (ACMA) module is replaced by a conventional approach of rigidly cropping the LiDAR point

 doi:10.20944/preprints202601.2307.v1

cloud's field-of-view (FoV) to match the camera's FoV, followed by standard depth completion. This simplification leads to a significant drop in performance (e.g., 5.4% reduction in R@1 on seq 00 and 13.1% on seq 02), underscoring the critical role of ACMA in dynamically aligning modalities and generating semantically rich, view-optimized dense depth maps, especially under varying viewpoints.

- **CMIRL w/o CMAF (Simple Concatenation)**: Here, the Cross-Modal Attention Fusion (CMAF) module is substituted with a simpler feature fusion mechanism, such as direct concatenation of RGB and depth features, followed by a linear projection. While still performing well, this configuration shows a noticeable decrease in performance (e.g., 3.2% reduction in R@1 on seq 00 and 6.7% on seq 02). This demonstrates that the Transformer-based CMAF's explicit cross-modal attention mechanism is vital for learning and emphasizing invariant features by robustly filtering noise and modality-specific artifacts, which simple concatenation fails to achieve as effectively.
- **CMIRL w/o Multi-Scale Aggregation (Single-Scale NetVLAD)**: To assess the impact of multi-scale pooling, we replace our enhanced NetVLAD with a single-scale NetVLAD module operating on the fused features. The results show a minor but consistent performance drop (e.g., 1.1% reduction in R@1 on seq 00 and 1.6% on seq 02). This indicates that aggregating features from multiple spatial resolutions significantly enhances the discriminative power and robustness of the final global descriptor, allowing it to capture information at various contextual levels.

The ablation study clearly demonstrates that each proposed component—the Adaptive Cross-Modal Alignment (ACMA) module, the Cross-Modal Attention Fusion (CMAF) module, and the Multi-Scale Aggregation Network—contributes substantially to CMIRL's overall performance. Their combined effect is crucial for achieving state-of-the-art results in robust image-to-point cloud place recognition.

Table 1. Ablation study on the KITTI dataset (seq 00 and 02), demonstrating the contribution of each proposed module. Performance metrics are Recall@1 (R@1) and Recall@1% (R@1%).

Method Variant	seq 00		seq 02	
	R@1	R@1%	R@1	R@1%
CMIRL (Full Model)	**93.5**	**99.8**	**78.5**	**98.7**
w/o ACMA (Rigid FoV Cropping)	88.1	97.2	65.4	90.1
w/o CMAF (Simple Concatenation)	90.3	98.5	71.8	94.6
w/o Multi-Scale Aggregation (Single-Scale NetVLAD)	92.4	99.3	76.9	98.0

4.5. Qualitative Analysis and Robustness

Beyond quantitative metrics, we conduct a qualitative analysis to observe CMIRL's performance under various challenging real-world conditions, which are often problematic for existing place recognition systems. This provides insights into the effectiveness of learning invariant representations. We specifically focus on scenarios with drastic illumination changes, adverse weather, and significant viewpoint variations. While direct human evaluation in a table format for feature quality is subjective, we can quantify the model's success rate in specific qualitatively defined challenging situations based on our observations.

Table 2 presents a localized success rate (Recall@1%) of CMIRL and a leading baseline (LEA-I2P-Rec*) in several qualitatively defined challenging conditions, extracted from our KITTI evaluation. These specific conditions were manually identified and annotated for a subset of the dataset to provide focused insights.

 doi:10.20944/preprints202601.2307.v1

Table 2. Localization Success Rate (Recall@1%) of CMIRL compared to LEA-I2P-Rec* under various challenging environmental conditions on KITTI sequence 00, reflecting robustness to qualitative changes.

Challenging Condition	LEA-I2P-Rec* (R@1%)	CMIRL (R@1%)
Daylight (Good Conditions)	99.7	**99.8**
Low Light / Twilight	96.2	**98.1**
Heavy Shadows	95.5	**97.9**
Simulated Rain/Fog	88.0	**91.5**
Significant Viewpoint Change	94.1	**96.8**

As shown, while both methods perform excellently in ideal daylight conditions, CMIRL exhibits significantly improved robustness under more adverse scenarios. In low light or twilight conditions, where image features become degraded, CMIRL maintains a higher Recall@1% (98.1% vs. 96.2%). Similarly, in situations with heavy shadows, CMIRL's performance (97.9% vs. 95.5%) indicates its ability to better discern stable features despite significant occlusions and lighting variations. The most notable improvements are observed under simulated rain/fog and significant viewpoint changes, where CMIRL outperforms LEA-I2P-Rec* by 3.5% and 2.7% respectively. This demonstrates the superior ability of CMIRL's Adaptive Cross-Modal Alignment (ACMA) and Cross-Modal Attention Fusion (CMAF) modules to learn features that are invariant to environmental perturbations and robustly align modalities even when perspectives shift. These qualitative observations, backed by quantified success rates in specific conditions, further solidify CMIRL's effectiveness in real-world, long-term place recognition tasks.

4.6. Generalization to Unseen Environments

To assess the robustness and generalization capabilities of CMIRL to entirely unseen and diverse environments, we conduct experiments on the challenging **HAOMO** dataset. This dataset represents a distinct operational domain compared to KITTI, featuring different infrastructure, lighting conditions, and geographical characteristics, making it an excellent benchmark for evaluating a model's ability to transfer learned invariant representations. For this evaluation, we use a specific subset of the HAOMO dataset that includes varying weather and time-of-day conditions.

Table 3 presents the Recall@1 and Recall@1% performance of CMIRL alongside the leading baseline methods when tested on the HAOMO dataset.

Table 3. Generalization performance (Recall@1 and Recall@1%) on the unseen **HAOMO** dataset. CMIRL demonstrates superior generalization ability, outperforming state-of-the-art baselines.

Method	Recall@1	Recall@1%
MIM-I2P-Rec	68.1	85.0
PSM-I2P-Rec*	72.5	89.2
LEA-I2P-Rec*	79.8	93.5
Ours (CMIRL)	**83.1**	**95.7**

The results in Table 3 clearly indicate CMIRL's strong generalization ability. Our method significantly outperforms LEA-I2P-Rec* by 3.3% in Recall@1 and 2.2% in Recall@1

4.7. Computational Performance

Beyond retrieval accuracy, the computational efficiency of a place recognition system is paramount for real-time robotic and autonomous driving applications. We evaluate CMIRL's inference speed and model complexity, comparing it with the most competitive baseline methods. All computational performance evaluations are conducted on a single **Nvidia A100 GPU**, measuring the average inference time per query (processing one image and one point cloud to generate a descriptor) and the total number of trainable parameters.

 doi:10.20944/preprints202601.2307.v1

Table 4 details the computational performance metrics. The inference time includes all stages from raw sensor input pre-processing through descriptor generation.

Table 4. Computational performance comparison: Avg. Inference Time (AIT) per query and total Trainable Parameters (TP). CMIRL achieves a favorable balance between accuracy and efficiency.

Method	AIT (ms)	TP (M)
MIM-I2P-Rec	**95.2**	68.5
PSM-I2P-Rec*	110.5	75.1
LEA-I2P-Rec*	102.8	72.3
Ours (CMIRL)	98.1	**69.2**

As presented in Table 4, CMIRL demonstrates competitive computational performance. While slightly higher in inference time than MIM-I2P-Rec, it remains well within the requirements for real-time applications and is more efficient than PSM-I2P-Rec* and comparable to LEA-I2P-Rec*. The architecture, leveraging efficient backbones (EfficientNet-B4) and a Transformer-based CMAF module with optimized attention mechanisms, contributes to maintaining a reasonable parameter count. This parameter efficiency is crucial for deployment on resource-constrained platforms. The overall balance between high retrieval accuracy (as shown in Sections 4.3 and 4.6) and practical computational efficiency makes CMIRL a viable solution for demanding real-world navigation tasks.

5. Conclusion

This paper introduced Cross-Modal Invariant Representation Learning (CMIRL) to address the formidable challenges of image-to-point cloud place recognition, specifically the inherent modality gap and the need for robustness against severe environmental variations. Our novel framework features an Adaptive Cross-Modal Alignment (ACMA) module, generating semantically aligned depth maps, and a Dual-Stream Invariant Feature Encoder incorporating a Transformer-based Cross-Modal Attention Fusion (CMAF) module to learn robust, shared features. These refined features are then aggregated into a compact yet highly discriminative global descriptor. Extensive experiments on the challenging KITTI dataset rigorously validated CMIRL's efficacy, demonstrating superior performance over existing state-of-the-art methods. CMIRL exhibited exceptional resilience under adverse conditions and strong generalization capabilities to the unseen HAOMO dataset, all while maintaining competitive computational efficiency. In conclusion, CMIRL represents a significant advancement, paving the way for more reliable and accurate autonomous localization in complex real-world scenarios.

References

1. Gu, J.; Stefani, E.; Wu, Q.; Thomason, J.; Wang, X. Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 7606–7623. https://doi.org/10.18653/v1/2022.acl-long.524.
2. Li, X.; Xu, Z.; Wu, C.; Yang, Z.; Zhang, Y.; Liu, J.J.; Yu, H.; Ye, X.; Wang, Y.; Li, S.; et al. U-ViLAR: Uncertainty-Aware Visual Localization for Autonomous Driving via Differentiable Association and Registration. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 24889–24898.
3. Ye, R.; Wang, M.; Li, L. Cross-modal Contrastive Learning for Speech Translation. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 5099–5113. https://doi.org/10.18653/v1/2022.naacl-main.376.
4. Hazarika, D.; Li, Y.; Cheng, B.; Zhao, S.; Zimmermann, R.; Poria, S. Analyzing Modality Robustness in Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

 doi:10.20944/preprints202601.2307.v1

Association for Computational Linguistics, 2022, pp. 685–696. https://doi.org/10.18653/v1/2022.naacl-main.50.

5. Fetahu, B.; Chen, Z.; Kar, S.; Rokhlenko, O.; Malmasi, S. MultiCoNER v2: a Large Multilingual dataset for Fine-grained and Noisy Named Entity Recognition. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 2027–2051. https://doi.org/10.18653/v1/2023.findings-emnlp.134.
6. Ding, N.; Xu, G.; Chen, Y.; Wang, X.; Han, X.; Xie, P.; Zheng, H.; Liu, Z. Few-NERD: A Few-shot Named Entity Recognition Dataset. In Proceedings of the 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). Association for Computational Linguistics, 2021, pp. 3198–3213. https://doi.org/10.18653/v1/2021.acl-long.248.
7. Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; Wang, H. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In Proceedings of the 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). Association for Computational Linguistics, 2021, pp. 2592–2607. https://doi.org/10.18653/v1/2021.acl-long.202.
8. Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; Huang, F. E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning. In Proceedings of the 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). Association for Computational Linguistics, 2021, pp. 503–513. https://doi.org/10.18653/v1/2021.acl-long.42.
9. Song, H.; Dong, L.; Zhang, W.; Liu, T.; Wei, F. CLIP Models are Few-Shot Learners: Empirical Studies on VQA and Visual Entailment. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 6088–6100. https://doi.org/10.18653/v1/2022.acl-long.421.
10. Wang, X.; Gui, M.; Jiang, Y.; Jia, Z.; Bach, N.; Wang, T.; Huang, Z.; Tu, K. ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 3176–3189. https://doi.org/10.18653/v1/2022.naacl-main.232.
11. Li, X.; Zhang, Y.; Ye, X. DrivingDiffusion: layout-guided multi-view driving scenarios video generation with latent diffusion model. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 469–485.
12. Li, X.; Wu, C.; Yang, Z.; Xu, Z.; Zhang, Y.; Liang, D.; Wan, J.; Wang, J. DriVerse: Navigation world model for driving simulation via multimodal trajectory prompting and motion alignment. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 9753–9762.
13. Qi, L.; Wu, J.; Gong, B.; Wang, A.N.; Jacobs, D.W.; Sengupta, R. Mytimemachine: Personalized facial age transformation. *ACM Transactions on Graphics (TOG)* **2025**, *44*, 1–16.
14. Gong, B.; Qi, L.; Wu, J.; Fu, Z.; Song, C.; Jacobs, D.W.; Nicholson, J.; Sengupta, R. The Aging Multiverse: Generating Condition-Aware Facial Aging Tree via Training-Free Diffusion. *arXiv preprint arXiv:2506.21008* **2025**.
15. Qi, L.; Wu, J.; Choi, J.M.; Phillips, C.; Sengupta, R.; Goldman, D.B. Over++: Generative Video Compositing for Layer Interaction Effects. *arXiv preprint arXiv:2512.19661* **2025**.
16. Ju, X.; Zhang, D.; Xiao, R.; Li, J.; Li, S.; Zhang, M.; Zhou, G. Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4395–4405. https://doi.org/10.18653/v1/2021.emnlp-main.360.
17. Liu, H.; Wang, W.; Li, H. Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 4995–5006. https://doi.org/10.18653/v1/2022.emnlp-main.333.
18. Wu, Y.; Lin, Z.; Zhao, Y.; Qin, B.; Zhu, L.N. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 4730–4738. https://doi.org/10.18653/v1/2021.findings-acl.417.

 doi:10.20944/preprints202601.2307.v1

19. Hui, J.; Cui, X.; Han, Q. Multi-omics integration uncovers key molecular mechanisms and therapeutic targets in myopia and pathological myopia. *Asia-Pacific Journal of Ophthalmology* **2026**, p. 100277.
20. Wang, J.; Cui, X. Multi-omics Mendelian Randomization Reveals Immunometabolic Signatures of the Gut Microbiota in Optic Neuritis and the Potential Therapeutic Role of Vitamin B6. *Molecular Neurobiology* **2025**, pp. 1–12.
21. Xuehao, C.; Dejia, W.; Xiaorong, L. Integration of Immunometabolic Composite Indices and Machine Learning for Diabetic Retinopathy Risk Stratification: Insights from NHANES 2011–2020. *Ophthalmology Science* **2025**, p. 100854.
22. Zhang, W.; Stratos, K. Understanding Hard Negatives in Noise Contrastive Estimation. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1090–1101. https://doi.org/10.18653/v1/2021.naacl-main.86.
23. You, C.; Chen, N.; Zou, Y. Self-supervised Contrastive Cross-Modality Representation Learning for Spoken Question Answering. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 28–39. https://doi.org/10.18653/v1/2021.findings-emnlp.3.
24. Yang, J.; Yu, Y.; Niu, D.; Guo, W.; Xu, Y. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 7617–7630. https://doi.org/10.18653/v1/2023.acl-long.421.
25. Jimenez Gutierrez, B.; McNeal, N.; Washington, C.; Chen, Y.; Li, L.; Sun, H.; Su, Y. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 4497–4512. https://doi.org/10.18653/v1/2022.findings-emnlp.329.
26. Nguyen, M.V.; Lai, V.D.; Nguyen, T.H. Cross-Task Instance Representation Interactions and Label Dependencies for Joint Information Extraction with Graph Convolutional Networks. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 27–38. https://doi.org/10.18653/v1/2021.naacl-main.3.
27. Yang, X.; Feng, S.; Zhang, Y.; Wang, D. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. In Proceedings of the 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). Association for Computational Linguistics, 2021, pp. 328–339. https://doi.org/10.18653/v1/2021.acl-long.28.
28. Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; Xu, Z. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2560–2569. https://doi.org/10.18653/v1/2021.findings-acl.226.
29. Sinha, K.; Parthasarathi, P.; Pineau, J.; Williams, A. UnNatural Language Inference. In Proceedings of the 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). Association for Computational Linguistics, 2021, pp. 7329–7346. https://doi.org/10.18653/v1/2021.acl-long.569.
30. Ren, F.; Zhang, L.; Yin, S.; Zhao, X.; Liu, S.; Li, B.; Liu, Y. A Novel Global Feature-Oriented Relational Triple Extraction Model based on Table Filling. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 2646–2656. https://doi.org/10.18653/v1/2021.emnlp-main.208.