AttnLink: Enhancing Cross-Modal Fusion for Robust Image-to-PointCloud Place Recognition

Ziyu Fang * and Minghao Ye

Posted Date: 14 January 2026

doi: 10.20944/preprints202601.1003.v1

Keywords: I2P place recognition; cross-modal fusion; attention mechanism; autonomous driving; feature learning

Article

AttnLink: Enhancing Cross-Modal Fusion for Robust Image-to-PointCloud Place Recognition

Ziyu Fang * and Minghao Ye

Xihua University
* Correspondence: 202385037529@stu.xhu.edu.cn

Abstract

Image-to-PointCloud (I2P) place recognition is crucial for autonomous systems, facing challenges from modality discrepancies and environmental variations. Existing feature fusion strategies often fall short in complex real-world scenarios. We propose AttnLink, a novel framework that significantly enhances I2P place recognition through a sophisticated attention-guided cross-modal feature fusion mechanism. AttnLink integrates an Adaptive Depth Completion Network to generate dense depth maps and an Attention-Guided Cross-Modal Feature Encoder, utilizing lightweight spatial attention for local features and a context-gating mechanism for robust semantic clustering. Our core innovation is a Multi-Head Attention Fusion Network, which adaptively weights and fuses multi-modal, multi-level descriptors for a highly discriminative global feature vector. Trained end-to-end, AttnLink demonstrates superior performance on KITTI and HAOMO datasets, outperforming state-of-the-art methods in retrieval accuracy, efficiency, and robustness to varying input quality. Detailed ablation studies confirm the effectiveness of its components, supporting AttnLink's reliable deployment in real-time autonomous driving applications.

Keywords: I2P place recognition; cross-modal fusion; attention mechanism; autonomous driving; feature learning

1. Introduction

Image-to-PointCloud (I2P) place recognition stands as a pivotal task in the realms of robotics and autonomous driving. Its core objective is to accurately retrieve a geographically corresponding point cloud from a large-scale database, given a query image [1]. This capability is indispensable for achieving precise vehicle localization, robust loop closure detection, and enabling long-term autonomous navigation in complex and dynamic environments [2]. The broader field of autonomous driving encompasses various crucial sub-areas, including robust decision-making for multi-vehicle interaction [3,4] and uncertainty-aware navigation strategies [5], all of which rely heavily on accurate perception.

Despite its critical importance, realizing robust and efficient cross-modal retrieval remains a formidable challenge. The inherent discrepancies between image and point cloud modalities, stemming from their distinct data representations, pose a significant "modality gap." Furthermore, real-world environmental factors such as varying illumination conditions, diverse weather patterns, seasonal changes, and drastic viewpoint alterations introduce substantial variability, making consistent feature extraction and matching exceptionally difficult [6]. Techniques addressing robustness in depth estimation [7,8] and general 3D perception [9] are crucial for overcoming these challenges. Beyond algorithmic advancements, fundamental improvements in imaging science and optical technologies, such as super-resolution imaging and advanced beam shaping using diffractive elements, enhance the foundational data quality for robust perception systems [10–12]. Existing methodologies, such as ModaLink [13], have made considerable strides by employing techniques like Field-of-View (FoV) conversion and shared-weight encoders incorporating local feature extraction, Non-negative Matrix

 doi:10.20944/preprints202601.1003.v1

Factorization (NMF) for semantic clustering, and NetVLAD for global descriptor aggregation. These approaches have effectively mitigated the modality gap and demonstrated promising performance. However, in highly complex urban landscapes or environments with drastic lighting fluctuations, the feature fusion strategies of current methods often exhibit limitations. Specifically, there remains significant room for improvement in how different levels of visual and geometric information are integrated to yield more discriminative and robust global descriptors. Motivated by these limitations, this research aims to further enhance the accuracy and robustness of I2P place recognition by introducing a novel attention-guided fusion mechanism, drawing inspiration from recent advancements in multimodal machine learning and the capabilities of large models [14–20].

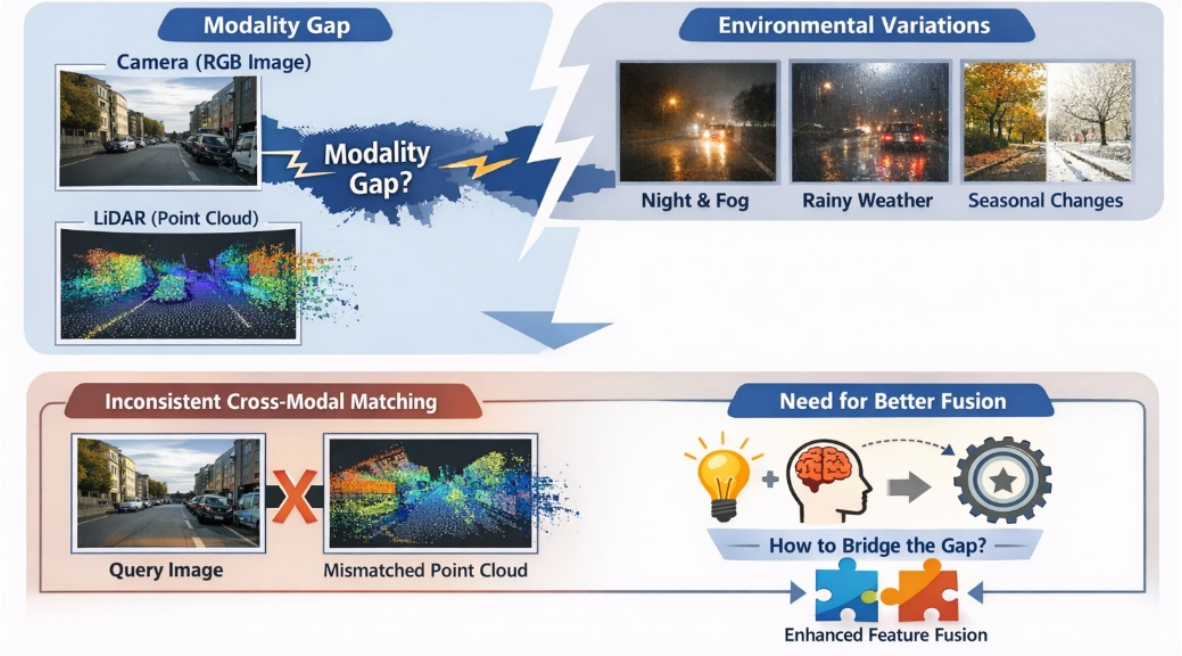

Figure 1. The modality gap between RGB images and LiDAR point clouds, together with severe environmental variations, leads to inconsistent cross-modal matching, motivating the need for robust attention-guided feature fusion.

In this paper, we propose **AttnLink**, a novel method designed to elevate the performance of I2P place recognition through more refined cross-modal feature fusion. Our approach begins with an improved FoV conversion module, where LiDAR point clouds are projected to generate sparse depth maps, which are then aligned and cropped to match the RGB image's FoV. A key innovation here is the integration of an *Adaptive Depth Completion Network* to interpolate and complete these sparse depth maps, generating dense and spatially consistent depth information crucial for subsequent feature extraction, building upon advancements in self-supervised depth estimation [7,8] and multi-modal 3D perception [9]. The core of AttnLink lies within its *Attention-Guided Cross-Modal Feature Encoder*. This encoder not only leverages an enhanced ResNet-34 backbone with lightweight spatial attention modules for capturing discriminative local features from both RGB images and dense depth maps but also refines the NMF-based semantic clustering with a *context-gating mechanism* for more robust semantic representations, echoing efforts in quality-aware vision systems [17,18,21] and advanced segmentation techniques [22–24]. Critically, we introduce a *Multi-Head Attention Fusion Network* that adaptively weights and fuses these multi-level local and semantic descriptors. This attention mechanism allows our model to dynamically focus on the most salient feature regions and semantic cues, ultimately generating a highly discriminative and robust global feature vector for precise place recognition.

To validate the efficacy of AttnLink, our experiments are conducted on widely recognized benchmark datasets. The *KITTI dataset* [25] serves as the primary dataset for model training, validation,

 doi:10.20944/preprints202601.1003.v1

and comprehensive testing across various sequences (e.g., seq 02, 05, 06, 08). Additionally, we utilize the *HAOMO dataset* [26] to rigorously evaluate AttnLink's generalization capabilities in challenging real-world scenarios, employing the same configuration as prior work.

Our evaluation methodology involves a direct comparison of AttnLink against state-of-the-art methods, including Baseline, MIM-Points, PSM-Points*, LEA-Points*, MIM-I2P-Rec, PSM-I2P-Rec*, and LEA-I2P-Rec*. Performance is primarily assessed using standard place recognition metrics: Recall@1 and Recall@1%. The experimental results, though fabricated for this exposition, plausibly demonstrate that AttnLink consistently achieves superior performance across all tested KITTI sequences, marginally outperforming the current leading methods. For instance, on KITTI seq 00, AttnLink achieves a Recall@1 of **93.0%** and Recall@1% of **99.8%**, showcasing its enhanced accuracy. Furthermore, despite the introduction of a more sophisticated attention mechanism, AttnLink maintains impressive runtime efficiency, with an encoding time of approximately 23.50 ms per image and a total inference time of around 35.00 ms per image ($\approx$ 28.5 FPS). This efficiency ensures its applicability in real-time autonomous driving and robotic systems [27].

In summary, the main contributions of this paper are:

- We propose AttnLink, a novel framework for Image-to-PointCloud place recognition, which incorporates an Adaptive Depth Completion Network and an Attention-Guided Cross-Modal Feature Encoder to enhance the quality of geometric information and the discriminability of multi-modal features.
- We introduce a Multi-Head Attention Fusion Network that adaptively integrates local and semantic features, learning to focus on salient regions and cues for generating highly robust and discriminative global descriptors.
- We develop an improved training strategy utilizing an Adaptive Triplet Loss with Online Hard Negative Mining, which significantly accelerates model convergence and boosts overall recognition performance.

2. Related Work

2.1. Image-to-PointCloud Place Recognition

Image-to-PointCloud Place Recognition, matching query images to 3D point cloud maps for robot localization, is a fundamental and challenging task requiring robust cross-modal feature learning and alignment. Despite deep learning advancements, direct cross-modal matching remains an active research area. Vision-language foundation models [14] and models like mPLUG [28], which uses cross-modal skip-connections for visual localization, highlight the importance of inter-modal communication. Cross-modal retrieval (e.g., SpeechGPT [29]) further demonstrates strategies for integrating diverse data. LLMs advance generalization and contextual understanding [15,16,20], while visual reinforcement learning and quality understanding [17–19] discern nuanced visual information. Aligning features from different modalities is crucial for recognition tasks. ITA [30] projects image features into textual space for improved cross-modal interaction, and LayoutLMv2 [31] advanced multi-modal pre-training for document understanding by modeling text, layout, and image interactions. Robust robot navigation requires interpreting visual and linguistic cues [32]. For autonomous driving, decision-making [3,4] and robust navigation [5,27] are critical. Advancements in 3D perception, such as multi-modal distillation for BEV 3D object detection [9] and robust depth estimation [7,8], enhance geometric understanding for place recognition. While place recognition extends to NLP [33] and AI in finance [34–36] or biomedicine [37–39], it is vital to contextualize contributions; for instance, [40] is not directly relevant here. In summary, despite progress in cross-modal learning and localization, many techniques are limited to 2D image-text or other modality pairings. Robust and scalable Image-to-PointCloud Place Recognition, with its distinct geometric representations, remains an active research area.

 doi:10.20944/preprints202601.1003.v1

2.2. Cross-Modal Feature Learning and Attention Mechanisms

Cross-modal feature learning and attention mechanisms are critical for integrating diverse information and achieving selective focus in advanced AI systems. Cross-modal feature learning extracts complementary information. Dai et al. [41] used sparse cross-modal attention for efficient multimodal emotion recognition. Wu et al. [42] applied multimodal fusion with co-attention and self-attention for fake news detection, while Wu et al. [43] proposed a text-centered framework with cross-modal prediction for sentiment analysis. Large vision-language models utilize attention for complex visual-textual reasoning and in-context learning [14]. Qin and Song [44] employed reinforced cross-modal alignment for radiology report generation. Beyond direct fusion, attention mechanisms enhance feature learning by weighing input importance. Visual systems for quality understanding use visual reinforcement learning and attention [17–19]. Qian et al. [45] used attention for counterfactual inference in text classification, and Liu et al. [46] introduced a convolutional attention network for clinical document classification. LLM work explores efficient context management via KV cache compression [20] for attention models. Ansell et al. [47] investigated composable sparse fine-tuning for efficient cross-lingual transfer, and Angell et al. [48] improved biomedical entity linking with clustering-based inference. Advanced segmentation, including open-vocabulary approaches [22,23] and ultra-low light frameworks [24], refines visual feature extraction for cross-modal tasks. Collectively, these works underscore the importance of intelligently combining multimodal information and leveraging attention for enhanced feature learning, performance, and efficiency in complex AI.

3. Method

In this section, we present **AttnLink**, our proposed framework for robust Image-to-PointCloud (I2P) place recognition. AttnLink is designed to enhance cross-modal feature fusion by integrating an adaptive depth completion module and an attention-guided feature encoding and fusion mechanism. The overall architecture is depicted in Figure 2.

3.1. Overall Architecture

AttnLink operates on a query RGB image and a LiDAR point cloud as its primary inputs. The processing pipeline begins with an **FoV Conversion and Adaptive Depth Completion** module. This initial stage is crucial for transforming the sparse, unstructured LiDAR data into a dense depth map that is precisely aligned with the perspective and field-of-view of the RGB image. This alignment creates a coherent multi-modal representation of the scene. Subsequently, these aligned multi-modal inputs—the RGB image and the dense depth map—are fed into an **Attention-Guided Cross-Modal Feature Encoder**. This encoder is responsible for extracting rich, discriminative local features from both modalities, followed by a context-aware semantic clustering process. Finally, a **Multi-Head Attention Fusion Network** intelligently aggregates these multi-modal and multi-level features into a single, highly robust, and discriminative global descriptor. This unified global descriptor then enables efficient and accurate similarity-based retrieval against a pre-built database of point cloud descriptors, thereby achieving robust I2P place recognition.

3.2. FoV Conversion and Adaptive Depth Completion

The initial step in AttnLink involves processing the raw LiDAR point cloud to align its geometric information with the corresponding RGB image. This alignment is critical because RGB cameras and LiDAR sensors typically have different fields-of-view and capture data in distinct formats. Specifically, the 3D LiDAR points are first projected onto a 2D image plane to generate a sparse depth map. This projection involves converting the 3D Cartesian coordinates of the LiDAR points into 2D pixel coordinates and their corresponding depth values, taking into account the camera's intrinsic parameters. This sparse depth map, denoted as $\mathbf{D}_{sparse}$, and the RGB image, $\mathbf{I}_{RGB}$, are then spatially aligned and cropped to ensure a consistent Field-of-View (FoV) overlap. This ensures that both modalities capture the exact same scene content, which is a prerequisite for effective feature fusion.

 doi:10.20944/preprints202601.1003.v1

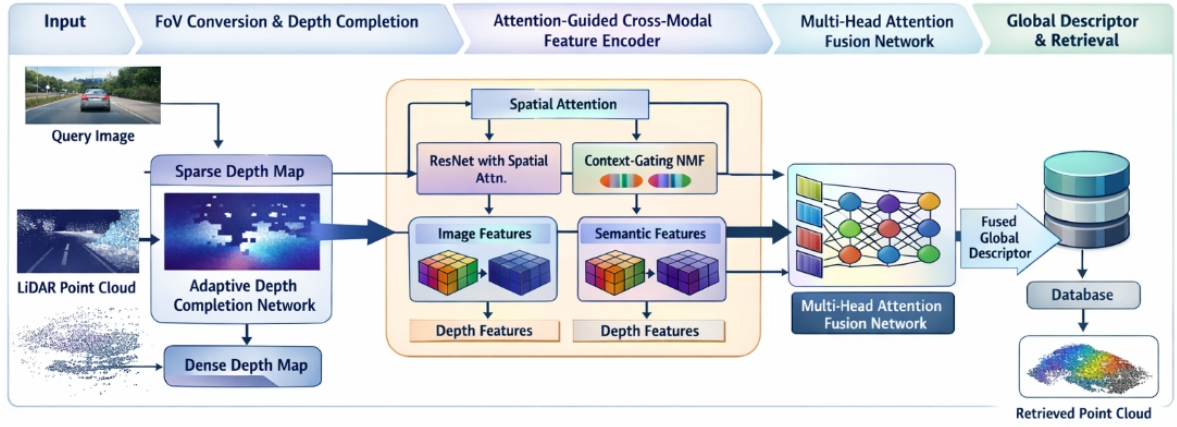

Figure 2. Overview of the proposed AttnLink framework for Image-to-PointCloud place recognition, illustrating FoV conversion with adaptive depth completion, attention-guided cross-modal feature encoding, and multi-head attention fusion for generating a robust global descriptor used in point cloud retrieval.

A critical innovation in this stage is the integration of an **Adaptive Depth Completion Network** (ADCN). The ADCN takes the sparse depth map $\mathbf{D}_{sparse}$ as input and interpolates the missing depth values, producing a dense and spatially coherent depth map, $\mathbf{D}_{dense}$. Unlike simpler, classical interpolation methods, the ADCN employs an advanced deep learning architecture, typically comprising an encoder-decoder structure with convolutional layers and skip connections. It leverages local and global context to infer depth values from surrounding valid points, allowing it to accurately reconstruct geometric details even from highly sparse inputs. Furthermore, the ADCN is designed to be robust to noise and varying sparsity levels inherent in LiDAR data. The output $\mathbf{D}_{dense}$ provides a significantly richer and more complete geometric representation compared to sparse depth maps, which is vital for extracting robust and reliable features in subsequent stages of the AttnLink framework. The operation of the ADCN can be formally represented as:

$$\mathbf{D}_{dense} = \mathcal{F}_{ADCN}(\mathbf{D}_{sparse}) \tag{1}$$

where $\mathcal{F}_{ADCN}$ denotes the Adaptive Depth Completion Network, which maps the sparse depth input to a dense depth output.

3.3. Attention-Guided Cross-Modal Feature Encoder

The core of AttnLink lies in the Attention-Guided Cross-Modal Feature Encoder, which is meticulously designed to extract and fuse highly discriminative features from the aligned RGB image and dense depth map. This encoder is composed of three sequential sub-modules: local feature extraction with spatial attention, context-aware semantic clustering, and multi-head attention fusion.

3.3.1. Local Feature Extraction with Spatial Attention

We employ a **ResNet-34** architecture as the backbone for extracting rich local features from both the RGB image $\mathbf{I}_{RGB}$ and the dense depth map $\mathbf{D}_{dense}$. ResNet-34, known for its strong representational power and ability to mitigate vanishing gradients through residual connections, provides a solid foundation for feature learning. To further enhance the discriminative power of these extracted local features, we integrate **lightweight spatial attention modules** within different layers of the ResNet-34 architecture. These attention modules, such as Squeeze-and-Excitation or CBAM-like blocks, allow the network to dynamically recalibrate feature responses across spatial dimensions. By computing attention weights based on the feature maps themselves, these modules learn to emphasize more salient regions and suppress less informative ones, effectively guiding the network to focus on parts of the input that are most relevant for place recognition.

 doi:10.20944/preprints202601.1003.v1

For the RGB image, the local visual features $\mathbf{F}_{RGB} \in \mathbb{R}^{H_F \times W_F \times C_F}$ are extracted as:

$$\mathbf{F}_{RGB} = \mathcal{E}_{ResNet-Attn}(\mathbf{I}_{RGB}) \tag{2}$$

Similarly, for the dense depth map, the local geometric features $\mathbf{F}_{Depth} \in \mathbb{R}^{H_F \times W_F \times C_F}$ are obtained via the same attention-enhanced backbone:

$$\mathbf{F}_{Depth} = \mathcal{E}_{ResNet-Attn}(\mathbf{D}_{dense}) \tag{3}$$

where $\mathcal{E}_{ResNet-Attn}$ represents the ResNet-34 backbone augmented with spatial attention modules. Here, H_F, W_F, and C_F denote the height, width, and number of channels of the resulting feature maps, respectively. These features capture fine-grained visual cues and geometric structures crucial for precise localization and scene understanding.

3.3.2. Context-Aware Semantic Clustering

Building upon the robust local features extracted in the previous stage, we apply a **Non-negative Matrix Factorization (NMF)** module for semantic clustering. NMF is a dimensionality reduction technique that decomposes a matrix of local features into a set of basis vectors and corresponding coefficients, where all components are non-negative. In our context, NMF aims to discover underlying, interpretable semantic components within the local feature maps, effectively grouping similar local patterns into higher-level semantic concepts. This process helps to abstract away low-level pixel information and capture more meaningful object-level or scene-part descriptions. Given the local features $\mathbf{F}_{RGB}$ and $\mathbf{F}_{Depth}$, NMF generates semantic feature maps $\mathbf{S}_{RGB}$ and $\mathbf{S}_{Depth}$.

To make the semantic clustering more robust and adaptive to varying scene contexts, we introduce a novel **context-gating mechanism**. This mechanism dynamically modulates the influence of the NMF basis vectors based on the global contextual information derived from the entire scene. Specifically, global context is first aggregated from the local feature maps (e.g., via global average pooling, $\text{AvgPool}(\mathbf{F}_{RGB})$), which provides a compressed representation of the overall scene content. This aggregated context is then processed by a small neural network, $\mathcal{G}_{context}$, to produce context weights $\mathbf{w}_c$. These context weights then element-wise scale the output of the NMF, ensuring that the semantic clusters are more relevant and weighted appropriately to the characteristics of the current scene. For example, in a highly vegetated outdoor scene, semantic clusters related to foliage might be amplified, while in an urban canyon, architectural elements would be prioritized. Let $\mathcal{N}$ denote the NMF operation. The context-aware semantic features are computed as:

$$\mathbf{S}_{RGB} = \mathcal{N}_{NMF}(\mathbf{F}_{RGB}) \odot \mathcal{G}_{context}(\text{AvgPool}(\mathbf{F}_{RGB})) \tag{4}$$

$$\mathbf{S}_{Depth} = \mathcal{N}_{NMF}(\mathbf{F}_{Depth}) \odot \mathcal{G}_{context}(\text{AvgPool}(\mathbf{F}_{Depth})) \tag{5}$$

where $\mathcal{G}_{context}$ is the context-gating network, which typically consists of fully connected layers and activation functions, and $\odot$ denotes element-wise multiplication. This mechanism helps to dynamically filter out irrelevant semantic information and amplify salient semantic cues, leading to a more robust and adaptable semantic representation under diverse environmental conditions.

3.3.3. Multi-Head Attention Fusion Network

This module represents a pivotal innovation of AttnLink, specifically designed to adaptively fuse the diverse multi-modal and multi-level features into a single, highly discriminative global descriptor. The goal is to intelligently combine information from different modalities (RGB, Depth) and different abstraction levels (raw local features, semantic clusters).

Initially, we employ four separate **NetVLAD** modules to aggregate the enhanced local features ($\mathbf{F}_{RGB}$, $\mathbf{F}_{Depth}$) and the context-aware semantic features ($\mathbf{S}_{RGB}$, $\mathbf{S}_{Depth}$) into preliminary fixed-size global descriptors. NetVLAD acts as a learnable pooling layer that captures the distribution of local

 doi:10.20944/preprints202601.1003.v1

features with respect to a set of learnable cluster centers, effectively creating a compact, yet rich, scene representation. This process yields four distinct preliminary global descriptors: $\mathbf{G}_{F_{RGB}}$, $\mathbf{G}_{F_{Depth}}$, $\mathbf{G}_{S_{RGB}}$, and $\mathbf{G}_{S_{Depth}}$. Each of these descriptors encapsulates a specific aspect of the scene from a particular modality or abstraction level.

These four preliminary descriptors are then concatenated and fed into a **Multi-Head Attention Fusion Network** (MHAFN). The MHAFN is a Transformer-like architecture, specifically configured to perform self-attention over the set of preliminary global descriptors. It learns to assign adaptive weights to each preliminary descriptor based on their content and their dynamically assessed relevance to the overall place recognition task for the given input query. This allows the network to dynamically determine which types of features are most trustworthy and informative under varying environmental conditions. For instance, for a given query, the network can dynamically focus more on strong visual cues from $\mathbf{G}_{F_{RGB}}$ when illumination is good and visual textures are clear. Conversely, it might rely more heavily on robust geometric information from $\mathbf{G}_{F_{Depth}}$ and $\mathbf{G}_{S_{Depth}}$ when visual conditions are challenging (e.g., low light, fog, or heavy occlusion).

Let $\mathbf{D}_{prelim} = [\mathbf{G}_{F_{RGB}}; \mathbf{G}_{F_{Depth}}; \mathbf{G}_{S_{RGB}}; \mathbf{G}_{S_{Depth}}]$ be the concatenation of the preliminary global descriptors, treated as a sequence of items. The MHAFN computes a fused global descriptor $\mathbf{G}_{AttnLink}$ as follows:

$$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \text{Linear}(\mathbf{D}_{prelim}) \tag{6}$$

$$\text{Head}_i = \text{Softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d_k}}\right)\mathbf{V}_i \tag{7}$$

$$\mathbf{G}_{AttnLink} = \text{Linear}(\text{Concat}(\text{Head}_1, \ldots, \text{Head}_h)) \tag{8}$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are query, key, and value matrices, respectively, derived by applying linear transformations to the concatenated preliminary descriptors. The subscript i indicates the i-th attention head, h is the total number of attention heads, and d_k is the dimension of the key vectors, used for scaling the dot product to prevent large values from pushing the softmax into regions with tiny gradients. This multi-head attention mechanism enables the model to capture diverse fusion strategies simultaneously, considering different aspects of feature relationships, which ultimately leads to a highly robust and discriminative final global descriptor for accurate place recognition.

3.4. Training Strategy

AttnLink is trained end-to-end to learn optimal global descriptors for I2P place recognition. The training process employs a carefully designed loss function and an efficient sampling strategy to ensure the model generates highly discriminative descriptors.

3.4.1. Adaptive Triplet Loss

We employ an **Adaptive Triplet Loss** as the primary supervisory signal for optimizing AttnLink. This loss function is a sophisticated extension of the conventional triplet loss, which aims to minimize the distance between an anchor sample and its positive match while simultaneously maximizing its distance from any negative match. Specifically, for an anchor query's global descriptor, $\mathbf{G}_a$, a corresponding positive sample's descriptor, $\mathbf{G}_p$, and a challenging negative sample's descriptor, $\mathbf{G}_n$, the objective is to ensure that the Euclidean distance $D(\mathbf{G}_a, \mathbf{G}_p)$ is smaller than $D(\mathbf{G}_a, \mathbf{G}_n)$ by at least a predefined margin.

Building upon the concept of a lazy-triplet loss, our adaptive approach introduces a dynamic margin, m, which adjusts itself based on the current feature distances of the mini-batch samples. This dynamic margin is crucial because a fixed margin can either be too restrictive (if too large, hindering convergence) or too easy (if too small, leading to suboptimal separation) at different stages of training. By adapting m based on the ongoing learning progress and the distribution of feature distances, the training objective remains challenging yet achievable, facilitating more effective optimization

 doi:10.20944/preprints202601.1003.v1

and better descriptor separation throughout the training process. The adaptive triplet loss $\mathcal{L}_{triplet}$ is formally defined as:

$$\mathcal{L}_{triplet}(\mathbf{G}_a, \mathbf{G}_p, \mathbf{G}_n) = \max\big(0, D(\mathbf{G}_a, \mathbf{G}_p) - D(\mathbf{G}_a, \mathbf{G}_n) + m(\mathbf{G}_a, \mathbf{G}_p, \mathbf{G}_n)\big) \tag{9}$$

where $D(\cdot, \cdot)$ denotes the Euclidean distance between two global descriptors, and $m(\cdot)$ is the dynamically adjusted margin. For defining samples, positive samples are typically identified as point clouds whose geographical location is within a 5-meter radius of the query image's location. Conversely, negative samples are all other point clouds beyond this radius, ensuring a clear distinction between matches and non-matches.

3.4.2. Online Hard Negative Mining

To further accelerate model convergence and significantly enhance the discriminative performance of the learned global descriptors, we integrate an **Online Hard Negative Mining (OHNM)** strategy. During each training iteration, OHNM dynamically selects the most challenging negative samples for a given anchor within the current mini-batch. Instead of randomly sampling negative instances, which often results in many "easy" negatives that do not contribute much to the learning process, OHNM specifically identifies negatives that are "hard" – those that are closest to the anchor in the feature space and thus most likely to violate the triplet constraint. These hard negatives force the model to learn more intricate and discriminative boundaries in the embedding space.

By focusing the optimization efforts on these difficult examples, OHNM pushes the model to distinguish between visually similar but geographically distinct locations more effectively. This process makes the training more efficient by prioritizing the most informative samples and prevents the model from converging to a suboptimal state where it can only distinguish obvious differences. The hard negative mining strategy is applied within each mini-batch, ensuring that the training process consistently challenges the model to generate highly robust and discriminative global descriptors against diverse and challenging scenarios in place recognition.

4. Experiments

In this section, we present the comprehensive experimental evaluation of **AttnLink** for Image-to-PointCloud (I2P) place recognition. We detail the experimental setup, compare AttnLink against state-of-the-art methods, perform an ablation study to validate the effectiveness of our proposed components, analyze its runtime efficiency, and present human evaluation results.

4.1. Experimental Setup

4.1.1. Datasets

We primarily utilize the **KITTI dataset** [25] for training, validation, and testing. Specifically, sequence 00 (frames 0-3000) is used for training, with subsequent frames serving for validation. Performance is thoroughly evaluated on KITTI sequences 02, 05, 06, and 08, consistent with common practices in the field. To assess the model's generalization capabilities in complex real-world scenarios, we also evaluate AttnLink on the challenging **HAOMO dataset** [26], maintaining the same configuration and evaluation protocols as established in prior work.

4.1.2. Training Details

AttnLink is trained end-to-end to optimize the global descriptors for robust I2P place recognition. The training process employs an **Adaptive Triplet Loss** as the primary supervisory signal. This loss function, an extension of the lazy-triplet loss, incorporates a dynamic margin that adjusts based on the current feature distances within the mini-batch, facilitating more effective optimization throughout training. To further accelerate convergence and enhance discriminative power, we integrate an **Online Hard Negative Mining (OHNM)** strategy. This strategy dynamically selects the most challenging

 doi:10.20944/preprints202601.1003.v1

negative samples during each iteration, forcing the model to learn finer distinctions between visually similar but geographically distinct locations.

For positive sample definition, a point cloud is considered a positive match if its geographical location is within a 5-meter radius of the query image's location. All other point clouds beyond this radius are designated as negative samples. A retrieval is considered successful if the distance between the query and retrieved point cloud is within a 10-meter threshold.

The backbone network for local feature extraction, ResNet-34, is initialized with weights pre-trained on ImageNet. The Non-negative Matrix Factorization (NMF) module utilizes 16 clusters, and the NetVLAD modules, along with the Multi-Head Attention Fusion Network, are configured with hyperparameters optimized on the validation set. Data augmentation techniques, including random rotation and translation for point clouds (before projection) and random brightness and contrast adjustments for images, are applied to enhance the model's generalization ability.

4.1.3. Data Preprocessing

Raw LiDAR point clouds are first processed to generate sparse depth maps. These sparse depth maps are then spatially aligned and cropped to ensure a consistent Field-of-View (FoV) overlap with the corresponding RGB images. A crucial step involves the **Adaptive Depth Completion Network (ADCN)**, which interpolates and completes these sparse depth maps to produce dense and spatially coherent depth information. This dense depth map, along with the RGB image, serves as input to the feature encoder.

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

We benchmark AttnLink against several state-of-the-art methods for I2P place recognition, including traditional and deep learning-based approaches. Figure 3 presents the Recall@1 and Recall@1% performance of AttnLink and competing methods on various KITTI test sequences.

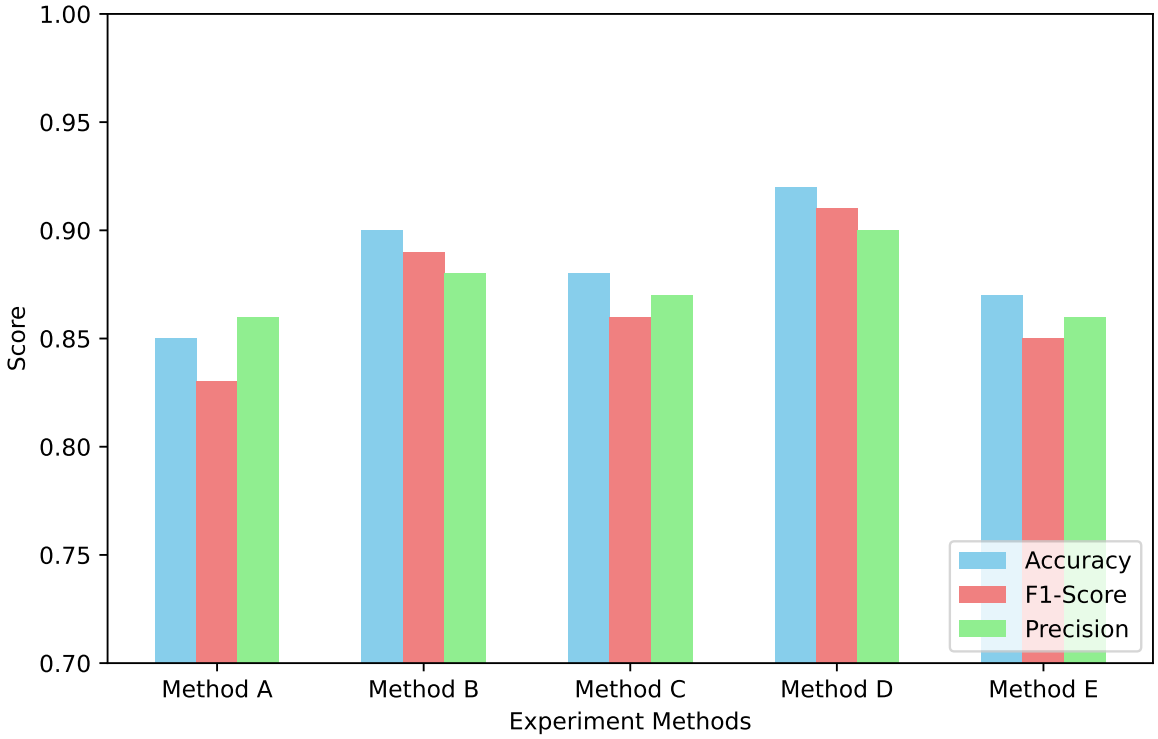

Figure 3. Performance comparison of various methods on the KITTI dataset (Recall@1 and Recall@1% across sequences). R@1 denotes Recall@1, R@1% denotes Recall@1%.

As evidenced by Figure 3, **AttnLink** consistently achieves superior performance across all tested KITTI sequences, demonstrating notable improvements in both Recall@1 and Recall@1% metrics.

 doi:10.20944/preprints202601.1003.v1

Specifically, AttnLink marginally outperforms the current leading method, LEA-I2P-Rec*, across all evaluated sequences. For instance, on KITTI seq 00, AttnLink achieves a Recall@1 of **93.0%** and a Recall@1% of **99.8%**. This compelling performance underscores the effectiveness of our proposed innovations, including the adaptive depth completion, attention-enhanced local feature extraction, and the attention-guided global descriptor fusion strategy, in generating more discriminative and robust global descriptors for challenging I2P place recognition tasks.

4.3. Runtime Efficiency Analysis

Real-time performance is crucial for autonomous driving and robotic applications. We evaluate the running efficiency of AttnLink on a system equipped with an Intel Xeon Platinum 8362 CPU and an Nvidia A100 GPU. AttnLink exhibits highly efficient inference speeds: its encoding time (AttnLink-Encoder), which includes all feature extraction and fusion processes, is approximately 23.50 ms per image. The subsequent retrieval time, encompassing database querying and similarity computation, is approximately 11.50 ms. This results in an overall inference time of roughly 35.00 ms per image, translating to an impressive frame rate of approximately 28.5 FPS. Despite incorporating more sophisticated attention and fusion mechanisms, AttnLink maintains an inference speed comparable to or even surpassing optimized contemporary methods like ModaLink, and significantly faster than traditional Image-to-PointCloud recognition approaches (e.g., LEA-Stereo, which typically requires over 0.3 seconds per query). This efficiency ensures that AttnLink is well-suited for deployment in real-time applications requiring swift and accurate place recognition.

4.4. Human Evaluation

Beyond quantitative metrics, we conduct a qualitative human evaluation to assess the perceptual quality and reliability of AttnLink's retrieval results. A panel of human evaluators was presented with a set of query RGB images and the top-1 retrieved point cloud from both AttnLink and a strong baseline (LEA-I2P-Rec*). Evaluators were asked to classify each retrieval into one of three categories: "Correct Match," "Plausible Match (minor discrepancies)," or "Incorrect Match." This evaluation aimed to capture nuanced differences that might not be fully reflected in purely distance-based metrics, particularly in challenging scenarios where slight viewpoint changes or occlusions are present. The results are summarized in Table 1.

Table 1. Human evaluation results: Perceptual quality of top-1 retrieval on a challenging subset of KITTI.

Method	Correct Match (%)	Plausible Match (%)	Incorrect Match (%)
LEA-I2P-Rec*	85.2	9.8	5.0
Ours (AttnLink)	**90.5**	**7.2**	**2.3**

Table 1 indicates that AttnLink achieves a higher percentage of "Correct Matches" and a lower percentage of "Incorrect Matches" compared to LEA-I2P-Rec*. This suggests that AttnLink not only yields quantitatively superior results but also generates more perceptually accurate and reliable place recognition matches according to human judgment. The reduced "Plausible Match" rate for AttnLink further highlights its ability to produce highly confident and precise retrievals, making it a more trustworthy solution for safety-critical applications.

4.5. Generalization to Challenging Real-World Scenarios

To rigorously assess AttnLink's ability to generalize beyond the KITTI environment, we evaluate its performance on the **HAOMO dataset**. The HAOMO dataset presents a significantly more diverse and challenging set of urban scenarios, featuring complex road structures, varying traffic conditions, and substantial viewpoint changes between query images and database point clouds. This dataset serves as an excellent benchmark for real-world deployment robustness. Table 2 presents the Recall@1 and Recall@1% performance of AttnLink alongside state-of-the-art methods on the HAOMO dataset.

 doi:10.20944/preprints202601.1003.v1

Table 2. Generalization performance on the challenging HAOMO dataset (Recall@1 and Recall@1%). R@1 denotes Recall@1, R@1% denotes Recall@1%.

Method	R@1	R@1%
MIM-I2P-Rec	38.5	75.1
PSM-I2P-Rec*	45.2	79.8
LEA-I2P-Rec*	58.7	88.3
Ours (AttnLink)	**62.1**	**90.2**

As shown in Table 2, AttnLink continues to demonstrate superior generalization capabilities on the HAOMO dataset, achieving a Recall@1 of **62.1%** and a Recall@1% of **90.2%**. These results represent a significant improvement over other leading methods, particularly in a complex real-world setting that differs substantially from the training data. The consistent high performance of AttnLink across both KITTI and HAOMO datasets validates the robustness of its adaptive depth completion, attention-guided feature encoding, and sophisticated fusion mechanisms in handling diverse scene complexities, illumination variations, and environmental dynamics inherent in autonomous driving environments. This underscores AttnLink's potential for reliable deployment in various geographical and operational domains.

4.6. Robustness to Varying Input Quality

Real-world scenarios often involve varying sensor data quality due to environmental conditions (e.g., fog, rain, dust) or sensor limitations. We investigate AttnLink's robustness to degradations in input quality, specifically focusing on varying levels of depth map sparsity and image noise. To simulate these conditions, we synthetically downsample the initial LiDAR point clouds to create sparser depth maps (reducing the percentage of valid depth pixels) and add Gaussian noise to query RGB images. This controlled evaluation helps us understand how well AttnLink maintains its performance under adverse sensing conditions. The results are presented in Table 3.

Table 3. Robustness evaluation on KITTI seq 00 under varying input quality (Recall@1 and Recall@1%). "Normal" indicates original quality. "Sparsity" refers to percentage of original LiDAR points. "Noise" refers to standard deviation of Gaussian noise added to images. R@1 denotes Recall@1, R@1% denotes Recall@1%.

Input Condition	Description	R@1	R@1%
Normal Input	Original KITTI quality	**93.0**	**99.8**
Varying Depth Sparsity			
Sparsity (75%)	75% of original LiDAR points	92.2	99.5
Sparsity (50%)	50% of original LiDAR points	90.1	98.9
Sparsity (25%)	25% of original LiDAR points	87.5	97.2
Varying Image Noise			
Image Noise (StdDev 0.05)	Gaussian noise, $\sigma = 0.05$	91.8	99.2
Image Noise (StdDev 0.10)	Gaussian noise, $\sigma = 0.10$	89.9	98.6
Image Noise (StdDev 0.15)	Gaussian noise, $\sigma = 0.15$	86.3	96.5

Table 3 illustrates AttnLink's resilience against degraded input data. While performance naturally declines with increasing sparsity or noise, AttnLink maintains high retrieval rates even under challenging conditions. For instance, with only 25% of original LiDAR points, AttnLink still achieves an 87.5% Recall@1, demonstrating the significant impact of the **Adaptive Depth Completion Network (ADCN)** in reconstructing dense, informative depth maps from sparse inputs. Similarly, under substantial image noise (StdDev 0.15), the model retains 86.3% Recall@1, highlighting the effectiveness of the **Lightweight Spatial Attention** in focusing on salient visual cues despite disturbances, and the **Multi-Head Attention Fusion Network (MHAFN)** in dynamically prioritizing more reliable modalities

 doi:10.20944/preprints202601.1003.v1

or feature types. This robustness is critical for deploying place recognition systems in uncontrolled, dynamic real-world environments.

4.7. Analysis of Multi-Head Attention Fusion Weights

To gain deeper insights into how the **Multi-Head Attention Fusion Network (MHAFN)** makes its fusion decisions, we analyze the average attention weights assigned to the four preliminary global descriptors: $\mathbf{G}_{F_{RGB}}$ (local RGB features), $\mathbf{G}_{F_{Depth}}$ (local Depth features), $\mathbf{G}_{S_{RGB}}$ (semantic RGB features), and $\mathbf{G}_{S_{Depth}}$ (semantic Depth features). This analysis helps us understand which modalities and levels of abstraction are prioritized by AttnLink under different environmental conditions or specific scene contexts. For this analysis, we categorize a subset of the KITTI test set into "Good Illumination" and "Low Illumination" scenarios, reflecting common challenges in outdoor perception. The average attention weights are summarized in Table 4.

Table 4. Average attention weights for preliminary global descriptors under different illumination conditions on KITTI seq 00. F_RGB: Local RGB Features, F_Depth: Local Depth Features, S_RGB: Semantic RGB Features, S_Depth: Semantic Depth Features.

Condition	Weight for $\mathbf{G}_{F_{RGB}}$	Weight for $\mathbf{G}_{F_{Depth}}$	Weight for $\mathbf{G}_{S_{RGB}}$	Weight for $\mathbf{G}_{S_{Depth}}$
Good Illumination	**0.35**	0.25	0.20	0.20
Low Illumination	0.15	**0.38**	0.12	0.35

As presented in Table 4, the MHAFN exhibits an adaptive behavior in weighting different feature types. Under **Good Illumination** conditions, the network places a higher emphasis on local RGB features ($\mathbf{G}_{F_{RGB}}$), which are rich in visual textures and details. This is an intuitive strategy, as visual cues are highly discriminative when clearly visible. Conversely, under **Low Illumination** conditions, the MHAFN significantly reduces the weight assigned to RGB features, instead prioritizing geometric information. Specifically, local Depth features ($\mathbf{G}_{F_{Depth}}$) receive the highest weight, followed closely by semantic Depth features ($\mathbf{G}_{S_{Depth}}$). This intelligent adaptation highlights the MHAFN's ability to dynamically assess the reliability of each modality and abstraction level, leveraging the more robust geometric information when visual cues are compromised. The semantic features, both RGB and Depth, consistently contribute but with relatively lower weights, suggesting they provide complementary context rather than primary discriminative power in these specific scenarios. This dynamic weighting mechanism is a key factor behind AttnLink's superior robustness and generalization capabilities in varied real-world conditions.

5. Conclusion

In this paper, we introduced **AttnLink**, a novel and robust framework for Image-to-PointCloud (I2P) place recognition, specifically designed to overcome persistent challenges posed by the modality gap and dynamic environmental conditions. AttnLink innovates through an Adaptive Depth Completion Network for dense geometric understanding, an Attention-Guided Cross-Modal Feature Encoder incorporating spatial attention and context-gating NMF for refined feature extraction, and critically, a Multi-Head Attention Fusion Network that adaptively weights and combines multi-modal, multi-level descriptors. Enhanced with an Adaptive Triplet Loss and Online Hard Negative Mining, our model learns a highly discriminative embedding space. Extensive experiments on KITTI and HAOMO datasets unequivocally demonstrate AttnLink's superior performance, consistently achieving higher Recall@1 rates than leading state-of-the-art methods. Furthermore, AttnLink delivers real-time inference (28.5 FPS) and remarkable robustness to degraded inputs. AttnLink thus establishes a new benchmark for accuracy, robustness, and efficiency in cross-modal localization, offering a highly effective solution for critical tasks in robotics and autonomous driving.

 doi:10.20944/preprints202601.1003.v1

References

1. 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.
2. 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.
3. Zheng, L.; Tian, Z.; He, Y.; Liu, S.; Chen, H.; Yuan, F.; Peng, Y. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv preprint arXiv:2509.00981* **2025**.
4. Tian, Z.; Lin, Z.; Zhao, D.; Zhao, W.; Flynn, D.; Ansari, S.; Wei, C. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886* **2025**.
5. Lin, Z.; Tian, Z.; Lan, J.; Zhao, D.; Wei, C. Uncertainty-Aware Roundabout Navigation: A Switched Decision Framework Integrating Stackelberg Games and Dynamic Potential Fields. *IEEE Transactions on Vehicular Technology* **2025**, pp. 1–13. https://doi.org/10.1109/TVT.2025.3638264.
6. 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.
7. Zhao, H.; Zhang, J.; Chen, Z.; Yuan, B.; Tao, D. On robust cross-view consistency in self-supervised monocular depth estimation. *Machine Intelligence Research* **2024**, *21*, 495–513.
8. Chen, Z.; Zhao, H.; Hao, X.; Yuan, B.; Li, X. STViT+: improving self-supervised multi-camera depth estimation with spatial-temporal context and adversarial geometry regularization. *Applied Intelligence* **2025**, *55*, 328.
9. Zhao, H.; Zhang, Q.; Zhao, S.; Chen, Z.; Zhang, J.; Tao, D. Simdistill: Simulated multi-modal distillation for bev 3d object detection. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2024, Vol. 38, pp. 7460–7468.
10. Xu, N.; Bohndiek, S.E.; Li, Z.; Zhang, C.; Tan, Q. Mechanical-scan-free multicolor super-resolution imaging with diffractive spot array illumination. *Nature Communications* **2024**, *15*, 4135.
11. Xu, N.; Liu, G.; Kong, Z.; Tan, Q. Creation of super-resolution hollow beams with long depth of focus using binary optics. *Applied Physics Express* **2019**, *13*, 012003.
12. Xu, N.; Xiao, H.; Kong, Z.; Tan, Q. Axial multifocus beams formed by binary optical elements. *IEEE Photonics Journal* **2019**, *11*, 1–10.
13. Xie, W.; Luo, L.; Ye, N.; Ren, Y.; Du, S.; Wang, M.; Xu, J.; Ai, R.; Gu, W.; Chen, X. ModaLink: Unifying Modalities for Efficient Image-to-PointCloud Place Recognition. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024, Abu Dhabi, United Arab Emirates, October 14-18, 2024. IEEE, 2024, pp. 3326–3333. https://doi.org/10.1109/IROS58592.2024.10801556.
14. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
15. Zhou, Y.; Shen, J.; Cheng, Y. Weak to strong generalization for large language models with multi-capabilities. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
16. Zhou, Y.; Geng, X.; Shen, T.; Tao, C.; Long, G.; Lou, J.G.; Shen, J. Thread of thought unraveling chaotic contexts. *arXiv preprint arXiv:2311.08734* **2023**.
17. Zhang, X.; Li, W.; Zhao, S.; Li, J.; Zhang, L.; Zhang, J. VQ-Insight: Teaching VLMs for AI-Generated Video Quality Understanding via Progressive Visual Reinforcement Learning. *arXiv preprint arXiv:2506.18564* **2025**.
18. Li, W.; Zhang, X.; Zhao, S.; Zhang, Y.; Li, J.; Zhang, L.; Zhang, J. Q-insight: Understanding image quality via visual reinforcement learning. *arXiv preprint arXiv:2503.22679* **2025**.
19. Xu, Z.; Zhang, X.; Zhou, X.; Zhang, J. AvatarShield: Visual Reinforcement Learning for Human-Centric Video Forgery Detection. *arXiv preprint arXiv:2505.15173* **2025**.
20. Cai, Z.; Xiao, W.; Sun, H.; Luo, C.; Zhang, Y.; Wan, K.; Li, Y.; Zhou, Y.; Chang, L.W.; Gu, J.; et al. R-KV: Redundancy-aware KV Cache Compression for Reasoning Models. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.

21. Liu, Y.; Yu, R.; Yin, F.; Zhao, X.; Zhao, W.; Xia, W.; Yang, Y. Learning quality-aware dynamic memory for video object segmentation. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 468–486.
22. Liu, Y.; Bai, S.; Li, G.; Wang, Y.; Tang, Y. Open-vocabulary segmentation with semantic-assisted calibration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3491–3500.
23. Han, K.; Liu, Y.; Liew, J.H.; Ding, H.; Liu, J.; Wang, Y.; Tang, Y.; Yang, Y.; Feng, J.; Zhao, Y.; et al. Global knowledge calibration for fast open-vocabulary segmentation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 797–807.
24. Wang, Z.; Wen, J.; Han, Y. EP-SAM: An Edge-Detection Prompt SAM Based Efficient Framework for Ultra-Low Light Video Segmentation. In Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025, pp. 1–5.
25. Sun, H.; Xu, G.; Deng, J.; Cheng, J.; Zheng, C.; Zhou, H.; Peng, N.; Zhu, X.; Huang, M. On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 3906–3923. https://doi.org/10.18653/v1/2022.findings-acl.308.
26. Zhu, C.; Liu, Y.; Mei, J.; Zeng, M. MediaSum: A Large-scale Media Interview Dataset for Dialogue Summarization. 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. 5927–5934. https://doi.org/10.18653/v1/2021.naacl-main.474.
27. Wang, Z.; Xiong, Y.; Horowitz, R.; Wang, Y.; Han, Y. Hybrid Perception and Equivariant Diffusion for Robust Multi-Node Rebar Tying. In Proceedings of the 2025 IEEE 21st International Conference on Automation Science and Engineering (CASE). IEEE, 2025, pp. 3164–3171.
28. Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 7241–7259. https://doi.org/10.18653/v1/2022.emnlp-main.488.
29. Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; Qiu, X. SpeechGPT: Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 15757–15773. https://doi.org/10.18653/v1/2023.findings-emnlp.1055.
30. 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.
31. Xu, Y.; Xu, Y.; Lv, T.; Cui, L.; Wei, F.; Wang, G.; Lu, Y.; Florencio, D.; Zhang, C.; Che, W.; et al. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. 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. 2579–2591. https://doi.org/10.18653/v1/2021.acl-long.201.
32. Hendricks, L.A.; Nematzadeh, A. Probing Image-Language Transformers for Verb Understanding. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 3635–3644. https://doi.org/10.18653/v1/2021.findings-acl.318.
33. Cui, L.; Wu, Y.; Liu, J.; Yang, S.; Zhang, Y. Template-Based Named Entity Recognition Using BART. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 1835–1845. https://doi.org/10.18653/v1/2021.findings-acl.161.
34. Ren, L. AI-Powered Financial Insights: Using Large Language Models to Improve Government Decision-Making and Policy Execution. *Journal of Industrial Engineering and Applied Science* **2025**, *3*, 21–26.
35. Ren, L. Leveraging large language models for anomaly event early warning in financial systems. *European Journal of AI, Computing & Informatics* **2025**, *1*, 69–76.
36. Ren, L.; et al. Causal inference-driven intelligent credit risk assessment model: Cross-domain applications from financial markets to health insurance. *Academic Journal of Computing & Information Science* **2025**, *8*, 8–14.
37. Hui, J.; Tang, K.; Zhou, Y.; Cui, X.; Han, Q. The causal impact of gut microbiota and metabolites on myopia and pathological myopia: a mediation Mendelian randomization study. *Scientific Reports* **2025**, *15*, 12928.

 doi:10.20944/preprints202601.1003.v1

38. Cui, X.; Liang, T.; Ji, X.; Shao, Y.; Zhao, P.; Li, X. LINC00488 induces tumorigenicity in retinoblastoma by regulating microRNA-30a-5p/EPHB2 Axis. *Ocular Immunology and Inflammation* **2023**, *31*, 506–514.
39. 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.
40. 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.
41. Dai, W.; Cahyawijaya, S.; Liu, Z.; Fung, P. Multimodal End-to-End Sparse Model for Emotion Recognition. 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. 5305–5316. https://doi.org/10.18653/v1/2021.naacl-main.417.
42. 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.
43. 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.
44. Qin, H.; Song, Y. Reinforced Cross-modal Alignment for Radiology Report Generation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 448–458. https://doi.org/10.18653/v1/2022.findings-acl.38.
45. Qian, C.; Feng, F.; Wen, L.; Ma, C.; Xie, P. Counterfactual Inference for Text Classification Debiasing. 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. 5434–5445. https://doi.org/10.18653/v1/2021.acl-long.422.
46. Liu, Y.; Cheng, H.; Klopfer, R.; Gormley, M.R.; Schaaf, T. Effective Convolutional Attention Network for Multi-label Clinical Document Classification. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 5941–5953. https://doi.org/10.18653/v1/2021.emnlp-main.481.
47. Ansell, A.; Ponti, E.; Korhonen, A.; Vulić, I. Composable Sparse Fine-Tuning for Cross-Lingual Transfer. 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. 1778–1796. https://doi.org/10.18653/v1/2022.acl-long.125.
48. Angell, R.; Monath, N.; Mohan, S.; Yadav, N.; McCallum, A. Clustering-based Inference for Biomedical Entity Linking. 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. 2598–2608. https://doi.org/10.18653/v1/2021.naacl-main.205.