A Hybrid Space Model for Misaligned Multi-modality Image Fusion

Yi Xiao[1*], Jia Wang[1*], Zhu Liu[1], Di Wang[2], Jinyuan Liu[1], Risheng Liu[1†]

[1]School of Software Technology, Dalian University of Technology, Dalian, China
[2]School of Computer Science and Artificial Intelligence, Civil Aviation University of China, Tianjin, China
xiaoyi@mail.dlut.edu.cn, jiawang0704@outlook.com, rsliu@dlut.edu.cn

Abstract

Infrared and visible image fusion aims to integrate complementary information, such as thermal saliency from infrared imagery and fine-grained texture details from visible imagery. However, real-world multi-modal misalignment and geometric deformation often introduce severe artifacts. Most existing methods focus on feature extraction within Euclidean space, thereby neglecting the inherent hierarchical structures embedded in multimodal representations. While Euclidean space excels at preserving local structural details and supporting efficient computation, hyperbolic space is naturally suited for modeling hierarchical relationships due to its geometric properties. Building upon these observations, this paper proposes a unified framework that jointly optimizes image registration and fusion through a dual-space architecture. This architecture synergistically combines the local fidelity of Euclidean geometry with the hierarchical modeling capability of hyperbolic geometry to enhance multimodal representation learning. Specifically, this paper introduces Hyperbolic Coupled Contrastive Learning Optimization (HCCLO), which aligns and optimizes the hierarchical structures of infrared and visible embeddings in hyperbolic space. Moreover, this paper designs a task-adaptive dual-space features fusion mechanism, which dynamically balances and fuses Euclidean local features with hyperbolic hierarchical representations, thereby improving adaptability for downstream tasks. Extensive experiments on misaligned multimodal datasets demonstrate that our method achieves state-of-the-art performance, while effectively capturing both spatial dependencies and hierarchical semantics.

Introduction

Due to limitations in illumination conditions and sensor hardware, a single imaging modality can only capture partial scene information. Multi-modality image fusion aims to integrate complementary information from multiple source images to generate a more informative and perceptually enhanced fused image. Among various fusion tasks, infrared and visible image fusion (IVIF) has attracted extensive research attention owing to its unique advantages in combining thermal saliency and detailed texture. (Liu et al. 2022b) The resulting fused images offer more comprehensive scene representations and improved visual perception, thereby benefiting downstream computer vision applications such as semantic segmentation (Li et al. 2023a; Liu et al. 2023a), object detection (Liu et al. 2022a; Zhao et al. 2023a), scene understanding (Huang et al. 2020), and autonomous driving systems.

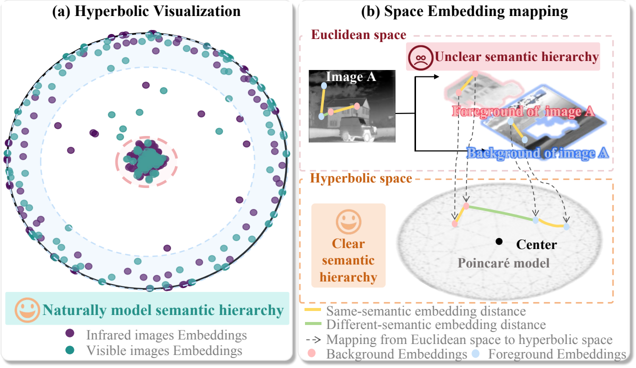
(a) Hyperbolic Visualization
Naturally model semantic hierarchy
Infrared images Embeddings
Visible images Embeddings
(b) Space Embedding mapping
Euclidean space
Unclear semantic hierarchy
Image A
Foreground of image A
Background of image A
Hyperbolic space
Clear semantic hierarchy
Center
Poincaré model
Same-semantic embedding distance
Different-semantic embedding distance
Mapping from Euclidean space to hyperbolic space
Background Embeddings
Foreground Embeddings

Figure 1: The superior capability of hyperbolic space in modeling semantic hierarchies. (a) Hyperbolic t-SNE visualizations of visible and infrared image embeddings reveal a clear hierarchical structure, particularly around the center and boundary regions of the Poincaré ball. (b) Hyperbolic visualizations of foreground and background features reveal a similar hierarchical organization, where features that are close in Euclidean space become distinctly separated in the Poincaré ball due to the semantic hierarchical capability of hyperbolic space.

Over the past decades, a variety of IVIF approaches have been proposed with the primary objective of enhancing fusion quality. Traditional methods are mainly categorized into five groups: Multi-Scale Transform (MST)-based methods (Li, Wu, and Kittler 2020), Sparse Representation (SR)-based methods (Liu et al. 2016), Subspace Decomposition-based methods (Lu et al. 2014), Saliency-Driven methods (Ma et al. 2017), and Optimization model-based methods (Ma et al. 2016), among others. However, these methods often rely on handcrafted designs and involve computationally intensive procedures. Recently, deep learning-based approaches have been introduced into IVIF, demonstrating

[*]These authors contributed equally.
[†]Corresponding author.