Hierarchical Network Fusion for Multi-Modal Electron Micrograph Representation Learning with Foundational Large Language Models

Sakhinana Sagar Srinivas Geethan Sannidhi Venkataramana Runkana

TRDDC & IIIT Pune

June 29, 2025

Outline

Motivation & Core Challenges

Context:

- Electron micrographs are essential for nanoscale inspection in semiconductor manufacturing and nanomaterials research.
- Sub-7nm technology nodes introduce complex, hierarchical visual patterns and structural heterogeneity.
- Conventional vision models (CNNs, ViTs) often fail under distributional shifts due to limited inductive priors and lack of domain knowledge integration.

Key Challenges:

- High Intra-Class Variability: Visual heterogeneity within the same nanomaterial category complicates generalization.
- High Inter-Class Similarity: Different nanomaterials exhibit overlapping structural motifs and textures.
- Spatial Heterogeneity: Salient patterns occur at multiple spatial resolutions, necessitating multi-scale representation learning.

Research Objective

Design a multi-modal, hierarchical representation learning framework that synergistically fuses patch-based and graph-based visual encodings with domain-specific language model insights to enable robust nanomaterial classification under complex visual distributions.

Core Insight: Multi-Modal Representation Learning

Electron micrographs contain hierarchical, multi-scale structures that cannot be fully captured using a single modality.

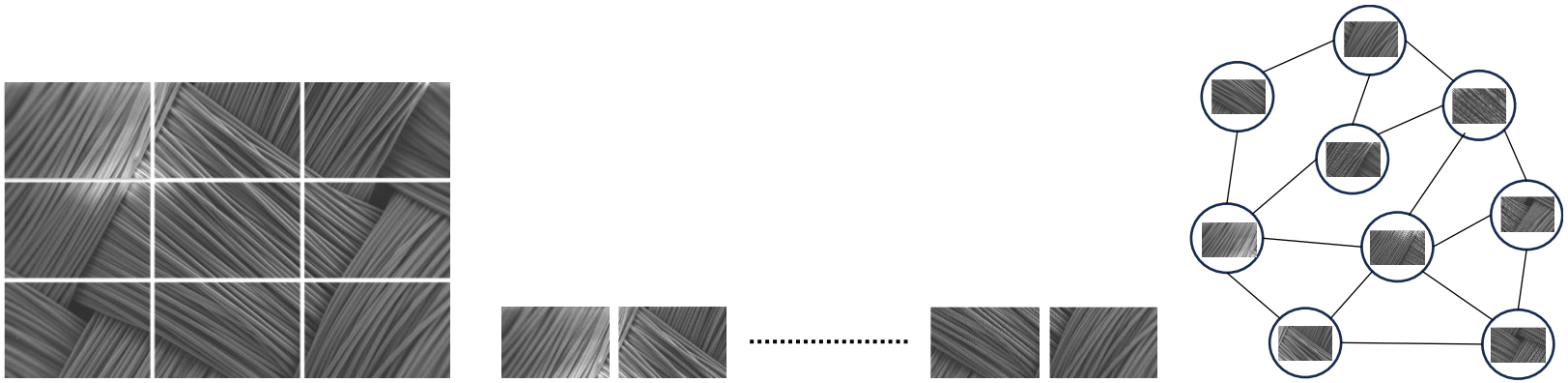
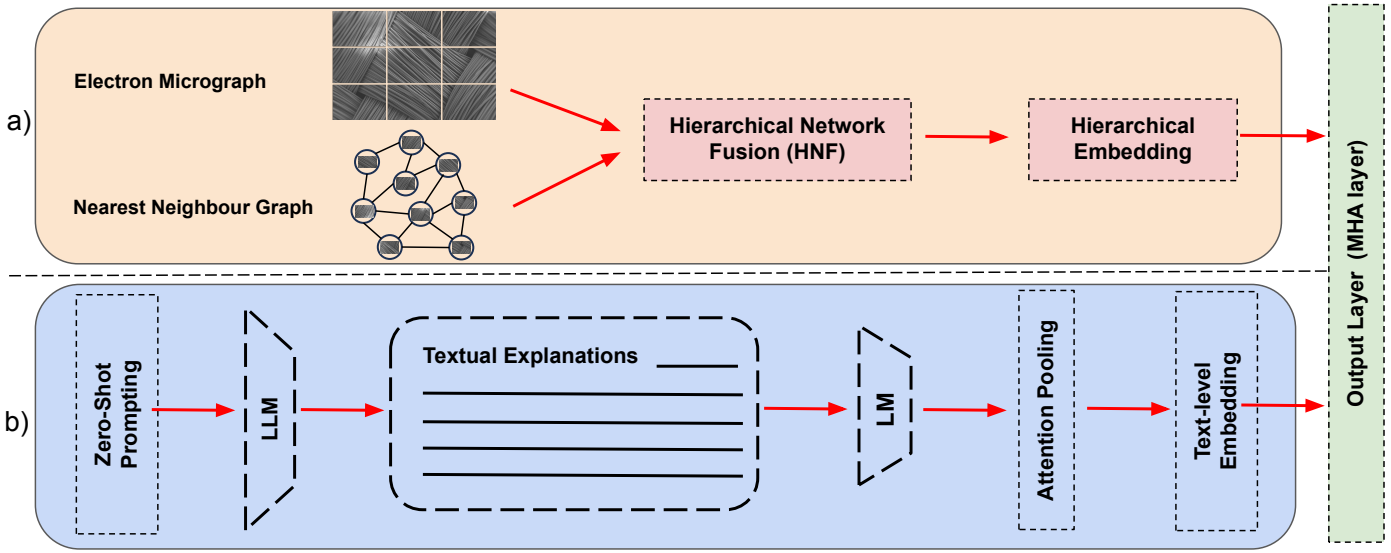


Figure: A micrograph is jointly represented as (a) a patch sequence for modeling spatial dependencies and (b) a vision graph encoding local structural relationships.

Key Design Hypothesis

By combining patch sequence representations (spatial layout) with vision graphs (structural priors), and grounding them in domain-specific technical knowledge from Large Language Models (LLMs), we can enable more discriminative and generalizable classification of electron micrographs.

Hierarchical Visual-Linguistic Fusion



- **Visual Stream**: Electron micrographs are tokenized into multi-resolution patch sequences and vision graphs. Hierarchical Network Fusion (HNF) uses *Neural ODEs* and *Graph Chebyshev Convolutions* across layers to generate cross-modal visual embeddings.
- **Textual Stream**: Zero-shot CoT prompts query LLMs (e.g., GPT-3.5) to generate technical nanomaterial descriptions, which are encoded by smaller language models (e.g., DeBERTa) using *Masked Language Modeling*.
- **Fusion Layer**: A Multi-Head Cross-Attention mechanism aligns and integrates the hierarchical visual embeddings with text-level embeddings for robust multi-class classification.

The Visual Processing Backbone

The Hierarchical Network Fusion (HNF) is a cascading, multi-layered visual backbone designed to capture both fine- and coarse-grained visual cues from electron micrographs through multi-scale representation learning.

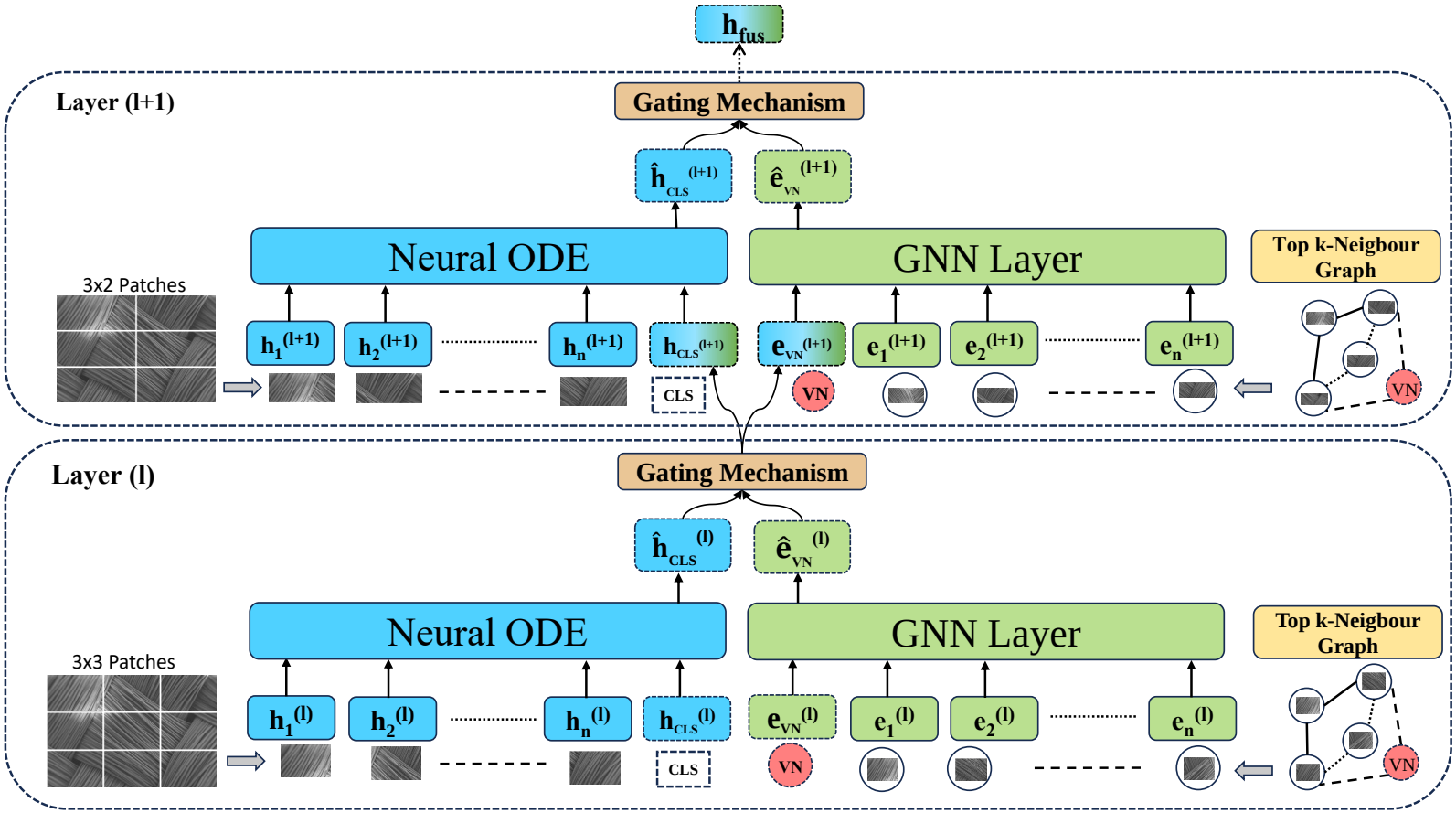


Figure: HNF constructs a hierarchical representation by processing both patch sequences and vision graphs at increasing patch resolutions across layers.

Multi-Scale Visual Backbone: Sequence-Graph Fusion

HNF employs an inverted pyramid architecture where each layer corresponds to a distinct patch size, progressively increasing in resolution to model hierarchical dependencies. At each layer:

- The electron micrograph is represented as:
  - A **patch sequence** (`<cls>` token included) for modeling long-range spatial dependencies.
  - A **vision graph**, constructed using k-nearest neighbor (kNN) similarity between patch embeddings, with a global *virtual node*.
- Patch embeddings are evolved using **bidirectional Neural ODEs**, modeling continuous inter-patch dynamics.
- Graph structure is encoded using **Graph Chebyshev Convolution** (GCC), capturing local structural context.
- A **Mixture-of-Experts (MoE)**-based gating mechanism integrates the CLS token and virtual node embedding, producing a unified hierarchical embedding.

Objective

Fuse spatial (sequence) and structural (graph) modalities to construct scale-aware, cross-domain embeddings for robust micrograph representation.

Why Involve Large Language Models?

The Limitation of Visual-Only Models

Purely visual models lack real-world context. They don't know that "nanowires" are thin, elongated structures or that "MEMS devices" have intricate, manufactured patterns. This ambiguity can lead to errors.

The Solution: Inject Domain Knowledge

Need: To ground visual features with explicit, high-level scientific and technical knowledge.

Methodology:

1. **Generate Knowledge:** Use a foundational LLM (e.g., GPT-3.5) to create detailed descriptions of each material class (synthesis, properties, applications, etc.).
2. **Embed Knowledge:** Fine-tune a smaller language model on this text to create powerful, domain-specific textual embeddings.

How to Fuse Vision and Text?

We have a powerful visual embedding from HNF and a powerful text embedding from the LLM. Simple concatenation is not enough.

The Multi-Head Attention (MHA) Layer

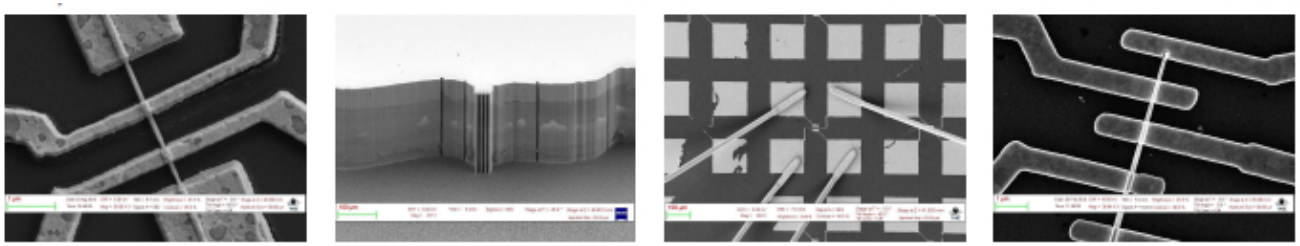
Need: A sophisticated mechanism that can find complex, context-dependent alignments between the two modalities. It needs to answer: "Which parts of the visual embedding are most relevant to the concept of 'porous sponge' described in the text?"

Methodology:

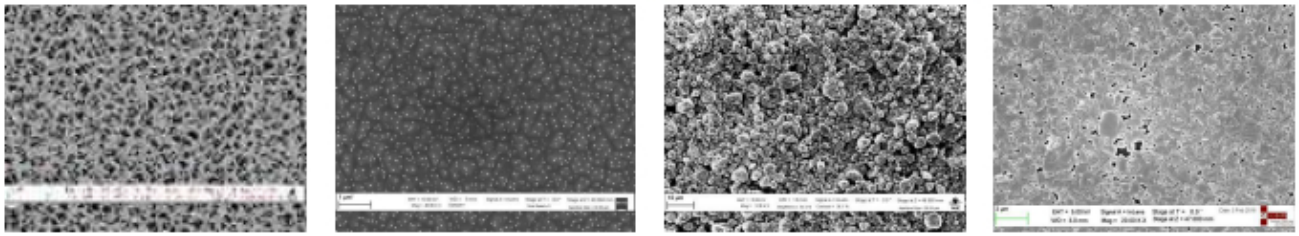
- The MHA layer lets the two modalities "talk" to each other.
- The visual embedding can **query** the text embedding to find relevant concepts and extract contextual information.
- Simultaneously, the text embedding can query the visual embedding to ground abstract scientific concepts in specific visual patterns.
- This creates a deeply integrated, unified representation for final classification.

Experimental Setup

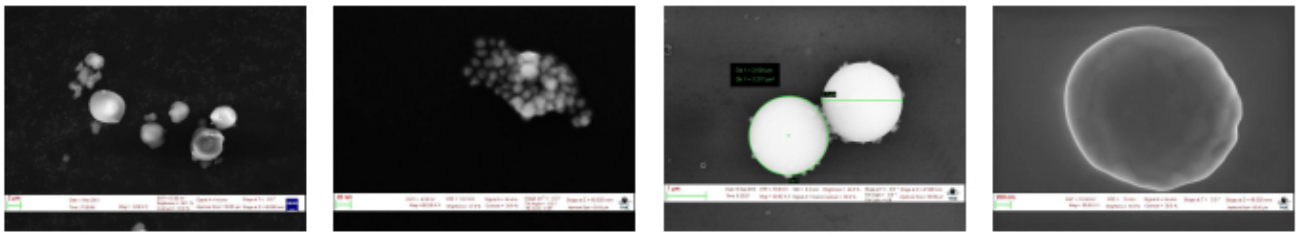
- **Primary Dataset**: SEM Dataset, containing 21k electron micrographs across 10 nanomaterial categories.



(a) High intra-class dissimilarity: The electron micrographs of the same nanomaterial (*MEMS* device) can exhibit a high degree of heterogeneity.



(b) High inter-class similarity: Electron micrographs across different nanomaterial categories (*listed from left to right as porous sponges, particles, powders, and films*) exhibit a noteworthy degree of similarity.



(c) Multi-spatial scales of patterns: The spatial heterogeneity of visual patterns in electron micrographs of *nanoparticles* is evident.

Comparison with Vision-Based Baselines

Our method significantly outperforms existing state-of-the-art models.

Table: Top-N accuracy on the SEM dataset.

Algorithm	**Top-1**	**Top-2**	**Top-3**	**Top-5**
ResNet	0.512	0.766	0.891	0.906
SwinT	0.675	0.766	0.891	0.938
T2TViT (Prev. SOTA)	0.702	0.859	0.906	0.938
MultiFusion-LLM (w/ BARD)	0.852	0.899	0.927	0.953
MultiFusion-LLM (w/ GPT-3.5)	**0.947**	**0.965**	**0.986**	**0.991**

Key Takeaway

Our model achieves a **25.8% relative improvement** in Top-1 accuracy over the next-best baseline, demonstrating the power of our multi-modal fusion approach.

Is Every Component Necessary?

We systematically removed key parts of our model to measure their impact.

Table: Performance impact of disabling model components.

Algorithm Variant	**Avg-Precision**	**Avg-Recall**	**Avg-F1 Score**
Full Model (Baseline)	**0.941**	**0.945**	**0.939**
w/o HNF (No visual hierarchy)	0.776	0.753	0.745
w/o LLMs (No text knowledge)	0.714	0.726	0.721
w/o MHA (Simple fusion)	0.827	0.831	0.823

Conclusion

Removing any component—the hierarchical visual network (HNF), the LLM-based knowledge, or the smart attention fusion—causes a significant drop in performance. This validates our design choices.

Conclusion

- We proposed **MultiFusion-LLM**, an end-to-end framework for nanomaterial classification that integrates hierarchical visual representations with LLM-derived domain knowledge.
- The **Hierarchical Network Fusion (HNF)** module captures both fine- and coarse-grained features via multi-scale patch sequences and vision graphs, refined through **Neural ODEs** and **Graph Chebyshev Convolutions**.
- Using **Zero-shot Chain-of-Thought (CoT)** prompting, we extract rich technical descriptions from foundational LLMs and distill them into compact **text embeddings** via masked language modeling with smaller LMs.
- A **Multi-Head Cross-Attention** mechanism fuses vision and text embeddings, enabling semantic alignment and improving classification under distributional shifts.
- The framework outperforms all baselines on the SEM dataset, achieving a **Top-1 accuracy of 94.7%**, with consistent improvements across multiple datasets and ablation settings.

Impact

Our approach enables robust, interpretable, and high-throughput characterization of nanomaterials from electron micrographs—advancing automated inspection in semiconductor manufacturing and materials discovery.

Thank You!

Questions?