Feature Hallucination for Self-supervised Action Recognition

Learning robust multi-modal representations from incomplete data

Lei Wang[1,2] Piotr Koniusz[2,3,4,1]

[1]Griffith University [2]Data61/CSIRO [3]UNSW [4]ANU

January 22, 2026

The Core AI Problem & Why Existing Methods Fall Short

- **Real-world AI is multimodal**: vision, motion, audio, skeletons, *etc.*
- **Complete data is rare**
  - Sensors fail
  - Modalities are missing or misaligned in time
  - Data quality varies
- **Expectation remains high**: generalize, reason, support decisions

How can AI learn robust representations when parts of the world are missing?

- Hidden assumption: all modalities are available – even at inference
- Common workarounds:
  - Drop samples
  - Fill missing modalities with averages/defaults
  - Hand-crafted heuristics
- **Two critical issues**:
  - Hallucinated features are treated as reliable (uncertainty ignored)
  - Motion, one of the strongest self-supervised signals, is under-exploited

Outcome: hallucination without trust $\rightarrow$ fragile representations

The One-Sentence Idea & A General AI Design Principle

Hallucination is useful only when a model knows how much it should trust what it hallucinates.

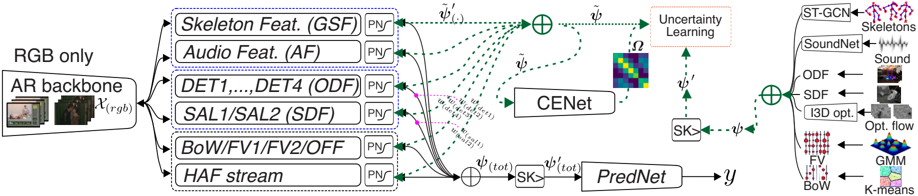
A General AI Design Principle: Hallucinate missing data within uncertainty, and anchor learning in reliable self-supervised signals (*e.g.*, motion).

Therefore, hallucination should be:

- Explicitly uncertainty-aware
- Grounded in reliable self-supervised signals (e.g., motion)

This principle applies **beyond computer vision and multimodal learning**

Framework Overview & Hallucination with Uncertainty



Unified multimodal self-supervised framework

- During **training**: Train with full modalities → learn cross-modal prediction
- During **testing**:
  - **Hallucinate missing modalities** when data is incomplete
  - **Estimate aleatoric uncertainty** for hallucinated features
  - **Use motion-aware descriptors** to stabilize learning
- **Practical advantages**:
  - Compatible with multiple backbones
  - Scales to large datasets

Uncertainty is used to:

- Down-weight unreliable hallucinations
- Stabilize representation learning

Why Motion Matters & What Breaks Without Our Ideas

Why Motion Matters

- Motion encodes what appearance cannot: structure, dynamics, and temporal consistency
- Remains informative even when **visual cues are weak or missing**

Motion as a Self-Supervised Anchor

- Improves **fine-grained recognition**
- Strengthens **cross-modal alignment**
- **Generalizes across domains**

Motion acts as a **reliable bridge between modalities**

	Without uncertainty	**With uncertainty**
Hallucination behavior	Blind trust	Confidence-weighted
System outcome	Failure	Graceful degradation

This gap is **critical for real-world deployment**, especially in autonomous and embodied systems

Scalability and Generality & Why This Belongs at AAAI

Framework Properties:

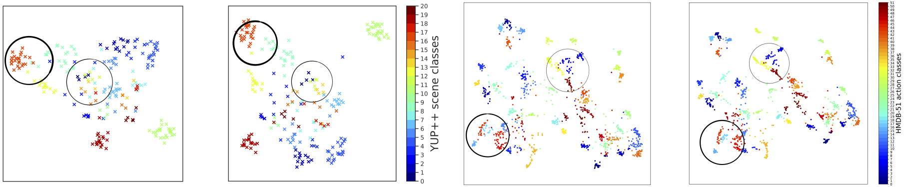
- Modality-agnostic; tested on action recognition
- Works with multiple architectures
- Scales to large datasets
- Does not require all modalities at inference ("one common modality for all")

Why Fits AAAI:

- Addresses a core AI challenge: robust representation learning from incomplete, heterogeneous data
- Contributions:
  - Representation learning
  - Multimodal reasoning
  - Embodied and autonomous AI
- Potential applications: robotics, HCI, sports analytics, surveillance

This is not an action recognition trick, it is a principle for learning under partial observability.

Key Takeaways & Closing



- Missing data is unavoidable in real AI systems
- Hallucination can help, when it is uncertainty-aware.
- Motion is a powerful **self-supervised signal**
- Combining these ideas leads to more **robust and generalizable AI**

Thank you!