Feature Hallucination for Self-supervised Action Recognition

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

[1]Griffith University, [2]Data61/CSIRO, [3]University of New South Wales, [4]Australian National University
{l.wang4, p.koniusz}@griffith.edu.au

Abstract

Understanding human actions in videos requires more than raw pixel analysis; it relies on high-level semantic reasoning and effective integration of multimodal features. We propose a deep translational action recognition framework that enhances recognition accuracy by jointly predicting action concepts and auxiliary features from RGB video frames. At test time, hallucination streams infer missing cues, enriching feature representations without increasing computational overhead. To focus on action-relevant regions beyond raw pixels, we introduce two novel domain-specific descriptors. *Object Detection Features* (ODF) aggregate outputs from multiple object detectors to capture contextual cues, while *Saliency Detection Features* (SDF) highlight spatial and intensity patterns crucial for action recognition. Our framework seamlessly integrates these descriptors with auxiliary modalities such as optical flow, Improved Dense Trajectories, skeleton data, and audio cues. It remains compatible with state-of-the-art architectures, including I3D, AssembleNet, Video Transformer Network, FASTER, and recent models like VideoMAE V2 and InternVideo2. To handle uncertainty in auxiliary features, we incorporate aleatoric uncertainty modeling in the hallucination step and introduce a robust loss function to mitigate feature noise. Our multimodal self-supervised action recognition framework achieves state-of-the-art performance on multiple benchmarks, including Kinetics-400, Kinetics-600, and Something-Something V2, demonstrating its effectiveness in capturing fine-grained action dynamics.

Broad Interest & AI Relevance

The presented framework advances multimodal and self-supervised learning, key themes in AI research with applications spanning robotics, human-computer interaction, sports analytics, surveillance, and embodied AI. By unifying hallucination of missing modalities, uncertainty-aware learning, and novel motion-aware descriptors, the work tackles the fundamental AI challenge of learning robust, generalizable models from incomplete and heterogeneous data. Its compatibility with multiple backbones, scalability to large datasets, and improved fine-grained recognition performance make it relevant not only to computer vision but also to broader AI areas, including representation learning, multimodal reasoning, and autonomous systems.

*Corresponding author.

Journal *vs.* Conference Contributions

The journal paper (Wang and Koniusz 2025) extends our ICCV 2019 (Wang, Koniusz, and Huynh 2019) and ACMMM 2021 (Wang and Koniusz 2021) in scope, methodology, and application breadth. While the conference papers focused on hallucinating improved dense trajectories and optical flow features, and two newly proposed descriptors, Object Detection Features (ODF) and Saliency Detection Features (SDF), from RGB inputs for action recognition, the journal article: (i) Expands the hallucination framework to handle multiple missing modalities, including skeleton and audio, to better capture spatial and motion cues. (ii) Introduces uncertainty-aware learning, modeling aleatoric uncertainty for robust fusion of noisy or incomplete features, paired with a robust loss function to stabilize training. (iii) Demonstrates architectural generality by integrating the approach with multiple modern video backbones (I3D, AssembleNet, Video Transformer Network, VideoMAE V2, and InternVideo2), significantly broadening applicability. (iv) Expands evaluation to large-scale experiments on Kinetics-400, Kinetics-600, and Something-Something V2, achieving new state-of-the-art results. These advances transform the original single-modality, supervised hallucination method into a unified, multimodal, uncertainty-aware, and self-supervised framework, with significantly stronger performance and broader impact in the field of AI.

References

Wang, L.; and Koniusz, P. 2021. Self-supervising action recognition by statistical moment and subspace descriptors. In *Proceedings of the 29th ACM international conference on multimedia*, 4324–4333.

Wang, L.; and Koniusz, P. 2025. Feature Hallucination for Self-supervised Action Recognition. *International Journal of Computer Vision*, 133: 7612 – 7646.

Wang, L.; Koniusz, P.; and Huynh, D. Q. 2019. Hallucinating IDT Descriptors and I3D Optical Flow Features for Action Recognition With CNNs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.