Feature Hallucination for Self-supervised Action Recognition (Abstract Reprint)

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

[1]School of Engineering and Built EnvironmentElectrical and Electronic Engineering, Griffith University, Brisbane, Australia
[2]School of Computing, College of Engineering, Computing, and Cybernetics, The Australian National University (ANU), Canberra, Australia
[3]Data61/CSIRO, Canberra, Australia
[4]School of Computer Science and Engineering, The University of New South Wales (UNSW), Sydney, Australia

Abstract Reprint. This is an abstract reprint of the journal article by Wang and Koniusz (2025).

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.

References

Wang, L.; and Koniusz, P. 2025. Feature Hallucination for Self-supervised Action Recognition. *International Journal of Computer Vision*, 133: 76127646.