Diversity of Auxiliary Loss Functions and Robustness in Video Representation Learning

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 14 peer-reviewed papers addressing the following research question: What is the correlation between the diversity of auxiliary loss functions integrated by MELTR and the robustness of video representations against temporal perturbations in Something-Something V2. 17 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Large-scale Robustness Analysis of Video Action Recognition Models. Research question: What is the correlation between the diversity of auxiliary loss functions integrated by MELTR and the robustness of video representations against temporal perturbations in Something-Something V2?.

2 Methodology

Systematic literature search across multiple databases yielded 14 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

14 papers retrieved. 17 claims extracted; 1 independently verified. Quality review score: 3.8/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The resolution of clips used for evaluation is 224$\times$224 for all datasets.	×	0.02
For Kinetics dataset, 10 uniform temporal crops are taken for each video and center crop is applied for each of these 10	×	0.03
For UCF101 and HMDB51, 5 uniform temporal crops are taken for each video and center crop is applied for each clip.	×	0.05
For SSv2, a single spatial crop is used and the number of frames is uniformly sampled as used in the original model impl	×	0.03
Two metrics are used to measure robustness: absolute accuracy drop and relative accuracy drop.	×	0.02
Absolute robustness γa is computed for each severity level s and perturbation p as γa_p,s = 1 - (Af_c - Af_p,s)/100.	×	0.02
Relative robustness γr is computed for each severity level s and perturbation p as γr_p,s = 1 - (Af_c - Af_p,s)/Af_c.	×	0.02
Spatial perturbations have the largest drop in performance as severity increases for Kinetics-P.	×	0.02
Transformer-based Timesformer and MViT models are typically more robust than CNN-based models for spatial perturbations.	✓	0.20
Performance of Timesformer drops by $\sim$5% and ResNet based R3D drops by $\sim$30% for spatial perturbations.	×	0.04
Models are more robust to variable rotation compared to a static rotation.	×	0.06
The mean performance on SSV2-P across perturbation types and severity is shown in Figure 5.	×	0.02
The robustness metrics for different models and perturbations are summarized in Tables 2 and 3.	×	0.07
The robustness of models against 5 different kinds of perturbations is analyzed for UCF101-P, Kinetics-P, HMDB51-P, and	×	0.10
The performance of different models on various classes and shifts is shown in Table (p4).	×	0.05
The robustness metrics (γa and γr) for different models and perturbation types are shown in Tables (p6), (p7), and (p8).	×	0.03
The performance of Timesformer, X3D, and Slowfast on different actions and perturbations is shown in Table (p8).	×	0.04

References

- http://arxiv.org/abs/2407.14834v1
- http://arxiv.org/abs/2207.01398v2
- http://arxiv.org/abs/2010.11757v4