MELTR-Enhanced Fine-Tuning for Robust Video-Based Pedestrian Attribute Recognition

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: Does MELTR-enhanced fine-tuning improve robustness in video-based pedestrian attribute recognition (PAR) on the RICAP-14 dataset when compared to traditional single-loss fine-tuning under motion blur. 16 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Spatio-Temporal Side Tuning Pre-trained Foundation Models for Video-based Pedestrian Attribute Recognition. Research question: Does MELTR-enhanced fine-tuning improve robustness in video-based pedestrian attribute recognition (PAR) on the RICAP-14 dataset when compared to traditional single-loss fine-tuning under motion blur perturbations?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

9 papers retrieved. 16 claims extracted; 2 independently verified. Quality review score: 4.2/10.

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
Existing Pedestrian Attribute Recognition (PAR) algorithms adopt CNN as the backbone network to extract feature represen	×	0.12
Existing Transformer-based models for attribute recognition support image-based attribute recognition only.	×	0.10
Chen et al. [16] formulate video-based pedestrian attribute recognition as a multi-task classification problem.	✓	0.15
Chen et al. [16] transform attribute labels into binary vectors for network optimization.	×	0.04
Current pedestrian attribute recognition is divided into RGB frame-based and video-based streams.	✓	0.15
Abdulnabi et al. [19] propose a multi-task learning approach utilizing multiple CNNs to learn attribute-specific feature	×	0.06
PromptPAR [41] adopts CLIP as the backbone and optimizes its parameters using prompt tuning.	×	0.05
Jin et al. [42] formulate attribute recognition as a phrase generation problem using pre-trained CLIP as visual and text	×	0.13
On the MARS-Attribute Dataset, the VTB method achieves an Accuracy of 90.37% and an F1 score of 78.32%.	×	0.03
On the DukeMTMC-VID-Attribute Dataset, the VTB method achieves an Accuracy of 84.24%.	×	0.06
On the MARS-Attribute Dataset, the TA(Video) method achieves an Accuracy of 87.01% and an F1 score of 72.04%.	×	0.03
On the MARS-Attribute Dataset, the 3DCNN method achieves an Accuracy of 81.95%.	×	0.04
For the 'gender' attribute on the MARS-Attribute Dataset, the VTFPAR++ method achieves an Accuracy of 91.32% and an F1 s	×	0.03
For the 'motion' attribute on the DukeMTMC-VID-Attribute Dataset, the VTFPAR++ method achieves an Accuracy of 97.65% and	×	0.07
The VTFormer model configuration with 5 components achieves an Accuracy of 92.90% and an F1 score of 81.94% on the MARS-	×	0.03
The VTFormer model configuration with 5 components has 157.53M parameters.	×	0.02

References

- http://arxiv.org/abs/2404.17929v1
- http://arxiv.org/abs/2504.10018v2
- http://arxiv.org/abs/1901.07474v2