Why can’t I dance in the mall?
Learning to Mitigate Scene Bias in Action Recognition

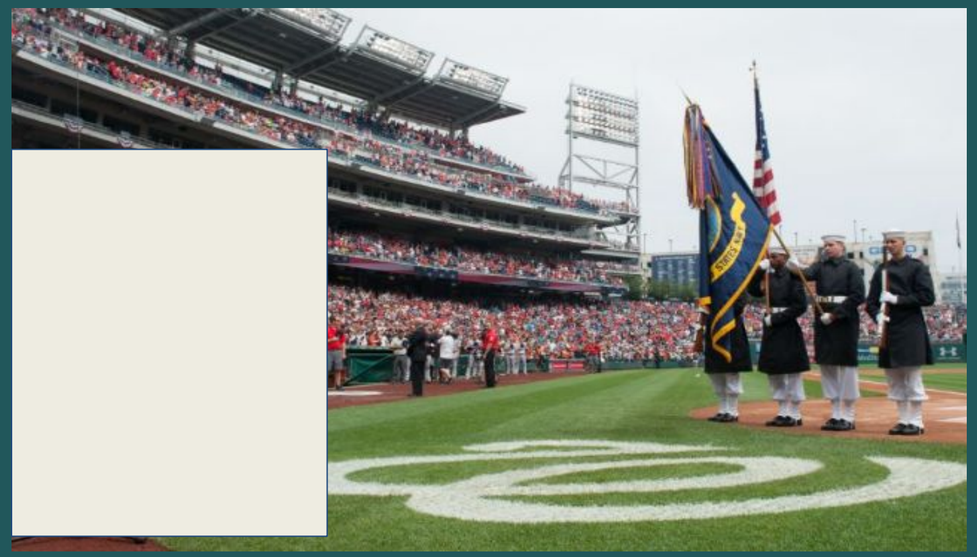
Reproducibility Study Presented By

Swabhi Papneja, Sai Rashwanth U M, Vishnu Teja

INTRODUCTION

- Action recognition models often focus on the **background or scenes** instead of the human activity itself.

- This is called **Scene Bias** and it leads to **"correct predictions for the wrong reasons"**



INTRODUCTION

- Action recognition models often focus on the **background or scenes** instead of the human activity itself.
- This is called **Scene Bias** and it leads to **"correct predictions for the wrong reasons"**

MOTIVATION

- This prevents models from performing well on new action classes or tasks, as they cannot adapt to different scenes.

OBJECTIVE

The objective of this study is to **mitigate scene bias** in action recognition models by introducing **novel loss functions** that encourage the model to focus on **human actions** rather than the background, **improving generalization** to new tasks and datasets.

SCENE BIAS RATIO

$$B_{\text{scene}} = \log[M(D, \phi_{\text{scene}})/M_{\text{rand}}].$$

For UCF101:

M(D,ϕscene) = 59.7%

Mrand = 1 / 101 = 1.0%

Bias ratio:

Bscene = log(59.7/1.0) = 4.09

The scene representation bias ratio Bscene quantifies **how much a dataset's action classification accuracy depends on scene information.**

- **ϕscene: Scene representation features**
- **M(D,ϕscene): The action classification accuracy achieved using ϕscene on dataset D**
- **Mrand: Random chance accuracy**

CORE IDEA

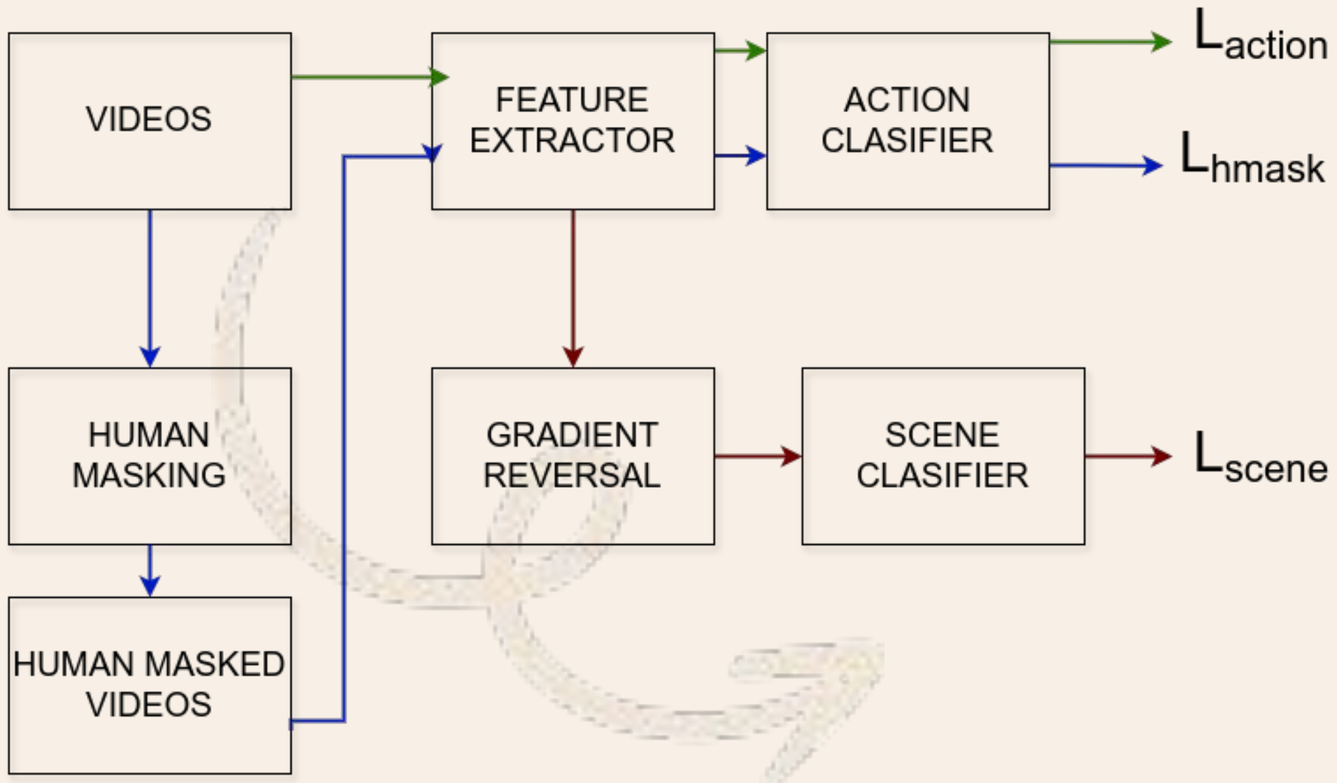
The paper introduces a **Debiaisng Algorithm** to **Migtigate the Scene Bias for Action Understanding Tasks** by introducing two additional Losses:

1. Adversarial Loss

2. Human Mask Confusion Loss

θ

Pre-Training MODEL ARCHITECTURE



Proposed Loss Functions

Adversarial Loss

This loss ensures that the **model learns features that are invariant to the scene** and focuses on human actions instead. This is achieved by adversarial training with pseudo-scene labels generated using Places365 dataset.

Human Mask Confusion Loss

This loss prevents the model from predicting actions when there’s no clear visual evidence of human activity. It ensures the model doesn’t make predictions when the **human evidence is removed**.

LOSS FUNCTIONS

01 CROSS ENTROPY LOSS

$$L_{CE} = -\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathbf{X},\mathbf{Y})} \sum_{k=1}^{N} y_k \log f_{\theta_A}(G_{\theta_f}(\mathbf{x})).$$

x: Input video.

y: True action label

N: Total Classes

Gθf(x): Features extracted by the feature extractor with parameters θf

fθA: Action classifier with parameters θA.

02 ADVERSARIAL LOSS

$$L_{\text{Adv}} = -\mathbb{E}_{(\mathbf{x},\mathbf{p})\sim(\mathbf{X},\mathbf{P})} \sum_{m=1}^{M} p_m \log f_{\theta_S}(G_{\theta_f}(\mathbf{x})).$$

x: Input video.

P: Pseudo scene label

M: Total Scene Classes

fθs: Scene classifier with parameters θs.

LOSS FUNCTIONS

03 HUMAN MASK CONFUSION LOSS

$$L_{\text{Ent}} = -\mathbb{E}_{\mathbf{x}_{\text{hm}} \sim \mathbf{X}_{\text{hm}}} \sum_{k=1}^{N} f_{\theta_A}(G_{\theta_f}(\mathbf{x}_{\text{hm}})) \log f_{\theta_A}(G_{\theta_f}(\mathbf{x}_{\text{hm}})).$$

xhm: Human-Masked video.

N: Total Action Classes

Gθ: Feature Extractor

OPTIMIZATION

$$L(\theta_f, \theta_S, \theta_A) = L_{CE}(\theta_f, \theta_A) - \lambda L_{\text{Adv}}(\theta_f, \theta_S),$$

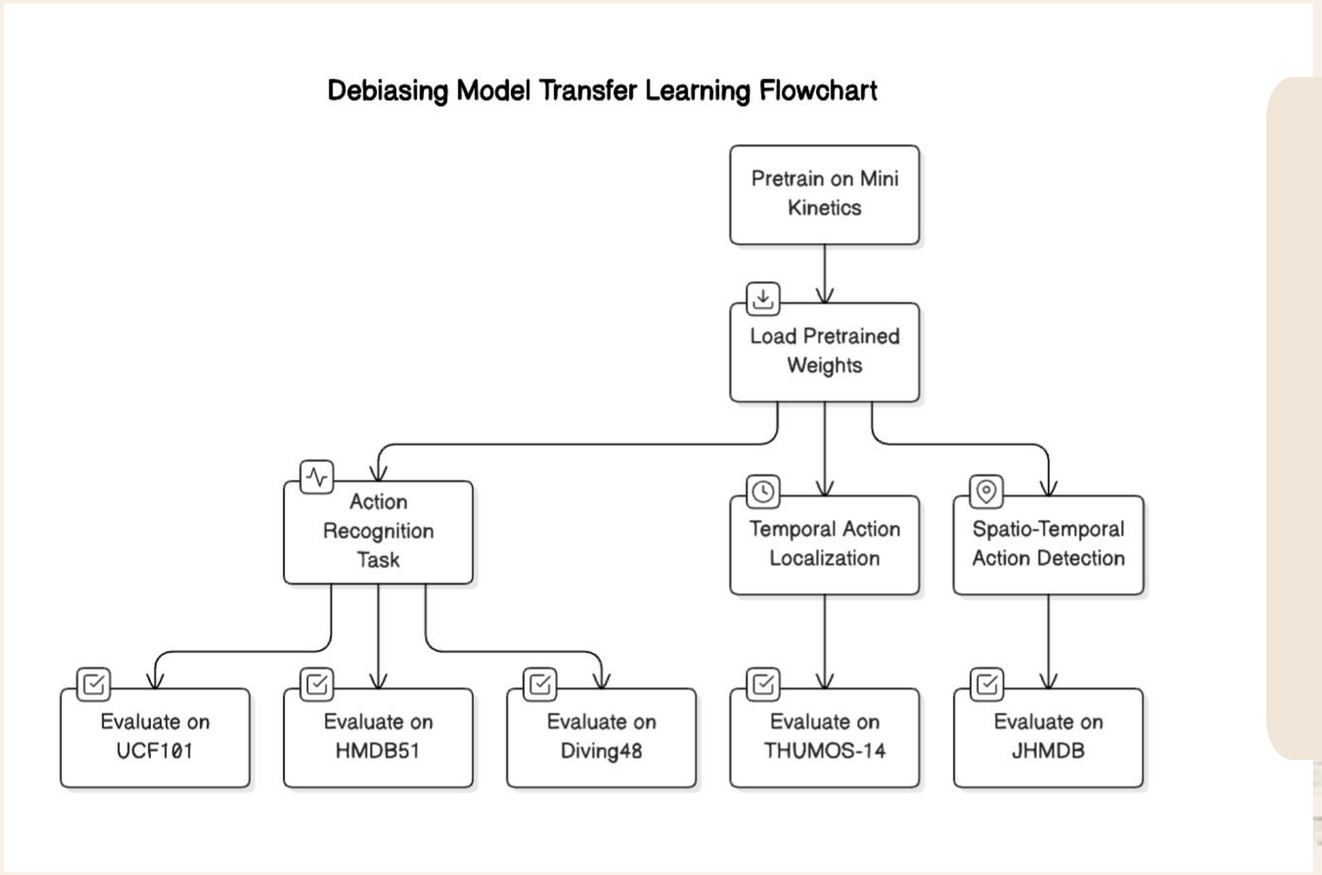
L: The total loss function, combining two losses for action classification and scene classification on original videos.

θf: Parameters of the Feature Extractor
θs: Parameters of the Scene Classifier
θa: Parameters of the Action Classifier

$$(\theta_f^*, \theta_A^*) = \underset{\theta_f, \theta_A}{\text{argmax}} L_{\text{Ent}}(\theta_f, \theta_A).$$

This equation determines the optimal parameters θf and θa by maximizing the entropy-based loss Lent

TRANSFER LEARNING



After pretraining the debiasing model with the Mini Kinetics Dataset, we can use the pre-trained weights to improve the performance on other datasets action understanding tasks such as:

1. **Action Recognition**
2. **Temporal Action Localization**
3. **Spatio -Temporal Action Detection**

DATASETS USED

01

MINI KINETICS 200

Number of videos: 80k

Number of classes: 200

02

HMDB-51

Number of videos: 6,766

Number of classes: 51

03

UCF101

Number of videos: 13,320

Number of classes: 101

04

DIVING48

Number of videos: 18k

Number of classes: 48

FRAME EXTRACTION

We worked on Frame Extraction of the following video datasets for our action understanding tasks:

Mini Kinetics 200

Size: 143 GB

Frames Extracted: 10,485,759

HMDB51

Size: 2.1 GB

Frames Extracted: 632,599

UCF101

Videos Size: 6.8 GB

Frames Extracted: 2,483,295

Diving48

Videos size: 9.7 GB

Frames Extracted: 2,465,263 + 216,808

CHALLENGES

1. Many youtube videos were not available.
2. No high quality videos available for few.
3. Only 57k videos were available out of total 80k videos from the Mini Kinetics Dataset.
4. Thus, we have 13 classes with no videos at all.

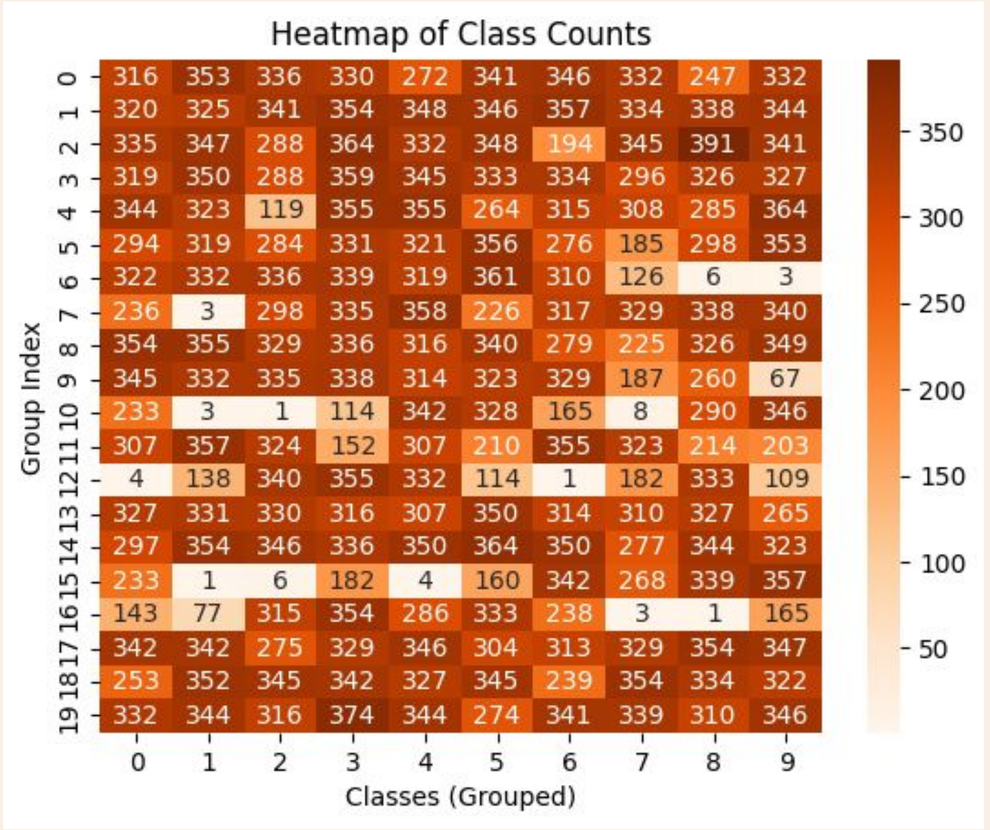
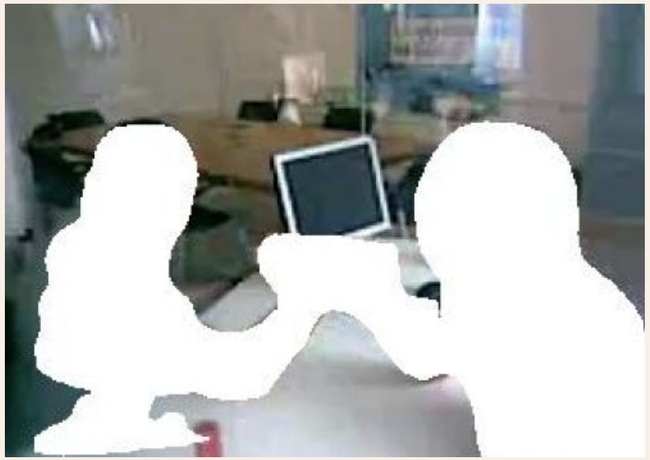


Fig: Class labels for MiniKinetics

DATASET PREPARATION

1. Collecting raw videos
   a. Youtube
   b. Available datasets
2. Frame extraction
   - Extracting each frame from the video to input to the 3D-CNN
3. Human masked video preparation
   - Masking human subjects in every frame for Human msk confusion loss



HYPERPARAMETERS AND MODEL ARCHITECTURES

Learning Rate	0.001
Momentum	0.9
Batch Size	32
Number of Action Classes	200
Number of Scene Classes	365
Alpha (Scene Adversarial)	0.5
Alpha (Human Mask)	0.5
Optimizer	SGD
Base model architecture	ResNet18
Human masking model	ResNet50 *
Pseudo Sence label generator	ReNet18 *

*: pretrained

EXPERIMENTS & RESULTS

Our Results				
	Mini Kinetics	UCF101	HMDB51	Diving48
Baseline with Pretrained Weights	--	72.4	32.4	
Baseline Pretrained from scratch	46.07	33.98	19.9	12.32
Debiasing Model Approach	53.42	In Progess	In Progess	In Progess

Published Results				
	Mini Kinetics	UCF101	HMDB51	Diving48
Baseline with Pretrained Weights	--	83.5	53.6	18.2
Debiasing Model Approach	53.42	84.5	56.7	20.5

1. For Pretraining the Baseline and the Debiasing Model, the Mini Kinetics 200 dataset was used that included labeled examples split into Training and Validation sets
2. We'll compare our baseline model with the proposed debiasing model to validate the reproducibility of the approach given for mitigating scene bias and Transfer Learning.

CONCLUSION

- We worked on re-implementation of a Model-Agnostic approach on MItigating scene bias for action classification.
- Implemented Adversarial and Human mask Confusion losses
- Testing the model on three transfer learning tasks

THANK YOU