ICLR

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

ALBAR: **A**dversarial **L**earning approach to mitigate **B**iases in **A**ction **R**ecognition



Joseph Fioresi, Ishan Rajendrakumar Dave, Mubarak Shah

ALBAR = “Guard of All”

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Biases in Video Action Recognition

- Model should use mainly *action-related* information

Main Forms of Action Recognition Bias

- *Background bias:*
  - Inferring action based on background cues
- *Foreground bias:*
  - Inferring action based on foreground cues
  - Harmful if using demographic information (skin color, gender, etc.)
- Neither of these are solved problems!

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Background Bias Examples (UCF-101)

Ground Truth: TennisSwing

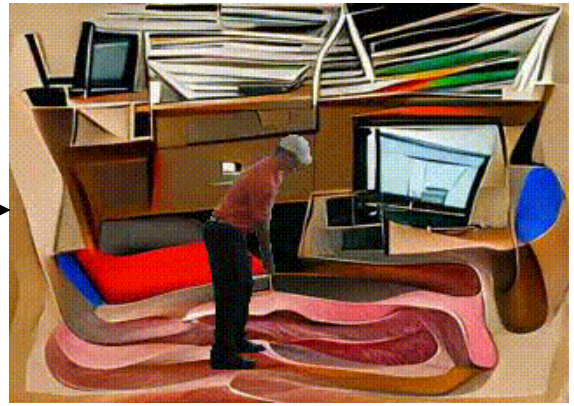
Background Change

Model Prediction: TennisSwing

Model Prediction: HorseRiding

Ground Truth: GolfSwing

Background Change



Model Prediction: GolfSwing

Model Prediction: HulaHoop

(best viewed as video)

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Foreground Bias

Prediction: passing American football (in game)
GT: archery

Prediction: finger snapping
GT: playing guitar

Prediction: eating spaghetti
GT: playing violin

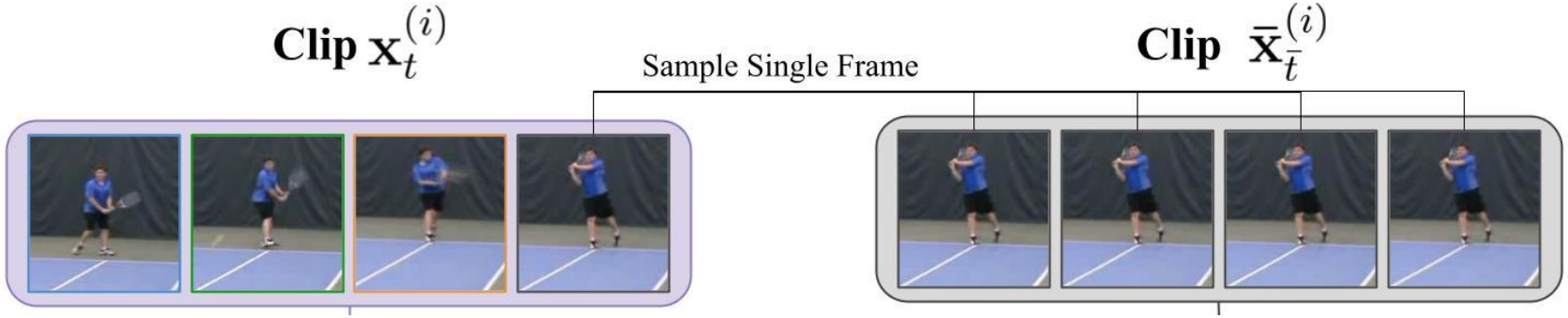
(best viewed as video)

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

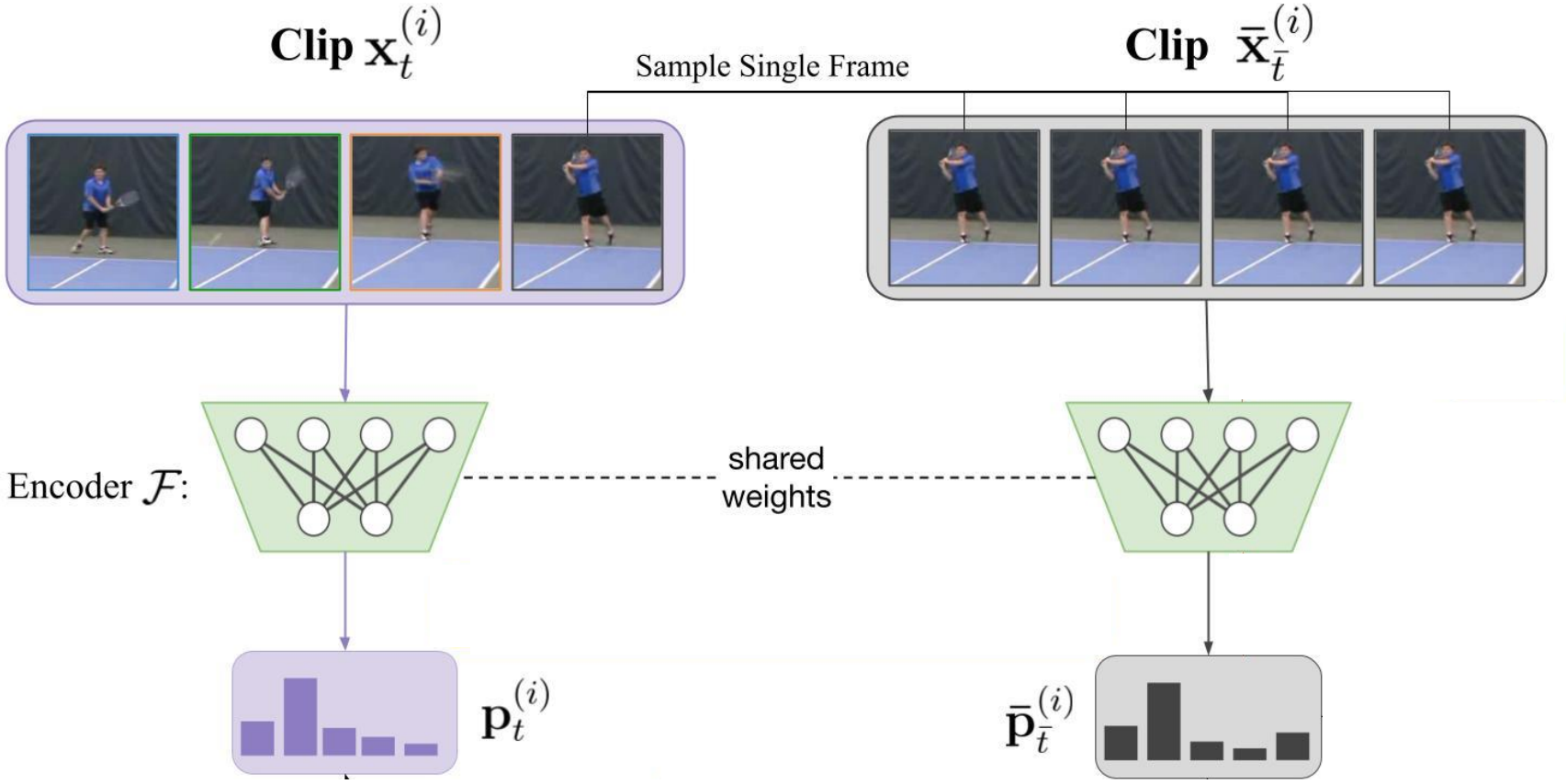
ALBAR framework

- Intuition: background/foreground biases are due to *static* spatial information
- Adversarial learning to mitigate biases
  - Adversary: the same network predicting actions with a static (no motion) clip
    - Cross-entropy of motion clip *minus* cross-entropy of static clip
  - Two supplementary objectives to stabilize training:
    - Static *entropy maximization* loss
    - Static *gradient penalty* loss
- Learns high-quality temporal motion features by penalizing spatial information usage

Framework Diagram

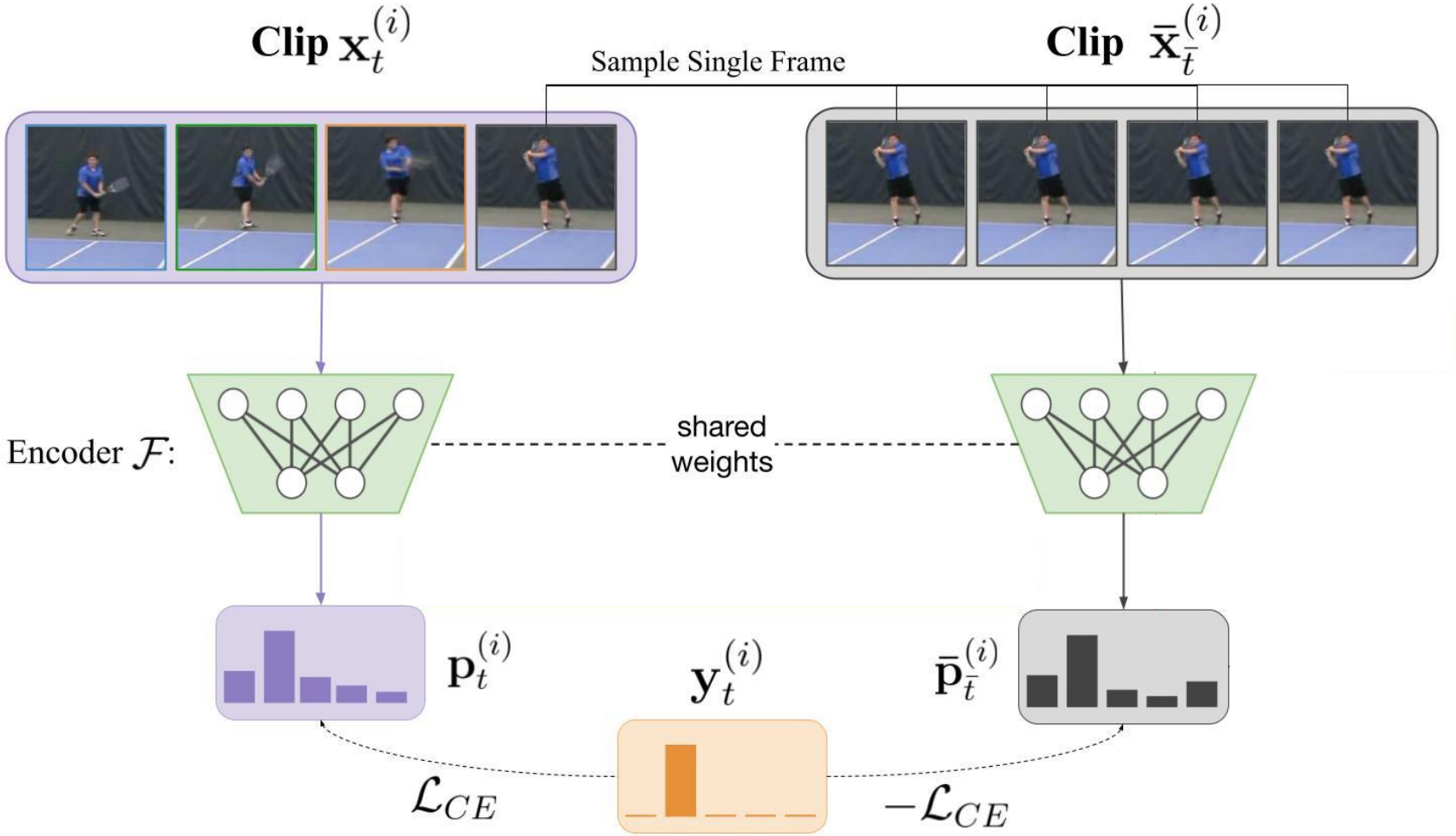


Framework Diagram



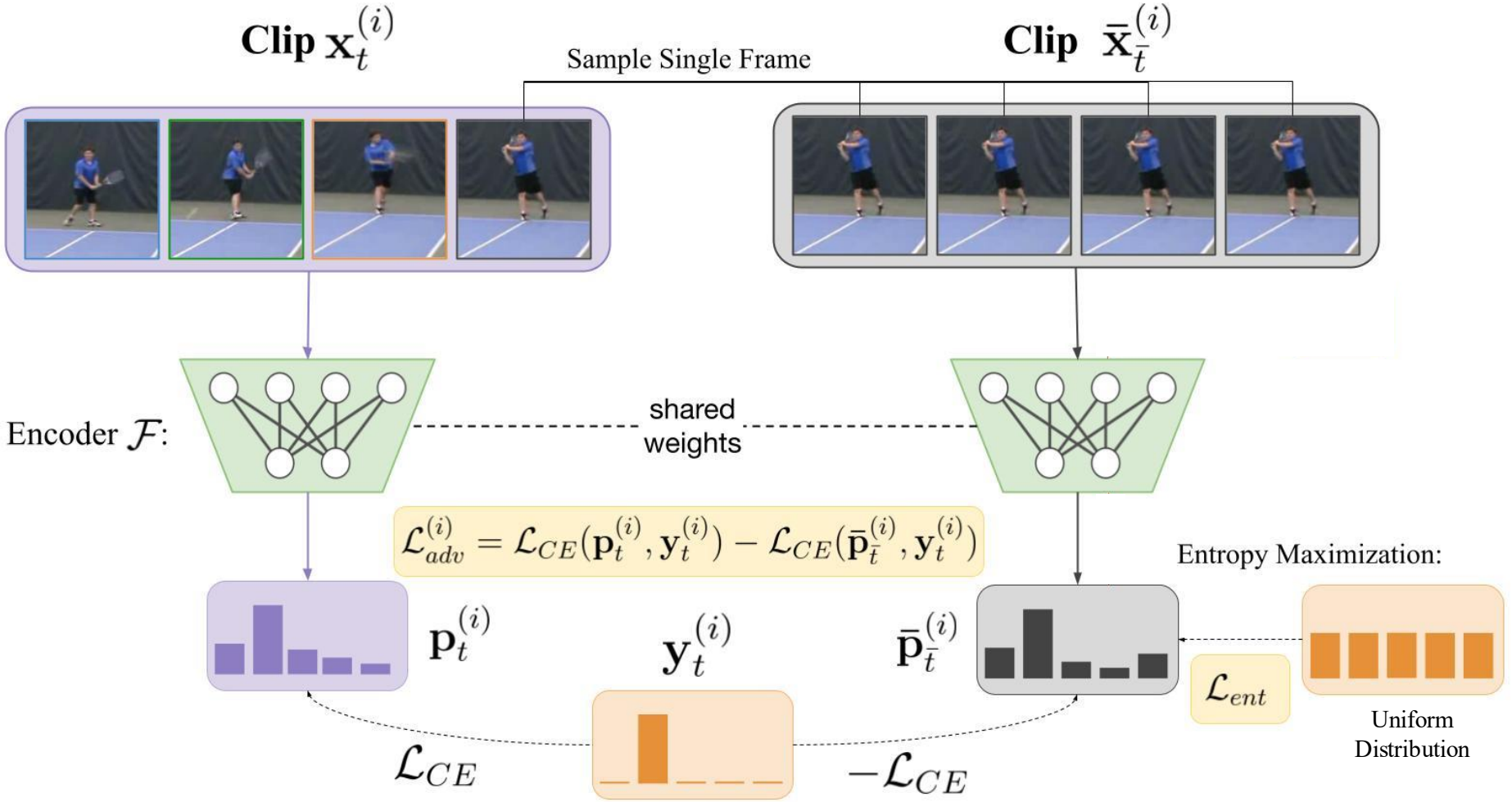
UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Framework Diagram



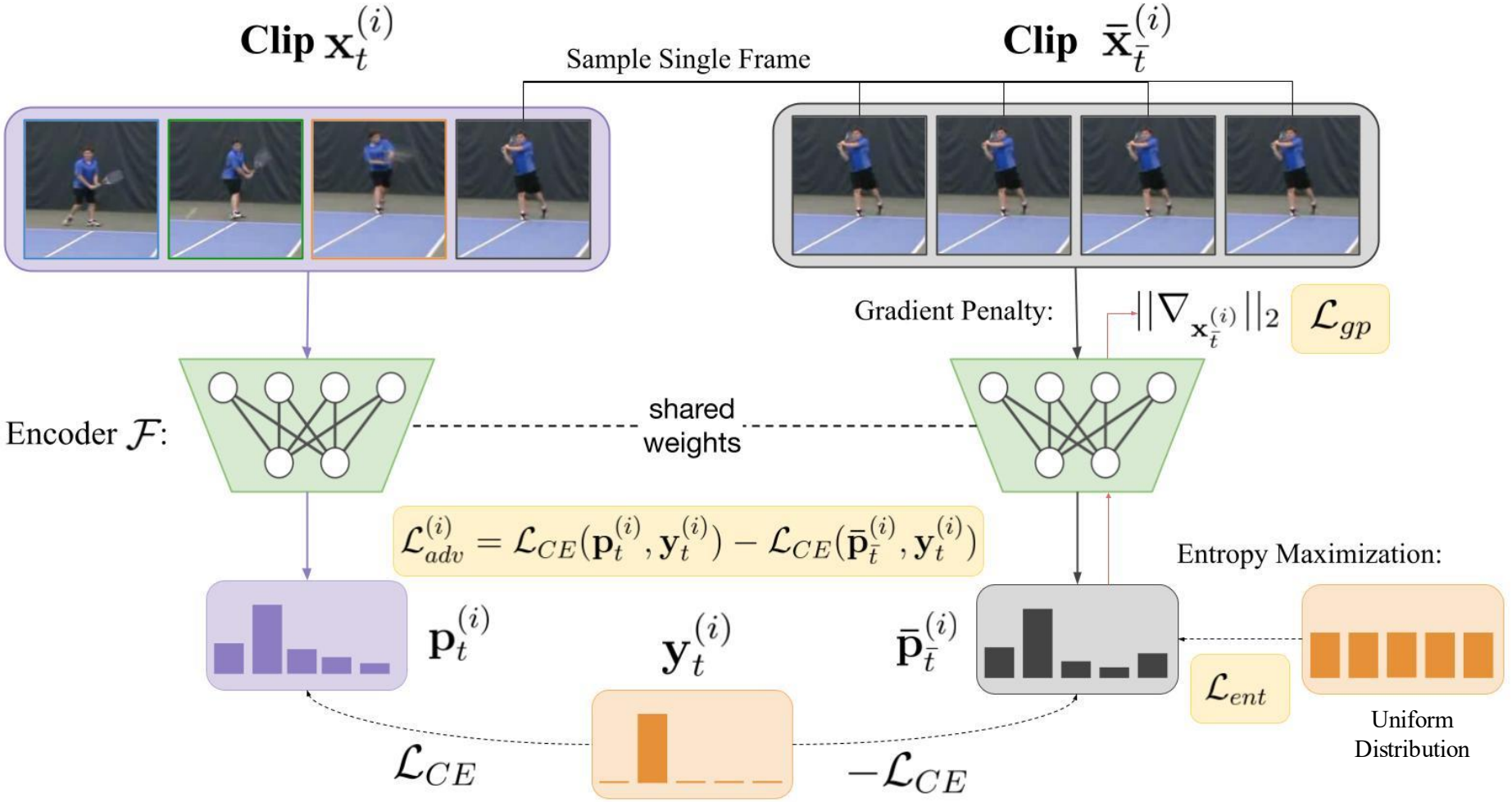
UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Framework Diagram



UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Framework Diagram



Overall Training Objective

- All that is required is two input clips: one with motion, one with no motion
  - One model, no bias labels/predictor networks

$$\mathcal{L}_{adv}^{(i)} = \mathcal{L}_{CE}(\mathbf{p}_t^{(i)}, \mathbf{y}^{(i)}) - \omega_{adv}\mathcal{L}_{CE}(\bar{\mathbf{p}}_{\bar{t}}^{(i)}, \mathbf{y}^{(i)})$$

$$\mathcal{L}_{\text{ent}}^{(i)} = \sum_{c=1}^{N_C} \bar{\mathbf{p}}_{\bar{t},c}^{(i)} \log(\bar{\mathbf{p}}_{\bar{t},c}^{(i)})$$

$$\mathcal{L}_{gp}^{(i)} = ||\nabla_{\bar{\mathbf{x}}_{\bar{t}}^{(i)}} \mathcal{F}(\bar{\mathbf{x}}_{\bar{t}}^{(i)})||_2$$

Combined Overall Objective: $$\mathcal{L}^{(i)} = \mathcal{L}_{adv}^{(i)} + \omega_{ent} * \mathcal{L}_{ent}^{(i)} + \omega_{gp} * \mathcal{L}_{gp}^{(i)}$$

Background/Foreground Bias Evaluation

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Original Video

[7] H. Li, Y. Liu, H. Zhang, and B. Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19911–19923, 2023.

Background/Foreground Bias Evaluation

Original Video

Static CUes in BAckground SCUBA

[7] H. Li, Y. Liu, H. Zhang, and B. Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19911–19923, 2023.

Background/Foreground Bias Evaluation

Original Video

Static CUes in BAckground
SCUBA

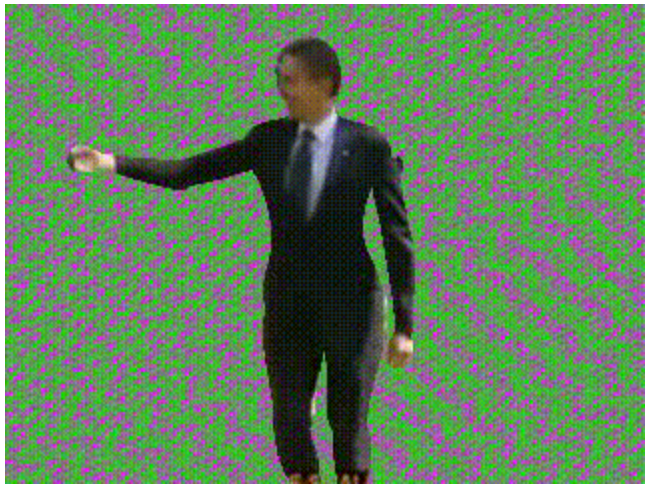
Select One Frame

Static CUes in FOreground
SCUFO

[7] H. Li, Y. Liu, H. Zhang, and B. Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19911–19923, 2023.

Conflicting Foreground (ConflFG)

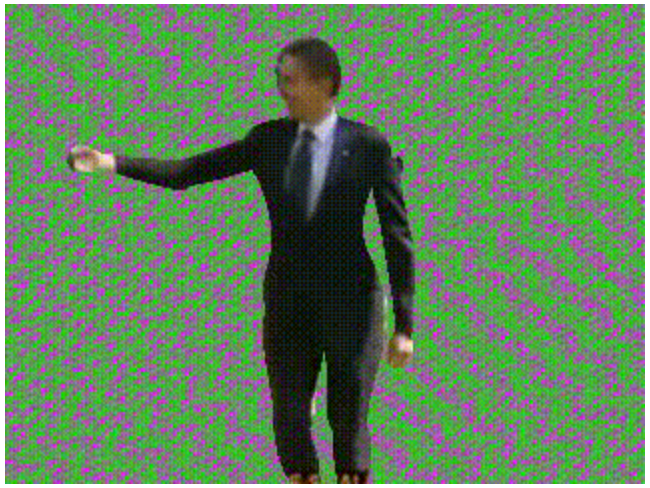
Random SCUBA Video



[7] H. Li, Y. Liu, H. Zhang, and B. Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19911–19923, 2023.

Conflicting Foreground (ConflFG)

Random SCUBA Video



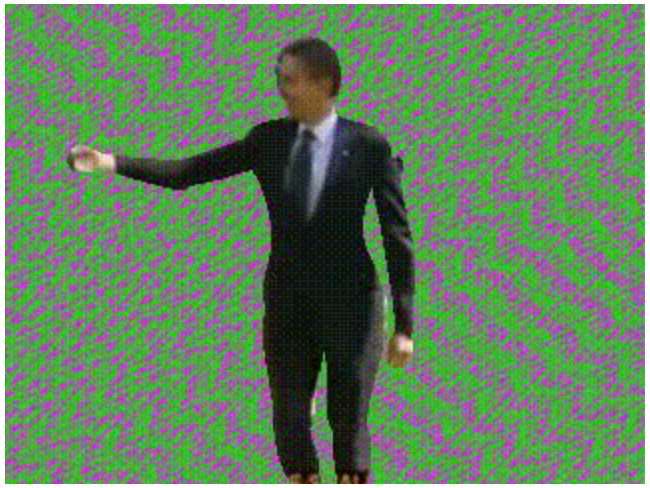
Different SCUFO Foreground

[7] H. Li, Y. Liu, H. Zhang, and B. Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19911–19923, 2023.

Conflicting Foreground (ConflFG)

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Random SCUBA Video



Different SCUFO Foreground

Combined ConflFG Video

[7] H. Li, Y. Liu, H. Zhang, and B. Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19911–19923, 2023.

HMDB51 Main Results

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Augmentation or Debiasing	IID	OOD			
		Avg SCUBA (↑)	Avg SCUFO (↓)	Confl-FG (↑)	Contra. Acc. (↑)
None	73.92	43.93	20.46	36.58	27.84

HMDB51 Main Results

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Augmentation or Debiasing	IID	OOD			
		Avg SCUBA (↑)	Avg SCUFO (↓)	Confl-FG (↑)	Contra. Acc. (↑)
None	73.92	43.93	20.46	36.58	27.84
Mixup [1] ICLR'18	74.58	43.10	21.17	36.62	26.09
VideoMix [2] arXiv'20	73.31	39.39	20.44	32.68	23.13
SDN [3] NeurIPS'19	<u>74.66</u>	40.02	20.22	34.87	22.88
BE [4] CVPR'21	74.31	43.56	19.96	35.99	27.84
ActorCutMix [5] CVIU'23	74.05	46.79	22.07	36.97	28.12
FAME [6] CVPR'22	73.79	51.40	26.92	39.61	29.66
StillMix [7] ICCV'23	**74.82**	51.81	13.39	47.38	40.28

HMDB51 Main Results

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Augmentation or Debiasing	IID	OOD			
		Avg SCUBA (↑)	Avg SCUFO (↓)	Confl-FG (↑)	Contra. Acc. (↑)
None	73.92	43.93	20.46	36.58	27.84
Mixup [1] ICLR'18	74.58	43.10	21.17	36.62	26.09
VideoMix [2] arXiv'20	73.31	39.39	20.44	32.68	23.13
SDN [3] NeurIPS'19	<u>74.66</u>	40.02	20.22	34.87	22.88
BE [4] CVPR'21	74.31	43.56	19.96	35.99	27.84
ActorCutMix [5] CVIU'23	74.05	46.79	22.07	36.97	28.12
FAME [6] CVPR'22	73.79	51.40	26.92	39.61	29.66
StillMix [7] ICCV'23	**74.82**	51.81	13.39	47.38	40.28
Ours	73.20	**53.20** ↑21.1%	**0.42** ↓98.0%	**49.84** ↑36.3%	**53.02** ↑90.5%

Kinetics400 Main Results

Augmentation or Debiasing	Original	Modified		
		Avg SCUBA (↑)	Avg SCUFO (↓)	Contra. Acc. (↑)
None	**68.13**	42.97	20.26	25.78
StillMix [7] ICCV'23	67.27	**45.83**	10.72	36.60
Ours	67.58	44.75 ↑4.1%	**0.11** ↓99.5%	**44.74** ↑73.6%

- ALBAR works on larger Kinetics400

[10] Haodong Duan, Yue Zhao, Kai Chen, Yuanjun Xiong, and Dahua Lin. Mitigating representation bias in action recognition: Algorithms and benchmarks. In European Conference on Computer Vision,pp. 557–575. Springer, 2022.

UCF101 Updated Protocol Results

UCF CENTER FOR RESEARCH IN COMPUTER VISION

Augmentation or Debiasing	Original	OOD			
		Avg SCUBA (↑)	Avg SCUFO (↓)	Confl-FG (↑)	Contra. Acc. (↑)
None	95.51	18.78	1.50	49.36	17.53
StillMix [7] ICCV'23	**96.14**	24.68	0.36	**55.03**	24.10
Ours	94.98	**26.25** ↑39.8%	**0.14** ↓90.7%	54.02 ↑9.4%	**26.23** ↑49.6%

Downstream Task Performance Results

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Method	Action Recognition HMDB51 (OOD)	Anomaly Detection UCF_Crime	Temporal Action Localization THUMOS14
Baseline	27.84	82.39	54.89
Ours	**53.02**	**84.91**	**55.20**

Reduced bias translates to improved performance across tasks!

Table Citations

[1] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[2] S. Yun, S. J. Oh, B. Heo, D. Han, and J. Kim. Videomix: Rethinking data augmentation for video classification. *arXiv preprint arXiv:2012.03457*, 2020.

[3] J. Choi, C. Gao, J. C. Messou, and J.-B. Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] J. Wang, Y. Gao, K. Li, Y. Lin, A. J. Ma, H. Cheng, P. Peng, F. Huang, R. Ji, and X. Sun. Removing the background by adding the background: Towards background robust self-supervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11804–11813, 2021.

[5] Y. Zou, J. Choi, Q. Wang, and J.-B. Huang. Learning representational invariances for data-efficient action recognition. *Computer Vision and Image Understanding*, 227:103597, 2023.

[6] S. Ding, M. Li, T. Yang, R. Qian, H. Xu, Q. Chen, J. Wang, and H. Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9716–9726, 2022.

[7] H. Li, Y. Liu, H. Zhang, and B. Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19911–19923, 2023.

UCF
CENTER FOR RESEARCH IN COMPUTER VISION

Thanks for listening!

Questions?

Webpage/Code: