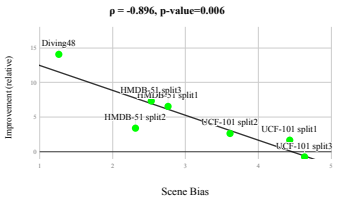
R1,R2,R3: *Improvement and scene bias:* We show a scatter plot on the relative improvement and scene representation bias of target datasets on the figure on the right side (best viewed with zoom). We measure the scene bias defined by Li et al.(referred to as "RESOUND"). We show the scene bias and the relative improvement of each split of the HMDB-51, UCF-101, Diving48 datasets. The Pearson correlation is $\rho = -0.896$ with a p-value 0.006, highlighting a strong negative correlation between the relative improvement and the scene bias.



R1,R2,R3: *Results on Diving48:* We compare our method with RESOUND C3D (L=16) on the Diving48 dataset. All the videos in the Diving48 dataset share similar scenes. Our proposed debiasing method shows a favorable result compared to RESOUND. We show a relatively larger relative improvement (14.1%) on the Diving48 dataset since it has a small scene bias of 1.26.

	Ours	RESOUND
w/o. debiasing	18.0	16.4
w. debiasing	20.5	N/A

R1: *Action-scene factorization vs. scene-invariance action features:* We believe both directions are worth pursuing. Both action-scene factorization and scene-invariant action features are effective in *typical* scenarios. For example, basketball players play basketballs in a basketball court. However, in *atypical* scenarios, learning a scene-invariant model is particularly important. For example, a singer is singing a song in a baseball field. Factorized models may incorrectly predict the actions because factorized models rely on the scene as well. On the other hand, scene-invariant models focus on the actual action taking place. Consequently, they can perform well, particularly for out-of-context actions. We will cite the missing papers and discuss the scene-action duality in the revision.

R1: *Methodology:* Our method differs from Ganin and Lempitsky, in that we consider another objective: *human mask entropy*. Specifically, we maximize the entropy of our model's action prediction when humans are masked out in the video. Our method differs from Wang and Hoai (referred to as "Factor") because we mask out humans by explicitly detecting them, while Factor temporally masks out actions (not humans) in the video by using conjugate examples. Note that there is no guarantee that the conjugate examples corresponding to the current action examples do not contain the actions, especially when the dataset is temporally trimmed e.g., UCF-101 and HMDB-51.

R1: *Information from the rest of the pixels (60%) within bounding boxes could leak into the action code?:* The information from the rest of the pixels (60%) within bounding boxes does not leak into the action code. Our training objective is to maximize the entropy of a model prediction, L_{Ent}, when detected bounding boxes are *masked out*. In this case, a model sees only the background context but not the actor. Maximizing the entropy of action prediction, therefore, prevents the background information from leaking into the action code.

	Ours	F-C3D	F-DTD	F-EigenTSN
Original	83.5	82.2	90.8	95.8
Proposed	84.5	84.5	91.3	95.8

R1: *Comparison with Factor:* We compare ours with Factor on the UCF-101 dataset. They show results on the Hollywood2 dataset on which we do not conduct experiments (instead we have results on HMDB-51). Note that the model used in Factor C3D (F-C3D) is 2.4× larger than ours. A comparison with Factor DTD+C3D (F-DTD) and Factor EigenTSN+C3D (F-EigenTSN) is not fair as they are both ensemble methods while ours is not.

R1, R2: *Weak baselines:* Unfortunately, we do not have access to computational resources for training deeper 3D networks such as 3D-ResNet-151 or I3D. We thus resort to using lighter backbones (3D-ResNet-18 and VGG-16). However, this does not undermine our core novelty as we focus on the improvement induced by our debiasing method. Here, we show the results using a deeper 3D-ResNet-50 backbone. On HMDB-51 dataset, without debiasing, the accuracy is 59.6%, and with debiasing, the accuracy is 60.1%. We also plan to plug-in our debiasing to the state-of-the-art backbone models in the future.

R1: *Scene classification accuracy:* On the Mini-Kinetics-200 validation set, without debiasing, the scene classification accuracy is 29.7%. With debiasing, the scene classification accuracy drops to 2.9%. The random chance is 0.3%. The proposed debiasing method indeed reduces the scene-dependent feature representation. Please note that there are no ground truth scene labels. Here we use pseudo labels to measure the scene classification accuracy.

R1,R3: *Missing citations, typos, and awkward sentences:* We will cite the missing related work (Factor, Zhao et al., He et al., Vu et al., and Khosla et al.) and discuss them in the revision. We will revise the paper accordingly.

L_{Adv}	L_{Ent}	frame mAP
×	×	32.5
✓	×	34.4
✓	✓	34.5

R3: *Why not using L_{Ent} for spatio-temporal detection?, Why ablation only on HMDB-51:* We show the result of using both L_{Adv} and L_{Ent} in the spatio-temporal action detection experiment in the table on the right side. Adding L_{Ent} improves the performance by 0.1 point. Due to the limited resources, we conducted the ablation experiments on the relatively smaller HMDB-51 dataset for the classification task.

R3: *Why debiasing is important for detection compared to classification:* We can imagine a scenario that an actor is running and the background scene is changing from urban to suburb. Debiased models can (spatio-)temporally localize the action correctly by focusing on the actual action. However, a non-debiased model, which is trained on the dataset where running only happens with the urban scene, might (spatio-)temporally localize the running action only in the urban scene and fail to recognize the running in the suburb scene.