Enabling Detailed Action Recognition Evaluation Through Video Dataset Augmentation

Jihoon Chung  
Princeton University  
jc5933@princeton.edu

Yu Wu  
Princeton University  
yuwu@princeton.edu

Olga Russakovsky  
Princeton University  
olgarus@princeton.edu

Abstract

It is well-known in the video understanding community that human action recognition models suffer from background bias, over-relying on scene cues in making their prediction. However, it is difficult to quantify this effect given the existing evaluation frameworks. We introduce the Human-centric Analysis Toolkit (HAT), which enable evaluation of the learned background bias without the need for new manual video annotation. It does so by automatically generating synthetically manipulated videos, leveraging the recent advances in image segmentation and video inpainting. Using HAT we perform an extensive analysis of 74 action recognition models trained on the Kinetics dataset. We confirm that all these models focus more on the scene background than on the human motion; further, we demonstrate that certain model design decisions (such as training with fewer frames per video or using dense as opposed to uniform temporal sampling) appear to worsen the background bias. We open-source HAT to enable the community to leverage its metrics to design more robust and generalizable human action recognition models.

1 Introduction

Human action recognition is about understanding what the human in the video is doing; however, human action recognition models frequently rely on background cues to make their predictions. Works such as [6, 33, 59, 60, 70] have leveraged visualization tools like GradCam [46] to demonstrate that the video background significantly influences the prediction of human action recognition models. This occurs due to representation bias in the dataset, where particular actions (e.g., eating) tend to occur in particular environments (e.g., kitchens). Such concerns limit the practical usability and generalizability of models despite the impressive overall progress in the field [36, 61, 67].

While it is known that this background bias phenomenon is occurring, quantifying the degree to which it is occurring is still necessary. Being able to accurately assess how much human action recognition models rely on human features rather than background scene cues would allow researchers to compare different model designs and select the ones that would be robust to their unique test domains. Efforts such as [34, 63] have introduced datasets specifically curated for quantifying background bias;
In this work, we introduce the Human-centric Analysis Toolkit (HAT) to measure background bias in human action recognition models without the need for costly human annotation. We leverage recent improvements in image segmentation [24, 32, 38] and video inpainting [29, 35, 37] to automatically synthesize counterfactual videos containing Human-Only (a spatio-temporal segmentation of the human figure against a gray background), Background-Only (the video with the human removed via inpainting) or Action-Swap (human figure against an unusual background). Examples are shown in Figure 1. This process is efficient and scalable, requiring no manual annotation. HAT thus enables us to evaluate the sensitivity of human action understanding models to the different visual cues by comparing the accuracy on the original and synthetically manipulated videos.

We demonstrate the capabilities of HAT by running extensive analysis of human action recognition models trained on the Kinetics-400 [28] dataset. Concretely, we evaluate 74 trained models, corresponding to 14 different model designs (TSN [61], I3D [5], Non-local Neural Networks [62], R(2+1)D [61], TSM [36], SlowFast/SlowOnly [15], CSN [58], TIN [49], TPN [68], X3D [14], OmniSource [12], TANet [39], and TimeSformer [4]) with varying hyperparameters and backbone architectures provided by the MMAction2 [8] implementation. We find, for example:

- All 74 models exhibit strong background bias. When evaluated on the Action-Swap videos, the 74 models predicted the action class of the human 16.8% of the time on average – but predicted the action class of the randomly-selected background 29.5% of the time!
- Models trained with fewer frames per video appear to be more prone to background bias. For example, the TSN-based models [61] trained with 8, 5 and 3 frames per video retain 0.679, 0.683 and 0.694 of their original accuracy respectively when evaluated on the Background-Only videos, demonstrating consistently high and somewhat increasing background bias.
- Models trained with dense temporal sampling around a single timestep appear to be more prone to background bias compared to models trained with uniform sampling throughout the video. For example, when evaluated on the Background-Only videos as above, TSM-based models [36] with dense sampling exhibit strong background bias by retaining 0.703 of the original accuracy compared to only 0.675 with uniform sampling.

Overall, we make three contributions. First, we develop and open-source the Human-centric Analysis Toolkit (HAT), which generates synthetic videos to evaluate the background bias learned by human action recognition models. Second, we demonstrate its capabilities through extensive evaluation of 74 released models. Finally, we show that HAT can identify the design choices that appear to influence the amount of background bias learned by the model, helping inform future model design.
2 Related Work

**Human Action Recognition Models.** Currently, human action recognition is largely dominated by deep learning methods. With strong success in image-based tasks [10, 20, 30, 51], CNN-based deep learning models [12, 14, 15, 36, 39, 49, 54, 58, 61, 62, 68] were the go-to method for human action recognition, with gradual improvements in the model structure going from 2D-CNN [20, 44, 61] to 3D-CNN [5, 14, 15, 54] to CNN models with specific temporal modeling [36, 39, 49]. A recent trend [2, 4] in human action recognition is to use a transformer module [57] as it has shown good performance [11] in image-based tasks. Another trend [4, 12, 58] is to incorporate large-scale datasets [16, 26, 65] into the training. In this work, we evaluate multiple action recognition models [4, 5, 12, 14, 15, 36, 39, 49, 54, 58, 61, 62, 68] in an effort to identify design decisions which appear to correlate with learned background bias.

**Human Action Recognition Dataset.** The early datasets [45, 66] offered a handful of human action classes that were collected in a controlled environment. UCF101 [52] and HMDB51 [31] were one of the first few datasets that were suitable for machine learning tasks. Although there are many different human action datasets [7, 9, 19, 47, 48, 72], these days, the most popular dataset must be Kinetics-400 [28], due to its large size and the variety of actions. However, due to the cost of collecting video datasets, the size of the dataset is still smaller compared to image datasets. Synthetic datasets [13, 17, 27, 41, 53], often used mixed with the real dataset, are popular methods of collecting data in an affordable manner. In human action recognition, the synthetic datasets are often used for training dataset [55, 56], and the model is tested on real videos. In this work, we generate synthetic counterfactual videos to enable detailed model evaluation without the need for costly annotation.

**Human-centric Analysis.** As the models have grown more complex, there has been an increased need for frameworks that provide insights into the model behavior beyond just a single accuracy number. Efforts have included model interpretability techniques [3, 46], detailed error analysis using additional manual annotations [1, 23, 33, 50], and (recently) stress-testing using automatically generated text or image data [25, 42]. There are a number of works studying specifically the impact of the human figure on human action recognition models. A common strategy employed by [6, 33, 59, 60, 70] is to use the GradCam [46] visualization to qualitatively demonstrate that the model’s attention is on the background cues rather than on the human in the video. Several of these papers [6, 59, 60, 70] propose methods to mitigate the effects of background bias during training; they evaluate their success both qualitatively through GradCam and quantitatively via accuracy on a downstream action recognition task (after fine-tuning the model trained with their new background-debiasing method). While this successfully demonstrates that their innovation is effective for model pre-training, it does not directly measure the learned background bias. The most natural analysis is to collect specific datasets [7, 18, 34, 48, 63], such that the trained models can have high accuracy on the dataset if and only if they can understand the human body movement. One such example is Mimetics [63] with 713 hand-collected videos of 50 human action classes from Kinetics-400 [28] happening against irrelevant backgrounds. However, scaling up or generalizing this effort would be extremely costly due to the need for manual annotation. In contrast, our toolkit provides quantitative metrics for directly measuring the effect of background bias without the need for manual annotation.

3 Human-centric Analysis Toolkit

Our Human-centric Analysis Toolkit (HAT) is a general framework that can be used to measure the amount of background bias learned by a human action recognition model. HAT takes two inputs: (1) a trained human action recognition model and (2) a set of validation videos each annotated with the human action class. HAT then proceeds in three steps. First, it leverages human segmentation models to separate the human visual cues from the background visual cues in the validation videos. Second, it generates six sets of counterfactual validation videos, including Human-Only, Background-Only, and four sets of Action-Swap videos (see Figure 1 for examples). Finally, it evaluates the trained model on these counterfactual videos and returns a set of ten metrics which quantify the different effects of background bias. This methodology can expand the dataset without any need to manually collect new data, allowing deeper analysis of human action recognition in an affordable manner.
3.1 Separating human from background

The first step of HAT is separating the visual cues corresponding directly to the human from the rest of the cues in the video. This can be done using a pre-trained human segmentation model. Interestingly, in our internal experiments we find that modern image-based segmentation models [24, 71, 69] tend to have better results than video-based segmentation models [40]. (We hypothesize that this might be due to the differences in training set size.) While older CNN-based image segmentation models [71, 69] suffer from low temporal consistency, missing human segments in some of the frames, the modern transformer-based SeMask [24] actually appears to have overcome this limitation. We use SeMask trained on ADE20K [74] in our implementation and include sample videos in the supplementary material.

One thing to note is that in the current instantiation of HAT we consider any objects that the human is interacting to be part of the background. Thus, for example, a person performing the “drinking coffee” action would be expected to be segmented separately from the coffee mug that they are holding (which becomes part of the background). One way of partially avoiding this would be to use a human bounding box instead of a segmentation mask – however, undesirable background cues would then also be included. Different tradeoffs can be considered in future instantiations of HAT.

3.2 Generating counterfactual validation videos

The core of our toolkit is generating synthetic validation videos with different visual cues, which allows us to investigate the effect of the different cues on human action recognition models.

The first two sets of videos are Background-Only (where only the background is shown and all human cues are removed) and Human-Only (where only the human cues are shown). For Background-Only, we leverage the video inpainting model [29] to remove all human pixels segmented by the model of Section 3.1. In contrast to prior works [6, 21] which fill the human pixels with a frame average color (e.g., grey), we use inpainting to generate a more realistic-looking video. For Human-Only, we instead keep only the segmented human pixels and fill in the rest with an average color. We use the dataset’s average color rather than the frame average, since that can reveal a lot about the background, e.g., green for a sports field or blue for a body of water.

The other four sets of videos are more complex Action-Swap videos, which combine different visual cues to investigate their additive effects. We synthesize these videos by combining the segmented human figure with the background from a different video, similar to [64]. While the Background-Only and Human-Only video sets are both decidedly outside the model’s training data distribution, these Action-Swap videos are arguably somewhat more realistic since they do contain a human figure against a viable background – although in an unexpected combination. Example frames are in Figure 1 and videos in supplementary material; more details on Action-Swap generation below.

3.2.1 Details of generating Action-Swap videos

HAT includes four different types of Action-Swap videos:

- **Random**: The background is swapped with a video from a different class.
- **Close**: The background is swapped with a video from a class with a similar background.
- **Far**: The background is swapped with a video from a class with a very different background.
- **Same**: The background is swapped with a video from the same class. This can be used as a theoretical upper bound of Action-Swap Accuracy.

To determine the appropriate classes for Close and Far Action-Swap videos, we need to determine how similar the backgrounds are across different classes. To do so, we first feed the frames from the original validation videos into a Places365 [73] trained scene classification model. For each action class, we then compute the average scene prediction vector by averaging the prediction probabilities from all frames of all videos of this class. We can then rank all the other classes according to the L1 distance in their average scene prediction vector. We consider the class to be “close” if it’s among the 5 classes with the smallest L1 distance and “far” if it’s among the 200 largest (of 399 classes total).
For generating an Action-Swap counterfactual video, we thus:

1. segment the human figure from the video using [24] as if creating a Human-Only video,
2. randomly sample a background action class, depending on the particular Action-Swap set,
3. randomly sample a video of the class from (2),
4. generate the Background-Only version of the video from (3),
5. paste in the human figure from (1) onto the video from (4)

One additional challenge is that we want to ensure that sufficient human and background cues are present in every generated Action-Swap video. Thus, we only consider videos where all frames have human masks taking up 5-50% of the pixels; when sampling background videos in step (3) we relax the lower bound to allow videos with few human pixels². Therefore, unlike the Background-Only and Human-Only sets, the Action-Swap sets have fewer video samples than the original dataset. In Kinetics-400, we end up with 5,631 videos, whereas the original validation set has 19,877 videos. To compensate for this, we run steps (2-5) three times for each video to generate three different videos.

3.3 Metrics

We use the generated counterfactual validation videos from Section 3.2 to evaluate the trained human action recognition models. We measure how much of the original recognition accuracy comes from the different cues:

\[
\text{Background-Only Ratio (BOR)} = \frac{\text{Background-Only Accuracy}}{\text{Original Accuracy}} \tag{1}
\]

\[
\text{Human-Only Ratio (HOR)} = \frac{\text{Human-Only Accuracy}}{\text{Original Accuracy}} \tag{2}
\]

If a model shows high BOR, i.e., a model can get close to the original accuracy with just the background cues, we see this as “right for the wrong reason.” In contrast, ideally models would have high HOR since they should be able to recognize the human action even without the background cues.

Finally, for Action-Swap videos recall that each counterfactual video is generated by combining the human figure foreground from class A with the background from a different class B. We then measure the Swap Human Accuracy (SHAcc) as the fraction of counterfactual videos the model predicts correctly as class A, and Swap Background Error (SBErr) as the fraction of times the model incorrectly predicts the video as the background class B. Human action recognition models that successfully rely on human motion cues would be expected to have high SHAcc; those that are driven primarily by background cues would be expected to have high SBErr.

4 Analyzing Action Recognition Models

We now demonstrate the capabilities of HAT by evaluating human action recognition models trained on the popular Kinetics-400 [28] dataset. We present the results on the different types of counterfactual videos in order (Background-Only in Section 4.2, Human-Only in Section 4.3, and Action-Swap in Section 4.4), along with discussing our findings and drawing conclusions about different model design decisions that appear to have contributed to the learned background bias. HAT is not limited to Kinetics-400, and can be used on other human action recognition datasets [19, 23, 52]. Please refer to the supplementary material for the experiments on UCF101.

4.1 Experimental Details

We test a number of different model designs, including TSN [61], I3D [5], Non-local Neural Networks [62], R(2+1)D [54], TSM [36], SlowFast, and SlowOnly [15], CSN [58], TIN [49].

²Please see visualization examples here: https://github.com/princetonvisualai/HAT/blob/main/doc/review_discussion.md#percentage-of-synthetic-pixels
We extract the videos in 30 FPS with original resolution. For other pre-processing, such as resizing and temporal sampling, we follow the configuration that each model specified. We list the details of the tested models and their configuration in the supplementary material. Within the scope of the paper, we chose not to retrain any models and rely on publicly released model weights. In drawing conclusions we try to do an apples-to-apples comparison whenever possible; however, we are not able to guarantee that all hyperparameter settings are directly comparable between the different models.

For image segmentation and video inpainting, we used 20 Nvidia RTX 3090 GPUs with 20 hours of forward pass to generate synthetic videos of the full Kinetics-400 validation set. See supplementary material for examples of the synthetic videos on Kinetics-400.

### 4.2 Background-Only Video

**Accuracy and Background-Only Ratio.** Table 1 tabulates the model accuracies in Background-Only Videos. For a fair comparison, we have separated the weights that use additional large-scale datasets [10, 12, 16, 26, 65]. Despite removing a human body from the video, thus removing any human action, all the models still show a strong tendency to predict the removed action. This hints at the possibility that the performance of the human action recognition models is highly dependent on the background, rather than the action itself.

Table 1 tabulates the Background-Only Ratio. It shows that on all the tested models, we see around 70% of the accuracy is coming from the non-human regions, revealing the problematic behavior, “right for the wrong reason”, is common in human action recognition. Next, we show examples of using Background-Only Ratio to analyze and improve the model design.

**Number of Frames used to Train.** The first three plots of Figure 2 visualize how the number of video frames used during the training can worsen the Background-Only Ratio. This shows that the models trained with fewer temporal frames tend to suffer more, with a lot of their accuracy coming from the background. A possible explanation is that when fewer frames are given, the model is not able to learn to understand temporal information, thus given a video with or without the human movement, the model will perform similarly, as they never learned to understand such complex human movement during training. Thus the accuracy would come from the temporally static background. While exact behavior can be different per model structure, we see this to be most severe on TSN [61] which lacks any sophisticated temporal modeling.

**Sampling Strategy.** We check if the frame sampling strategy can affect the Background-Only Ratio. The results are visualized on the last plot of Figure 2. Unlike uniform sampling, i.e., getting uniformly distributed frames, dense sampling strategy, i.e., sampling frames with a specified stride, shows higher BOR in general. We believe this is due to the dense sampling strategy having a smaller temporal
Table 1: Accuracies over Background-Only Videos. OAcc and BAcc denote original accuracy and accuracy on Background-Only Videos, respectively. Models using additional large-scale data are tabulated separately below. We only tabulate the setting with the highest OAcc per backbone. Check supplementary material for the experiment results of all 74 weights.

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>Pre-trained</th>
<th>OAcc (%)</th>
<th>BAcc (%)</th>
<th>BAcc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Normal-scale dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSM [36]</td>
<td>MNetV2 [44]</td>
<td>ImageNet</td>
<td>69.87</td>
<td>48.84</td>
<td>0.6990</td>
</tr>
<tr>
<td>R(2+1)D [54]</td>
<td>ResNet34</td>
<td>-</td>
<td>74.22</td>
<td>52.99</td>
<td>0.7140</td>
</tr>
<tr>
<td>TSN [49]</td>
<td>ResNet50</td>
<td>ImageNet</td>
<td>71.75</td>
<td>49.02</td>
<td>0.6835</td>
</tr>
<tr>
<td>TSN [61]</td>
<td>ResNet50</td>
<td>TSM-Kinetics400</td>
<td>70.89</td>
<td>48.32</td>
<td>0.6816</td>
</tr>
<tr>
<td>I3D [5]</td>
<td>ResNet50</td>
<td>ImageNet</td>
<td>73.57</td>
<td>52.26</td>
<td>0.7053</td>
</tr>
<tr>
<td>NL-TSM [62]</td>
<td>ResNet50</td>
<td>ImageNet</td>
<td>74.09</td>
<td>52.25</td>
<td>0.7053</td>
</tr>
<tr>
<td>NL-I3D [62]</td>
<td>ResNet50</td>
<td>ImageNet</td>
<td>74.91</td>
<td>52.84</td>
<td>0.7054</td>
</tr>
<tr>
<td>NL-SlowOnly [62]</td>
<td>ResNet50</td>
<td>ImageNet</td>
<td>75.78</td>
<td>53.51</td>
<td>0.7062</td>
</tr>
<tr>
<td>CSN [58]</td>
<td>ResNet50</td>
<td>-</td>
<td>73.22</td>
<td>51.97</td>
<td>0.7098</td>
</tr>
<tr>
<td>TPN [58]</td>
<td>ResNet50</td>
<td>ImageNet</td>
<td>76.16</td>
<td>54.40</td>
<td>0.7143</td>
</tr>
<tr>
<td>SlowOnly [15]</td>
<td>ResNet50</td>
<td>ImageNet</td>
<td>75.35</td>
<td>53.97</td>
<td>0.7163</td>
</tr>
<tr>
<td>SlowFast [15]</td>
<td>ResNet50</td>
<td>-</td>
<td>76.61</td>
<td>53.46</td>
<td>0.6978</td>
</tr>
<tr>
<td>SlowOnly [15]</td>
<td>ResNet101</td>
<td>-</td>
<td>76.26</td>
<td>54.38</td>
<td>0.7131</td>
</tr>
<tr>
<td>SlowFast [15]</td>
<td>ResNet101+50</td>
<td>-</td>
<td>76.55</td>
<td>55.19</td>
<td>0.7210</td>
</tr>
<tr>
<td>SlowFast [15]</td>
<td>ResNet101</td>
<td>-</td>
<td>78.10</td>
<td>56.14</td>
<td>0.7189</td>
</tr>
<tr>
<td>CSN [58]</td>
<td>ResNet152</td>
<td>-</td>
<td>77.62</td>
<td>54.33</td>
<td>0.6999</td>
</tr>
<tr>
<td>SlowFast [15]</td>
<td>ResNet152+50</td>
<td>-</td>
<td>77.24</td>
<td>55.46</td>
<td>0.7179</td>
</tr>
<tr>
<td>X3D [14]</td>
<td>X3D_S</td>
<td>-</td>
<td>72.67</td>
<td>50.61</td>
<td>0.6964</td>
</tr>
<tr>
<td>X3D [14]</td>
<td>X3D_M</td>
<td>-</td>
<td>75.55</td>
<td>52.47</td>
<td>0.6944</td>
</tr>
<tr>
<td>TANet [39]</td>
<td>TANet</td>
<td>ImageNet</td>
<td>76.10</td>
<td>53.71</td>
<td>0.7059</td>
</tr>
<tr>
<td><strong>Large-scale dataset</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSN [61]</td>
<td>ResNet50</td>
<td>IG-1B [65]</td>
<td>70.96</td>
<td>49.05</td>
<td>0.6912</td>
</tr>
<tr>
<td>Omni-TSN [12]</td>
<td>ResNet50</td>
<td>IG-1B [65]</td>
<td>74.70</td>
<td>52.09</td>
<td>0.6973</td>
</tr>
<tr>
<td>Omni-SlowOnly [12]</td>
<td>ResNet50</td>
<td>-</td>
<td>76.49</td>
<td>55.00</td>
<td>0.7190</td>
</tr>
<tr>
<td>CSN [58]</td>
<td>ResNet50</td>
<td>IG65M [16]</td>
<td>79.09</td>
<td>55.83</td>
<td>0.7059</td>
</tr>
<tr>
<td>Omni-SlowOnly [12]</td>
<td>ResNet101</td>
<td>-</td>
<td>80.00</td>
<td>58.05</td>
<td>0.7255</td>
</tr>
<tr>
<td>CSN [58]</td>
<td>ResNet152</td>
<td>IG65M [16]</td>
<td>82.38</td>
<td>58.97</td>
<td>0.7159</td>
</tr>
</tbody>
</table>

window so that the model was not able to learn the body movement sufficiently. Surprisingly, the effect of the sampling strategy would have not been clear if we only used original accuracy alone (see x-axis), showing a clear benefit of using BOR for model training analysis.

4.3 Human-Only Video

Accuracy and Human-Only Ratio. We plot HOR in left of Figure 3. We tabulated the evaluation results in the supplementary material. Given only the human action, all the models suffer significantly with an accuracy of around 20%. Despite Human-Only modification keeping the human action intact, the ratio is far lower than Background-Only Accuracy. By comparing BOR (with around 0.7) and HOR (with around 0.3), we quantitatively measure the well-believed problem of the current state of human action recognition, that most existing methods are all highly influenced by the background, more than the foreground human action.

Thankfully, we see a strong correlation between Human-Only Ratio and the original accuracy. This could hint that the performance improvement of the action recognition model is benefited from a better understanding of the human body, showing the important direction of where the human action recognition field needs to focus. Next, we show one example case where HOR can be used to evaluate different model structures.

TSN vs TSM. While the original paper on TSM [36] claims +4% accuracy improvements over TSN [61] on Kinetics-400, using different training and testing conditions, MMAction2 [61] shows
that the accuracy of TSN can be achieved on par with TSM, as shown in the x-axis of left of Figure 3. However, using Human-Only Ratio as a metric, we show that TSM does indeed show superior performance over TSN when a non-human region is removed. One possible explanation is that, as TSM design makes use of temporal difference, e.g., human body movement, it can capture the information of the human body better, as TSN cannot distinguish between human and background using its basic temporal modeling design.

4.4 Action-Swap Videos

Accuracy on Action-Swap. Table 2 right of Figure 3 details the performance of different models over the Action-Swap Videos. It shows that when we randomly swap background with other videos, all the models lean towards predicting the class of the background, rather than the foreground human action. Swapping between classes that are similar/different shows a gain/drop in SHAcc, showing that the output of a human action recognition model is largely dependent on the background.

Original Accuracy vs. Action-Swap Accuracy. Among models using normal-scale datasets, SlowFast-Res101 [15] shows the best accuracy when the background is relevant to the foreground action, on both original Kinetics accuracy (See Tab. 1 and Same Swap (See Tab. 2). However, given counterfactual videos that have irrelevant backgrounds, their performance drops to 18%, while the model falsely predicts 34 percent of the validation videos as their background class, one of the highest among the models we have tested. Such low performance on human action could be due to its reliance on the background, as models with better Random Swap SHAcc (CSN-Res152, SlowFast-Res152+50, etc.) show fewer background errors. Such experiment shows that models showing good accuracy in original Kinetics-400, might not be a good human action recognition model, due to their reliance on the background.

Use of Non-local Module. To demonstrate the evaluation of a model design using Action-Swap, we select Non-local [62] module as an example. Table 3 tabulates the evaluation results on Random Swap. We see that the Non-local module not only improves the original accuracy, but also drops the background error on all the tested models, showing reduced background bias. However, Non-local module do not always improve the focus on the human body, as for I3D [5] models, we see that SHAcc tends to drop.

Use of Large-scale Dataset for Pre-training. Table 4 tabulates the performance of models where we compare trained weight with/without additional large-scale pre-training. It shows that in all the cases, using a large-scale dataset improves the original accuracy and Random Swap Human Accuracy. However, as CSN shows an increase in the Background Error, this does not necessarily mean that the model is being better at recognizing the human. We expect the model is recognizing the image feature better when pre-trained with large-scale dataset, regardless of the scene or the person.
We test Human-centric Analysis Toolkit on the Kinetics-400 dataset and evaluate the generated Table 2: Action-Swap experiment results. We average the numbers from 3 random runs. We show standard deviation as well. See supplementary material for the full experiments.

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>Pre-trained</th>
<th>SHAcc↑</th>
<th>SHAcc↑</th>
<th>SHErr↓</th>
<th>SHAcc↑</th>
<th>SHErr↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal-scale dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TSM [56]</td>
<td>MNetV2</td>
<td>ImgNet</td>
<td>62.2±0.3</td>
<td>13.9±0.1</td>
<td>29.8±2</td>
<td>24.4±4</td>
<td>26.4±4</td>
</tr>
<tr>
<td>R2+1ID [54]</td>
<td>Res34</td>
<td>-</td>
<td>65.5±3</td>
<td>15.8±3</td>
<td>30.3±3</td>
<td>26.6±3</td>
<td>27.1±4</td>
</tr>
<tr>
<td>TSN [61]</td>
<td>Res50</td>
<td>ImgNet</td>
<td>60.2±0.3</td>
<td>13.3±0.1</td>
<td>28.1±2</td>
<td>23.4±4</td>
<td>26.7±4</td>
</tr>
<tr>
<td>TIN [49]</td>
<td>Res50</td>
<td>Kin400</td>
<td>58.6±1</td>
<td>18.3±2</td>
<td>20.8±1</td>
<td>27.1±2</td>
<td>21.0±2</td>
</tr>
<tr>
<td>TSM [56]</td>
<td>Res50</td>
<td>ImgNet</td>
<td>66.6±4</td>
<td>17.2±5</td>
<td>33.7±5</td>
<td>27.8±1</td>
<td>29.2±2</td>
</tr>
<tr>
<td>I3D [51]</td>
<td>Res50</td>
<td>ImgNet</td>
<td>64.9±4</td>
<td>17.0±2</td>
<td>29.9±1</td>
<td>27.4±3</td>
<td>26.6±5</td>
</tr>
<tr>
<td>NL-TSM [62]</td>
<td>Res50</td>
<td>ImgNet</td>
<td>58.6±4</td>
<td>16.5±2</td>
<td>21.8±2</td>
<td>25.9±6</td>
<td>21.7±2</td>
</tr>
<tr>
<td>NL-I3D [62]</td>
<td>Res50</td>
<td>ImgNet</td>
<td>64.9±4</td>
<td>16.2±2</td>
<td>30.0±4</td>
<td>27.0±2</td>
<td>26.6±1</td>
</tr>
<tr>
<td>NL-SlowOnly [62]</td>
<td>Res50</td>
<td>ImgNet</td>
<td>63.8±1</td>
<td>17.5±2</td>
<td>28.5±5</td>
<td>27.0±2</td>
<td>25.6±3</td>
</tr>
<tr>
<td>CSN [58]</td>
<td>Res50</td>
<td>-</td>
<td>65.9±3</td>
<td>17.9±2</td>
<td>31.6±2</td>
<td>28.2±2</td>
<td>27.6±5</td>
</tr>
<tr>
<td>TPN [63]</td>
<td>Res50</td>
<td>ImgNet</td>
<td>69.3±2</td>
<td>18.8±2</td>
<td>33.2±5</td>
<td>29.0±3</td>
<td>29.0±6</td>
</tr>
<tr>
<td>SlowOnly [15]</td>
<td>Res50</td>
<td>ImgNet</td>
<td>68.2±1</td>
<td>17.5±2</td>
<td>32.8±5</td>
<td>28.1±2</td>
<td>28.7±2</td>
</tr>
<tr>
<td>SlowFast-Fast [15]</td>
<td>Res50</td>
<td>-</td>
<td>68.4±3</td>
<td>18.0±3</td>
<td>33.7±6</td>
<td>28.8±2</td>
<td>29.7±5</td>
</tr>
<tr>
<td>SlowFast-Slow [15]</td>
<td>Res101</td>
<td>-</td>
<td>69.4±4</td>
<td>19.8±3</td>
<td>31.1±6</td>
<td>31.0±2</td>
<td>28.1±3</td>
</tr>
<tr>
<td>SlowFast-Fast [15]</td>
<td>Res101+50</td>
<td>-</td>
<td>67.9±2</td>
<td>17.5±3</td>
<td>31.9±4</td>
<td>28.4±2</td>
<td>29.0±2</td>
</tr>
<tr>
<td>SlowFast-Slow [15]</td>
<td>Res101</td>
<td>-</td>
<td>69.6±3</td>
<td>18.2±2</td>
<td>33.6±6</td>
<td>29.2±3</td>
<td>29.4±3</td>
</tr>
<tr>
<td>CSN [58]</td>
<td>Res152</td>
<td>-</td>
<td>67.8±3</td>
<td>20.4±5</td>
<td>30.1±3</td>
<td>30.8±2</td>
<td>26.3±3</td>
</tr>
<tr>
<td>SlowFast-Fast [15]</td>
<td>Res152+50</td>
<td>-</td>
<td>69.3±5</td>
<td>20.3±6</td>
<td>31.9±7</td>
<td>31.0±1</td>
<td>28.5±3</td>
</tr>
<tr>
<td>X3D [14]</td>
<td>X3D-S</td>
<td>-</td>
<td>60.8±3</td>
<td>13.9±3</td>
<td>26.7±7</td>
<td>24.2±2</td>
<td>24.7±3</td>
</tr>
<tr>
<td>X3D [14]</td>
<td>X3D-M</td>
<td>-</td>
<td>64.3±3</td>
<td>15.6±2</td>
<td>27.3±1</td>
<td>26.5±4</td>
<td>25.0±1</td>
</tr>
<tr>
<td>TANet [49]</td>
<td>TANet</td>
<td>ImgNet</td>
<td>67.1±3</td>
<td>18.3±3</td>
<td>30.5±4</td>
<td>28.5±2</td>
<td>27.0±3</td>
</tr>
</tbody>
</table>

| Large-scale dataset | | | | | | | |
| TSN [61] | Res50 | IG-1B [63] | 57.7±5 | 12.0±5 | 27.4±3 | 21.4±3 | 25.7±1 | 10.1±3 | 32.1±2 |
| Omni-TSN [12] | Res50 | IG-1B [63] | 63.9±6 | 13.8±4 | 30.7±1 | 24.4±1 | 27.9±2 | 11.8±2 | 36.8±6 |
| Omni-Slow [12] | Res50 | - | 69.5±3 | 18.0±6 | 34.4±5 | 29.1±2 | 29.8±2 | 15.0±2 | 40.8±2 |
| CSN [58] | Res50 | IG65M [16] | 70.4±3 | 22.1±5 | 32.7±2 | 32.4±4 | 28.9±4 | 18.8±1 | 38.7±2 |
| T5former [41] | T5former | Img21K [10] | 65.3±3 | 15.6±3 | 28.8±1 | 25.8±1 | 27.4±3 | 13.0±3 | 33.2±5 |
| Omni-Slow [12] | Res101 | - | 73.3±4 | 22.6±2 | 33.5±4 | 33.4±5 | 30.1±5 | 19.4±2 | 39.2±3 |
| CSN [58] | Res152 | IG65M [16] | 72.9±1 | 25.2±4 | 32.2±5 | 35.6±3 | 28.4±6 | 22.1±3 | 38.0±3 |

5 Conclusion

We introduce a general framework for human-centric analysis for human action recognition models. We test Human-centric Analysis Toolkit on the Kinetics-400 dataset and evaluate the generated dataset on a number of existing action recognition models.

Through extensive experiments over 74 trained models, we find that all the models we tested have stronger background bias. However, we found that the background bias can be mitigated when more frames are fed during the training, the temporal stride between frames is increased, and temporal/spatial modeling is improved using Non-local module. Moreover, we see that the original accuracy do not fully represent the human understanding as the accuracy cannot differentiate TSN and TSM, large-scale dataset and Non-local module improves original accuracy but not necessarily SHAcc.

From our findings, we suggest the future researchers to (1) not rely on the accuracy as the only metric, as original accuracy do not fully represent the performance of the model based on the human action; (2) carefully select the temporal hyper-parameters, as temporal parameters can improve/worsen the background bias of human action recognition models; and (3) use HAT toolkit to see if the model design (e.g., as Non-local) can improve your model on accuracy and reduce the background bias. We hope that this tool can be adopted by future researchers for a better human-centric analysis of human action recognition models.
Table 3: Performance comparison when using a Non-local module \[62\]. NL-EG, NL-G, and NL-Dot denote Non-local method using embedded Gaussian, Gaussian, and dot product, respectively. Numbers are bolded when the Non-local module improves the metric.

<table>
<thead>
<tr>
<th>Model</th>
<th>frames</th>
<th>OA++</th>
<th>SHAcc</th>
<th>SBErr</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSM</td>
<td>8</td>
<td>72.89</td>
<td>16.55</td>
<td>22.64</td>
</tr>
<tr>
<td>TSM + NL-EG</td>
<td>8</td>
<td>74.06</td>
<td>16.54</td>
<td>21.77</td>
</tr>
<tr>
<td>TSM + NL-G</td>
<td>8</td>
<td>72.61</td>
<td>17.52</td>
<td>20.07</td>
</tr>
<tr>
<td>TSM + NL-Dot</td>
<td>8</td>
<td>73.52</td>
<td>17.27</td>
<td>20.91</td>
</tr>
<tr>
<td>I3D</td>
<td>32</td>
<td>75.33</td>
<td>17.05</td>
<td>31.78</td>
</tr>
<tr>
<td>I3D + NL-EG</td>
<td>32</td>
<td>76.90</td>
<td>16.23</td>
<td>30.04</td>
</tr>
<tr>
<td>I3D + NL-G</td>
<td>32</td>
<td>75.96</td>
<td>17.22</td>
<td>30.94</td>
</tr>
<tr>
<td>I3D + NL-Dot</td>
<td>32</td>
<td>76.17</td>
<td>15.63</td>
<td>30.14</td>
</tr>
<tr>
<td>SlowOnly</td>
<td>4</td>
<td>75.28</td>
<td>14.75</td>
<td>33.27</td>
</tr>
<tr>
<td>SlowOnly + NL-EG</td>
<td>4</td>
<td>76.10</td>
<td>15.46</td>
<td>30.21</td>
</tr>
<tr>
<td>SlowOnly</td>
<td>8</td>
<td>75.18</td>
<td>16.12</td>
<td>31.49</td>
</tr>
<tr>
<td>SlowOnly + NL-EG</td>
<td>8</td>
<td>77.74</td>
<td>17.54</td>
<td>28.48</td>
</tr>
</tbody>
</table>

Table 4: Performance when using a large-scale dataset. We compare the same settings except for the initial weight. Numbers are bolded when the large-scale dataset improves the metric.

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>Pre-trained</th>
<th>OA++</th>
<th>SHAcc</th>
<th>SBErr</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSN [61]</td>
<td>ResNet50</td>
<td>ImageNet</td>
<td>72.55</td>
<td>11.34</td>
<td>28.62</td>
</tr>
<tr>
<td>TSN [61]</td>
<td>ResNet50</td>
<td>IG-1B</td>
<td>73.39</td>
<td>11.96</td>
<td>27.45</td>
</tr>
<tr>
<td>ir-CSN [58]</td>
<td>ResNet50</td>
<td>None</td>
<td>75.51</td>
<td>17.88</td>
<td>31.58</td>
</tr>
<tr>
<td>ir-CSN [58]</td>
<td>ResNet50</td>
<td>IG65M</td>
<td>81.46</td>
<td>22.05</td>
<td>32.68</td>
</tr>
<tr>
<td>ir-CSN [58]</td>
<td>ResNet152</td>
<td>None</td>
<td>78.08</td>
<td>19.51</td>
<td>30.76</td>
</tr>
<tr>
<td>ir-CSN [58]</td>
<td>ResNet152</td>
<td>Sports1M</td>
<td>78.98</td>
<td>20.52</td>
<td>31.14</td>
</tr>
<tr>
<td>ir-CSN [58]</td>
<td>ResNet152</td>
<td>IG65M</td>
<td>83.17</td>
<td>25.25</td>
<td>32.07</td>
</tr>
<tr>
<td>ip-CSN [58]</td>
<td>ResNet152</td>
<td>None</td>
<td>79.26</td>
<td>20.37</td>
<td>30.11</td>
</tr>
<tr>
<td>ip-CSN [58]</td>
<td>ResNet152</td>
<td>Sports1M</td>
<td>79.38</td>
<td>20.37</td>
<td>32.06</td>
</tr>
<tr>
<td>ip-CSN [58]</td>
<td>ResNet152</td>
<td>IG65M</td>
<td>83.92</td>
<td>25.19</td>
<td>32.16</td>
</tr>
</tbody>
</table>

6 Discussion

Limitation As we use an off-the-shelf image semantic segmentation model and a video inpainting model, the quality of the synthetic dataset is limited by the performance of the aforementioned models.

Ethical Concerns Our tool requires the use of image segmentation and inpainting tool to generate a dataset, requiring computation cost for the initial setup. However, as human-centric analysis using our tool does not require any new training, we believe our tool is more environmentally friendly than the existing methods. Moreover, as our tool is automated, human labor for data collection is not required. Also, as we generate a dataset from an existing dataset, we show fewer concerns about privacy issues when a new video dataset is generated.

License MMAction2 \[8\] and SeMask \[24\] follow Apache License 2.0. We used author-released code for Deep Video Inpainting \[29\] which did not specify any license. Kinetics-400 annotation data is licensed under a Creative Commons Attribution 4.0 International License, but some of the video sources do not specify any license. Please refer to the individual licenses when using our released code.

Acknowledgements. We are grateful for the support from the National Science Foundation under Grant No. 2112562, Microsoft, Princeton SEAS Project X Innovation Fund, and Princeton First Year Ph.D. Fellowship to JC.
References


[34] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In ECCV, 2018.


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See appendix.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.1

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 6
(b) Did you mention the license of the assets? [Yes] See Section 6

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]