Lightweight Probability-Level Fusion for Occlusion-Robust Human Activity Recognition from RGB and Skeleton Data

Marcus Guo[1]

Received April 13, 2026
Accepted July 17, 2026
Electronic access August 15, 2026

Human activity recognition (HAR) models often degrade when people or body parts are occluded. This study evaluates whether a lightweight probability-level fusion of RGB and skeleton predictions can improve action recognition under synthetic occlusion. We use frozen pretrained C3D and PoseC3D expert streams on UCF-101 split 1. Each stream produces a video-level 101-dimensional class-probability vector, and fusion is performed only over these cached outputs. We compare RGB-only, skeleton-only, non-learned averaging, and learned linear fusion across 0%, 10%, 25%, 40%, 50%, and 70% occlusion. Two protocols are tested. In the train-clean/test-occluded protocol, the fusion head is trained only on 0% occlusion outputs. Here, non-learned fusion is most reliable, with average-log-probability fusion performing best from 0% to 50% occlusion. Average-probability fusion is the most reliable at 70% occlusion. In the same-occlusion protocol, the fusion head is trained and tested at the same occlusion level. Learned log-probability fusion becomes the strongest here, reaching 79.12% top-1 accuracy at 40% occlusion and 72.46% at 50% occlusion with single-modal baselines 62% and 52% respectively. These results show that probability-level RGB-skeleton fusion can improve robustness under synthetic occlusion, but learned fusion works best when trained under matched occlusion conditions.

Keywords: human activity recognition; occlusion robustness; synthetic occlusion; RGB-skeleton fusion; late fusion; probability-level fusion; PoseC3D; UCF-101

Introduction

One compelling part of smart spaces is human activity recognition (HAR) models. HAR models are increasingly deployed in smart spaces, physical environments with sensors or computing that can perceive activity and context through cameras, microphones, or IoT devices and respond in real time. HAR models are designed to classify a human's action by analyzing data from cameras. Two common HAR models are RGB video and skeleton-based pose representations. RGB models can accurately use scene context and human activity to predict the action, but their performance can degrade when parts of the frame become occluded. Skeleton-based models represent human motion as a sequence of joint coordinates or pose heatmaps, but they depend on accurate pose estimation.

A major challenge for both approaches is occlusion. In real environments, people may be partially hidden by furniture, objects, camera angles, other people, or the edge of the frame. RGB models can lose visual evidence when body parts or objects are masked. Skeleton models can also fail because occlusion often causes missing or inaccurate joint estimates. Since RGB and skeleton streams fail in different ways, combining them may improve robustness when one stream loses useful information.

In this work, we combine two independently trained streams as fixed experts: PoseC3D and a C3D. Occlusion is simulated through random joint drops, and random masking of the RGB frame. Each stream produces a 101-dimensional class-probability vector from each occluded input video. Instead of designing a computationally expensive cross-modal fusion, we apply late fusion over probability vectors and log-probability vectors. This allows us to test whether complementary evidence from RGB and skeleton streams can improve robustness under controlled synthetic occlusion.

This study asks: under controlled synthetic occlusion, does late fusion of RGB and skeleton class-probability outputs improve video-level human activity recognition compared with single-modality baselines and simple non-learned fusion rules? We hypothesize that RGB and skeleton streams will degrade differently under occlusion because RGB models retain appearance and scene context, while skeleton models emphasize body motion but depend on reliable pose estimates. Therefore, a lightweight fusion head trained on the two streams' class-probability distributions may recover complementary evidence that is unavailable to either stream alone.

This paper makes three contributions. First, it evaluates pretrained RGB and skeleton action-recognition streams un-

[1] *Woodside Priory High School, CA, USA*

der controlled synthetic occlusion on UCF-101 split 1. Second, it compares lightweight probability-level fusion against RGB-only, skeleton-only, average-probability, average-log-probability, and matched-capacity single-stream baselines. Third, it analyzes robustness across multiple occlusion levels rather than reporting only one heavy-occlusion setting.

Related Work

Human activity recognition has been studied using RGB video, skeleton representations, and multimodal combinations of visual and pose-based information. This section reviews three areas most relevant to this study: skeleton-based action recognition, RGB-skeleton multimodal fusion, and occlusion-robust HAR. The goal is to position this work as a controlled evaluation of lightweight probability-level late fusion under synthetic occlusion, rather than as a new backbone architecture.

RGB and Skeleton-Based Action Recognition

RGB action-recognition models classify actions from appearance and motion in video frames. An early design is the two-stream network proposed by Simonyan and Zisserman, which trains separate spatial and temporal networks and combines their class scores using late fusion [1]. C3D is an early 3D convolutional model that learns spatiotemporal features by applying 3D convolutional filters over consecutive frames [2]. Because C3D processes both spatial and temporal dimensions, it provides a useful RGB expert for testing how visual action recognition changes when parts of the frame are masked. Later 3D-CNN backbones improved on C3D. I3D inflates pretrained 2D image-classification filters into 3D and benefits from large-scale pretraining on Kinetics [3], while SlowFast networks process video at two temporal rates to capture slow semantic content and fast motion [4]. The SlowOnly pathway from this family also serves as the backbone of the PoseC3D skeleton expert used in this study [4,5]. C3D is kept as the RGB expert here because it is a standard pretrained UCF-101 checkpoint and this study focuses on fusion behavior rather than backbone strength.

Skeleton-based action recognition represents human motion through joint trajectories or pose heatmaps rather than raw pixels. Progress in skeleton-based recognition has been supported by large-scale benchmarks such as NTU RGB+D, which provides synchronized RGB, depth, and 3D skeleton data for more than 56,000 clips across 60 action classes [6]. ST-GCN introduced graph convolution over body joints and temporal connections, making graph-based skeleton modeling a standard approach for skeleton HAR [7]. Later methods such as 2s-AGCN and CTR-GCN improved graph modeling by learning adaptive or channel-wise graph structures [8,9]. All of these skeleton pipelines depend on an upstream pose estimator such as HRNet, which maintains high-resolution feature maps throughout the network to produce accurate 2D keypoints [10]. Errors in pose estimation therefore propagate directly into skeleton-based recognition, which is one reason occlusion is especially harmful for this modality. PoseC3D takes a different approach by converting pose sequences into stacked 3D heatmaps and processing them with a 3D CNN [5]. Because this study uses PoseC3D as the frozen skeleton expert, it evaluates how PoseC3D class-probability outputs behave under synthetic keypoint dropout rather than proposing a new skeleton backbone.

Occlusion-Robust Skeleton Recognition

Occlusion is a major challenge for HAR because it removes or corrupts action-relevant evidence. For RGB inputs, synthetic masking methods such as Random Erasing and Cutout approximate occlusion by removing rectangular image regions [11,12]. These methods support rectangular masking as a controlled corruption strategy, but rectangular masks are still only a simplified approximation of real occlusion.

Skeleton-based methods have also studied missing or corrupted body information. Dual Inhibition Training reduces overreliance on specific body parts for occluded skeleton recognition [13]. RA-GCN takes a related approach for incomplete skeleton data: it trains multiple graph-convolutional streams that are encouraged to activate different, complementary groups of joints, so recognition degrades more gracefully when some joints are missing or occluded [14]. Peng et al. studied one-shot skeleton recognition, where a model must recognize action classes from only one labeled skeleton example, under both random joint occlusions and object-like occlusions generated by projecting 3D furniture models into skeleton scenes [15]. Other studies simulate skeleton occlusion by removing body parts from skeleton representations, reconstructing missing body parts before recognition, or designing realistic data augmentation for noisy pose estimates [16–18]. These studies show the importance of testing skeleton recognition under missing body evidence, but they mainly focus on skeleton-only recognition rather than probability-level RGB-skeleton fusion.

RGB-Skeleton Multimodal Fusion

RGB and skeleton streams provide complementary cues. RGB models capture appearance, objects, and scene context, while skeleton models emphasize body pose and motion. Prior multimodal HAR methods combine these cues through early fusion, feature-level fusion, attention-based fusion, or late fusion. The question of where to fuse separate video streams predates RGB-skeleton models. Feichtenhofer et al. systemat-

ically compared spatial and temporal fusion locations in two-stream networks and showed that the choice of fusion point has a large effect on accuracy[19]. For example, Baradel et al. used pose-driven attention to guide RGB-based action recognition[20]. VPN learns a joint video-pose embedding in which pose features spatially attend to RGB features, improving recognition of activities of daily living[21]. MMFF combines skeleton sequences with RGB frames using skeleton-guided attention and cross-attention[22]. BPAN uses bilinear pooling and attention to combine RGB and skeleton features[23]. VT-BPAN uses transformer-based feature fusion for RGB and skeleton action recognition[24].

This study is narrower and lighter-weight than those methods. Instead of designing a new feature-level or attention-based architecture, it freezes both expert streams and combines only their video-level class-probability vectors. This design makes the fusion model much smaller and easier to analyze, but also less expressive. Therefore, this paper compares learned fusion against simple averaging, matched-capacity single-stream controls, and shuffled-skeleton controls to test whether gains come from complementary RGB-skeleton information rather than extra parameters or dataset-level bias.

Position of This Work

This study takes a different direction. Instead of designing a new attention or transformer architecture, it asks how far a simple late-fusion strategy can go when the base RGB and skeleton models are frozen. The fusion head operates only on video-level class-probability vectors produced by pretrained C3D and PoseC3D streams. Combining classifier probability outputs also has a long history outside deep learning. Kittler et al. formalized common combination rules such as the sum rule and the product rule and analyzed how they behave under estimation errors[25]. The average-probability and average-log-probability baselines in this study correspond to the sum rule and the product rule respectively, so the non-learned baselines used here are grounded in classical classifier-combination theory. This makes the method much lighter than feature-level cross-modal models, but also less expressive. For that reason, this paper compares learned fusion against simple averaging and matched-capacity single-stream controls to test whether the gain comes from complementary multimodal information rather than from extra trainable parameters alone.

Methods

Dataset: UCF-101 with Occlusion

We use UCF-101, a standard benchmark for video action recognition with 101 action classes. In the official split, there are 9537 training videos and 3783 videos used for model evaluation[26]. We generate cached RGB and skeleton predictions at each occlusion level for both training and evaluation splits. The pretrained expert streams remain frozen; only the fusion head is trained on cached outputs. We apply synthetic occlusion during inference for both streams at 0%, 10%, 25%, 40%, 50%, and 70%. For the RGB stream, we apply a rectangular mask covering a specific fraction of the frame. For the skeleton stream, we apply keypoint dropout after pose extraction and before heatmap generation.

RGB Video Stream: C3D 3D-CNN Model

For the RGB stream, we use a pretrained 3D convolutional action recognizer, C3D, which learns spatiotemporal filters over consecutive video frames[2]. Unlike 2D CNNs that operate only over spatial width and height, 3D CNNs add a temporal dimension, allowing the model to learn motion patterns across frames. The RGB expert uses an MMAction2 C3D checkpoint pretrained on Sports-1M and fine-tuned on UCF-101[27]. Each RGB input is processed as a 16-frame clip, resized to 128 $\times$ 171, cropped to 112 $\times$ 112, and normalized. Synthetic RGB occlusion is applied during cached expert-output generation, not during C3D backbone training in this study. A random rectangular region of each sampled clip is erased to simulate partial visual occlusion.

Architecture

The C3D is composed of five convolutional stages, two fully connected layers, and a final softmax classifier. All layers use 3x3x3 kernels and pooling operations to progressively reduce spatial and temporal resolution. Then, the model outputs a probability distribution over 101 action classes for UCF-101.

Skeleton Pose Stream: PoseC3D Model

For the skeleton-based model, we used PoseC3D. Rather than modeling joints as a graph like previous skeleton-based models, PoseC3D converts pose sequences into a sequence of heatmaps. First, a pose estimator extracts the human joint coordinates from each frame. Then, each joint is represented as a Gaussian blob on a 2D grid (heat map). A 3D CNN then processes these spatiotemporal features, as these heat maps are stacked over time. This representation retains spatial joint layout and temporal motion patterns[5]. Each video contains 17 keypoints represented on a 56 $\times$ 56 spatial grid over 48 frames, yielding a spatiotemporal tensor of size 17 $\times$ 48 $\times$ 56 $\times$ 56. This is a spatio-temporal representation of skeletal motion. Skeletons were extracted from the original UCF-101 pose annotations and then artificially corrupted before heatmap generation; they were not re-extracted from RGB frames after masking. The RGB and skeleton streams are treated as frozen expert models. We do not fine-tune either expert during this study.

Table 1 Related Works

Work	Main Modality	Fusion Type	Occlusion Handling	Difference from this study
ST-GCN[7]	Skeleton	None	Not primary focus	No multimodal fusion
2s-AGCN[8]	Skeleton	None	Not primary focus	No multimodal fusion
PoseC3D[5]	Skeleton / RGB-pose possible	Heatmap-based pose representation	Heatmap-based skeleton representation	Used here as a frozen skeleton expert
Dual Inhibition Training[13]	Skeleton	None	Input and prediction inhibition	Occlusion-focused, but no multimodal fusion
Trans4SOAR[15]	Skeleton	Transformer / mixed attention	Designed for diverse skeleton occlusions	Strong architecture, but not multimodal fusion
MMFF[22]	RGB frame + skeleton sequence	Early attention + cross-attention	Not primary focus	Uses advanced fusion, no occlusion testing included
BPAN[23]	RGB + skeleton	Bilinear pooling + attention	Not primary focus	Occlusion not tested
VT-BPAN[24]	RGB + skeleton	Transformer + bilinear attention	Not primary focus	Strong architecture, high complexity, but not occlusion tested
Mathe et al.[16]	Skeleton	None	Artificial body-part occlusion	No multimodal fusion, but supports synthetic skeleton occlusion
Vernikos and Spyrou[17]	Skeleton	Reconstruction before recognition	Missing body-part reconstruction	Completes missing skeletons rather than fusing different streams
This Study	RGB + skeleton	Probability-level late fusion	RGB masking and skeleton keypoint dropout across multiple severities	Tests whether a lightweight late fusion head can improve robustness over fixed expert streams and whether it performs better under occlusion

Architecture

The heatmap is fed into a SlowOnly ResNet-50 3D-CNN backbone. Although the backbone architecture is similar to video-recognition models used for RGB, it operates over pose heatmaps rather than RGB frames. This is a standard 3D ResNet backbone that processes pose heatmaps over time. This network was trained on Kinetics-400[28], then fine-tuned on UCF-101 skeleton data using the MMAction2 PoseC3D checkpoint (split 1)[27]. The PoseC3D model has substantially fewer parameters than the RGB backbone, which is one efficiency advantage of skeleton-based approaches. The PoseC3D checkpoint produces a 101-dimensional pred-score vector for UCF-101. In the cached MMAction2 outputs used for fusion, this vector is probability-like rather than a pre-softmax logit vector.

Synthetic Occlusion Protocol

We apply synthetic occlusion separately to the RGB and skeleton streams. For the RGB stream, a rectangular mask is applied to each sampled video clip. The mask covers a specified fraction of the frame area and is kept fixed across the clip to approximate a persistent visual obstruction. For the skeleton stream, occlusion is applied after pose extraction and before heatmap generation. At occlusion level p, each keypoint observation is independently removed with probability p by setting both its coordinate and confidence score to zero. No coordinate jitter is used in the main severity-curve experiment. This design isolates the effect of missing skeletal evidence rather than mixing occlusion with pose-estimation noise.

We evaluate occlusion levels of 0%, 10%, 25%, 40%, 50%, and 70%. These occlusions are synthetic and should not be interpreted as complete simulations of real-world occlusion. Rectangular RGB masks may hide background instead of the actor, and independent keypoint dropout does not fully represent object-based or multi-person occlusion. The purpose of this protocol is to provide a controlled stress test for comparing single-modality and fusion models.

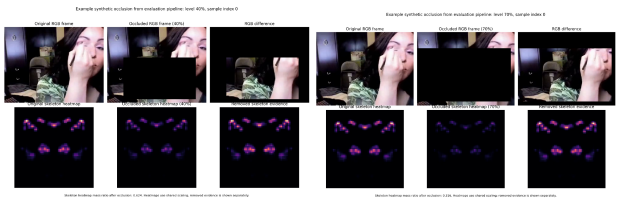

Figure. 1 Example synthetic occlusion from the evaluation pipeline. The RGB stream is occluded using a rectangular mask. The skeleton stream is corrupted after pose extraction and before heatmap generation by setting selected keypoint coordinates and confidence scores to zero. The first image shows 40% occlusion, the second image shows 70% occlusion.

Experiments

For each video, the frozen RGB expert produces a class-probability vector of dimension 101, and the frozen skeleton expert also produces a vector of dimension 101. Each vector contains one score for each UCF-101 action class. We evaluate several probability-level fusion methods. We report video-level top-1 and top-5 accuracy. Top-1 accuracy measures whether the highest-scoring predicted class is correct, while top-5 accuracy measures whether the true class appears among the five highest-scoring classes.

For each occlusion level, we evaluate RGB-only, skeleton-only, average-probability late fusion, average-log-probability

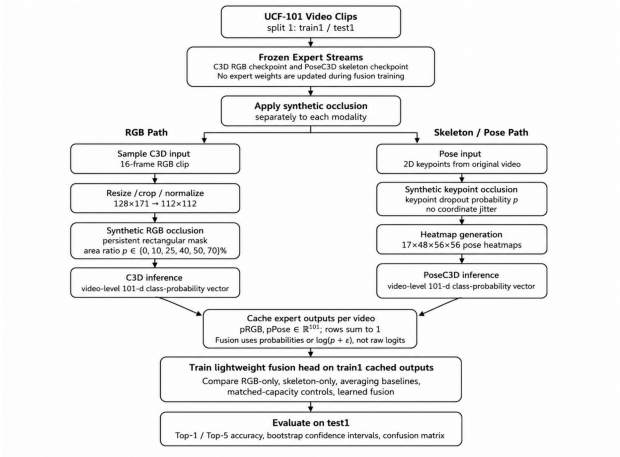

Figure. 2 Overview of the occlusion-aware multimodal fusion pipeline steps.

fusion, learned probability fusion, and learned log-probability fusion. These vectors are nonnegative and sum to approximately 1, so they are treated as class-probability vectors rather than raw pre-softmax logits. We therefore evaluate fusion over probability vectors and log-probability vectors.

Average Probability is defined as

$$p_i^{\text{avg}} = \frac{1}{2}\left(p_i^{\text{RGB}} + p_i^{\text{pose}}\right). \tag{1}$$

Average Log Probability is defined as

$$s_i^{\text{avglog}} = \log\left(p_i^{\text{RGB}} + \varepsilon\right) + \log\left(p_i^{\text{pose}} + \varepsilon\right), \tag{2}$$

For probability linear fusion, the model receives the concatenated vector

$$x_i = \left[p_i^{\text{RGB}} \;\; p_i^{\text{pose}}\right] \in R^{202}. \tag{3}$$

For log-probability linear fusion, the model receives the concatenated vector

$$x_i = \left[\log\left(p_i^{\text{RGB}} + \varepsilon\right); \log\left(p_i^{\text{pose}} + \varepsilon\right)\right] \in R^{202} \tag{4}$$

where ε (epsilon) is a small constant for numerical stability. A linear fusion head computes

$z_i = Wx_i + b, \quad z_i \in R^{101}, \quad W \in R^{101\times 202}, \quad x_i \in R^{202}, \quad b \in R^{101}.$

The final prediction is obtained by applying softmax to z_i. The learned linear fusion head contains approximately $101 \times 202 + 101 = 20{,}503$ trainable parameters. It is lightweight relative to the frozen C3D and PoseC3D expert backbones.

During fusion training, the RGB and skeleton-based models are frozen, and we first run each pretrained model and cache their outputs. The fusion head produces a 101-dimensional fused vector, which is converted to class probabilities with a softmax layer and trained using cross-entropy loss. We optimize the fusion head using Adam, an adaptive gradient-based optimizer that maintains running estimates of the first and second moments of the gradients, using a learning rate of 1×10^{-3}, weight decay of 1×10^{-4}, and a batch size of 256 and 20 epochs. Although this fusion head is much simpler than other methods that combine modalities, it introduces few extra parameters.

We include control experiments to test whether improvements from fusion are really caused by complementary RGB-skeleton information. We include matched-capacity RGB-only and skeleton-only controls. These controls use the same linear-head input size and parameter count, but receive duplicated single-modality inputs instead of both modalities. If true fusion outperforms the matched-capacity controls, the gain is less likely to be explained only by extra trainable parameters. We include shuffled-skeleton controls. In the shuffled-test control, the model is trained normally with correctly paired RGB and skeleton predictions, but skeleton predictions are randomly permuted only during evaluation. In the shuffled-train control, skeleton predictions are randomly permuted during fusion-head training, so the model learns from mismatched RGB-skeleton pairs.

We evaluate two fusion protocols. In the same-occlusion protocol, the fusion head is trained and evaluated using cached outputs from the same occlusion level. This tests how well a lightweight fusion head can adapt when the corruption level is known during training. In the train-clean/test-occluded protocol, the fusion head is trained only on 0% occlusion train outputs and evaluated across all test occlusion levels. This is the stricter robustness protocol because the fusion head cannot adapt separately to each occlusion severity. Average probability and average log probability fusion are non-learned baselines. They do not use the training split, and are applied directly to the RGB and skeleton probability vectors. Therefore, their values are identical in the train-clean/test-occluded and same-occlusion protocols.

We also compute 95% bootstrap confidence intervals over test videos. These confidence intervals estimate uncertainty from the finite UCF-101 test set, but they do not estimate variation across different random occlusion seeds.

Finally, we perform a conditional recovery analysis to examine whether fusion corrects different types of single-stream failures. Each video is assigned to one of four conditions: RGB wrong/Skeleton wrong, RGB right/Skeleton wrong, RGB wrong/Skeleton right, RGB right/Skeleton right. For each condition, we report how often the fused model produces the correct class. This analysis helps us distinguish whether fusion recovers cases where skeleton information corrects RGB failures, whether RGB information corrects skele-

ton failures, or whether fusion recovers cases where both single-stream top-1 predictions were wrong.

Table 2 Implementation and Reproducibility

Component	Value
Dataset	UCF-101 Split 1
Train / Test videos	9537 / 3783
RGB Expert	C3D Sports-1M → UCF-101
Skeleton Expert	PoseC3D SlowOnly R50 Kinetics-400 → UCF-101
Framework	MMAction2 1.2.0
PyTorch	2.3.1 + cpu
Hardware	AWS EC2 m4.2xlarge, 8 vCPUs, 31 GiB RAM, no CUDA
Output	video-level pred_score, probability-like vector

We used UCF-101 split 1, consisting of 9,537 training videos and 3,783 test videos. RGB predictions were generated using the MMAction2 C3D Sports-1M-pretrained UCF-101 checkpoint, and skeleton predictions were generated using the MMAction2 PoseC3D SlowOnly R50 UCF-101 split-1 checkpoint. The SHA-256 hashes of the checkpoints were recorded for reproducibility. Cached MMAction2 pred_score outputs were video-level 101-dimensional class-probability vectors rather than raw logits, as the scores were nonnegative and summed to approximately 1 across classes. Fusion was therefore evaluated over probability vectors and log-probability vectors.

Results

This section evaluates whether probability-level RGB-skeleton fusion improves action recognition under controlled synthetic occlusion. Results are reported on UCF-101 split 1 using top-1 and top-5 accuracy. We first report the no-occlusion baseline, then evaluate performance across the two occlusion protocols: train-clean/test-occluded and same-occlusion training. Finally, we analyze control experiments, how information is recovered, and per-class gains and failures.

Table 3 No Occlusion Baseline. No-occlusion video level baseline results on UCF-101 split 1

Method	Top-1	95% CI	Top-5	95% CI
RGB Only	0.8308	[0.8189, 0.8430]	0.9593	[0.9527, 0.9656]
Skeleton Only	0.8681	[0.8570, 0.8789]	0.9715	[0.9662, 0.9765]
Average Probabilities	0.9262	[0.9180, 0.9350]	0.9855	[0.9815, 0.9892]
Average log-probabilities	0.9408	[0.9334, 0.9482]	0.9900	[0.9868, 0.9931]
Linear fusion probabilities	0.9241	[0.9162, 0.9321]	0.9812	[0.9767, 0.9855]
Linear fusion log-probabilities	0.8964	[0.8866, 0.9062]	0.9767	[0.9720, 0.9812]

Table 3 reports the no-occlusion baseline results. Without any occlusion, both single-stream models perform strongly. RGB-only achieves 83.08% top-1 accuracy, while skeleton-only achieves 86.81% top-1 accuracy. Combining the two modalities in any way improves performance beyond either individual stream. These results, especially the average probabilities, show that the RGB and skeleton experts contain complementary information before occlusion is applied. However, the learned linear fusion heads do not outperform average-log-probability fusion in the clean setting. This suggests that a simple non-learned score combination is already a strong baseline and should be tested as an important comparison for other learned linear fusion.

Table 4 Protocol 1, Train-clean/Test-occluded Severity Curve. The fusion head is trained on 0% occlusion train outputs and evaluated across synthetic occlusion levels.

Test Occlusion	RGB only	Skeleton only	Avg prob	Avg log-prob	Linear prob	Linear log-prob
0%	0.8308	0.8681	0.9262	0.9408	0.9241	0.8964
10%	0.7629	0.8454	0.9059	0.9231	0.9033	0.8636
25%	0.5948	0.7436	0.8208	0.8538	0.8197	0.7608
40%	0.3682	0.6238	0.6944	0.7412	0.6849	0.5982
50%	0.2580	0.5271	0.5784	0.6106	0.5726	0.4544
70%	0.1084	0.3164	0.3323	0.3154	0.3230	0.2313

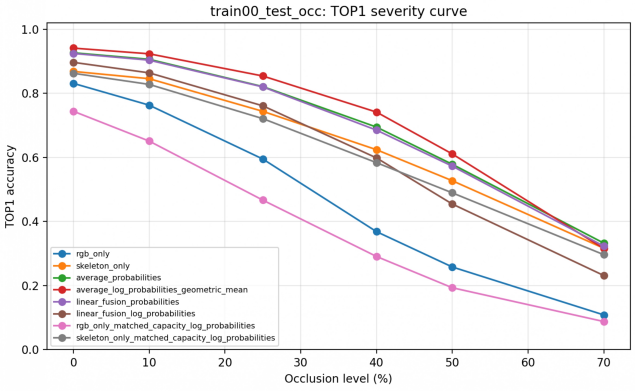

Figure. 3 Top-1 Severity Curve Across Methods.

Table 4 and Figure 3 show the train-clean/test-occluded protocol. Fusion heads are trained on 0% occlusion outputs and then evaluated across all occlusion levels. This tests whether a fusion head trained on clean data can generalize to increasingly corrupted outputs. As occlusion increases, RGB-only and skeleton-only performance declines sharply. However, RGB-only decreases to just 10% at 70% occlusion, while skeleton-only degrades to 31%. This supports the hypothesis that RGB and skeleton streams degrade differently under occlusion.

Under this stricter train-clean/test-occluded protocol, non-learned fusion is the strongest approach. Average-log-probability fusion is best from 0% through 50% occlusion, reaching 74.12% top-1 accuracy at 40% occlusion, and 61.06% at 50% occlusion. Learned linear fusion stays close to average log-probabilities throughout the occlusion levels, while learned linear log-probability fusion drops to just

23.13% top-1 at 70% occlusion. This indicates that a learned fusion head trained only on clean data does not automatically generalize to severe occlusion.

Table 5 reports the train-occluded/test-occluded protocol, where a separate fusion head is trained and evaluated at each occlusion level. This tests whether a lightweight learned fusion can adapt when corruption level is known during training. The train-occluded/test-occluded results show a different pattern compared to protocol 1. At low-occlusion, average log-probability fusion remains the strongest. However, as occlusion increases, learned log-probability fusion reaches 79% at 40% occlusion, compared to 74% for average log-probabilities. At 70% occlusion, learned log-probability fusion reaches 53% in contrast to 33% for average probabilities. These results suggest that learned fusion is most useful when trained under matched occlusion conditions. However, because this protocol gives the model access to occlusion-matched training outputs, it is less strict than train-clean/test-occluded evaluation.

Table 6 reports matched-capacity and shuffled-skeleton controls for the same-occlusion learned log-probability fusion model. The matched-capacity controls test whether the gain from fusion is caused only by additional parameters. The shuffled controls test whether performance depends on correctly paired RGB and skeleton predictions.

True RGB-skeleton fusion outperforms the RGB matched-capacity control at every occlusion level. True fusion also outperforms the skeleton matched-capacity control at most occlusion levels, although the gap becomes increasingly small at 70% occlusion. This suggests that under extreme occlusion, the model relies more heavily on skeleton information. The shuffled controls show that correct RGB-skeleton pairing matters, dropping to very low levels at 70% occlusion. This large decrease suggests that the fusion head is not simply exploiting dataset-level class priors.

Tables 7 and 8 analyze fusion behavior by separating test videos according to whether the RGB-only and skeleton-only predictions were correct. This analysis helps identify whether a fusion method is using complementary information from one stream when the other stream fails, or whether it is preserving already-correct predictions.

Table 7 reports the same-occlusion learned log-probability fusion model. At 40% occlusion, learned log-probability fusion correctly classifies 1330/1479 of "RGB wrong, skeleton right". At 50%, it correctly classifies 1321/1499 of these videos. Even at 70% occlusion, it correctly classifies 841/1080 of these videos. At 40% occlusion, learned log-probability fusion also correctly classifies 424/512 of RGB right, skeleton wrong videos. At 50% occlusion and 70% occlusion, the model correctly classifies 375/481 and 201/293 of RGB right, skeleton wrong videos. This suggests that the learned fusion head often uses skeleton information or RGB information to correct the other model's wrong inference. The learned fusion model also recovers many cases where both single stream predictions were wrong. At 70% occlusion, the learned log-probability fusion correctly classifies 881/2293 RGB-wrong/skeleton-wrong cases. This is much higher than average-log-probability fusion at just 6.8%. This suggests that same-occlusion learned fusion is not simply copying the top-1 prediction of either expert. Rather, it uses lower-ranked classes as evidence from the RGB and skeleton predictions to predict the final class scores.

Table 8 shows why average-log-probability fusion remains strong. It's more conservative than learned fusion when one or both streams are already correct. For example, when both RGB and skeleton are correct, average-log-probability fusion preserves the correct answer in 100.0% of cases at 40%, 50%, and 70% occlusion. In contrast, learned log-probability fusion preserves only 98.0%, 96.8%, and 79.5% of those cases in Table 7. This shows that the learned fusion head can sometimes override correct single-stream predictions, especially under 70% occlusion.

Overall, Tables 7 and 8 show the tradeoff between learned fusion and non-learned average. Average-log-probability fusion is safer when one or both streams are already correct. Same-occlusion learned log-probability fusion is more aggressive as it sometimes breaks correct single-stream predictions. However, it recovers far more hard cases where both RGB and skeleton top-1 predictions are wrong. This explains why learned log-probability fusion becomes strongest under same-occlusion training, while average-log-probability fusion remains the more reliable method under clean-trained evaluation.

Table 9 reports per-class examples at 70% occlusion. Several motion-heavy classes improve under same-occlusion learned log-probability fusion. For example, Horse Riding improves from 8.16% RGB-only accuracy and 6.12% skeleton-only accuracy to 97.96% under learned fusion. Soccer Penalty, Salsa Spin, Soccer Juggling, and Tennis-related motion classes also show large gains.

However, Table 8 also shows that learned fusion can hurt some classes where the skeleton stream alone is strong. Typing falls from 79.07% skeleton-only accuracy to 30.23% under learned fusion. Playing Daf, Hula Hoop, Breast Stroke, and Playing Tabla also show negative gains relative to the best single stream. These failures suggest that the fusion head can overcorrect class scores under severe occlusion, especially when one modality already provides a reliable prediction.

Overall, the results show that probability-level RGB-skeleton fusion improves robustness under controlled synthetic occlusion, but the best fusion method depends on the evaluation protocol. When the fusion head is trained on only clean outputs, simple average-probability and average-log-probability fusion are more reliable. When the fusion head

Table 5 Protocol 2, Train-occluded/Test-occluded Severity Curve

Train/Test Occlusion	RGB only	Skeleton Only	Avg prob	Avg log-prob	Linear Prob	Linear log-prob
0%	0.8308	0.8681	0.9262	0.9408	0.9241	0.8964
10%	0.7629	0.8454	0.9059	0.9231	0.9062	0.8937
25%	0.5948	0.7436	0.8208	0.8538	0.8438	0.8559
40%	0.3682	0.6238	0.6944	0.7412	0.7135	0.7912
50%	0.2580	0.5271	0.5784	0.6106	0.5895	0.7246
70%	0.1084	0.3164	0.3323	0.3154	0.3288	0.5329

Table 6 Control Experiments

Occlusion	True Linear log Fusion	RGB matched capacity	Skeleton matched capacity	Shuffled skeleton test	Shuffled skeleton train
0%	0.8964	0.7439	0.8623	0.3814	0.6862
10%	0.8937	0.6936	0.8417	0.3521	0.6479
25%	0.8559	0.5897	0.7917	0.2448	0.5350
40%	0.7912	0.4526	0.7251	0.1353	0.3960
50%	0.7246	0.3283	0.6714	0.0806	0.2855
70%	0.5329	0.1406	0.5297	0.0320	0.1049

Table 7 Conditional Recovery Analysis for Protocol 2 learned log-prob fusion

Occlusion	RGB wrong, Skeleton wrong	RGB wrong, Skeleton right	RGB right, Skeleton wrong	RGB right, Skeleton right
40%	376/911 (41.3%)	1330/1479 (89.9%)	424/512 (82.8%)	863/881 (98.0%)
50%	566/1308 (43.3%)	1321/1499 (88.1%)	375/481 (78.0%)	479/495 (96.8%)
70%	881/2293 (38.4%)	841/1080 (77.9%)	201/293 (68.6%)	93/117 (79.5%)

Table 8 Conditional Recovery Analysis for Average Log-probs

Occlusion	RGB wrong, Skeleton wrong	RGB wrong, Skeleton right	RGB right, Skeleton wrong	RGB right, Skeleton right
40%	191/911 (21.0%)	1295/1479 (87.6%)	437/512 (85.4%)	881/881 (100.0%)
50%	200/1308 (15.3%)	1226/1499 (81.8%)	389/481 (80.9%)	495/495 (100.0%)
70%	156/2293 (6.8%)	714/1080 (66.1%)	206/293 (70.3%)	117/117 (100.0%)

Table 9 Per-class gain/loss Examples For 70% Occlusion

Class	RGB	Skeleton	Linear Log Fusion	Gain vs. Best single
Horse Riding	8.16%	6.12%	97.96%	+89.80%
Playing Sitar	0%	15.91%	95.45%	+79.55%
Soccer Penalty	4.88%	0%	82.93%	+78.05%
Salsa Spin	9.3%	6.98%	74.42%	+65.12%
Soccer Juggling	7.69%	2.56%	71.79%	+64.1%
Typing	4.65%	79.07%	30.23%	-48.84%
Playing Daf	26.83%	65.85%	17.07%	-48.78%
Hula Hoop	0%	61.76%	17.65%	-44.12%
Breast Stroke	17.86%	82.14%	46.43%	-35.71%
Playing Tabla	3.23%	61.29%	29.03%	-32.26%

is trained under matched occlusion conditions, learned log-probability fusion provides the strongest performance. The control and recovery analysis suggest that this improvement depends on correctly paired RGB-skeleton information and often reflects skeleton-assisted recovery.

Discussion

This study asked whether lightweight probability-level fusion of RGB and skeleton predictions can improve human activity recognition under controlled synthetic occlusion. The results show that many types of fusion can improve robustness, but not in the same way. When the fusion head is trained only on clean 0% occlusion outputs, non-learned averaging is the most reliable strategy. However, when the fusion head is trained and tested under the same occlusion level, learned log-probability fusion becomes the strongest method. This means the main hypothesis was partially supported: RGB and skeleton streams

do provide complementary information, but learned fusion only becomes superior when trained under matched occlusion conditions.

The most important implication is that occlusion robustness depends not only on the two modalities being fused, but also how the fusion rule is trained. A learned fusion head trained on clean data does not automatically generalize to severe occlusion. This suggests that the reliability of RGB and skeleton predictions changes as occlusion increases. In contrast, average fusion is less flexible and more conservative, which helps it preserve correct predictions when one stream remains reliable. Same-occlusion learned fusion is more adaptive. It can recover difficult examples where RGB and skeleton top-1 predictions both fail, but it can also override correct single-stream predictions. The results show a tradeoff between conservative score averaging and more dynamic learned linear fusion.

The control experiments strengthen the interpretation that gains from the fusion are not only caused by adding more trainable parameters. At 40% and 50% occlusion, true RGB-skeleton fusion outperforms the skeleton matched-capacity by several percentage points. However, at 70% occlusion, true fusion is very close to the skeleton matched-capacity control. This suggests that when RGB evidence is extremely degraded, the model depends mostly on skeleton-side information, and the learned head behaves largely like a recalibration model for the skeleton stream.

These findings contribute to HAR research by showing that a small late-fusion head can reveal important differences between RGB and skeleton reliability under occlusion. Many multimodal HAR methods focus on more complex feature-level or attention-based fusion, but this study shows that a lightweight probability-level approach is still useful for analyzing robustness. The results also show why simple averaging should not be treated as a weak baseline. It can outperform a learned fusion head trained only on clean data.

Limitations and Future Work

This study has several important limitations. First, evaluation is limited to UCF-101 split 1. Although UCF-101 is a standard action-recognition benchmark, results from one dataset and one split do not establish cross-dataset generalization. Second, the occlusion protocol uses one random seed. The confidence intervals reported in this paper are computed over test videos, so they estimate uncertainty caused by the test set, but do not measure variation across different randomly generated occlusion patterns. As a result, the severity curves should be interpreted as controlled single-seed robustness comparisons.

Third, the occlusions are synthetic. Rectangular RGB masks may obscure background rather than the actor, and independent keypoint dropout does not fully reproduce real pose-estimator failures caused by furniture, objects, or multi-person overlap. Fourth, the expert streams are frozen checkpoints, so the study evaluates lightweight late fusion over end-to-end multimodal adaptation.

Future work should repeat the experiment with multiple occlusion seeds, evaluate on an additional dataset such as NTU RGB+D, and test more realistic occlusion models such as person-centered object masks or structured body part removal. Future work should also compare probability-level late fusion to feature-level fusion, cross-modal attention, and uncertainty-aware fusion under the same occlusion protocol.

Conclusion

This study evaluates lightweight probability-level fusion for occlusion-robust human activity recognition using frozen RGB and skeleton expert streams on UCF-101 split 1. The results show that RGB and skeleton predictions contain complementary information, but the best way to combine them depends on the training protocol. When learned fusion is trained only on clean 0% occlusion outputs, simple average-probability and average-log-probability fusion are more reliable under increasing occlusion. When the fusion head is trained and tested under the same occlusion level, learned log-probability fusion becomes the strongest method at moderate and severe occlusion levels. The strongest multimodal benefit appears at 40–50% occlusion, where learned log-probability fusion outperforms both non-learned averaging and matched-capacity single-stream controls.

These findings suggest that lightweight fusion is a useful and computationally simple approach for improving HAR under controlled synthetic occlusion, but they do not establish full real-world occlusion robustness. The results are limited to one dataset split, one occlusion seed, frozen expert checkpoints, and synthetic RGB masking and keypoint dropout. Future work should repeat the experiment with multiple occlusion seeds, evaluate on additional datasets such as NTU RGB+D, and test more realistic occlusion models such as person-centered object masks, structured body-part removal, and multi-person occlusion. Future work should also compare probability-level late fusion with feature-level fusion, uncertainty-aware fusion, and cross-modal attention under the same occlusion protocol.

References

1 K. Simonyan and A. Zisserman, *Two-Stream Convolutional Networks for Action Recognition in Videos*, 2014.

2 D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, *Learning Spatiotemporal Features with 3D Convolutional Networks*, 2015.

3 J. Carreira and A. Zisserman, *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*, 2017.

4 C. Feichtenhofer, H. Fan, J. Malik and K. He, *SlowFast Networks for Video Recognition*, 2019.

5 H. Duan, Y. Zhao, K. Chen, D. Lin and B. Dai, *Revisiting Skeleton-Based Action Recognition*, 2022.

6 A. Shahroudy, J. Liu, T.-T. Ng and G. Wang, *NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis*, 2016.

7 S. Yan, Y. Xiong and D. Lin, *Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition*, 2018.

8 L. Shi, Y. Zhang, J. Cheng and H. Lu, *Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition*, 2019.

9 Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng and W. Hu, *Channel-Wise Topology Refinement Graph Convolution for Skeleton-Based Action Recognition*, 2021.

10 K. Sun, B. Xiao, D. Liu and J. Wang, *Deep High-Resolution Representation Learning for Human Pose Estimation*, 2019.

11 Z. Zhong, L. Zheng, G. Kang, S. Li and Y. Yang, *Random Erasing Data Augmentation*, 2020.

12 T. DeVries and G. W. Taylor, *Improved Regularization of Convolutional Neural Networks with Cutout*, 2017.

13 Z. Chen, H. Wang and J. Gui, *Occluded Skeleton-Based Human Action Recognition with Dual Inhibition Training*, 2023.

14 Y.-F. Song, Z. Zhang, C. Shan and L. Wang, *Richly Activated Graph Convolutional Network for Robust Skeleton-Based Action Recognition*, 2021.

15 K. Peng, A. Roitberg, K. Yang, J. Zhang and R. Stiefelhagen, *Delving Deep into One-Shot Skeleton-Based Action Recognition with Diverse Occlusions*, 2023.

16 E. Mathe, I. Vernikos, E. Spyrou and P. Mylonas, *Leveraging Artificial Occluded Samples for Data Augmentation in Human Activity Recognition*, 2025.

17 I. Vernikos and E. Spyrou, *Skeleton Reconstruction Using Generative Adversarial Networks for Human Activity Recognition Under Occlusion*, 2025.

18 M. Cormier, Y. Schmid and J. Beyerer, *Enhancing Skeleton-Based Action Recognition in Real-World Scenarios Through Realistic Data Augmentation*, 2024.

19 C. Feichtenhofer, A. Pinz and A. Zisserman, *Convolutional Two-Stream Network Fusion for Video Action Recognition*, 2016.

20 F. Baradel, C. Wolf, J. Mille and G. W. Taylor, *Human Activity Recognition with Pose-Driven Attention to RGB*, 2018.

21 S. Das, S. Sharma, R. Dai, F. Brémond and M. Thonnat, *VPN: Learning Video-Pose Embedding for Activities of Daily Living*, 2020.

22 X. Zhu, Y. Zhu, H. Wang, H. Wen, Y. Yan and P. Liu, *Skeleton Sequence and RGB Frame Based Multi-Modality Feature Fusion Network for Action Recognition*, 2022.

23 W. Xu, M. Wu, M. Zhao and T. Xia, *Fusion of Skeleton and RGB Features for RGB-D Human Action Recognition*, 2021.

24 Y. Sun, W. Xu, X. Yu, J. Gao and T. Xia, *Integrating Vision Transformer-Based Bilinear Pooling and Attention Network Fusion of RGB and Skeleton Features for Human Action Recognition*, 2023.

25 J. Kittler, M. Hatef, R. P. W. Duin and J. Matas, *On Combining Classifiers*, 1998.

26 K. Soomro, A. R. Zamir and M. Shah, *UCF101: A Dataset of 101 Human Actions Classes From Videos in the Wild*, 2012.

27 MMAction2 Contributors, *OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark*, 2020.

28 W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman and A. Zisserman, *The Kinetics Human Action Video Dataset*, 2017.