SBF: Augmenting Skeleton for Effective Video-based Human Action Recognition

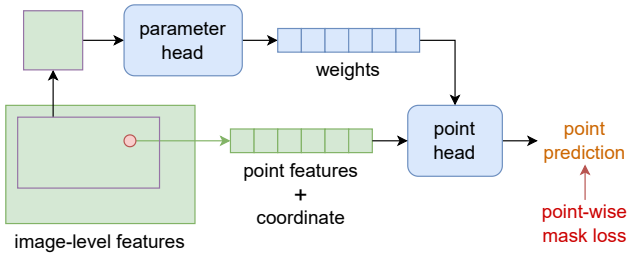
Supplementary Material

8. More Implementation Details

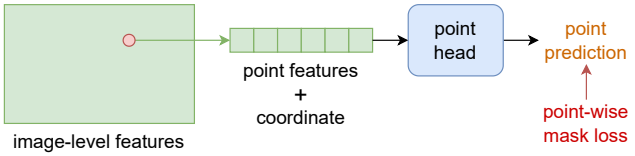
In this section, we elaborate more details of our implementations of SFSNet and SBFConv3D. All our experiments are conducted on two hardware platforms, one with 8 NVIDIA GeForce 2080Ti GPUs and 16 CPUs and the other with 4 3090Ti GPUs and 40 CPUs.

8.1. SFSNet

Fig. 8 illustrates the detailed structure of our Simplified PointRend module compared to Implicit PointRend [6]. Simplified PointRend retains the overall structure of Implicit PointRend, but the dynamic point head is replaced by a common 3-layer perceptron with ReLU as the activation function. During the process of point annotation generation, we use $\rho = 10$, $N_{pos} = 32, N_{neg} = 128, N_{body} = N_{flow} = 256$, $\alpha = 19$ and $\beta = 0.8, \gamma = 0.2$. The loss weights are set to $\lambda_{joint} = 0.025$ and $\lambda_{body} = 1$. Both training stages use Adam as the optimizer with a learning rate $1e-2$, and the batch size is set to 128.



(a) Implicit PointRend



(b) Simplified PointRend (Ours)

Figure 8. Comparison of Implicit PointRend [6] and our Simplified PointRend architectures.

8.2. SBFConv3D

We select Adam as the optimizer and CosineAnnealing as the learning rate scheduler with a maximum learning rate 0.2. The batch size is set to 128.

9. More Quantitative Results

In this section, we presents more experimental results to further validate the effectiveness of our SBF and SFSNet.

9.1. Results of the "Limb" Variant of SBF

Our SBF has two variants, "joint" and "limb". While previous discussions focus on the "joint" variant, this section presents the performance of SBFConv3D based on SBF of the "limb" variant. We compare it to the limb stream of PoseConv3D [11]. Tab. 8 shows that SBFConv3D outperforms PoseConv3D significantly in all settings on NTU [29] and NTU120 [23].

9.2. Comparison on Different Action Categories

Fig. 9 lists the action-specific accuracy difference between SBFConv3D and PoseConv3D [11] in action categories with the "Medium" difficulty level from the NTU120 X-Sub setting [23]. SBFConv3D achieves higher or equal accuracy compared to PoseConv3D in 28 of 31 "Medium" actions. Besides, SBFConv3D outperforms PoseConv3D in 63 out of 74 "Easy" actions and in 105 out of 120 total actions, with a maximum improvement of 20.9%.

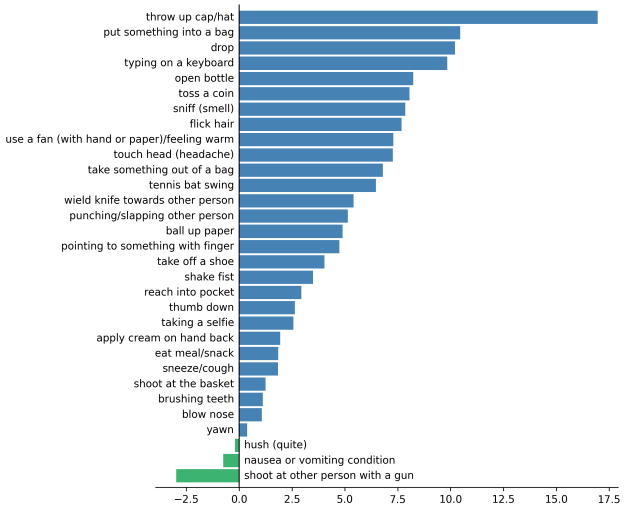


Figure 9. The accuracy difference (%) between our SBFConv3D and PoseConv3D for "Medium" actions on NTU120 X-Sub.

9.3. Comparison with 3D Skeleton

Our SBF captures depth information by using the scale of each joint. One might question why we do not use 3D skeletons, which inherently include depth information. To address this, we evaluate CTR-GCN [5] using either ground-truth 3D skeletons (GT 3D Skl) or 3D skeletons predicted by a state-of-the-art 2D-to-3D lifting method (Pred 3D Skl). As shown in Tab. 10, using ground-truth 3D skeletons results in reduced performance compared to 2D skeletons, due

Method	Category	Clips	NTU		NTU120	
			X-Sub	X-View	X-Sub	X-Set
PoseConv3D	Skeleton (Limb)	1	93.2	95.7	85.7	89.4
		10	93.4	96.0	85.9	89.7
SBFConv3D	SBF	1	94.4	96.4	88.5	91.3
		10	**94.4**	**96.6**	**88.6**	**91.6**

Table 8. Comparison of accuracy (%) with PoseConv3D using only the "limb" variant.

μ	Acc (%)
0	94.5 (.46)
0.05	94.5 (.48)
0.1	**94.6**
0.5	93.9
1	93.3

Table 9. Ablation Study on μ.

Method	Category	Add. Anno.	Acc (%)
CTR-GCN	2D Skl	✗	93.6
	GT 3D Skl	✓	92.1
	Pred 3D Skl	✓	63.9
SBFConv3D	SBF	✗	**95.0**

Table 10. Comparison of accuracy (%) with skeleton-based methods trained using 3D skeletons.

Method	Category	Clip Len.	X-Sub	X-View
CTR-GCN	Skeleton	48	92.6	97.8
PoseConv3D	Skeleton	48	94.1	97.1
SBFConv3D	SBF	48	**95.0**	**98.1**

Table 11. Comparison of accuracy (%) with skeleton-based methods with the same number of frames per clip.

Skeleton	$\mathcal{S}$	$\mathcal{B}$	$\mathcal{F}$	Acc (%)
✓				93.5
	✓			90.3
		✓		86.7
			✓	87.8
✓	✓			93.6
✓		✓		93.6
✓			✓	94.3
	✓	✓		91.1
	✓		✓	92.3
		✓	✓	89.8
✓	✓	✓		93.8
✓	✓		✓	94.5
✓		✓	✓	94.5
	✓	✓	✓	93.3
✓	✓	✓	✓	**94.6**

Table 12. Complete Ablation Study on SBF components on NTU X-Sub.

to the inferior quality of 3D data, as noted in [11]. Predicted 3D skeletons further degrade accuracy because joint depth cannot be accurately inferred from flat 2D skeletons. This validate that using joint scale to capture depth information is more reliable than directly predicting 3D skeleton.

9.4. Ablation Studies

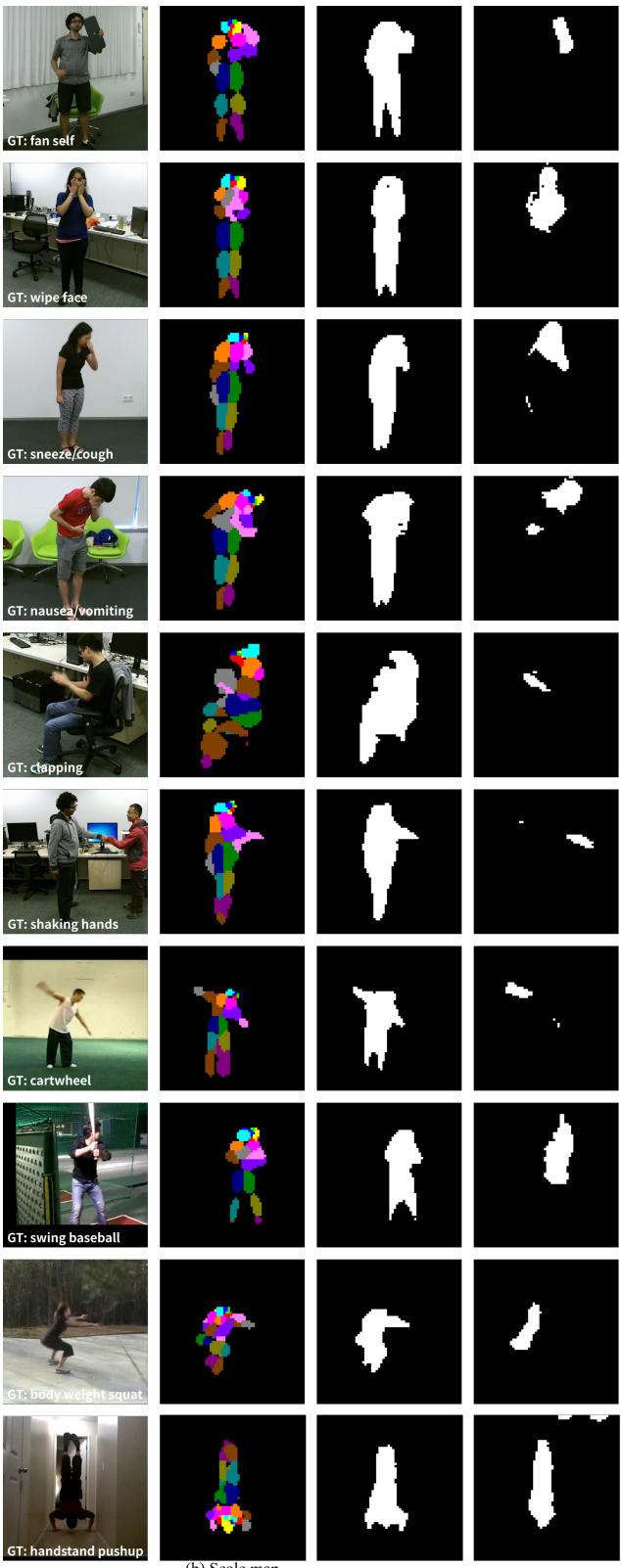
Effects of SBF Components: We validate the effectiveness of all three components of SBF, scale map volume $\mathcal{S}$, body map $\mathcal{B}$ and flow map $\mathcal{F}$, on NTU X-Sub. Results in Tab. 12 demonstrate that all components collectively enhance the performance of our skeleton+SBF-based HAR. Notably, the flow map contributes the most, underscoring the importance of incorporating human-object interaction into the HAR pipeline.

Effects of μ in the joint map volume. The parameter μ mentioned in **??** balances the skeleton and scale map volume $\mathcal{S}$. According to the results in Tab. 9, SBF achieves optimal performance when $\mu = 0.1$ among the choices. This result aligns with our expectation that the depth information in $\mathcal{S}$ improves the HAR accuracy in some challenging scenarios where using only skeleton can lead to ambiguity.

Effects of Clip Length: In previous experiments, GCN-based methods process 100 frames per clip, whereas PoseConv3D and SBFConv3D uses only 48 frames. This allows GCN-based methods to leverage more temporal information, reducing the ambiguity in action understanding. To fairly compare SBFConv3D and GCN-based methods using the same temporal information, we evaluate CTR-GCN [5] with 48 frames. Results in Tab. 11 demonstrate that when the clip length is 48 for all methods, our SBFConv3D achieves the highest accuracy. Notably, although SBFConv3D obtains lower performance than CTR-GCN on NTU X-View with a smaller clip length, it outperforms CTR-GCN when both methods use the same clip length.

10. More Visualization Results

This section presents more visualizations of our predicted SBF in various datasets.



(a) Image (b) Scale map volume (c) Body map (d) Flow map

Figure 10. Visualization of SBF components predicted by SF-SNet on NTU120 [23] (row 1-6), HMDB51 [19] (row 7-8) and UCF101 [33] (row 9-10). Each joint is depicted in a distinct color.