Activity-Biometrics: Person Identification from Daily Activities

Supplementary Material

The supplementary is organized as follows:

- Section A provides detailed description of the gallery probe setup of the datasets.
- Section B provides comparison of ABNet with state-of-the-art methods on the cross-activity evaluation protocol.
- Section C provides ablations and some more discussion and analysis
- Section D provides some more qualitative samples of retrieval results of ABNet.

The code and all the datasets used for this work will be made publicly available.

A. Gallery probe setup

We evaluate the performance in terms of same activity and cross activity. In the same activity evaluation protocol, probe and gallery contains all the activities, however, probe contains a smaller subset of samples and the rest are placed in gallery. In the cross activity evaluation protocol, probe and gallery contains mutually exclusive activities, where probe contains a smaller subset of samples and rest of the samples from those activities are discarded; on the contrary the gallery contains all samples from a certain activity. Here for each actor there are multiple activity samples, and each activity again has different view-point or setup variation (for NTU RGB-AB and PKU MMD-AB). The samples are randomly selected for gallery and probe sets. For NTU RGB-AB and PKU MMD-AB two variations are checked - probe view included in gallery (View^{+}) and probe view excluded from gallery (View^{-}) in case of both same activity and cross activity protocol. However, since Charades and ACC-MM1-Activities does not contain multiple view points, the evaluation protocol with inclusion/exclusion of probe view from gallery is not relevant in these case. Table 1 illustrates a detailed description of all the datasets.

B. Comparison with state-of-the-art methods

We present the comparison of different state-of-the-art methods against our proposed ABNet to show its effectiveness across NTU RGB-AB, PKU MMD-AB, Charades-AB and ACM-MM1-Activities datasets on the cross-activity View^{+} evaluation protocol in Table 2 which corresponds to Table 1 of the main paper. Similar to the quantitative comparisons presented in the main paper, in case of cross-activity evaluation protocol as well, ABNet outperforms all the existing methods and baselines by a competitive margin in terms of both evaluation metrics. This shows the robustness of our method against same or cross activity evaluation.

Table 1. ***Dataset statistics***

Dataset	Split	#actors	#activities	#samples
NTU RGB-AB	train	85	94	70952
	gallery	21		14192
	probe			3548
PKU MMD-AB	train	53	41	13634
	gallery	13		2727
	probe			681
Charades-AB	train	214	157	45111
	gallery	53		9022
	probe			2256
ACC-MM1-Activities	train	182	7	7717
	gallery	45		1543
	probe			386
BRIAR-BGC3	train	870	3	20000
	gallery	130		4171
	probe			922

C. More analysis and discussion

Ablations on cross-activity evaluation protocol. Table 3 illustrates the effect of each component of our proposed ABNet on NTU RGB-AB dataset on the cross-activity evaluation protocol. This table is an extension of Table 4 of the main paper and similar to the same-activity evaluation protocol, the performance of the model remains stable in case of cross-activity and also each modification component gives a performance boost to the model, which finally contributes to the overall model's performance. Now, some activities might be easier to recognize and hence, we perform an experiment on top 5 best and top 5 worst performing activities with and without the activity prior (AP) to see whether the easily recognizable activities introduce any bias through the activity information. In Figure 1 we see that the performance pattern remains consistent across activities with or without AP which indicates that AP consistently helps and the difficulty level of activities do not introduce any bias.

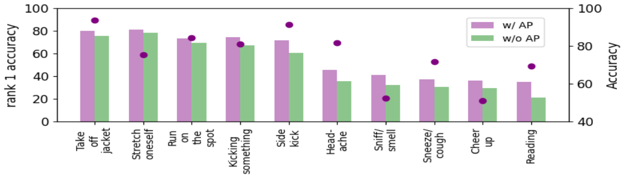

Figure 1. ***Performance analysis w/ and w/o activity prior***; bars represent biometrics rank 1 and dots represent activity accuracy.

Effect of distortion: Table 4 reports the effect of distortion on cross-activity evaluation protocol on the NTU RGB-AB

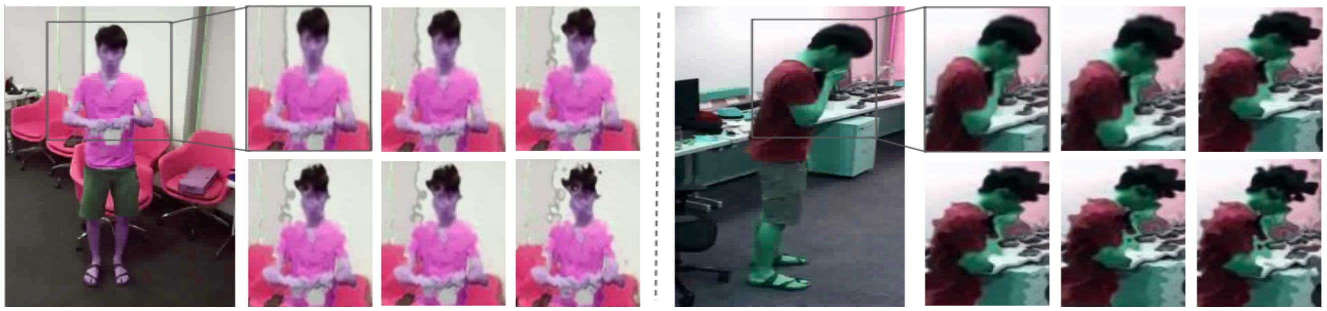

Figure 2. ***Effect of distortion amount*** Original sample zoomed in to show effect of $\alpha = 50, 100, 150$ (top) and $\alpha = 200, 250, 300$ (bottom). As α increases, the distortion keeps increasing.

Table 2. ***Comparison with state-of-the-art person identification methods*:** Evaluation shown on NTU RGB-AB, PKU MMD-AB, Charades-AB, and ACC-MM1-Activities on cross-activity, View$^+$ evaluation protocol . †: this model was trained on silhouettes.

	Methods	Venue	NTU RGB-AB		PKU MMD-AB		Charades-AB		ACC-MM1-Activities	
			Rank 1	mAP	Rank 1	mAP	Rank 1	mAP	Rank 1	mAP
Image	CAL [16]	CVPR22	70.31	24.08	78.31	43.43	40.13	21.23	67.33	38.21
	PSTR [5]	CVPR22	68.34	32.54	77.98	41.23	35.12	20.32	53.46	30.18
	SCNet [17]	ACM MM23	68.82	26.31	73.91	39.65	27.42	17.61	55.38	32.42
	AIM [43]	CVPR23	72.79	30.21	79.22	44.90	35.56	26.36	66.81	38.14
Video	TSF [23]	AAAI20	67.81	26.88	71.61	33.22	30.21	18.29	41.31	21.43
	VKD [35]	ECCV20	66.33	31.46	72.19	34.34	31.89	18.81	51.26	22.16
	BiCnet-TKS [21]	CVPR21	69.13	30.21	77.13	33.32	38.33	23.34	58.41	30.21
	STMN [12]	ICCV21	70.21	30.13	71.53	42.21	33.89	20.81	57.61	37.61
	PSTA [40]	ICCV21	65.13	31.42	72.43	47.42	38.72	24.84	67.31	37.33
	SINet [4]	CVPR22	66.21	27.81	74.11	26.21	37.31	21.90	61.32	36.41
	Video-CAL [16]	CVPR22	73.31	31.73	77.34	45.72	41.50	25.81	67.48	38.23
Baselines	GaitGL [28] †	-	57.04	27.13	61.22	27.84	14.51	4.85	35.13	16.31
	ResNet3D-50 [18]	-	62.80	23.52	65.12	29.41	27.35	14.89	39.89	19.83
	MViTv2 [26]	-	59.27	21.38	61.40	25.31	21.89	12.79	37.31	17.80
	ABNet (ours)	-	**77.01**	**37.64**	**81.44**	**51.79**	**44.82**	**28.78**	**68.31**	**38.83**

Table 3. ***Ablation studies*** of each component of ABNet on NTU RGB-AB on cross activity evaluation protocol

B/L	K/D	A/P	F/D	View$^+$		View$^-$	
				R@1	mAP	R@1	mAP
✓				62.80	23.52	61.71	21.41
✓	✓			66.90	23.94	63.03	22.01
✓		✓		66.24	23.81	64.61	22.48
✓	✓	✓		69.21	31.01	66.41	30.43
✓	✓		✓	74.33	33.79	72.85	31.68
✓	✓	✓	✓	**77.01**	**37.64**	**76.43**	**36.14**

Table 4. ***Effect of distortion*** on model performance for NTU RGB-AB on the cross activity evaluation protocol

Distortion amount	View$^+$		View$^-$	
	R@1	mAP	R@1	mAP
$\alpha = 200$	75.91	37.04	75.12	35.83
$\alpha = 250$	**77.01**	**37.64**	**76.43**	**36.14**
$\alpha = 300$	72.70	29.01	71.03	28.94

Table 5. ***Effect of face restriction*** on model performance for NTU RGB-AB on cross activity evaluation protocol

Face Restricted	View$^+$		View$^-$	
	R@1	mAP	R@1	mAP
Yes	77.01	37.64	76.43	36.14
No	77.70	39.01	76.98	38.84

dataset which is an extension of Table 5 of the main paper. Figure 3 illustrates the t-SNE plots of the biometrics and appearance feature space for ten individuals from two of the challenging datasets; Charades-AB and BRIAR-BGC3. It is observed from this figure that the effect of α is consistent and our choice of α is applicable for even these challenging datasets as well. Figure 2 illustrates the qualitative samples of the effect of distortion. In Figure 2, we zoomed in the face portion of the actor for better visualization and it is observed that as α increases gradually, the distortion amount increases and by $\alpha = 300$, the sample is so distorted that it becomes unsuitable for our purpose.

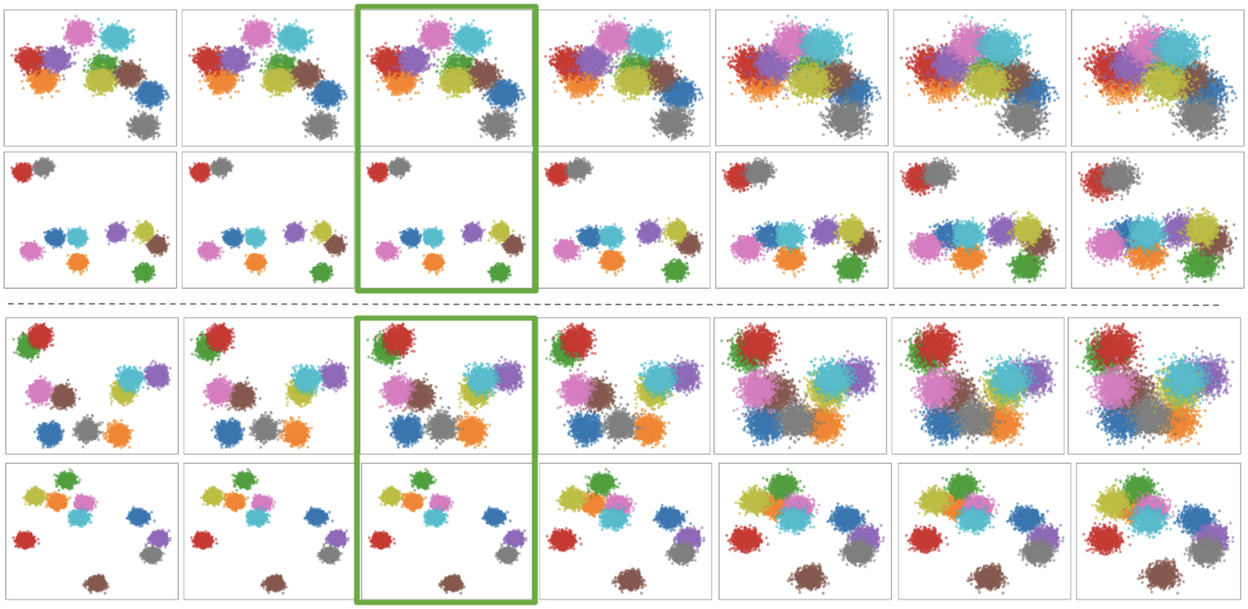

Figure 3. ***Effect of distortion on feature space.*** The t-SNE plots illustrate the impact of varying distortion amount $\alpha \in [200, 225, 250, 275, 300, 325, 350]$ on biometrics (top) and appearance (bottom) features of ABNet for ten random identities for Charades-AB (top) and BRIAR-BGC3 (bottom). As α increases from left to right, the optimal results occur at $\alpha = 250$ (shown in square) where biometrics changes while appearance remains consistent. Beyond $\alpha = 250$, appearance gets distorted too, making it unsuitable for disentanglement

Table 6. ***Activity recognition*** performance of different datasets on ABNet. x-sub and x-view respectively denote cross-subject and cross-view evaluation protocols for its corresponding dataset, if applicable.

Dataset	x-sub	x-view
NTU RGB-AB	88.71	89.50
PKU MMD-AB	91.42	94.21
Charades-AB	41.31	-
ACC-MM1-Activities	71.08	-
BRIAR-BGC3	79.31	-

Effect of face restriction on cross-activity evaluation protocol is reported in Table 5 on the NTU RGB-AB dataset. Similar to the results reported in the main paper, even in case of the cross-activity evaluation protocol, the model performance remains stable even when faces are restricted showing the learning of non-facial cues across cross-activity evaluation protocol.

Choice of backbone. The performance comparison of different backbone networks is shown in Table 7, where the backbone model takes the silhouette/RGB video frames as input respectively for the teacher/student network for the task of person identification. Here this experiment is run only on the baseline where none of the modification components are present. This selection of backbones ensures that the teacher network contributes its expertise to the specific task it is designed for in the student network. Moreover, similar to existing recent work [5, 16, 21, 43] in person identification, in our case also CNN based backbones

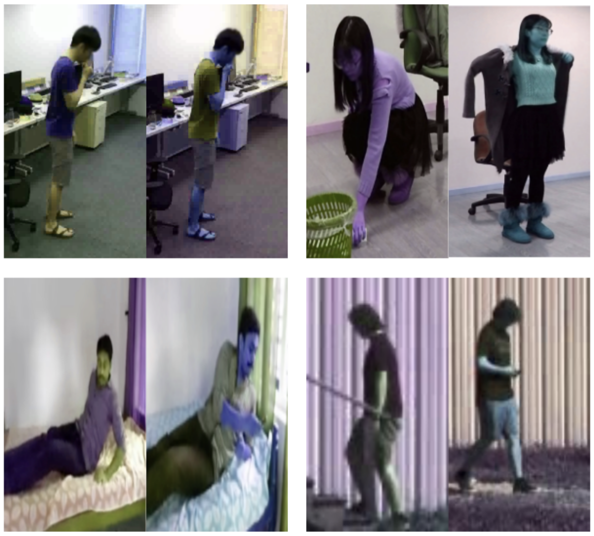

Figure 4. ***Dataset samples.*** Here each two samples show different values of hue shifting for the same video of NTU RGB-AB (top-left), PKU MMD-AB (top-right), Charades-AB (bottom-left) and ACC-MM1-AB (bottom-right). All the samples have their faces blurred.

Table 7. ***Choice of Backbone.*** Performance comparison of different backbones on NTU RGB-AB

Network	Backbone	Same activity				Cross activity			
		View$^+$		View$^-$		View$^+$		View$^-$	
		R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP
Teacher	**GaitGL**[28]	**61.51**	**28.89**	57.78	**26.78**	57.04	**27.13**	55.80	**26.41**
	GaitPart[13]	54.79	16.73	53.93	15.91	52.18	15.01	46.89	13.84
	GaitBase[14]	60.21	28.02	**59.04**	26.76	**59.90**	26.31	**57.91**	25.96
Student	MViT v2[26]	63.87	26.41	61.01	**23.81**	59.27	21.38	59.16	20.01
	ViViT[3]	58.81	20.41	57.10	16.42	57.30	12.41	52.01	9.68
	Swin[32]	59.20	21.68	58.41	19.41	58.70	16.91	54.31	11.47
	ResNet3D-50[18]	**64.23**	**26.89**	**62.10**	22.45	**62.80**	**23.52**	**61.71**	**21.41**
	ResNet3D-34[18]	63.90	25.93	60.45	21.87	60.21	22.74	59.79	20.47

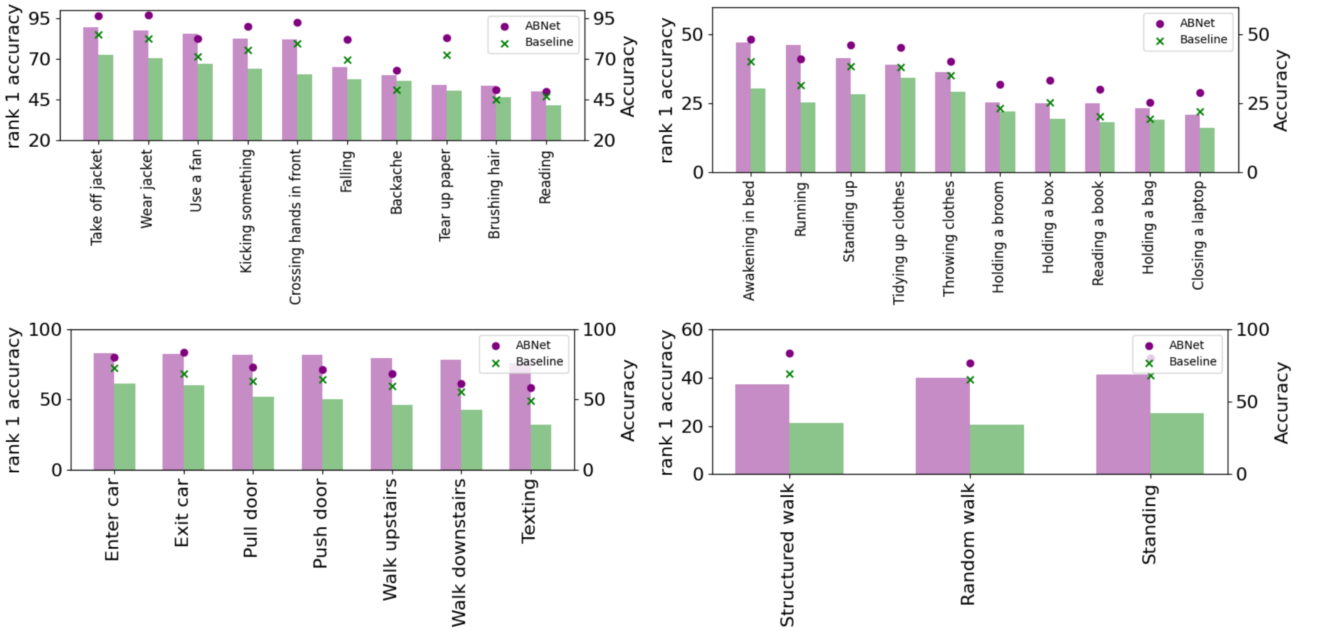

Figure 5. ***Performance analysis across activities.*** The bar plot on left axis shows rank 1 identification accuracy for given activity of ABNet against baseline PKU MMD-AB (top-left), Charades AB (top-right), ACC-MM1-Activities (bottom-left) and BRIAR-BGC3 (bottom-right) datasets. The scatter plot with markers on right axis shows activity recognition accuracy for corresponding classes.

outperform transformer based ones. From this experiment, we pick the best performing backbone for both networks.

Performance of action recognition: Table 6 reports the performance of ABNet on activity recognition results for different datasets. Here the reported evaluation metric is accuracy on cross-subject and cross-view evaluation protocol. NTU RGB-AB and PKU MMD-AB are evaluated on these two protocols, however, since there is no explicit view information for rest of the three datasets, the accuracies are reported in terms of cross-subject because the test and train split contains mutually exclusive actors/subjects.

Figure 5 compare ABNet and the baseline across the top five best and bottom five worst performing activities in person identification for PKU MMD-AB and Charades (top row). The bottom row shows person identification performance across all 7 activities of the ACC-MM1-Activities dataset and all 3 activities of BRIAR-BGC3 dataset. It is observed that activities with minimal overall body movement pose greater challenges for individual identification, whereas more overall body movement contribute to higher person identification accuracy. This highlights the significance of incorporating activity prior in our model. Moreover, it also emphasizes the importance of activity cues demonstrating the efficacy of our joint training approach in effectively learning such cues.

Accuracy of silhouette extractor and effectiveness of silhouettes: The accuracy of the silhouette extraction process will indeed affect model's performance and to explore that we perform an experiment using Grounded-SAM [37] which is an open-world segmentation model. The results

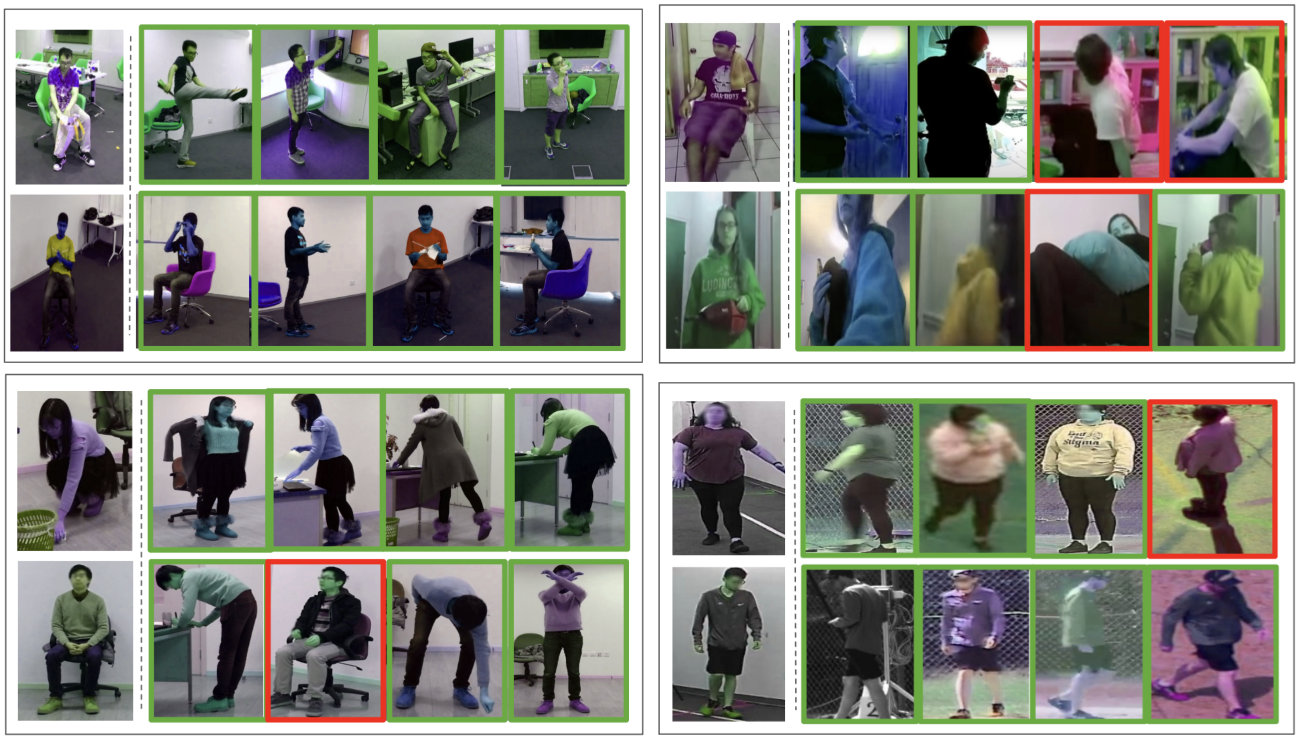

Figure 6. ***Top 4 rank retrieval*** samples for ABNet on NTU RGB-AB (top-left), Charades-AB (top-right), PKU MMD-AB (bottom-left) and BRIAR-BGC3 (bottom-right). The left most columns for each dataset samples hold the probe samples and the following four columns to that probe are its retrieval list. Accurate retrieval is shown with green box and inaccurate with red.

Table 8. ***Performance with varying silhouette extractors***

Silhouette extractor	Rank 1	mAP
Mask2Former [8]	85.2	87.3
Grounded-SAM [37]	87.8	88.5

Table 9. ***Performance of disentangled features***

Feature	Rank 1	mAP
Biometrics	**45.8**	**31.6**
Non-biometrics	2.8	0.4
Biometrics w/ distorted sils	21.4	10.5

are reported on a small subset (10 action classes) of the NTU-RGB-AB dataset on the same activity View$^+$ setting in Table 8. It is observed that with Grounded-SAM as silhouette extractor the performance does go up, which can be attributed to it being an open-world model and thus being more robust. Similarly, a $\mathbf{3}\%$ rank 1 accuracy gain is seen in case of a small subset of the Charades-AB dataset when using Grounded-SAM as opposed to Mask2Former. Nevertheless, even with a weaker silhouette extractor our model still performs well and since this extraction process is not part of the inference stage, training the model with a better silhouette extractor will provide some benefits. The main motivation behind using silhouette features is to distill appearance-less knowledge, e.g. purely biometrics information that not only contains gait; but also pose, body shape, structure etc information to aid disentanglement. The recognition performance of the two decoupled features is reported in Table 9 for the Charades-AB dataset. The huge performance gap between the biometrics and non-biometrics features shows that the non-biometrics features do not have meaningful information to perform person identification; essentially proving the effectiveness of the disentanglement process. To demonstrate the effectiveness of using silhouette features in our method, we distort the silhouettes and distill that knowledge to the biometrics features, which resulted in a huge performance drop (about $\mathbf{24}\%$) (row 3 of Table 9). This shows that even in case of activities beyond walking, the silhouette-based biometrics features contribute to a great extent in accurate recognition. We specifically select the Charades-AB dataset for this experiment as it is a real-world dataset encompassing a diverse range of appearance variations.

D. Qualitative analysis

Figure 4 illustrates examples of different values of hue-shifting, from which it can be observed that the color profile for each frame is distinct from the other. Figure 6 illustrates the top 4 rank retrieval results for a given probe for NTU RGB-AB, PKU MMD-AB and Charades-AB datasets. Some of the failure cases is seen for having difficulty performing accurate retrieval due to the absence of a lot of overall body movement (e.g. probe activity is sitting in first sample of Charades-AB and second sample of PKU MMD-AB). Moreover, another failure case is seen in case of the second sample of Charades-AB which shows the inherent challenges present in the dataset, e.g. data quality, no standard way of performing an activity etc. Despite these challenges, from the figure it is observed that accurate retrieval is done in most cases irrespective of view-point, activity and appearance, which shows the effectiveness of ABNet.