APPENDIX

A BROADER IMPACT

The study of skeleton-based human action recognition is of great practical significance. It is not only computationally more efficient to use skeletons instead of raw videos, but it also resolves the special concern for privacy in the applications of human action recognition. For example, our model can be deployed for violence detection, at the same time keeping the crowds' identities anonymous.

B REPRODUCIBILITY - EXPERIMENT DETAILS

In order to ensure reproducibility, we provide all training hyperparameters for our method for all datasets in the following.

All experiments are conducted on a single Tesla V100 GPU with the PyTorch deep learning framework. A total number of 110 epochs is chosen for all the experiments and the warmup method in He et al. (2016) is adopted in the first 5 epochs for a more stable training process. We train the model using Stochastic Gradient Descent (SGD) with Nesterov momentum (0.9) and weight decay (0.0004 for NTU RGB+D and NTU RGB+D 120, 0.0001 for Northwestern-UCLA) for optimization. We apply cross-entropy loss as the objective function. The learning rate is initialized to 0.1 for NTU RGB+D and NTU RGB+D 120 and is reduced by a factor of 10 at epoch 90 and 100, following InfoGCN Chi et al. (2022). For Northwestern-UCLA, we adopt a smaller learning rate of 0.05 and the same decay schedule. For NTU RGB+D and NTU RGB+D 120, the batch size is set to 64, each sample is resized to 64 frames, and we follow the data pre-processing in Zhang et al. (2020). For Northwestern-UCLA, we use a batch size of 16, and adopt the data pre-processing as in Cheng et al. (2020); Chen et al. (2021). Our code is based on the official implementation of Chen et al. (2021) and Zhang et al. (2020) and will be fully released upon acceptance.

We show in Tab. 1 the default hyperparameters for training our ToANet on NTU RGB+D, NTU RGB+D 120 and Northwestern-UCLA.

Config.	NTU RGB+D and NTU RGB+D 120	Northwestern-UCLA
random choose	False	True
random rotation	True	False
window size	64	52
weight decay	4e-4	1e-4
base lr	0.1	0.05
lr decay rate	0.1	0.1
lr decay epoch	90, 100	90, 100
warm up epoch	5	5
batch size	64	16
num. epochs	110	110
optimizer	Nesterov Accelerated Gradient	Nesterov Accelerated Gradient

Table 1: Default hyperparameters for our ToANet on NTU RGB+D, NTU RGB+D 120 and Northwestern-UCLA.

C MORE EXPERIMENT RESULTS

We also provide the experiment results for each modality on different benchmarks in detail, see Tab. 2 and Tab. 3.

Modality	NTU-RGB+D 120		NTU-RGB+D	
	X-Sub(%)	X-Set(%)	X-Sub(%)	X-View(%)
Joint	85.7	87.3	90.2	94.6
Bone	86.6	88.6	90.4	95.6
Motion	82.7	84.2	88.2	92.9
Bone Motion	82.6	84.1	88.1	92.5

Table 2: Classification accuracy of our ToANet using different modalities on the NTU RGB+D and NTU RGB+D 120 dataset.

Modality	Northwestern-UCLA (%)
Joint	93.3
Bone	92.0
Motion	92.0
Bone Motion	90.3

Table 3: Classification accuracy of our ToANet using different modalities on the Northwestern-UCLA dataset.

REFERENCES

Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13359–13368, 2021.

Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 183–192, 2020.

Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20186–20196, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Pengfei Zhang, Cuiling Lan, Wenjun Zeng, Junliang Xing, Jianru Xue, and Nanning Zheng. Semantics-guided neural networks for efficient skeleton-based human action recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1112–1121, 2020.