FSAR: Federated Skeleton-based Action Recognition with Adaptive Topology Structure and Knowledge Distillation – Supplemental Material –

Jingwen Guo[1] Hong Liu[1*] Shitong Sun[2] Tianyu Guo[1] Min Zhang[3] Chenyang Si[4*]
[1]Peking University Shenzhen Graduate School [2]Queen Mary University of London
[3]Harbin Institute of Technology Shenzhen Graduate School [4]Singapore Nanyang Technological University
{jingwenguo, levigty}@stu.pku.edu.cn hongliu@pku.edu.cn
shitong.sun@qmul.ac.uk zhangminmt@hotmail.com chenyang.si.mail@gmail.com

1. Additional Dataset Information

Under the federated-by-dataset scenario, each client constructs its own dataset for local training. Five datasets with significant variances in skeleton video amounts and category numbers are selected in Table 1. These variances between clients lead to huge domain gaps among each other and simulate the imbalance of data across clients to mimic the statistical heterogeneity in the real-world scenario.

Table 1. The scales of the datasets used in the federated-by-dataset scenario. Each client is trained on an individual dataset.

Clients	*Train.*	*Test.*	*Classes.*
PKU MMD I [7]	18,835	2,704	51
PKU MMD II [7]	5,294	1,610	41
NTU RGB+D 60 [9]	40,091	16,487	60
NTU RGB+D 120 [8]	63,026	50,919	120
UESTC [5]	11,361	14,240	40

2. Overall Optimization Algorithm

The optimization strategy algorithm of FSAR is demonstrated in Algorithm 1, where the client-server collaborative learning is iteratively performed to achieve a globally generalized server model. Specifically, in each round $r \in \{0, 1, ..., R\}$, the central server qualifies the server update direction as:

$$\Delta^r = -(\mathcal{W}_g^r - \mathcal{W}_g^{r-1}), \tag{1}$$

where $\mathcal{W}_g^r$ and $\mathcal{W}_g^{r-1}$ are the parameters of central server in current round r and the previous round $r-1$, respectively. The accelerated global model is defined as:

$$\mathcal{W}_g^r = \mathcal{W}_g^r - \xi\Delta, \tag{2}$$

which is broadcast to each client as parameters re-initialization for local training in the next round. Δ^r keeps past global gradient information and serves as taking lookahead parameters during the client-server communication. Each local client model is optimized with its own data following $\mathcal{W}_{i,k}^r = \arg\min_{\mathcal{W}_{i,k}^r} \mathcal{L}_{All}$. The central server then aggregates the local model parameters as follows:

$$\mathcal{W}_g^{r+1} = \sum \mathcal{W}_i^r + (1-\tau)(\mathcal{W}_g^r - \xi\Delta^r). \tag{3}$$

The above is the manner the server updates. Here, hyperparameters are set as $\xi = 0.8$, and $\tau = 0.8$ in our experiments as default.

3. Additional Ablation Studies

We analyze the effect of different implementation detail settings in the federated learning paradigms under the federated-by-dataset scenario, like the loss weights for local training of each client, the number of local training epochs, the number of selected clients in each client-server communication round.

Loss Weights. The default loss weights in Sec 3.5 are set as $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = 0.1$ to make our FSAR more general, even though other values can bring more gains. Table 2 presents the test accuracy with different parameter value combinations of the loss weights. Overall, the FSAR architecture is not sensitive to these parameters.

Selected clients. In each communication round, the server will randomly select S clients to aggregate the local model parameters for the global model. The default setting is $S = 5$, which means the central server aggregates the model parameters of all five participating clients. Table 3 compares FSAR with different settings of *Selected Clients* S. We can see that updating with an arbitrary client ($S = 1$) in each round is superior for small-scale datasets (*e.g.* PKU MMD I), while inferior for large-scale datasets (*e.g.* NTU RGB+D 120). That is because the collaboration of multiple

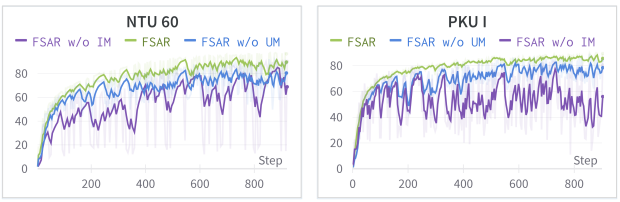
Algorithm 1: FSAR

Central Server Executing:
Init $\mathcal{W}_g^0$, init ξ, init $\Delta^0 \leftarrow 0$;
for *each round* $r \leftarrow 0$ **to** R **do**
  $S_r \leftarrow$ (Random selection of clients);
  $\mathcal{W}_g^r \leftarrow \mathcal{W}_g^r - \xi\Delta^r$;
  for *client* $i \in S_r$ *in parallel* **do**
    $\mathcal{W}_{i,K}^r \leftarrow$ Local Client Training $(\mathcal{W}_g^r, r)$;
  end
  $\mathcal{W}_g^{r+1} \leftarrow \sum_{m \in S_r} \frac{n_m}{n} \mathcal{W}_i^r + (1-\tau)(\mathcal{W}_g^r - \xi\Delta^r)$;
  $\Delta^{r+1} \leftarrow -(\mathcal{W}_g^{r+1} - \mathcal{W}_g^r)$;
end
Local Client Training $(\mathcal{W}_g^r, r)$:
for *local epoch* $k \leftarrow 0$ **to** K **do**
  $\mathcal{W}_{i,k}^r \leftarrow \mathcal{W}_g^r$;
  for *data batch* $b \in \mathcal{B}$ **do**
    $\mathcal{L}_{CE} \leftarrow CE(\Phi_i(\mathbf{h}), y) +$
      $\sum_{m=1}^{\bar{m}} CE(\Phi_i(\bar{\mathbf{h}}_m), y)$;
    $\mathcal{L}_{KD} \leftarrow \sum_{m=1}^{\bar{m}} KL(\Phi_i(\bar{\mathbf{h}}_m), \Phi_i(\mathbf{h}))$;
    $\mathcal{L}_{Reg} \leftarrow \frac{1}{2}||\mathcal{W}_{i,k}^r - \mathcal{W}_g^r||^2$;
    $\mathcal{L}_{All} \leftarrow \lambda_1\mathcal{L}_{CE} + \lambda_2\mathcal{L}_{KD} + \lambda_3\mathcal{L}_{Reg}$;
    $\mathcal{W}_{i,k}^r \leftarrow \arg\min_{\mathcal{W}_{i,k}^r} \mathcal{L}_{All}$;
  end
end
Return $\mathcal{W}_{i,K}^r$

clients makes large-scale clients contribute more knowledge to the central server.

Local epochs. The number of local epochs in FSAR represents the trade-off between communication cost and performance. Fig. 3 compares the test accuracy of local epochs $K = 1$, $K = 2$, and $K = 3$ with a total of 300 training rounds, where the larger K potentially promotes the communication efficiency and reduces the communication cost. The results demonstrate that the performance of FSAR gradually decreases when K increases, which is even more noticeable in the small-scale dataset (*e.g.* PKU MMD I). This is caused by the accumulation of bias in each local client, which further indicates the trade-off between performance and communication cost in FSAR.

Parameters Sensitivity. Table 4 shows the model sensitivity of our FSAR to the hyperparameters τ and ξ. Parameter ξ controls the acceleration toward momentum of the central server model and local client models, and parameter τ is the server learning rate. Both parameters control the impact degree of the previous global updates of the central model on the current client-server transmission. Table 4 illustrates that the larger τ or ξ is not friendly to small-scale datasets (*e.g.* PKU MMD I and PKU MMD II), since it imposes oscillation for the global optimization procedure. $\tau = 0.8$ and $\xi = 0.6$ are thus set as default to balance the accuracy across clients.



(a) NTU 60 (b) PKU I

Figure 1. Illustration of the different effects of IM and UM. IM is shared across clients to improve the stability of FL training. UM is client-specific to prevent current clients from being affected by other clients with different dataset scales.

KD Loss. For the proposed MKD, the knowledge distillation is jointly trained with the classification. To impose better supervision on the model, we remove the Cross-Entropy loss, which may cause interference with KD loss, and optimize the model by only the KD loss. The results in Table 5 reflect the interference is minor.

Results for Cross-View settings. We also evaluate the performance of FSAR under different dataset evaluation metrics. Specifically, NTU datasets have two standard evaluation metrics: Cross-View and Cross-Subject. The training and test sets are divided based on the views of cameras and person identities for these two metrics, respectively. To validate the effectiveness of FSAR, we report the additional results on Corss-View metrics in Table 6.

4. Additional Visualization

Confusion Matrix. We supplement the visualization of the confusion matrix of the IM and UM in Figs. 4 to 7, to compare their variations under different ablation study settings: 1) accession of IM only, 2) accession of UM only, and 3) accession of both IM and UM. Without loss of generality, we visualize the matrices in PKU I, NTU 60, and UESTC datasets for clarity. There exist huge variations of similarities in UM between clients at the beginning of training (R1) and at the end of training (R300), while slight for IM. These results are consistent with our conclusions in Sec 4.4, which further indicates the rationality of our FSAR.

Different effects of IM and UM. Apart from the manually set and static matrix, the two matrices, IM and UM, collaborate with each other, but each has its own role to play. As illustrate in Fig. 1, IM is shared across clients to learn domain-invariant topology, which improves the stability of FL training (Fig. 1 UM is further designed to maintain client-specific topology to prevent current clients from being affected by other clients with different dataset scales. Unlike IM, updated with FL strategy, UM updates its parameters individually without communication or aggregation at the server. In other words, parameters of IM participate in the client-server communication, while parameters

Table 2. The effect of different loss weights on performance with respect to test accuracy (%). Here, λ_1, λ_2 and λ_3 are the weights of loss $\mathcal{L}_{CE}$, loss $\mathcal{L}_{KD}$ and loss $\mathcal{L}_{Reg}$, respectively. Even though other combinations bring gains, we set $\lambda_1 = 1$, $\lambda_2 = 1$, and $\lambda_3 = 0.1$ as default to make our FSAR more general.

Settings			PKU MMD I		PKU MMD II		NTU RGB+D 60		NTU RGB +D 120		UESTC	
λ_1	λ_2	λ_3	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ
1	1	0.1	81.96	-	56.30	-	91.30	-	84.31	-	91.88	-
1	5	0.1	**86.36**	+4.40	54.60	-1.70	**93.53**	+2.23	85.79	+1.48	90.22	-1.66
1	0.5	0.1	83.95	+1.99	56.61	+0.31	92.72	+1.42	85.27	+0.96	91.84	-0.04
1	1	0.5	81.27	-0.69	**56.97**	+0.67	91.10	-0.20	**85.92**	+1.61	92.67	+0.79
1	1	0.05	82.48	+0.52	55.10	-1.20	92.11	+0.81	83.24	-1.07	**93.01**	+1.13

of UM is kept in local for each client.

5. Proof of Privacy

Differential privacy [4, 3, 2] is a privacy-preserving technology that aims to protect sensitive data by adding noises. It constitutes a strong standard for privacy guarantees for algorithms on aggregate datasets. This standard is defined in terms of the concept of adjacent databases, which is specific to each application. For example, in our experiments, each training dataset consists of a collection of video-label pairs. Two of these sets are considered adjacent if they differ by only one entry, which means that one video-label pair is present in one set but absent in the other.

Definition 1. Randomized mechanism $\mathcal{A} : \mathcal{D} \rightarrow \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies $\varepsilon-$ differential privacy if for any two inputs from two adjacent domains, namely $d \in \mathcal{D}$, $d' \in \mathcal{D}'$. For any subset of outputs $\mathcal{O} \subseteq \mathcal{R}$, it satisfies the following formula:

$$\exp(-\varepsilon) \leq \frac{\Pr[\mathcal{A}(d) = \mathcal{O}]}{\Pr[\mathcal{A}(d') = \mathcal{O}]} \leq \exp(\varepsilon). \tag{4}$$

Specifically, suppose $\mathcal{A}$ is our model, $\mathcal{D}$ is one of the datasets that participated in FL training, and $\mathcal{D}'$ is the neighboring dataset of $\mathcal{D}$ (where samples of one random category are replaced with data from other datasets). We plot the probability distribution curves in Fig. 2 and find these two curves are close ($\varepsilon = 0.1$). It satisfies the inequality in Eq. (4), namely, meets the $\varepsilon-$ differential privacy preservation. Additionally, for ε-differential privacy preservation analysis [1, 6], our FSAR provides stronger privacy, with $\varepsilon = 0.1$, compared to centralized methods with $\varepsilon = 0.7$.

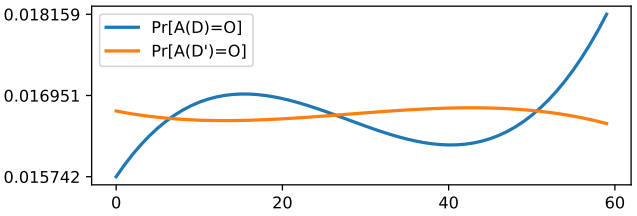


Figure 2. Comparison of the Probability Distribution Curves, when the model $\mathcal{A}$ is evaluated on $\mathcal{D}$ and $\mathcal{D}'$ (take NTU 60 as an example). $\varepsilon = 0.1$ can be calculated following Eq. (4).

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (ACM CCS)*, pages 308–318, 2016. 3

[2] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Proceedings of the Annual International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*, pages 486–503. Springer, 2006. 3

[3] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Theory of Cryptography Conference (TCC)*, pages 265–284, 2006. 3

[4] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014. 3

[5] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. A large-scale RGB-D database for arbitrary-view human action recognition. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 1510–1518, 2018. 1

[6] Honglu Jiang, Jian Pei, Dongxiao Yu, Jiguo Yu, Bei Gong, and Xiuzhen Cheng. Applications of differential privacy in social network analysis: A survey. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 35(1):108–127, 2023. 3

[7] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 1

[8] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 42(10):2684–2701, 2019. 1

[9] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016. 1

Table 3. The effect of *Selected Clients* S on performance with respect to test accuracy (%), which is the number of clients participating in aggregation in each communication round ($S = 5$ is set as default). For large-scale datasets (*e.g.* NTU RGB+D 120), the larger S brings more gains. While for small datasets (*e.g.* PKU MMD I), the smaller S bring more gains.

Settings	PKU MMD I		PKU MMD II		NTU RGB+D 60		NTU RGB +D 120		UESTC	
Selected Clients S	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ
5	81.96	-	56.30	-	91.30	-	84.31	-	91.88	-
4	83.94	+1.98	56.10	-0.20	89.91	-1.39	**83.98**	-0.33	92.80	+0.92
3	85.74	+3.78	55.13	-0.17	**90.09**	-1.21	83.42	-0.89	92.69	+0.81
2	86.06	+4.10	55.98	-0.32	88.07	-3.23	81.26	-3.05	92.24	+0.36
1	**86.84**	+4.88	**58.97**	+2.67	89.04	-2.26	78.44	-5.87	**93.87**	+1.99

Table 4. The effect of different hyperparameter settings τ and ξ on performance with respect to test accuracy (%), which are defined in Algorithm 1 for model parameters aggregation ($\tau = 0.8$ and $\xi = 0.8$ are set as default).

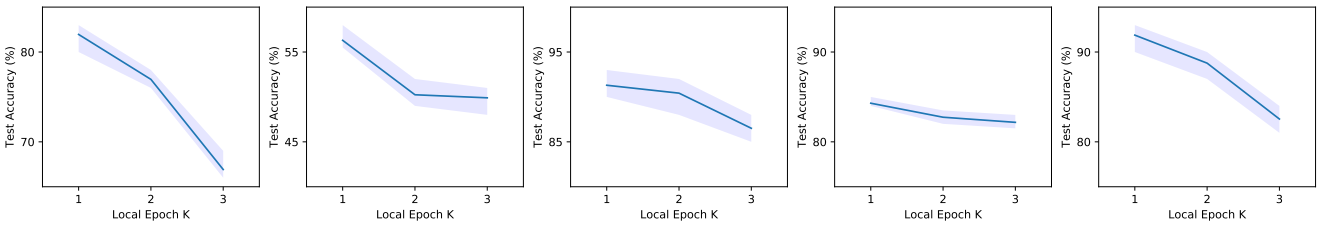
Settings		PKU MMD I		PKU MMD II		NTU RGB+D 60		NTU RGB +D 120		UESTC	
τ	ξ	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ
0.8	0.8	81.96	-	56.30	-	**91.30**	-	**84.31**	-	91.88	-
0.8	0.6	82.21	+0.25	57.87	+1.57	90.10	-1.20	80.70	-3.61	90.67	-1.21
0.6	0.8	82.36	+0.40	54.77	-1.53	87.70	-3.60	78.89	-5.42	90.62	-1.26
0.4	0.8	**84.07**	+2.11	59.59	+3.29	87.97	-3.33	79.26	-5.05	**92.29**	+0.41
0.8	0.4	82.52	+0.56	**60.12**	+3.82	86.23	-5.07	79.73	-4.58	91.80	-0.08

Table 5. The effect of *KD Loss* on performance with respect to test accuracy (%). The elimination of KD loss reduces the performance of the model. Compared with traditional CE loss for classification, KD loss alleviates the drift between the clients and the server.

Models	PKU MMD I		PKU MMD II		NTU RGB+D 60		NTU RGB +D 120		UESTC	
	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ
w KD loss (FSAR)	81.96	-	56.30	-	91.30	-	84.31	-	91.88	-
w/o KD loss	80.76	-1.20	56.20	-0.33	88.84	-2.46	82.75	-1.56	89.90	-1.98

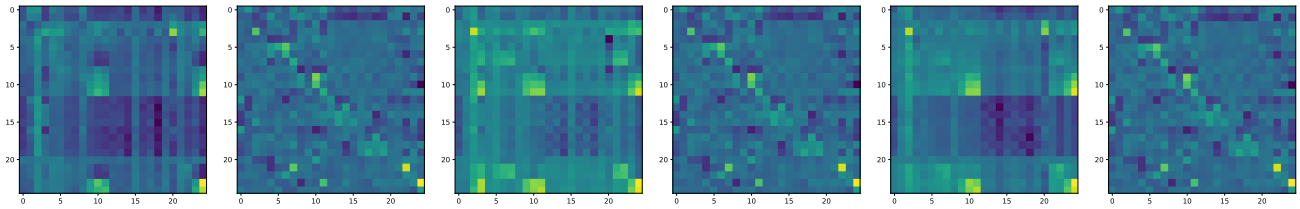
Table 6. The performance of FSAR on different evaluation metrics. PKU MMD I (Cross-View), PKU MMD II (Cross-View), NTU 60 (Cross-View), NTU 120 (Cross-Setup) are chosen, respectively.

Models	PKU MMD I		PKU MMD II		NTU RGB+D 60		NTU RGB +D 120		UESTC	
	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ	acc.	Δ
Vanilla FSAR	79.93	-	50.19	-	79.12	-	76.11	-	83.39	-
FSAR	**83.39**	+3.66	**57.03**	+6.84	**88.36**	+9.24	**84.72**	+8.61	**92.86**	+9.47



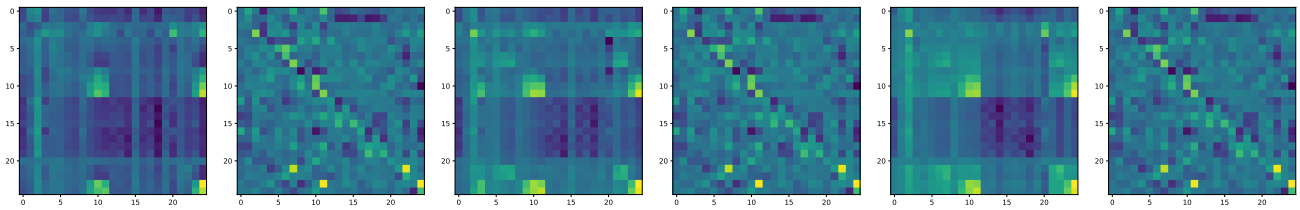
(a) PKU I (b) PKU II (c) PKU NTU 60 (d) NTU 120 (e) UESTC

Figure 3. The effect of *Local Epoch* K on performance with respect to test accuracy (%), which is the number of epochs for clients local training in each communication round ($K = 1$ is set as default). The larger *Local Epoch* K is detrimental to performance, especially on small-scale datasets (*e.g.* PKU MMD I and UESTC), since it brings more bias in clients.



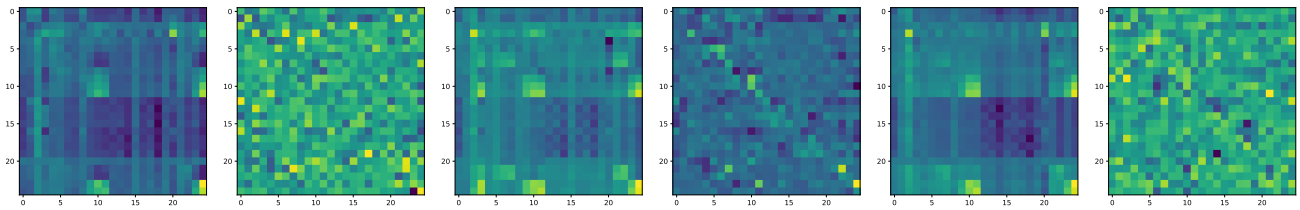
(a) PKU I, R1 (b) PKU I, R300 (c) NTU 60, R1 (d) NTU 60, R300 (e) UESTC, R1 (f) UESTC, R300

Figure 4. Visualization of IM under the FSAR (A + IM + UM) settings for PKU MMD I, NTU RGB+D 60, and UESTC datasets, at the beginning of the training (R1) and at the end of the training (R300).



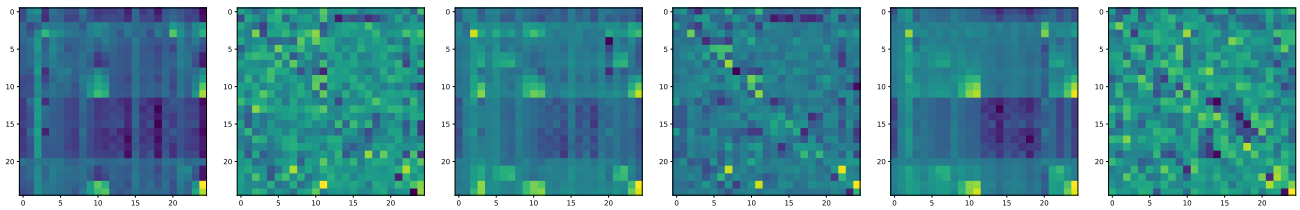
(a) PKU I, R1 (b) PKU I, R300 (c) NTU 60, R1 (d) NTU 60, R300 (e) UESTC, R1 (f) UESTC, R300

Figure 5. Visualization of IM under the FSAR (A + IM) settings for PKU MMD I, NTU RGB+D 60, and UESTC datasets, at the beginning of the training (R1) and at the end of the training (R300).



(a) PKU I, R1 (b) PKU I, R300 (c) NTU 60, R1 (d) NTU 60, R300 (e) UESTC, R1 (f) UESTC, R300

Figure 6. Visualization of UM under the FSAR (A + IM + UM) settings for PKU MMD I, NTU RGB+D 60, and UESTC datasets, at the beginning of the training (R1) and at the end of the training (R300).



(a) PKU I, R1 (b) PKU I, R300 (c) NTU 60, R1 (d) NTU 60, R300 (e) UESTC, R1 (f) UESTC, R300

Figure 7. Visualization of UM under the FSAR (A + UM) settings for PKU MMD I, NTU RGB+D 60, and UESTC datasets, at the beginning of the training (R1) and at the end of the training (R300).