A Cross-Modal Knowledge Distillation Approach for RGB-to-Infrared Video Action Recognition

Zhenzhen Quan$^{1,2[0000-0002-3981-7128]}$, Daisuke Deguchi$^{2[0000-0003-0603-8790]}$, Jialei Chen$^{2[0009-0005-0654-4281]}$, Chenkai Zhang$^{2[0000-0002-7258-272X]}$, Yujun Li$^{1,*[0000-0003-4455-5991]}$, Seigo Ito$^{2[0009-0003-8544-0135]}$, and Hiroshi Murase$^{2[0000-0002-8103-9294]}$

[1] Shandong University, Qingdao 266237, Shandong, China
[2] Nagoya University, Nagoya 464-8601, Japan
{quan.zhenzhen.g7, chen.jialei.s6,
zhang.chenkai.d4}@s.mail.nagoya-u.ac.jp
{ddeguchi, murase}@nagoya-u.jp
liyujun@sdu.edu.cn
iseigo@vislab.is.i.nagoya-u.ac.jp
$*$ Corresponding author

Abstract. In the domain of human action recognition (HAR), infrared data has emerged as a pivotal sensing technology in robotic applications due to its robust performance under low-light or rapidly changing lighting conditions. However, HAR methods that depend exclusively on infrared data fall short due to the lack of color and texture features. Combining other modal data, such as RGB modality, can alleviate this problem. In this paper, we introduce a cross-modal knowledge distillation approach that achieves knowledge transfer by using a teacher network with RGB data input to guide the recognition of an infrared data student network. The RGB data is exclusively utilized during the training phase. To make full use of RGB information, firstly, we construct a multi-scale graph cross-attention module between different convolutional layers of the teacher and student networks to reduce the modality difference between infrared data and RGB data modalities. Secondly, we employ a decoupled knowledge distillation module (DKD) to focus on more dark knowledge, i.e., knowledge related to similar behaviors, thereby enhancing the network's robustness. We prove the effectiveness of the proposed approach on two datasets, i.e., NTU RGB+D and PKU-MMD datasets, providing strong support for the intelligent behavior of robots in various environments.

Keywords: Human action recognition · RGB videos · Infrared videos.

1 Introduction

As robotics technology rapidly progresses, human action recognition (HAR) systems are being increasingly adopted in areas such as security monitoring, medical care, and human-computer interaction [4],[9],[10]. By recognizing and understanding human actions, robots can carry out tasks with enhanced intelligence and autonomy. However, traditional HAR systems predominantly depend on RGB data, which offers rich visual information under optimal lighting conditions. Yet, their performance deteriorates significantly in low-light or rapidly changing lighting environments, thereby restricting the system's range of applications. Infrared data, i.e., near infrared data, serves as an excellent alternative or supplement since it relies on thermal radiation information and is unaffected by visible light conditions. This allows it to operate reliably across a wide range of lighting environments.

However, HAR methods that depend exclusively on infrared data also have limitations. In scenarios where color and texture features are essential, infrared images sometimes fail to accurately analyze actions due to their absence of these features. Furthermore, the acquired infrared images frequently suffer from coarse resolution due to the intrinsic noise interference of the infrared image receiver. In contrast, RGB data provides clear textures and high-quality images, addressing the shortcomings of infrared images in human action recognition. By combining RGB and infrared data, the complementary information from both modalities can guide infrared action recognition and enhance the performance.

Therefore, we raise a cross-modal knowledge distillation approach that utilizes a teacher-student network to transfer cross-modal knowledge from an RGB data teacher network to an infrared data network. Notably, the RGB data is involved solely during the training phase. To make RGB data better guide infrared data recognition, firstly, we construct a multi-scale graph cross-attention module (MGCAM) in different intermediate convolutional layers to reduce the modality difference between infrared data and RGB data modalities and obtain the fused information between heterogeneous data. Then, we employ a decoupled knowledge distillation module (DKD) to obtain the output layer information of the teacher and student networks according to its relevance to the target, focusing on more dark knowledge and improving the network's generalization capabilities.

The primary contributions of this work are as follows:

1) We construct a multi-scale graph cross-attention module across various intermediate convolutional layers to learn similar features from different modalities.

2) We construct teacher and student networks and use decoupled knowledge distillation to transfer complementary information from the RGB modality.

3) We showcase the efficacy and viability of our approach via experimental validation on both the NTU RGB+D and PKU-MMD datasets.

The structure of this paper is outlined as follows: Section II provides a review of related work. Section III introduces the raised approach. Section IV details datasets, experimental setup, and results. Lastly, Section V concludes the paper.

2 Related Work

2.1 Action Recognition based on RGB and Infrared Data

In recent years, there has been a growing interest among researchers in the field of HAR, largely driven by advancements in deep learning. Extensive research has been conducted not only in HAR based on RGB data [1],[12],[25],[26] but also in infrared human action recognition [11],[14]. Nie et al. [16] introduced an innovative 3D ConvNet featuring a deep architecture and residual structure to enhance the classification performance of infrared videos. Meng et al. [3] utilized convolutional neural networks to develop a framework for analyzing human behavior around parked aircraft. Additionally, there are also some studies based on both RGB and infrared data. Zhu et al. [29] explored a feature mapping method for the same action from the visual spectrum to the infrared range. Hilsenbeck et al. [7] recorded infrared and visible spectra through Hough forest and constructed a multispectral behavioral dataset. Sun et al. [22] proposed a robust feature matching strategy utilizing feature matching and multi-object tracking in complex, non-flat environments of infrared-RGB videos. Piao et al. [19] introduced the use of convolutional neural networks (CNNs) to create weight maps expressing the significance of individual pixels, which are then used for weighted fusion in action analysis. Quan et al. [20] employed attention relation matching and activation domain consistency constraints to minimize the modality differences between RGB and infrared data. Unlike the aforementioned methods, we consider constructing graph cross-view attention map to learn similar features across different modalities.

2.2 Knowledge Distillation

Knowledge distillation has received widespread attention due to its ability to transfer knowledge from larger networks to smaller ones [8]. Park et al. [17] considered both distance and angular losses to reduce the differences between different data structures. Tung et al. [24] focused solely on student networks learning the representational space information of teacher networks rather than the representational information. Touvron et al. [23] created distillation tokens leveraging the transformer architecture to enable knowledge distillation. Zhao et al. [28] decomposed the knowledge distillation loss based on its relevance to the target and introduced a decoupled knowledge distillation method. Gou et al. [5] examined the cross-channel features of different samples as well as the various channels of a single sample. Guo et al. [6] highlighted the significance of class attention transfer in enhancing the performance of convolutional networks. Unlike previous knowledge distillation methods, we raise MGCAM across various feature layers to transfer informations and DKD to focus on more dark knowledge.

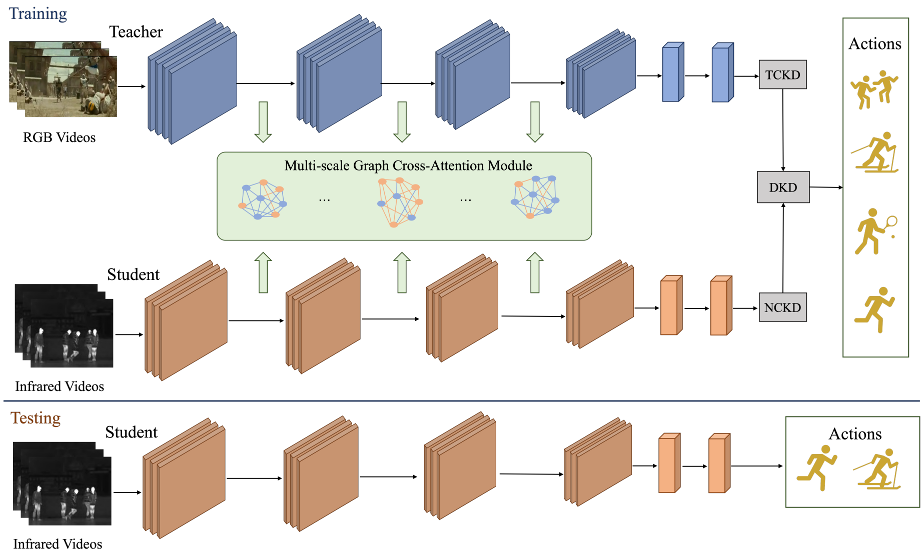


Fig. 1. The framework of the raised approach. Where TCKD, NCKD, and DKD denote target class knowledge distillation, non-target class knowledge distillation, and decoupled knowledge distillation. Notably, during the testing stage, only the student network is utilized.

3 Methodology

3.1 Overview

The schematic of our raised approach is depicted in Fig. 1. In our approach, the teacher network (TN) takes RGB video as input, whereas the student network (SN) takes infrared video as input. And the TN is involved solely during the training phase, while only the SN is utilized during the testing phase. To make full use of RGB information, we construct the MGCAM to merge features from the intermediate layer of both networks, allowing the integration of different modalities at the intermediate layers. Moreover, we leverage the DKD module to extract both positive and negative information from the TN, effectively guiding the SN in action analysis.

3.2 Multi-scale Graph Cross-Attention Module

To reduce the discrepancies between the modalities of RGB data and infrared data, we consider the information fusion between feature layers of different scales from both modalities. Therefore, we raise MGCAM, as shown in Fig. 2. By constructing graph structures of different modal features, we utilize graph-based information to perform cross-view attention calculations, thus capturing the intermodal information.

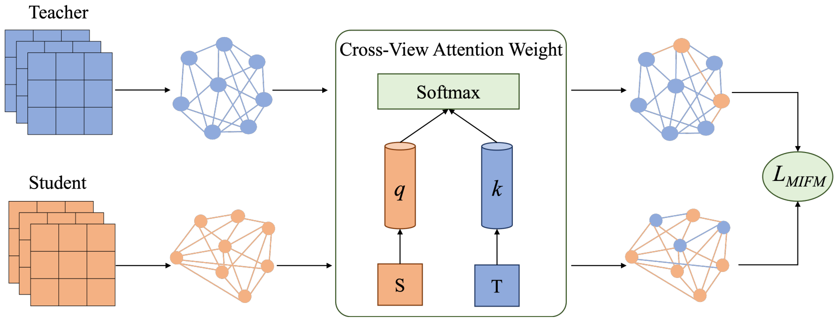


Fig. 2. The design of the multi-scale graph cross-attention module.

We assume that the ith feature layer of the SN is $F_i^S \in \mathbb{R}^{b\times c\times h\times w}$, where b denotes the batch size, c represents the channel dimension, and h, w are the height and width of the spatial dimension. The feature layer j of the TN is $F_j^T \in \mathbb{R}^{b\times c\times h\times w}$. Nie et al. [15] demonstrated that manifold learning uses dimensionality reduction to extract the intrinsic structure of the data and reduce the vibration of high-dimensional data. Inspired by [15], we construct graphs representing the feature F_i^S from the ith feature layer of the SN and the feature F_j^T from the jth feature layer of the TN.The features extracted from the feature layers are viewed as vertices in a graph, where the connections between features obtained by Gaussian similarity define the edges of the graph. We calculate the edge weights $E_{ij} \in \mathbb{R}^{b\times b}$ between feature data through gaussian similarity:

$$E_{ij} = \exp\left(-\frac{\left\|F_i^S - F_j^T\right\|^2}{2}\right). \tag{1}$$

And then we get the graph matrix $Q \in \mathbb{R}^{b\times b}$ through graph Laplacian normalization [2].

$$Q = D^{-\frac{1}{2}} E_{ij} D^{-\frac{1}{2}}, \tag{2}$$

$D \in \mathbb{R}^{b\times b}$ is a diagonal matrix, with each diagonal element (n, n) representing the sum of the corresponding row n in the edge weight matrix.

Then, we use the graph matrices of SN and TN to calculate cross-view attention weights A_{ij}^S, which capture the similarity between diverse forms of knowledge:

$$A_{ij}^S = softmax\left(\frac{Q_i^S \left(Q_j^T\right)^T}{\sqrt{d}}\right), \tag{3}$$

where Q_i^S represents the graph matrix of the i-th feature layer of the SN, and Q_j^T represents the graph matrix of the j-th feature layer of the TN. Q_i^S and Q_j^T are obtained by Eq. (2).

The fused feature map $W_j^T \in \mathbb{R}^{b\times c\times h\times w}$ for the j-th feature layer of the TN is obtained by combining the feature activation map A_j with the cross-view attention weight A_{ij}^{S} using a linear operation.

$$W_j^T = A_{ij}^S A_j. \tag{4}$$

Similarly, we can get the fused feature map $W_i^S \in \mathbb{R}^{b\times c\times h\times w}$ for the i-th feature layer of the SN is:

$$W_i^S = A_{ij}^S A_i. \tag{5}$$

We calculate the mean square error (MSE) loss between the fused feature maps of the SN and those of the TN. This loss helps in learning the knowledge of the middle layers:

$$L_{MGCAM} = \frac{1}{k}\left(\sum_{i,j=1}^{k}\left\|\frac{W_j^T}{\left\|W_j^T\right\|_2} - \frac{{W_i}^S}{\left\|{W_i}^S\right\|_2}\right\|_2^2\right), \tag{6}$$

where $\|.\|_2^2$ is the MSE calculation, and k is the number of the feature layers for graph cross-attention computation. To obtain more fusion information of multi-scale features from RGB modality and infrared modality, we incorporate graph cross-attention computation into each residual block of both the TN and SN.

3.3 Decoupled Knowledge Distillation Module

During the distillation process, some information is relevant to the target, while some is not. Therefore, the distillation loss can also be decomposed in terms of its relevance to the target [28]:

$$L_{DKD} = \alpha_T L_{TCKD} + \alpha_N L_{NCKD}, \tag{7}$$

where α_T, α_N represent distinct hyperparameters. L_{TCKD} refers to target class knowledge distillation, while L_{NCKD} is non-target class knowledge distillation.

$$L_{TCKD} = KL\left(p^T \middle\| p^S\right), \tag{8}$$

$$L_{NCKD} = KL\left(\hat{p}^T \middle\| \hat{p}^S\right), \tag{9}$$

where $p = \left[p_m, p_{\backslash m}\right] \in \mathbb{R}^{1\times 2}$ denotes the binary probability p_m of the target class and the binary probabilities $p_{\backslash m}$ of all other non-target classes. $\hat{p} = [\hat{p}_1, \ldots, \hat{p}_{m-1}, \hat{p}_{m+1}, \ldots, \hat{p}_C] \in \mathbb{R}^{1\times(C-1)}$ is independently modeling the probabilities between non-target classes without considering the m-th class. After decoupling, the knowledge distillation loss can obtain more dark knowledge from nontarget classes by L_{NCKD}.

The total loss for training the SN comprises both the decoupled knowledge distillation loss and the cross-entropy loss L_{CE} between the SN's output and the ground truth label. The complete loss function is:

$$L = L_{CE} + \alpha L_{DKD} + \beta L_{MGCAM}, \tag{10}$$

where α, β are the hyperparameters, and L_{DKD} is the decoupled knowledge loss.

The steps of the raised approach are shown in Alg. 1.

Algorithm 1

```
INPUT: Action categories, teacher network, student network, the cross-entropy loss $L_{CE}, L_{CE}^T$, the decoubled knowledge distillation loss $L_{DKD}$, the multi-scale graph cross-attention loss $L_{MGCAM}$, learning rate, total iterations and hyperparameters $\alpha, \beta, \alpha_T, \alpha_N$
OUTPUT: Student network
INITIALIZE: Learning rate, total iterations and hyperparmeters $\alpha, \beta, \alpha_T, \alpha_N$
PRETRAIN TEACHER NETWORK:
    for $i \leftarrow 0$ to total iterations do
        train the teacher network with $L_{CE}^T$
    update: The teacher network parameters
    end for
TRAIN STUDENT NETWORK:
    for $j \leftarrow 0$ to total iterations do
        train the student network based on $L = L_{CE} + \alpha L_{DKD} + \beta L_{MGCAM}$
    update: The student network parameters
    end for
```

4 Experiments

4.1 Datasets

1) The NTU RGB+D dataset [21] comprises various data types, such as RGB, skeleton, depth map, and infrared data. It encompasses a total of 60 action classes and contains over 56,000 videos. In this paper, we select ten daily-life related categories (drop, throw, stand up, pick up, clap, brush teeth, drink, wash hair, sit, and eat) to evaluate our approach, as shown in Fig. 3.

2) The PKU-MMD dataset [13] contains data modalities such as RGB, infrared data, depth, and 3D joint. It consists of 51 action categories and includes over 1,000 lengthy videos. For our evaluation, we focus on all RGB and infrared data categories, assessing our approach's effectiveness using the mean average precision (mAP) across cross-view and cross-subject settings.

4.2 Experimental Setup

The teacher network employs ResNet50 as its backbone, whereas the student network uses ResNet18 as its backbone. For both datasets, the segment is 8. The batch size of the pretrained teacher network is 16, while that of the student network is set to 8. For the NTU RGB+D dataset, the learning rate is 0.001, the hyperparameters $\alpha, \beta, \alpha_T, \alpha_N$ values are $\{0.1, 1, 0.7, 0.35\}$ and the number of epoch is 80. Similarly, for the PKU-MMD dataset, the learning rate is 0.001, the values of the hyperparameters $\{\alpha, \beta, \alpha_T, \alpha_N\}$ are $\{0.01, 1, 0.7, 0.35\}$ and the number of epoch is 120. All experimental procedures are carried out using the PyTorch framework on a single NVIDIA RTX 3090 GPU.

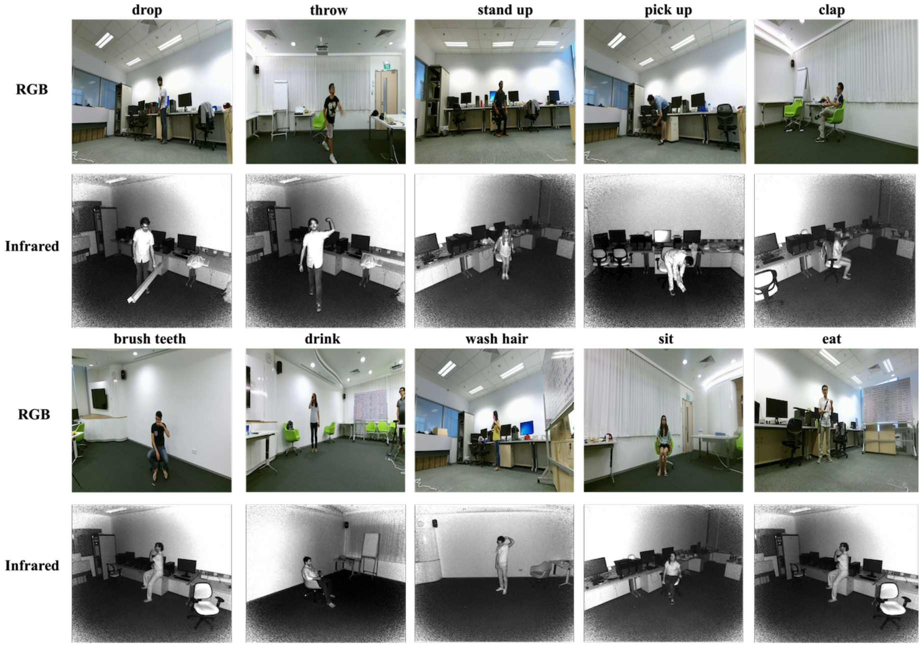


Fig. 3. The examples of NTU RGB+D dataset.

Table 1. The results of the raised approach and the existing methods on the NTU RGB+D dataset.

Method	Teacher	Student	Teacher Input	Student Input	Accuracy (%)
TSTDDs [14]	-	-	-	IR Videos	72.36
ST [8]	ResNet50	ResNet18	RGB Videos	IR Videos	74.69
AT [27]	ResNet50	ResNet18	RGB Videos	IR Videos	74.58
SP [24]	ResNet50	ResNet18	RGB Videos	IR Videos	76.03
CC [18]	ResNet50	ResNet18	RGB Videos	IR Videos	75.21
RKD [17]	ResNet50	ResNet18	RGB Videos	IR Videos	74.38
CCSKD [5]	ResNet50	ResNet18	RGB Videos	IR Videos	76.24
Ours	ResNet50	ResNet18	RGB Videos	IR Videos	**80.86**

4.3 Comparison with Existing Methods

To assess the performance of our raised approach, we benchmark it against state-of-the-art approaches such as TSN [25], TSTDDs [14], ST [8], AT [27], SP [24], CC [18], RKD [17], CCSKD [5] on the NTU RGB+D and PKU-MMD datasets.

As illustrated in Table 1 and 2, our approach outperforms existing methods on both datasets. Our approach achieves an accuracy of 80.86% on the NTU RGB+D dataset, 4.62% higher than CCSKD [5] method. On the PKU-MMD dataset, our approach achieves mean Average Precision (mAP) scores of 0.588 and 0.586 for the cross-view and cross-subject strategies, respectively. These

Table 2. The results of the raised approach and the existing methods on the PKU-MMD dataset.

Method	Teacher Input	Student Input	Cross-View	Cross-Subject
TSN [25]	-	IR Videos	0.539	0.567
ST [8]	RGB Videos	IR Videos	0.554	0.574
AT [27]	RGB Videos	IR Videos	0.562	0.584
SP [24]	RGB Videos	IR Videos	0.582	0.570
CC [18]	RGB Videos	IR Videos	0.580	0.567
RKD [17]	RGB Videos	IR Videos	0.579	0.578
CCSKD [5]	RGB Videos	IR Videos	0.556	0.565
Ours	RGB Videos	IR Videos	**0.588**	**0.586**

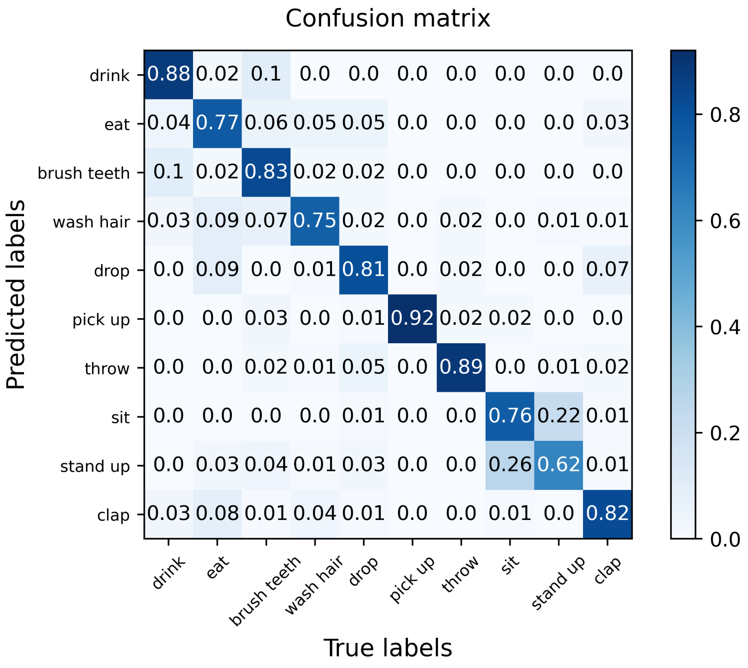
findings suggest that our approach, by combining the MGCAM with the DKD, successfully mitigates the modality discrepancies and utilizes valuable knowledge from various teacher network feature layers to enhance the performance.

Moreover, we showcase the confusion matrix generated from the NTU RGB+D dataset, depicted in Fig. 4. According to Fig. 4, the recognition performance for the 10 selected actions from the NTU RGB+D dataset notably improves with the aid of RGB data. Actions like 'pick up', 'throw', and 'drink' show significantly enhanced recognition performance. However, the action 'stand up' shows relatively lower recognition rate. Among these, actions like 'pick up' and 'throw', known for their extensive motion ranges, exhibit distinct characteristics that aid in their identification. Unfortunately, there is a 22% probability of misclassification, where 'sit' is erroneously categorized as 'stand up'. This confusion arises because these two actions are opposites and have a similar intermediate phase, leading to ambiguity.

4.4 Ablation Study

To evaluate the contributions of the MGCAM and DKD in our approach, we constructed 3 variations based on our raised approach: 1) Student network baseline: The student network is used exclusively, with training and testing conducted using only infrared data, shown in the 2nd line of Table 3 and Table 4. 2) Ours (w/o MGCAM): Our raised approach without the multi-scale graph cross-attention module, shown in the 3rd line of Table 3 and Table 4. 3) Ours (w/o DKD): Our raised approach without the decoupled knowledge distillation module, shown in the 4th line of Table 3 and Table 4.

As depicted in Table 3 and Table 4, the MGCAM and DKD modules play crucial roles in our approach's performance. From the data presented in the 2nd and 3rd rows of Table 3 and Table 4, it is evident that MGCAM plays a crucial role. This indicates that bridging the modality gap between infrared data and RGB data is crucial. Additionally, both Ours (w/o MGCAM) and Ours (w/o DKD) achieve better performance than the student network baseline. This demonstrates that the MGCAM and DKD are helpful for information fusion



	drink	eat	brush teeth	wash hair	drop	pick up	throw	sit	stand up	clap
drink	0.88	0.02	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
eat	0.04	0.77	0.06	0.05	0.05	0.0	0.0	0.0	0.0	0.03
brush teeth	0.1	0.02	0.83	0.02	0.02	0.0	0.0	0.0	0.0	0.0
wash hair	0.03	0.09	0.07	0.75	0.02	0.0	0.02	0.0	0.01	0.01
drop	0.0	0.09	0.0	0.01	0.81	0.0	0.02	0.0	0.0	0.07
pick up	0.0	0.0	0.03	0.0	0.01	0.92	0.02	0.02	0.0	0.0
throw	0.0	0.0	0.02	0.01	0.05	0.0	0.89	0.0	0.01	0.02
sit	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.76	0.22	0.01
stand up	0.0	0.03	0.04	0.01	0.03	0.0	0.0	0.26	0.62	0.01
clap	0.03	0.08	0.01	0.04	0.01	0.0	0.0	0.01	0.0	0.82

Fig. 4. The confuse matrix of NTU RGB+D dataset.

Table 3. The results on the NTU RGB+D dataset. w/o stands for without.

MGCAM	DKD	Teacher Input	Student Input	Accuracy (%)
-	-	-	IR Videos	71.20
-	✓	RGB Videos	IR Videos	78.60
✓	-	RGB Videos	IR Videos	80.04
✓	✓	RGB Videos	IR Videos	**80.86**

Table 4. The results on the PKU-MMD dataset. w/o stands for without, CV stands Cross-View and CS stands for Cross-Subject.

MGCAM	DKD	Teacher Input	Student Input	CV	CS
-	-	-	IR Videos	0.539	0.567
-	✓	RGB Videos	IR Videos	0.557	0.571
✓	-	RGB Videos	IR Videos	0.575	0.574
✓	✓	RGB Videos	IR Videos	**0.588**	**0.585**

of different modalities and extraction of complementary knowledge to enhance infrared data analysis.

5 Conclusion

In this paper, we raise a decoupled knowledge distillation approach using graph cross-attention. To reduce the modality gap between infrared data and RGB data modalities, we construct a multi-scale graph cross-attention module that operates across various convolutional layers of both teacher and student networks, enabling the learning of analogous features across modalities. Additionally, to enhance the network's robustness by leveraging more dark knowledge, we employ a decoupled knowledge distillation loss. The effectiveness of our approach is validated on two datasets.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number 23K28164 and 24H00733, and JST CREST Grant Number JPMJCR22D1. We employed the NTU RGB+D Action Recognition Dataset, which was developed by the ROSE Lab at Nanyang Technological University in Singapore.

References

1. Chen, T., Yu, H., Yang, Z., Li, Z., Sun, W., Chen, C.: Ost: Refining text knowledge with optimal spatio-temporal descriptor for general video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18888–18898 (2024). https://doi.org/10.48550/arXiv.2312.00096
2. Chung, F.R.: Spectral graph theory, vol. 92. American Mathematical Soc. (1997)
3. Ding, M., Ding, Y., Wei, L., Xu, Y., Cao, Y.: Individual surveillance around parked aircraft at nighttime: Thermal infrared vision-based human action recognition. IEEE Trans. Syst., Man, Cybern.: Systems **53**(2), 1084–1094 (2022). https://doi.org/10.1109/TSMC.2022.3192017
4. Frank, A.E., Kubota, A., Riek, L.D.: Wearable activity recognition for robust human-robot teaming in safety-critical environments via hybrid neural networks. In: Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 449–454. IEEE (2019). https://doi.org/10.1109/IROS40897.2019.8968615
5. Gou, J., Xiong, X., Yu, B., Zhan, Y., Yi, Z.: Channel correlation-based selective knowledge distillation. IEEE Trans. Cogn. Dev. Syst. (2022). https://doi.org/10.1109/TCDS.2022.3232569
6. Guo, Z., Yan, H., Li, H., Lin, X.: Class attention transfer based knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11868–11877 (2023). https://doi.org/10.48550/arXiv.2304.12777
7. Hilsenbeck, B., Münch, D., Grosselfinger, A.K., Hübner, W., Arens, M.: Action recognition in the longwave infrared and the visible spectrum using hough forests. In: 2016 IEEE International Symposium on Multimedia (ISM). pp. 329–332. IEEE (2016). https://doi.org/10.1109/ISM.2016.0072
8. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015), http://arxiv. org/abs/1503.02531

9. Iqbal, T., Rack, S., Riek, L.D.: Movement coordination in human–robot teams: a dynamical systems approach. IEEE Trans. Robot. **32**(4), 909–919 (2016). https://doi.org/10.1109/TRO.2016.2570240
10. Iqbal, T., Riek, L.D.: Coordination dynamics in multihuman multirobot teams. IEEE Robot. Autom. Lett. **2**(3), 1712–1717 (2017). https://doi.org/10.1109/LRA.2017.2673864
11. Jiang, Z., Rozgic, V., Adali, S.: Learning spatiotemporal features for infrared action recognition with 3d convolutional neural networks. 2017. arXiv preprint arXiv:1705.06709 (2017), http://arxiv. org/abs/1705.06709
12. Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., Wang, L.: Tea: Temporal excitation and aggregation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 909–918 (2020). https://doi.org/10.48550/arXiv.2004.01398
13. Liu, C., Hu, Y., Li, Y., Song, S., Liu, J.: Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. arXiv preprint arXiv:1703.07475 (2017), http://arxiv. org/abs/1703.07475
14. Liu, Y., Lu, Z., Li, J., Yang, T., Yao, C.: Global temporal representation based cnns for infrared action recognition. IEEE Signal Process. Lett. **25**(6), 848–852 (2018). https://doi.org/10.1109/LSP.2018.2823910
15. Nie, F., Xu, D., Tsang, I.W.H., Zhang, C.: Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. IEEE Trans. Image Process. **19**(7), 1921–1932 (2010). https://doi.org/10.1109/TIP.2010.2044958
16. Nie, J., Yan, L., Wang, X., Chen, J.: A novel 3d convolutional neural network for action recognition in infrared videos. In: 2021 4th International Conference on Information Communication and Signal Processing (ICICSP). pp. 420–424. IEEE (2021). https://doi.org/10.1109/ICICSP54369.2021.9611896
17. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3967–3976 (2019). https://doi.org/10.48550/arXiv.1904.05068
18. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5007–5016 (2019)
19. Piao, J., Chen, Y., Shin, H.: A new deep learning based multi-spectral image fusion method. Entropy **21**(6), 570 (2019). https://doi.org/10.3390/e21060570
20. Quan, Z., Chen, Q., Li, Y., Liu, Z., Cui, Y.: Arctic: A knowledge distillation approach via attention-based relation matching and activation region constraint for rgb-to-infrared videos action recognition. Comput. Vis. and Image Und. **237**, 103853 (2023). https://doi.org/10.1016/j.cviu.2023.103853
21. Shahroudy, A., Liu, J., Ng, T.T., Wang, G.: Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1010–1019 (2016)
22. Sun, X., Xu, T., Zhang, J., Zhao, Z., Li, Y.: An automatic multi-target independent analysis framework for non-planar infrared-visible registration. Sensors **17**(8), 1696 (2017). https://doi.org/10.3390/s17081696
23. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning. pp. 10347–10357. PMLR (2021)
24. Tung, F., Mori, G.: Similarity-preserving knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1365–1374 (2019)

25. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European Conference on Computer Vision. pp. 20–36. Springer (2016)
26. Wu, W., Sun, Z., Ouyang, W.: Revisiting classifier: Transferring vision-language models for video recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2847–2855 (2023). https://doi.org/10.1609/aaai.v37i3.25386
27. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In: Proceedings of the International Conference on Learning Representations. pp. 1–13 (2017). https://doi.org/10.48550/arXiv.1612.03928
28. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11953–11962 (2022). https://doi.org/10.48550/arXiv.2203.08679
29. Zhu, Y., Guo, G.: A study on visible to infrared action recognition. IEEE Signal Process. Lett. **20**(9), 897–900 (2013). https://doi.org/10.1109/LSP.2013.2272920