# RightFeatKD: Selective Feature-based Knowledge Distillation

Syed Tousiful Haque[1], Yan Yan[2], and Anne Hee Hiong Ngu[1]

[1] Texas State University, San Marcos Texas, USA
{bgu9,angu}@txstate.edu
[2] University of Illinois Chicago, IL, USA
yyan55@uic.edu

**Abstract.** Knowledge distillation (KD) has emerged as a widely adopted approach for cross-modal knowledge transfer, where transferring knowledge from a semantically rich teacher model to a lightweight student model enhances the performance of the student model. However, existing KD approaches rely on logit-based distillation, which suffers from limited logit diversity inherent in binary classification tasks such as fall detection. Moreover, existing KD methods transfer knowledge indiscriminately from both correct and incorrect predictions of the teacher model, which can undermine the reliability of the student model. We propose RightFeatKD, a novel feature-based distillation framework that captures rich semantic knowledge from the intermediate representations of the teacher model using a specialized loss function. Importantly, RightFeatKD focuses more on distilling knowledge from the teacher's correct predictions, ensuring reliable and meaningful supervision. Experimental evaluations of RightFeatKD with two public fall detection datasets SmartFallMM and Up-Fall respectively demonstrate 4.04% and 12.59% gain in F1-score over the baseline model. Furthermore, real-time evaluation of the fall detection model trained with RightFeatKD validates its practical utility.

**Keywords:** Multimodal Learning, Cross-Modal Learning, Knowledge Distillation, Fall Detection, Wearable Devices.

## 1 Introduction

Wearable systems that enable human activity recognition (HAR) by leveraging information from diverse visual (e.g., skeleton, RGB video) and non-visual (e.g., accelerometer, gyroscope) modalities hold significant societal value, particularly in addressing the needs of an aging population [22]. For example, an accurate wearable sensor-based HAR system (WSHAR) can detect falls and monitor fall risks, enabling older adults to live independently.

However, due to energy constraints, wearable sensors can only capture non-visual data, which can have limited accuracy for fall detection on its own. For example, models trained with the Up-Fall [17] dataset demonstrated that the model's performance with only IMU (inertial measurement unit) data had an

Table 1: Comparison of knowledge distillation methods. ✓ indicates the method uses this component; ✗ indicates it does not.

| Method | Logit-Based | Feature-Based | Weighting | Real-Time |
|---|---|---|---|---|
| LSTMKD [27] | ✓ | ✗ | ✗ | ✗ |
| PreFallKD [3] | ✓ | ✗ | ✗ | ✗ |
| FitNet [23] | ✗ | ✓ | ✗ | ✗ |
| LightHART [6] | ✓ | ✗ | ✗ | ✓ |
| DMFT [12] | ✓ | ✗ | ✗ | ✗ |
| **RightFeatKD (Ours)** | ✗ | ✓ | ✓ | ✓ |

F1-score of 70.31%.Another work [4], which combined data from three wearable sensor-based datasets, only achieved F1-score of 69.93%.

On the contrary, multimodal learning algorithms can improve fall detection performance by leveraging multiple sources of information for decision-making. Experimental results in [1] show that fusing image (visual) and IMU (non-visual) data can achieve an F1-score of 88% on the Up-Fall dataset. Despite these advantages, implementing them in wearable devices is challenging due to hardware limitations and the inability to acquire the visual modality continuously with on-body sensors.

Knowledge Distillation (KD) [7] has emerged as a promising technique for transferring knowledge to lightweight deployable wearable-based models from complex visual or multimodal-based models. In HAR involving vision modalities, KD has been effectively applied to address challenges like occlusion. For instance, the MMAct framework [11] leverages multimodal KD by integrating diverse sensor streams. Similarly, cross-modal knowledge distillation has been explored in [20] to transfer knowledge from multimodal teacher networks to unimodal student networks. In the domain of fall detection, distillation-based approaches have been proposed in [3] and [16] for both pre-fall impact and fall detection scenarios.

Despite these advancements, existing methods face key limitations in fall detection, a binary classification task where the teacher logit offers limited distributional richness, which is necessary for effective knowledge transfer. In addition, indiscriminately distilling knowledge from all teacher outputs without weighting correct predictions, can propagate errors to the student, and hinder their learning. Table 1 highlights that current fall detection distillation methods mostly rely on logit-based approaches and fail to prioritize the teacher's correct predictions during knowledge transfer.

To overcome these challenges, we propose RightFeatKD, a feature-based cross-modal knowledge distillation method tailored for fall detection. Unlike prior approaches, our method selectively assigns higher weights to the visual teacher's correct predictions during distillation and aligns the wearable student's intermediate representation with the teacher's rich internal representations. Additionally, our method employs a focal loss mechanism to emphasize

challenging samples—particularly those misclassified both by the teacher and student—improving robustness. This dual focus not only mitigates the drawbacks of logit-based KD in binary settings but also ensures that the student learns from the most informative and reliable cues.

We validate RightFeatKD in three stages. First, we train a baseline model of each modality without KD. Next, we evaluate RightFeatKD on the multimodal SmartFallMM and Up-Fall [17] datasets, against other state-of-the-art KD methods. Finally, we deploy the SmartFallMM-trained RightFeatKD model on a smartwatch running the SmartFall App [18] for real-world evaluation. In addition, we also perform ablation studies to further substantiate the effectiveness of our RightFeatKD method.

The main contributions of our paper are as follows: We propose **RightFeatKD**, a novel selective feature-based distillation method that captures rich semantic knowledge from the intermediate representations of the teacher model, emphasizing correct prediction using a specialized loss function. We demonstrate that the distilled fall detection model shows a performance gain of 4.04% on SmartFallMM and 12.59% on Up-Fall datasets. RightFeatKD also outperforms four other logit-based KD methods.We evaluate the RightFeatKD distilled student's fall detection model in the real world with five participants and show that its performance is consistently better than the non-distilled model.

## 2 Related Work

Knowledge distillation (KD) [7] transfers knowledge from high-capacity teacher models to compact student models, typically through softened logits. Beyond logit matching, recent work has explored richer forms of supervision, including intermediate representations [23, 26], attention maps [15, 10], and layer-wise activations across diverse domains such as computer vision [26], natural language processing [9], and human activity recognition [13, 5]. These advances have established KD as an effective and widely adopted paradigm for model compression and knowledge transfer.

Extending beyond shared input spaces, cross-modal distillation aims to transfer knowledge across heterogeneous modalities and domains. Recent work has explored such transfers in various settings, including audio-to-visual distillation [8] and cross-domain time-series distillation [25]. In human activity recognition (HAR), several approaches distill knowledge from vision-based modalities (e.g., RGB, skeleton) to wearable sensors. Methods such as VSKD [21] introduce explicit cross-modal alignment objectives, while PSKD [19] adopts progressive multi-teacher supervision strategies. Transformer-based approaches, including LightHART [6] and MMAct [11], further advance cross-modal representation learning. More recently, COMODO [2] employs contrastive learning to enrich student features with teacher-derived information.

KD has also been applied to deployment-constrained HAR tasks such as fall detection, where latency and reliability are critical. PreFallKD [3] combines logit distillation with focal loss to train lightweight CNNs for resource-constrained de-

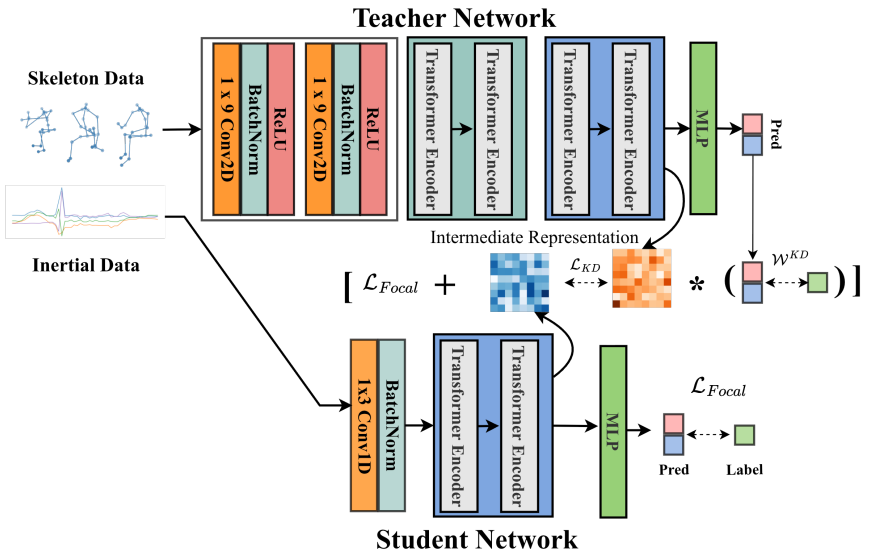



Fig. 1: RightFeatKD method. The teacher (top) processes skeleton data; the student (bottom) processes inertial data. Intermediate representation are aligned via distillation loss.

vices. MLMEC [16] utilizes multi-layer distillation to improve frontend networks with supervision from backend networks.

Despite these advances, existing KD approaches face two critical limitations. First, they struggle to bridge modality gaps in cross-modal settings, particularly for imbalanced tasks such as fall detection where logit distributions exhibit low entropy and minimal inter-class separation. Second, current methods indiscriminately transfer knowledge from all teacher predictions, including erroneous ones, thereby propagating noise and degrading student performance.

## 3 Selective Feature-based Knowledge Distillation

The architecture of RightFeatKD comprises three main components: the teacher network, the student network, and the RightFeatKD mechanism, as shown in Fig. 1. Code is available at `https://github.com/tousifulhaque/RightFeatKD`.

### 3.1 Teacher Network

We employ a Transformer-based [24] teacher network that learns a mapping from skeleton sequences to binary predictions. Formally, let $\mathbf{x}_t \in \mathcal{X} = \mathbb{R}^{T \times J \times C}$ denote the skeleton input for the teacher, where $T$, $J$, and $C$ represent the number of

timestamps, joints, and coordinate dimensions, respectively. The teacher network computes predictions as:

$$\hat{y}_t = f_t(g_t(\mathbf{x}_t)) \in \{0, 1\}, \tag{1}$$

where $g_t : \mathcal{X} \to \mathbb{R}^{T \times E}$ is the encoder that produces an intermediate representation $\mathbf{z}_t = g_t(\mathbf{x}_t)$, with $E$ denoting the encoder output dimension. The encoder adopts a hierarchical architecture that progressively extracts spatial and temporal features: it begins with two convolutional blocks, each comprising a 2D convolutional layer, batch normalization, and ReLU activation, to capture local spatial dependencies among neighboring joints; subsequently, two Transformer encoder layers model global inter-joint relationships across the skeleton structure; finally, two additional Transformer encoder layers capture temporal dependencies across the sequence. The function $f_t : \mathbb{R}^{T \times E} \to \{0, 1\}$ is an MLP classifier that maps $\mathbf{z}_t$ to the final prediction $\hat{y}_t$.

### 3.2 Student Network

We employ a lightweight student model based on LightHART [6] for resource-constrained devices. Let $\mathbf{x}_s \in \mathcal{X} = \mathbb{R}^{T \times C}$ denote the inertial input sample, where $T$ and $C$ are the number of timestamps and channels. The student network computes the output as:

$$\hat{y}_s = f_s\big(g_s(\mathbf{x}_s)\big) \in \{0, 1\}. \tag{2}$$

The encoder $g_s : \mathcal{X} \to \mathbb{R}^{T \times E}$, produces an intermediate representation $\mathbf{z}_s = g_s(\mathbf{x}_s)$. The encoder adopts a minimal design for efficiency: a 1D convolutional layer first embeds the inertial signals, followed by two Transformer encoder layers that capture temporal dependencies across timestamps. The MLP head $f_s : \mathbb{R}^{T \times E} \to \{0, 1\}$, maps intermediate representation $\mathbf{z}_s$ from encoder to the final output $\hat{y}_s = f_s(\mathbf{z}_s)$. The entire model contains only **9,857** parameters.

### 3.3 RightFeatKD

**Softened Feature Matching via KL Divergence**: Instead of distilling final logits, we align intermediate feature representations from the teacher ($\mathbf{z_t}$) and student ($\mathbf{z_s}$) using KL divergence:

$$\mathcal{L}_{KD} = \mathrm{KL}\left(\phi\left(\frac{\mathbf{z}_t}{\tau}\right) \,\|\, \phi\left(\frac{\mathbf{z}_s}{\tau}\right)\right) \tag{3}$$

where $\phi(\cdot) = \mathrm{softmax}(\cdot)$ and $\tau > 1$ is the temperature parameter.

**Teacher Confidence Weighting**: To prioritize reliable teacher knowledge, we introduce an adaptive weighting mechanism $\mathcal{W}^{KD}$:

$$\mathcal{W}^{KD} = \beta \times (1 - |\hat{y}_t - y|), \quad \text{where} \quad \beta = \begin{cases} 0.66 & \text{if } \hat{y}_t = y \\ 0.34 & \text{otherwise} \end{cases} \tag{4}$$

where $\hat{y}_t \in [0, 1]$ is the teacher's predicted probability, $y \in \{0, 1\}$ is the ground-truth label, and $\beta$ adaptively assigns higher weight to correctly predicted samples.

**Focal Loss for Hard Examples**: To emphasize challenging instances, we employ binary focal loss:

$$\mathcal{L}_{\text{Focal}} = -\lambda y(1 - \hat{y}_s)^{\gamma} \log(\hat{y}_s) - (1 - \lambda)(1 - y)\hat{y}_s^{\gamma} \log(1 - \hat{y}_s) \tag{5}$$

where $\hat{y}_s \in [0, 1]$ is the student's predicted probability, $\gamma > 0$ is the focusing parameter, and $\lambda \in [0, 1]$ is the class weight.

**Combined Objective**: The total loss combines weighted feature distillation and focal loss:

$$\mathcal{L}_{\text{Total}} = \mathcal{W}^{KD} \cdot \alpha \cdot \mathcal{L}_{KD} + (1 - \alpha) \cdot \mathcal{L}_{\text{Focal}} \tag{6}$$

where $\alpha \in [0, 1]$ balances both components. We set $\alpha = 0.60$ and $\gamma = 2.0$ for both datasets, with $\tau = 4.50$ for SmartFallMM and $\tau = 4.00$ for Up-Fall.

## 4 Experiments

### 4.1 Datasets

**SmartFallMM** is a multimodal fall detection dataset comprising younger participants (median age 23) and older adults (median age 65.5). It contains inertial sensor readings and 32-keypoint skeleton sequences. We use data from younger participants only, as older adults did not perform fall activities.

**Up-Fall** [17] contains vision (RGB video), ambient, and wearable sensor data from 17 participants performing 11 activities across 14 devices. We extract 17-keypoint skeleton sequences from RGB videos using GAST-Net [14]. For both datasets, we use accelerometer data from a left-wrist-mounted sensor as the inertial modality.

### 4.2 Preprocessing

We synchronize inertial and skeleton modalities using Dynamic Time Warping and apply a Butterworth filter to denoise accelerometer signals. For SmartFallMM, we extract sliding windows of 128 timestamps (≈4 s) with 50% overlap (≈2 s). For Up-Fall, we use 18-timestamp windows (≈1 s) due to its shorter activity duration.

### 4.3 Evaluation Protocol

To assess the effectiveness of our proposed RightFeatKD method, we follow a three-stage evaluation protocol:

Table 2: Performance comparison of knowledge distillation methods on SmartFallMM. Arrows indicate improvement (↑) or degradation (↓) relative to their baseline student model trained w/o KD.

| | **SmartFallMM** | | | |
|---|---|---|---|---|
| **Method** | **Accuracy** | **F1-score** | **Precision** | **Recall** |
| KD [7] | 75.91 (↓0.42%) | 70.83 (↓4.19%) | 76.85 (↓5.46%) | 69.00 (↓1.58%) |
| LSTMKD [27] | 61.22 (↑0.36%) | 48.63 (↑2.37%) | 47.23 (↑3.97%) | 53.34 (↑0.06%) |
| PreFallKD [3] | 75.63 (↓3.77%) | 69.98 (↓4.49%) | 70.60 (↓5.20%) | 71.66 (↓3.69%) |
| FitNet [23] | 81.00 (↑4.67%) | 76.23 (↑1.21%) | 82.50 (↑0.19%) | 71.40 (↑0.82%) |
| **RightFeatKD** | **82.66** (↑**2.33%**) | **79.06** (↑**4.04%**) | **82.31** (**0.00%**) | **76.21** (↑**5.63%**) |

- **Baseline Setup (Independent Training):** In the first stage, we train the teacher model independently using only skeleton data to leverage its rich structural information. Simultaneously, we train the student models on wrist accelerometer data without teacher supervision. These independently trained inertial models serve as baselines to evaluate the standalone performance of unimodal inertial data.
- **RightFeatKD Distillation Setup:** In the second stage, we apply our RightFeatKD method by using a pretrained skeleton-based teacher model to guide the training of the student model for transferring knowledge by aligning intermediate representations.
- **Real-Time Evaluation:** In the final stage, we deploy both the baseline and distilled models trained on SmartFallMM on a smartwatch to evaluate performance and inference time. We recruit five student participants under IRB 9461 to perform the activities. We measure inference time by averaging 1000 predictions, following the protocol in MLP-HAR [28].

We evaluate all trained models using Leave-One-Subject-Out (LOSO) cross-validation ($k$ equals the number of subjects in the dataset) to ensure robustness and generalizability. We evaluate performance using accuracy, precision, recall, and F1-score, and assessed model efficiency in terms of FLOPs and parameter count. We implement the system in PyTorch on an NVIDIA A100 GPU, with all models trained using the AdamW optimizer with a learning rate of $1 \times 10^{-4}$.

## 5 Results

As shown in Table 2, RightFeatKD achieves 4.04% improvement over the baseline on SmartFallMM, outperforming FitNet [23] by 2.83% and LSTMKD [27] by 1.67%. Conversely, logit-based approaches PreFallKD [3] and KD [7] degrade performance by 4.19% and 4.49%, demonstrating their ineffectiveness for limited distribution diversity in logits in fall detection.

We observe a similar trend on the Up-Fall dataset as shown in Table 3, where RightFeatKD achieves a slightly lower F1-score of 77.46% than PreFallKD due

Table 3: Performance comparison of knowledge distillation methods on Up-Fall. Arrows indicate improvement (↑) or degradation (↓) relative to baseline student model trained without KD.

| | Up-Fall | | | |
|---|---|---|---|---|
| **Method** | **Accuracy** | **F1-Score** | **Precision** | **Recall** |
| KD [7] | 77.55 (↑10.06%) | 74.17 (↑9.43%) | 73.53 (↑12.20%) | 77.29 (↑5.80%) |
| LSTMKD [27] | 65.67 (↓0.56%) | 52.71 (↑1.12%) | 51.39 (↓2.95%) | 56.53 (↓8.53%) |
| PreFallKD [3] | 82.96 (↓0.90%) | 79.87 (↑1.56%) | 81.80 (↓2.87%) | 80.37 (↑0.79%) |
| FitNet [23] | 77.41 (↑9.85%) | 74.07 (↑9.20%) | 72.85 (↑11.52%) | 77.58 (↑6.09%) |
| **RightFeatKD** | **81.36** (↑**13.88%**) | **77.46** (↑**12.59%**) | **77.60** (↑**16.27%**) | **79.41** (↑**8.39%**) |

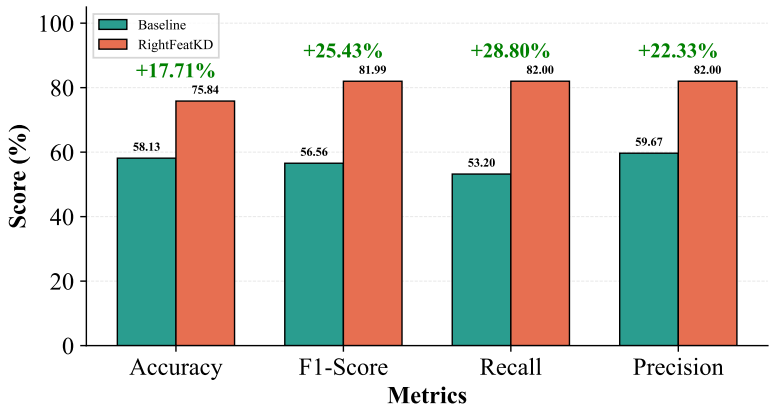



Fig. 2: Comparison of real-time performance between baseline student model and RightFeatKD in terms of F1-score, precision, accuracy, and recall.

to the shorter 1 s input window. RightFeatKD delivers the largest improvement, boosting the baseline by 12.59%, and outperforming KD, LSTMKD, PreFallKD, and FitNet by 3.16%, 11.47%, 11.03%, and 3.39%, respectively.

In the real-time evaluation (Fig. 2), the baseline student model's F1-score drops from 75.0% to 56.56%, whereas the RightFeatKD-distilled model achieves 81.99%, a 25.43% improvement, further underscoring the premise that RightFeatKD can transfer vital knowledge that can improve the model's performance in real-time also. Fig. 3 shows that RightFeatKD detects 57 of 75 falls compared to 23 by the baseline, and reduces false positives by 12 for activities with high vertical movement, such as sit and stand, highlighting the benefit of transferring visual knowledge. In Table 4, LSTMKD [27] shows the highest overhead with 76,417 parameters, 141M FLOPs, and the slowest inference ($13.88 \pm 29.28$ ms). PreFallKD offers the fastest inference ($2.17 \pm 4.20$ ms) but with 74,756 parameters. In contrast, RightFeatKD strikes the best balance, using only 9,857

Table 4: Comparison of model complexity and efficiency in terms of size, FLOPs (in millions), and inference time.

| Method | Params | FLOPs (M) | Inference (ms) |
|---|---|---|---|
| LSTMKD [27] | 76,417 | 141.68 | $13.88 \pm 29.28$ |
| PreFallKD [3] | 74,756 | 15.09 | $2.17 \pm 4.20$ |
| KD [7] | 9,857 | 20.44 | $6.11 \pm 8.21$ |
| FitNet [23] | 9,857 | 20.44 | $6.11 \pm 8.21$ |
| **RightFeatKD (Ours)** | 9,857 | 20.44 | $6.11 \pm 8.21$ |

parameters—an order of magnitude fewer—while maintaining 20.44M FLOPs and real-time inference ($6.11 \pm 8.21$ ms). KD [7] and FitNet [23] has the same number of parameters and inference time as RightFeatKD, as we use the same student architecture; however, they have lower performance than RightFeatKD. This compact yet efficient design makes RightFeatKD well-suited for resource-constrained deployment.

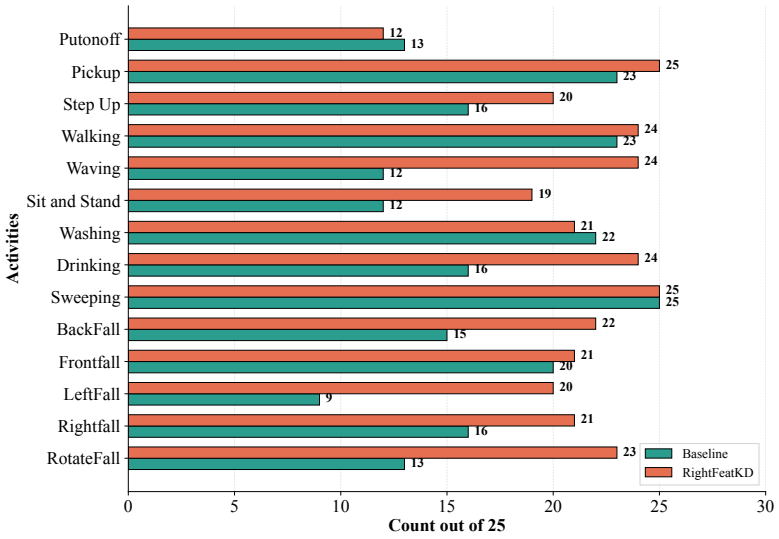



Fig. 3: Comparison of real-time performance between baseline student model and RightFeatKD in terms of correct predictions per activity.

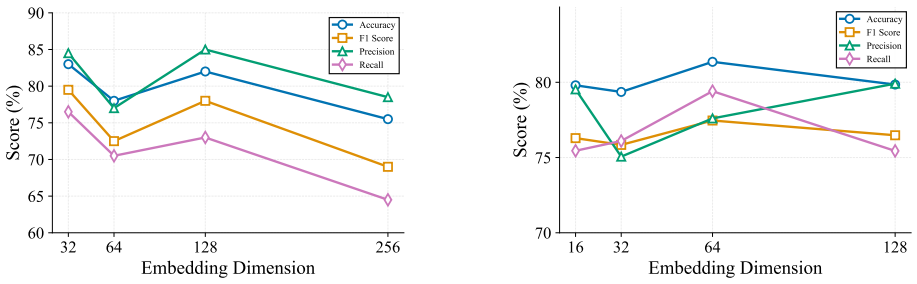
## 6 Ablation Study

We conduct comprehensive ablation experiments to analyze the contribution of individual components and validate our design choices.

Table 5: Ablation study on weighting and feature-based distillation loss on the SmartFallMM dataset.

| Configuration | Accuracy | F1-score | Recall |
|---|---|---|---|
| With Weight | 82.66 | 79.06 | 76.21 |
| Without Weight | 80.27 | 76.11 | 71.05 |
| Feature-Based | 82.66 | 79.06 | 76.21 |
| Logit-Based | 78.42 | 73.95 | 70.67 |

**Impact of weighting function**: To assess the importance of our proposed weighting scheme, we evaluate the distillation method on SmartFallMM after removing the weighting function that assigns higher importance to correct predictions. Table 5 shows that removing this component results in a 2.95% drop in F1-score, confirming that the weighting mechanism plays a critical role in guiding the student model toward reliable predictions.
**Impact of feature-based loss**: We compare our feature-based KL-divergence loss against a standard logit-based variant. Replacing the feature-based loss with logit-based distillation degrades performance by 1.17% below the baseline, demonstrating that distilling intermediate representations rather than final logits is essential for stable performance gains. These results confirm that the combination of the weighting scheme and feature-based loss is necessary for effective knowledge transfer.



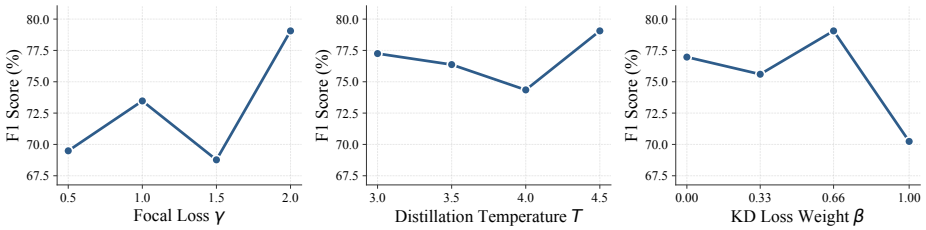

(a) SmartFallMM (b) Up-Fall

Fig. 4: Effect of teacher's intermediate representation size on student's performance.

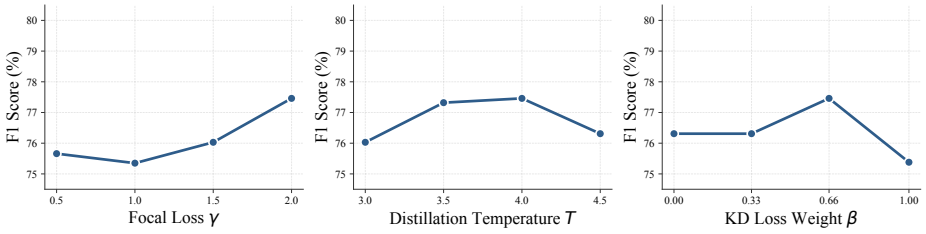
**Effect of teacher intermediate representation size:** We analyze how the teacher's intermediate representation size affects performance. Fig. 4 presents results for both SmartFallMM and Up-Fall datasets. Both datasets exhibit optimal generalization at relatively small dimensions—32 for SmartFallMM and 64 for Up-Fall. This can be attributed to the limited dataset size, where smaller rep-

resentations prevent overfitting. Larger representation size lead to performance fluctuations and reduced generalization, suggesting that the teacher's capacity should be calibrated to the dataset scale.

**Hyperparameter sensitivity:** We examine three key hyperparameters: the focal loss focusing parameter $\gamma$, distillation temperature $\tau$, and the weighting parameter $\beta$ in $\mathcal{W}^{KD}$. Fig. 5 shows consistent trends across both datasets. Performance improves with increasing $\gamma$, indicating that emphasizing hard examples benefits distillation. Both datasets require relatively high temperature values ($\tau \in [4.0, 4.5]$), which we attribute to the need for softer supervision when distilling intermediate representations—low temperatures may cause the student to overfit to modality-specific patterns from the teacher, undermining the benefits of knowledge distillation. For the weighting parameter $\beta$, we observe that extremely high values (assigning nearly all weight to correct predictions) can destabilize training and impede gradient flow. A balanced weighting that incorporates some signal from incorrect predictions ensures smoother convergence.




(a) SmartFallMM




(b) UpFall

Fig. 5: Impact of Focal loss $\gamma$, distillation temperature $\tau$ and $\beta$ hyperparameter of $\mathcal{W}^{KD}$ on student .

## 7 Conclusion

In this work, we propose **RightFeatKD**, a selective feature-based knowledge distillation method that effectively transfers knowledge from correct predictions of the teacher. RightFeatKD builds on the idea that logit KD lacks diversity in probability distribution and passes knowledge indiscriminately. Our method

achieves an F1-score of 79.06% on the SmartFallMM dataset and 77.46% on the Up-Fall dataset. Compared to other KD methods, our method shows the highest improvement from baseline models for both datasets. We also validate the real-time performance of RightFeatKD. We show that our distilled model achieves a much higher performance than the baseline model with low computation cost and inference time. These results underscore the effectiveness of RightFeatKD for fall detection.

## References


1. Iqra Aijaz Abro and Ahmad Jalal. Multi-modal sensors fusion for fall detection and action recognition in indoor environment. In *2024 3rd International Conference on Emerging Trends in Electrical, Control, and Telecommunication Engineering (ETECTE)*, pages 1–6. IEEE, 2024.
2. Baiyu Chen, Wilson Wongso, Zechen Li, Yonchanok Khaokaew, Hao Xue, and Flora Salim. COMODO: Cross-Modal Video-to-IMU Distillation for Efficient Egocentric Human Activity Recognition, March 2025. arXiv:2503.07259 [cs].
3. Tin-Han Chi, Kai-Chun Liu, Chia-Yeh Hsieh, Yu Tsao, and Chia-Tai Chan. Prefallkd: Pre-impact fall detection via cnn-vit knowledge distillation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
4. Vanilson Fula and Plinio Moreno. Wrist-based fall detection: towards generalization across datasets. *Sensors*, 24(5):1679, 2024.
5. Nuno Cruz Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. Distillation Multiple Choice Learning for Multimodal Action Recognition. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2754–2763, Waikoloa, HI, USA, January 2021. IEEE.
6. Syed Tousiful Haque, Jianyuan Ni, Jingcheng Li, Yan Yan, and Anne Hee Hiong Ngu. Lighthart: Lightweight human activity recognition transformer. In *Pattern Recognition: 27th International Conference, ICPR 2024, Kolkata, India, December 1–5, 2024, Proceedings, Part XV*, page 425–441, Berlin, Heidelberg, 2024. Springer-Verlag.
7. Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
8. Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. $C^2$ KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16006–16015, Seattle, WA, USA, June 2024. IEEE.
9. Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for Natural Language Understanding, October 2020. arXiv:1909.10351 [cs].
10. Heegon Jin, Seonil Son, Jemin Park, Youngseok Kim, Hyungjong Noh, and Yeonsoo Lee. Align-to-Distill: Trainable Attention Alignment for Knowledge Distillation in Neural Machine Translation, March 2024. arXiv:2403.01479 [cs].
11. Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8657–8666, 2019.

12. Jingcheng Li, Lina Yao, Binghao Li, and Claude Sammut. Distilled mid-fusion transformer networks for multi-modal human activity recognition. *arXiv preprint arXiv:2305.03810*, 2023.
13. Jingcheng Li, Lina Yao, Binghao Li, and Claude Sammut. Distilled mid-fusion transformer networks for multi-modal human activity recognition. *Knowledge-Based Systems*, 327:114160, October 2025.
14. Junfa Liu, Juan Rojas, Yihui Li, Zhijun Liang, Yisheng Guan, Ning Xi, and Haifei Zhu. A graph attention spatio-temporal convolutional network for 3d human pose estimation in video. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3374–3380, 2021.
15. Amir M. Mansourian, Arya Jalali, Rozhan Ahmadi, and Shohreh Kasaei. Attention-guided Feature Distillation for Semantic Segmentation, March 2025. arXiv:2403.05451 [cs].
16. Wei-Lung Mao, Chun-Chi Wang, Po-Heng Chou, Kai-Chun Liu, and Yu Tsao. Meckd: Deep learning-based fall detection in multi-layer mobile edge computing with knowledge distillation. *IEEE Sensors Journal*, 2024.
17. Lourdes Martínez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Núñez-Martínez, and Carlos Peñafort-Asturiano. Up-fall detection dataset: A multimodal approach. *Sensors*, 19(9), 2019.
18. Anne Hee Hiong Ngu, Vangelis Metsis, Shuan Coyne, Priyanka Srinivas, Tarek Salad, Uddin Mahmud, and Kyong Hee Chee. Personalized watch-based fall detection using a collaborative edge-cloud framework. *International Journal of Neural Systems*, 32(12):2250048, 2022. PMID: 35972790.
19. Jianyuan Ni, Anne HH Ngu, and Yan Yan. Progressive cross-modal knowledge distillation for human action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5903–5912, 2022.
20. Jianyuan Ni, Raunak Sarbajna, Yang Liu, Anne Hee Hiong Ngu, and Yan Yan. Cross-modal knowledge distillation for vision-to-sensor action recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4448–4452. IEEE, 2022.
21. Jianyuan Ni, Raunak Sarbajna, Yang Liu, Anne H.H. Ngu, and Yan Yan. Cross-modal knowledge distillation for vision-to-sensor action recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4448–4452, 2022.
22. Jianyuan Ni, Hao Tang, Syed Tousiful Haque, Yan Yan, and Anne HH Ngu. A survey on multimodal wearable sensor-based human action recognition. *arXiv preprint arXiv:2404.15349*, 2024.
23. Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.
24. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems(NeurIPS)*, 30, 2017.
25. Qing Xu, Min Wu, Xiaoli Li, Kezhi Mao, and Zhenghua Chen. Reinforced Cross-Domain Knowledge Distillation on Time Series Data. *Advances in Neural Information Processing Systems*, 37:47217–47241, December 2024.
26. Zhendong Yang, Zhe Li, Ailing Zeng, Zexian Li, Chun Yuan, and Yu Li. ViTKD: Feature-based Knowledge Distillation for Vision Transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1379–1388, Seattle, WA, USA, June 2024. IEEE.

27. Hannah Zhou, Allison Chen, Celine Buer, Emily Chen, Kayleen Tang, Lauryn Gong, Zhiqi Liu, and Jianbin Tang. Fall detection using knowledge distillation based long short-term memory for offline embedded and low power devices. *arXiv preprint arXiv:2308.12481*, 2023.
28. Yexu Zhou, Tobias King, Haibin Zhao, Yiran Huang, Till Riedel, and Michael Beigl. Mlp-har: Boosting performance and efficiency of har models on edge devices with purely fully connected layers. In *Proceedings of the 2024 ACM International Symposium on Wearable Computers*, ISWC '24, page 133–139, New York, NY, USA, 2024. Association for Computing Machinery.