This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

View-aware Cross-modal Distillation for Multi-view Action Recognition

Trung Thanh Nguyen[1,2*], Yasutomo Kawanishi[2,1], Vijay John[3], Takahiro Komamizu[1], Ichiro Ide[1]
[1]Nagoya University, Japan [2]Guardian Robot Project, R-IH, RIKEN, Japan
[3]Lawrence Technological University, USA

Abstract

The widespread use of multi-sensor systems has increased research in multi-view action recognition. While existing approaches in multi-view setups with fully overlapping sensors benefit from consistent view coverage, partially overlapping settings where actions are visible in only a subset of views remain underexplored. This challenge becomes more severe in real-world scenarios, as many systems provide only limited input modalities and rely on sequence-level annotations instead of dense frame-level labels. In this study, we propose ***Vi****ew-aware* ***Cro****ss-modal* ***K****nowledge* ***D****istillation (ViCoKD), a framework that distills knowledge from a fully supervised multi-modal teacher to a modality- and annotation-limited student. ViCoKD employs a cross-modal adapter with cross-modal attention, allowing the student to exploit multi-modal correlations while operating with incomplete modalities. Moreover, we propose a View-aware Consistency module to address view misalignment, where the same action may appear differently or only partially across viewpoints. It enforces prediction alignment when the action is co-visible across views, guided by human-detection masks and confidence-weighted Jensen–Shannon divergence between their predicted class distributions. Experiments on the real-world MultiSensor-Home dataset show that ViCoKD consistently outperforms competitive distillation methods across multiple backbones and environments, delivering significant gains and surpassing the teacher model under limited conditions.*

1. Introduction

The increasing use of multi-sensor systems in smart homes [25], surveillance [17], and assistive robotics [1] has driven growing interest in multi-view action recognition. By aggregating observations from multiple viewpoints, such systems capture richer spatio-temporal cues than single-view setups and improve robustness to occlusion, viewpoint changes, and background clutter [18, 26, 33, 37]. Most existing methods, however, are designed for fully overlapping sensor setups. In such settings, sensors cover the same region of the scene, and each action instance is fully visible across all views. Datasets such as Nanyang Technological University (NTU) RGB+D [18, 33] and NorthWestern-University of California at Los Angeles (NW-UCLA) Multiview Action 3D [37] adopt this assumption, enabling methods [4, 19, 32, 39] to exploit contrastive or disentanglement learning to enforce viewpoint invariance. In contrast, partially overlapping settings are more representative of real-world deployments. These setups feature non-aligned sensor coverage, where an action may be visible in some views but completely occluded or out of scope in others. Recent methods [23, 25, 40, 41] address this challenge by proposing fusion strategies to integrate fragmented evidences or recover missing observations. However, they focus primarily on feature and view fusion, without explicitly enforcing consistency across views when actions are only partially observed. Therefore, their effectiveness in realistic partially overlapping scenarios is limited, and explicit modeling of view-aware consistency remains underexplored.

On the other hand, multi-view action recognition models achieve their best performance when all modalities are fully available and dense frame-level annotations are provided [25]. However, in real-world multi-sensor deployments, these conditions are often unmet. First, modality-limited scenarios occur when some input streams (e.g., audio or depth) are unavailable due to hardware or environmental constraints [13]. Second, annotation-limited scenarios arise as frame-level labeling is prohibitively expensive for multi-view videos, leaving only weak sequence-level tags in many datasets [40, 41]. Although prior works [12, 23, 24] address learning under limited conditions, they primarily rely on weak supervision, remaining sensitive to fragmented observations and noisy labels.

To handle this limitation, the potential direction is to transfer supervision from a stronger model to a weaker one. Knowledge Distillation (KD) [11] offers a promising solution by allowing a fully supervised multi-modal teacher trained with strong labels to guide a modality- and

*Corresponding author: nguyent@cs.is.i.nagoya-u.ac.jp.

annotation-limited student. Multi-modal KD (MKD) [3, 8, 10, 28] extends this concept to cases where the teacher benefits from multi-modal inputs and frame-level labels, while the student operates with incomplete modalities or only coarse sequence-level labels. However, applying KD to partially overlapping multi-view action recognition remains challenging, as it requires transferring knowledge while preserving consistency across views where the action may appear in some views but be absent in others.

To address these challenges, we propose **Vi**ew-aware **Cro**ss-modal **K**nowledge **D**istillation (ViCoKD), a method that distills knowledge from a fully supervised multi-modal teacher to a modality- and annotation-limited student. ViCoKD employs feature-level and logit-level KD, leveraging a cross-modal adapter that enables the student to exploit audio-visual correlations through Cross-modal Attention while operating solely on visual input. In addition, we introduce a View-aware Consistency module that explicitly handles partially overlapping sensor setups by encouraging the model to produce the same prediction across different views whenever the action is co-visible. This is achieved through a confidence-weighted Jensen–Shannon divergence [22] between the predicted class distributions of co-visible views, ensuring that supervision is concentrated on frames with reliable, view-consistent evidence. The main contributions of this study are as follows:

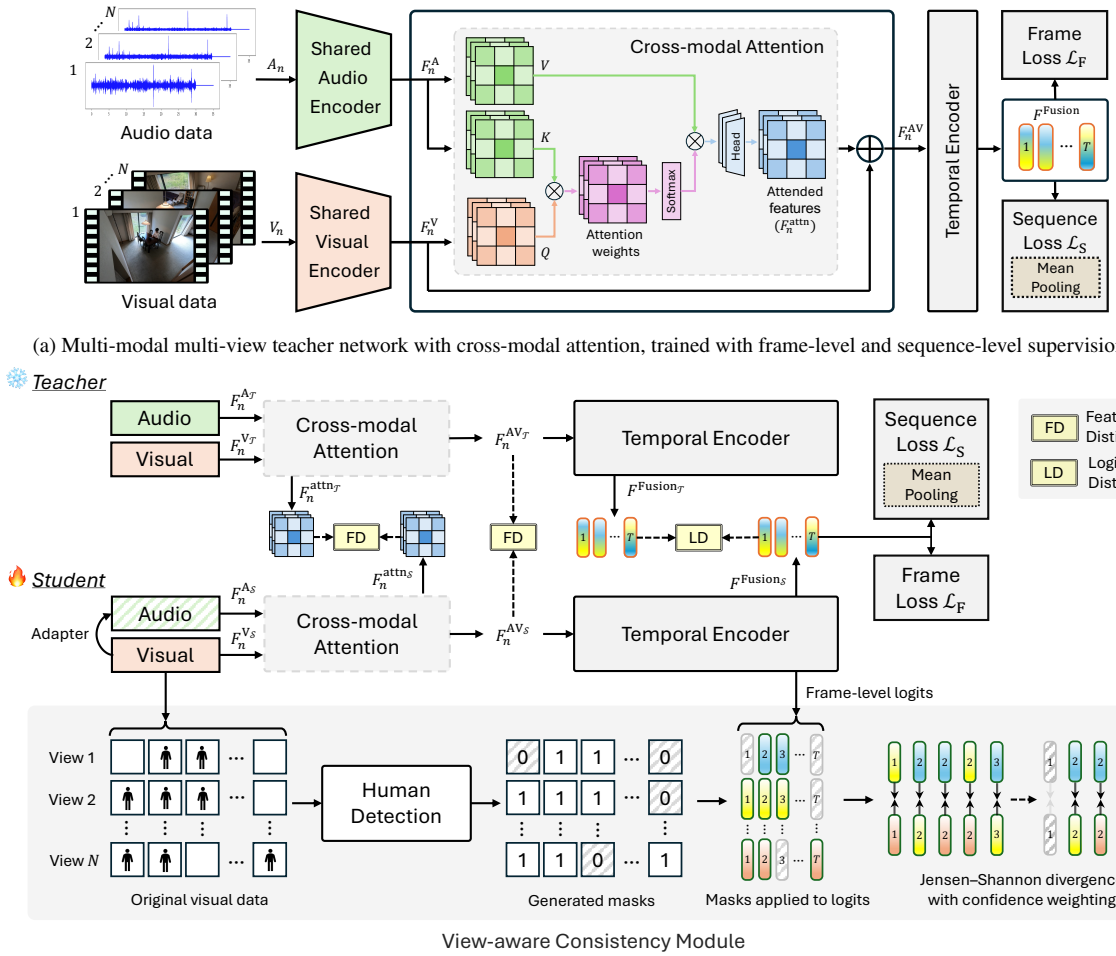
- **ViCoKD method:** A cross-modal attention-based KD method that transfers knowledge from a fully supervised multi-modal teacher to a modality- and annotation-limited student for multi-view action recognition in partially overlapping sensor settings.
- **View-aware Consistency module:** A consistency module that leverages human-detection masks and confidence-weighted Jensen–Shannon divergence [22] to align predictions only on view pairs with reliable and visible action evidence.
- **Comprehensive evaluation:** Experiments on the real-world MultiSensor-Home dataset [25] across diverse backbones and home environments show that ViCoKD consistently outperforms competitive distillation methods and surpasses the teacher under limited conditions.

The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 introduces the proposed ViCoKD. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes the paper.

2. Related Work

Multi-view Action Recognition. Depending on sensor coverage, multi-view action recognition tasks are studied under either fully overlapping or partially overlapping settings. Most prior work focuses on the former, with datasets such as NTU RGB+D [18, 33], NW-UCLA [37], and Toyota Smarthome [5] where all views capture the same subject. Building on these datasets, ViewCLR [4] employs view generation to generalize to unseen viewpoints, ViewCon [32] applies supervised contrastive learning to produce viewpoint-invariant embeddings, and Dual-Recommendation Disentanglement Network (DRDN) [19] leverages disentanglement learning to decouple action-specific and view-specific cues. In contrast, real-world applications involve wide-area surveillance with only partially overlapping views, where actions may be visible from some sensors but entirely occluded in others. Recent work has introduced datasets tailored to this challenging setting [25, 41]. Yasuda et al. [40] introduce Multi-Trans, which models inter-sensor relationships to integrate multi-view cues, and later propose Guided Masked sELf-Distillation (Guided-MELD) [41] to handle fragmented observations by reducing redundancy and filling missing sensor data for coherent event-level representations. John and Kawanishi [12] present a weakly supervised latent embedding framework that learns from sequence-level labels while enabling frame-level action detection. Recently, Nguyen et al. [23, 25] proposed MultiASL for weakly supervised action selection across sensors and MultiTSF for robust cross-sensor temporal fusion. Despite these advances, most existing work focuses on feature and view fusion strategies, while explicit modeling of view-aware consistency in partially overlapping settings is still limited.

Multi-modal Knowledge Distillation (MKD). KD [11] trains a smaller or weaker student model to mimic the predictions or intermediate representations of a teacher model, improving efficiency or performance in limited conditions. Beyond matching final logits [11], KD has evolved to include feature-level supervision [30, 43] and relational knowledge transfer [27] to strengthen teacher–student alignment. MKD [3, 44] extends KD to settings where the teacher and student operate over multiple modalities. The teacher typically leverages full-modality inputs, such as RGB, audio, depth, or skeleton data, while the student learns to replicate its performance under restricted modalities. Prior works have demonstrated the benefits of MKD for RGB-depth [10], RGB-flow [8], and audio-visual [3, 28] tasks. In video understanding, MKD has been shown to improve recognition under missing modalities [29] by transferring temporal and cross-modal cues from the teacher to the student. In this work, we study MKD in the context of partially overlapping multi-view action recognition under missing-audio conditions. Unlike these methods, our student operates solely with visual inputs and is trained under frame-level or sequence-level supervision, the latter representing a more challenging weakly supervised setting. To address this gap, we propose the ViCoKD method with View-aware Consistency, enabling students to learn robust cross-modal and cross-view representations under the limited conditions.



(a) Multi-modal multi-view teacher network with cross-modal attention, trained with frame-level and sequence-level supervisions.

(b) Knowledge distillation pipeline with a frozen teacher and a student trained via feature and logit distillation. A cross-modal adapter generates pseudo-audio features when audio is unavailable, enabling cross-modal attention. The view-aware consistency module applies human-detection-based masks and Jensen–Shannon divergence with confidence weighting for view-consistent supervision.

Figure 1. Overview of the proposed ViCoKD method: (a) A multi-modal multi-view teacher with cross-modal attention, and (b) a knowledge distillation pipeline where the student trained using feature-level and logit-level distillation under view-aware consistency supervision.

3. ViCoKD: View-aware Cross-modal Distillation Framework

We address the task of multi-view action recognition in partially overlapping settings within a KD framework. Actions are captured from N synchronized viewpoints. In this study, we address the multi-modal setting as comprising audio and visual inputs. At each time step $t \in \{1, \dots, T\}$, the input from view $n \in \{1, \dots, N\}$ of teacher comprises two modalities: audio segment $a_t^n \in \mathbb{R}^F$, represented as a time-aligned spectrogram with F frequency bins per frame, and visual frame $v_t^n \in \mathbb{R}^{H \times W \times C}$, where H and W denote spatial resolution, and C is the number of color channels. Audio and visual inputs from view n are denoted as $A^n = \{a_t^n\}_{t=1}^T$ and $V^n = \{v_t^n\}_{t=1}^T$, respectively. The objective is to train a student model for multi-label action classification by transferring knowledge from a fully supervised multi-modal teacher model. Here, we focus on the audio-missing setting, where the student lacks access to the audio modality. Two supervision scenarios are considered for the student: (1) Frame-level supervision, where the student is trained with frame-level and sequence-level labels, and (2) Sequence-level supervision, where the student is trained using only video-level labels. The goal is to exploit the feature representations learned by the teacher (Figure 1a) and distill this knowledge into a student model (Figure 1b) under missing-modality conditions.

3.1. Teacher Model with Cross-modal Attention

Figure 1a illustrates the teacher model, which is designed to learn spatio-temporal representations from synchronized multi-modal multi-view inputs, trained with strong supervision at the frame and sequence levels.

Modality-specific Encoders. Each input sequence consists of audio data A_n and visual data V_n from view n, which are processed by modality-specific encoders. Specifically, we use Audio Spectrogram Transformer (AST) [9] as the audio encoder E_{A} and Vision Transformer (ViT) [6] as the visual encoder E_{V} as:

$$F_n^{\mathrm{A}} = E_{\mathrm{A}}(A_n), \quad F_n^{\mathrm{V}} = E_{\mathrm{V}}(V_n), \tag{1}$$

where $F_n^{\mathrm{A}} \in \mathbb{R}^{T \times D_a}$ and $F_n^{\mathrm{V}} \in \mathbb{R}^{T \times D_v}$ are temporally aligned audio and visual feature sequences, respectively. The AST and ViT encoders share parameters across all views to ensure consistent representation learning.

Cross-modal Attention Module. To integrate audio-visual information, we adopt a cross-modal attention mechanism [38, 42] where audio features are used as keys and values and visual features serve as queries. Positional encodings are added to both modalities to retain temporal and spatial ordering before cross-modal attention is applied as:

$$F_n^{\mathrm{attn}} = \mathrm{CrossAttn}(Q = F_n^{\mathrm{V}}, K = F_n^{\mathrm{A}}, V = F_n^{\mathrm{A}}). \tag{2}$$

The attended features F_n^{attn} are fused with the visual features via element-wise addition as:

$$F_n^{\mathrm{AV}} = F_n^{\mathrm{V}} + F_n^{\mathrm{attn}}. \tag{3}$$

This fusion strategy preserves the spatial-temporal structure of the visual representation while enhancing it with complementary audio information.

Temporal Encoding. The fused representation $F_n^{\mathrm{AV}} \in \mathbb{R}^{T \times D}$ from each view n is fed into a Transformer-based [35] temporal encoder to capture sequential dependencies across time as:

$$F_n^{\mathrm{Fusion}} = \mathrm{TemporalEncoder}(F_n^{\mathrm{AV}}), \tag{4}$$

where TemporalEncoder is a multi-layer Transformer that models self-attention across frames to learn long-range temporal patterns within the sequence. Output $F_n^{\mathrm{Fusion}} \in \mathbb{R}^{T \times D'}$ represents a temporally enriched feature sequence encoding spatio-temporal cues.

Supervision Objective. The teacher is optimized with frame-level and sequence-level classification losses. At the frame-level, we predict per-frame logits and compute a Two-way loss [15] that combines sample-wise and class-wise objectives. At the sequence-level, we apply temporal average pooling followed by classification, using the same loss formulation. The overall training objective for the teacher is calculated as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{F}} + \lambda \mathcal{L}_{\mathrm{S}}, \tag{5}$$

where $\mathcal{L}_{\mathrm{F}}$ and $\mathcal{L}_{\mathrm{S}}$ are the frame-level and sequence-level classification losses, respectively, and λ is a balancing hyperparameter.

3.2. Teacher-to-student View-aware Cross-modal Knowledge Distillation

Figure 1b illustrates the knowledge distillation pipeline, where the student ($\mathcal{S}$) adopts the same architecture as the teacher ($\mathcal{T}$) while operating with a missing audio modality. The student is trained to mimic the teacher's knowledge through the supervision objective in Eq. (5), combined with a KD loss that transfers feature- and logit-level representations from the teacher. In addition, the student benefits from the View-aware Consistency loss, which enforces consistent predictions across partially overlapping views when the action is co-visible.

Cross-modal Adapter. When audio is missing, a cross-modal adapter $\mathcal{A}_\psi$ is employed to synthesize pseudo-audio features $\hat{F}_n^{\mathrm{A}} = \mathcal{A}_\psi(F_n^{\mathrm{V}})$ from visual features F_n^{V}, where $\mathcal{A}_\psi$ is a lightweight feed-forward network with a non-linear activation. The pseudo-audio feature is then used in cross-modal attention as:

$$F_n^{\mathrm{attn}_{\mathcal{S}}} = \mathrm{CrossAttn}(Q = F_n^{\mathrm{V}}, K = \hat{F}_n^{\mathrm{A}}, V = \hat{F}_n^{\mathrm{A}}). \tag{6}$$

Feature Distillation (FD). We distill knowledge at two stages: (1) The attended audio-visual features $F_{n,t}^{\mathrm{attn}}$ from the cross-modal attention, which transfer audio cues into the visual stream, and (2) The fused representations $F_{n,t}^{\mathrm{AV}}$, which integrate multi-modal context. Formally, we minimize the frame-wise Euclidean distance between the teacher's and the student's features as:

$$\mathcal{L}_{\mathrm{FD}} = \frac{1}{T} \sum_{t=1}^{T} \left(\left\| F_{n,t}^{\mathrm{attn}_{\mathcal{T}}} - F_{n,t}^{\mathrm{attn}_{\mathcal{S}}} \right\|_2^2 + \left\| F_{n,t}^{\mathrm{AV}_{\mathcal{T}}} - F_{n,t}^{\mathrm{AV}_{\mathcal{S}}} \right\|_2^2 \right). \tag{7}$$

This alignment enhances the student's ability to inherit the teacher's cross-modal attention behavior and multi-modal feature integration.

Logit Distillation (LD). We apply logit-level distillation at the frame-level on the fused features F^{Fusion} from multiple views. Let $p_t^{\mathcal{T}}$ and $p_t^{\mathcal{S}}$ denote the temperature-scaled softmax outputs of the teacher and student, respectively. The loss is computed as the Kullback–Leibler (KL) divergence [16] between these distributions as:

$$\mathcal{L}_{\mathrm{LD}} = \frac{\tau^2}{T} \sum_{t=1}^{T} \mathrm{KL}\big(p_t^{\mathcal{T}} \,\|\, p_t^{\mathcal{S}}\big), \tag{8}$$

where τ is the temperature parameter.

View-aware Consistency Module. To enforce consistent predictions across views, we introduce a view-aware consistency loss applied at the frame-level. Figure 1b shows the details: For each view n at time step t, we obtain probabilities $p_{n,t} \in \mathbb{R}^C$. A binary visibility mask $m_{n,t} \in \{0, 1\}$ is generated using human detection [36], where $m_{n,t} = 1$

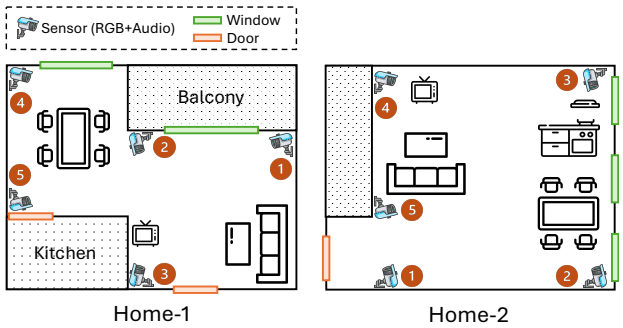


Figure 2. Room layouts and sensor views for the MultiSensor-Home dataset [25] used in the experiments. Each home environment is equipped with multiple RGB and Audio sensors, capturing scenes from different viewpoints with partial overlaps.

indicates that a human is visible in view n at frame t, suggesting a high potential for action occurrence. Given a view pair (i, j), we compute the Jensen–Shannon (JS) divergence [22] between their prediction distributions as:

$$\mathrm{JS}(p_i, p_j) = \frac{1}{2}\mathrm{KL}(p_i \,\|\, m) + \frac{1}{2}\mathrm{KL}(p_j \,\|\, m)\,, \tag{9}$$

where $m = \frac{1}{2}(p_i + p_j)$. The divergence is computed only on frames where $m_{i,t} = m_{j,t} = 1$, and is weighted by the product of the maximum class probabilities from both views to emphasize confident predictions. The View-aware Consistency loss is defined as:

$$\mathcal{L}_{\mathrm{VC}} = \frac{1}{N_{\mathrm{pairs}} T} \sum_{(i,j)} \sum_{t=1}^{T} w_t^{(i,j)}\, m_{i,t} m_{j,t}\, \mathrm{JS}(p_{i,t}, p_{j,t})\,, \tag{10}$$

where $w_t^{(i,j)} = \max_c p_{i,t,c} \cdot \max_c p_{j,t,c}$ serves as a confidence weight, encouraging consistency when both views are individually certain in their predictions, and N_{pairs} is the number of view pairs. This loss aligns predictions where the action is visible in both views and places greater emphasis on high-confidence predictions.

4. Evaluation

4.1. Experimental Conditions

Data Preparation. We conduct the experiments on the publicly available MultiSensor-Home dataset [25], which includes two home environments: Home-1 and Home-2, as shown in Figure 2. Following the official experimental setup, we use a 7:3 train/test split within each environment. To ensure a balanced class distribution, the iterative stratification strategy proposed by Sechidis et al. [31] is employed. For all experiments, we extract a fixed number of T synchronized frames per sample. During training, we apply uniform sampling with slight random perturbations to generate frame sequences that cover the entire video while preserving the fixed length T. This sampling strategy also augments temporal data, enhancing robustness by introducing sequence variability. For evaluation, we adopt deterministic uniform sampling without perturbation to ensure reproducibility across runs.

Evaluation Metrics. Following [15, 23, 25], we evaluate multi-label action recognition performance using mean Average Precision (mAP), a macro-averaged metric widely adopted in multi-label classification tasks.

Comparison Methods. To evaluate the effectiveness of the proposed ViCoKD method, we compare against several representative KD methods under a consistent experimental setup. All methods adopt the same teacher–student configuration for a fair comparison. The teacher network is trained with frame-level audio-visual modalities ($\mathrm{AV_F}$), while the student is trained with visual-only inputs under either frame-level or sequence-level supervision ($\mathrm{V_F}$ or $\mathrm{V_S}$). The teacher and student models use the architectures from the original MultiTrans [40], MultiASL [23], and MultiTSF [25] methods.

- Student/Teacher: The teacher is a full-modality model with RGB and audio modalities. The student uses only the RGB modality, without KD, serving as the lower-bound baseline.
- LogitKD [29, 34]: Logit-level KD methods, in which the student learns from the teacher's softened output logits using Kullback–Leibler divergence [11].
- ModalKD [20]: Feature-level KD method, in which the teacher network distills knowledge to the visual and pseudo-audio features of the student network.
- DualKD [7]: Feature-level and logit-level KD method, in which the teacher network distills knowledge to features and logits of the student network.
- CoKD (ViCoKD w/o Vi): The proposed ViCoKD method without the View-aware Consistency module.

Models & Hyperparameters. We adopt the backbone architecture and hyperparameters of the teacher and student networks from the original work. Following Nguyen et al. [25], we sample input videos at 2.5 FPS and fix the sequence length to $T = 70$ frames, based on the average video duration in the dataset. The cross-modal adapter is implemented as a lightweight feed-forward module with two linear layers and a Rectified Linear Unit (ReLU) [2] activation in between. The human-detection masks are generated by a You Only Look Once (YOLO) v10-based [36] detector. All loss terms are equally weighted with coefficients set to 1.0 for simplicity. Optimization is performed using the Adaptive moment estimation (Adam) [14] optimizer with an initial learning rate of 10^{-4}, weight decay of 5.0×10^{-4}, and a batch size of 12 for 100 training epochs. The learning rate is scheduled using cosine annealing [21] for gradual decay over the course of training.

Table 1. Comparison of the proposed ViCoKD method with baseline distillation methods in terms of mAP [%]. Here, $AV_F \rightarrow V_F$ and $AV_F \rightarrow V_S$ denote distillation from an audio-visual teacher to a visual-only student with frame-level and sequence-level supervisions, respectively. Numbers in parentheses show the difference from the non-distilled student baseline. The best is emphasized using **bold**.

Environment	Method	MultiTrans [40]		MultiASL [23]		MultiTSF [25]	
		$AV_F \rightarrow V_F$	$AV_F \rightarrow V_S$	$AV_F \rightarrow V_F$	$AV_F \rightarrow V_S$	$AV_F \rightarrow V_F$	$AV_F \rightarrow V_S$
Home-1	Teacher	61.40		73.81		76.12	
	Student	60.77	57.59	63.24	55.91	73.17	61.17
	LogitKD	61.28 (+0.51)	60.48 (+2.89)	65.26 (+2.02)	58.78 (+2.87)	70.08 (−3.09)	60.67 (−0.50)
	ModalKD	58.22 (−2.55)	54.04 (−3.55)	69.50 (+6.26)	58.92 (+3.01)	73.67 (+0.50)	62.88 (+1.71)
	DualKD	58.84 (−1.93)	55.57 (−2.02)	68.42 (+5.18)	60.45 (+4.54)	73.12 (−0.05)	63.22 (+2.05)
	CoKD	61.34 (+0.57)	60.65 (+3.06)	75.33 (+12.09)	64.52 (+8.61)	76.89 (+3.72)	63.27 (+2.10)
	ViCoKD (Ours)	**66.27** (+5.50)	**62.66** (+5.07)	**76.83** (+13.59)	**65.23** (+9.32)	**82.91** (+9.74)	**65.28** (+4.11)
Home-2	Teacher	86.60		90.14		92.12	
	Student	84.06	80.83	82.25	81.67	86.89	79.16
	LogitKD	86.70 (+2.64)	85.55 (+4.72)	87.59 (+5.34)	84.94 (+3.27)	88.60 (+1.71)	84.73 (+5.57)
	ModalKD	85.43 (+1.37)	83.35 (+2.52)	91.45 (+9.20)	87.98 (+6.31)	90.92 (+4.03)	87.56 (+8.40)
	DualKD	86.41 (+2.35)	84.63 (+3.80)	85.58 (+3.33)	81.88 (+0.21)	87.71 (+0.82)	85.36 (+6.20)
	CoKD	88.64 (+4.58)	86.12 (+5.29)	91.77 (+9.52)	86.97 (+5.30)	88.42 (+1.53)	87.75 (+8.59)
	ViCoKD (Ours)	**90.86** (+6.80)	**89.31** (+8.48)	**91.80** (+9.55)	**89.89** (+8.22)	**91.27** (+4.38)	**89.21** (+10.05)

4.2. Quantitative Results

Table 1 presents the mAP results of the proposed ViCoKD method compared with competitive KD baselines across three backbones: MultiTrans [40], MultiASL [23], and MultiTSF [25]. The evaluation covers KD supervision scenarios: $AV_F \rightarrow V_F$ and $AV_F \rightarrow V_S$, which denote distillation from an audio-visual teacher to a visual-only student with frame-level and sequence-level supervisions, respectively.

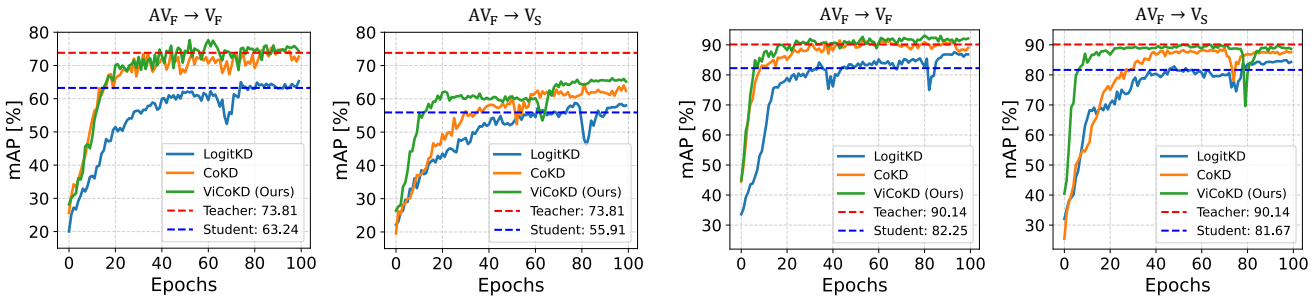
Home-1 Environment. This environment is particularly challenging due to its greater view disparity and weaker inter-view correlation. ViCoKD consistently delivered the highest mAP across all backbones and both KD supervision levels, with particularly large gains in the sequence-level regime. Using the MultiASL [23] backbone, ViCoKD improved mAP over the non-distilled student by +13.59 ($AV_F \rightarrow V_F$) and +9.32 ($AV_F \rightarrow V_S$), surpassing the teacher in both cases. The gains over CoKD (ViCoKD w/o Vi) across all experiments for both frame-level and sequence-level supervisions highlight the contribution of the proposed View-aware Consistency module in leveraging spatially localized cues under partial overlaps. In contrast, methods such as ModalKD and DualKD often underperformed the teacher, while LogitKD occasionally performed worse than the non-distilled student, underscoring their inability to adapt to scenarios where cross-view correspondence is weak or sparse.

Home-2 Environment. In this environment, the sensor views are closer in what they capture, which makes the setting less challenging than Home-1. Under these conditions, ViCoKD still secured top performance across all backbones. It achieved mAP improvements of up to +9.55 ($AV_F \rightarrow V_F$) and +10.05 ($AV_F \rightarrow V_S$) over the non-distilled student baseline. While LogitKD, ModalKD, and DualKD performed competitively in $AV_F \rightarrow V_F$, their advantages diminished under sequence-level supervision ($AV_F \rightarrow V_S$), reflecting the difficulty of exploiting weak labels. The consistent margins over CoKD reaffirm the effectiveness of the View-aware Consistency module in integrating human-centric spatial constraints into partially overlapping setups.

Table 2. Mean ± standard deviation of mAP [%] over three runs for the proposed ViCoKD method using different backbones.

Backbone	Home-1		Home-2	
	$AV_F \rightarrow V_F$	$AV_F \rightarrow V_S$	$AV_F \rightarrow V_F$	$AV_F \rightarrow V_S$
MultiTrans	66.27 ± 0.59	62.66 ± 0.77	90.86 ± 0.36	89.31 ± 0.40
MultiASL	76.83 ± 0.80	65.23 ± 0.74	91.80 ± 0.85	89.89 ± 0.57
MultiTSF	82.91 ± 0.65	65.28 ± 0.21	91.27 ± 0.79	89.21 ± 0.70

Key Findings. Across environments and backbones, ViCoKD consistently attained the highest performance, with particularly large gains in sequence-level supervision ($AV_F \rightarrow V_S$), where label sparsity is most severe. Moreover, as shown in Table 2, ViCoKD achieved high mAP while maintaining low variance (≤ 0.85 of mAP) over three runs, indicating stable convergence and robustness. Figure 3 further demonstrates that ViCoKD converged faster and reached higher mAP than LogitKD and CoKD across all supervision settings and environments. The clearest margin was observed under $AV_F \rightarrow V_S$ in both environments, where ViCoKD surpassed the non-distilled student early in train-



(a) Home-1 environment.

(b) Home-2 environment.

Figure 3. mAP [%] curves on the test set using the MultiASL [23] backbone under different distillation settings.

Table 3. Impact of the View-aware Consistency module when applied to the teacher network. $a \rightarrow b$ denotes the change in mAP from the original teacher (a) to the teacher with view-aware consistency (b).

Backbone	Home-1	Home-2
MultiTrans [40]	61.40 → 66.31 (+4.91)	86.60 → 88.88 (+2.28)
MultiASL [23]	73.81 → 75.99 (+2.18)	90.14 → 91.78 (+1.64)
MultiTSF [25]	76.12 → 82.16 (+6.04)	92.12 → 92.61 (+0.49)

ing and maintained high performance, while LogitKD and CoKD achieved lower scores or exhibit fluctuations. In addition, Table 3 shows that applying the View-aware Consistency module to the teacher network itself yielded consistent mAP improvements. These results indicate the general effectiveness of the module as a principled mechanism for improving multi-view representation learning.

4.3. Ablation Studies

We conducted ablation studies to quantify the contribution of each component in the proposed ViCoKD method, with the experimental results shown in Table 4. Note that some changes resulted in differences smaller than 0.85 of mAP, which can be considered minor and likely within run-to-run variation, while others caused substantial drops, as indicated in Table 2.

View-aware Consistency Module. Eliminating Confidence Weighting (w/o ConfW) or the human-detection mask (w/o Mask) yielded notable drops of up to -4.50 of mAP, particularly in the Home-1 environment within $AV_F \rightarrow V_S$. This indicates that weighting supervision by prediction reliability and masking irrelevant views are both crucial for robust cross-view alignment. Replacing the divergence from JS to KL (JS → KL) also caused a small performance drop, showing the benefit of using a symmetric divergence.

Knowledge Distillation Strategies. Removing Feature-level Distillation (w/o FD) resulted in small drops in most settings, whereas omitting Logit-level Distillation (w/o LD) caused a large decrease in $AV_F \rightarrow V_S$. This suggests that

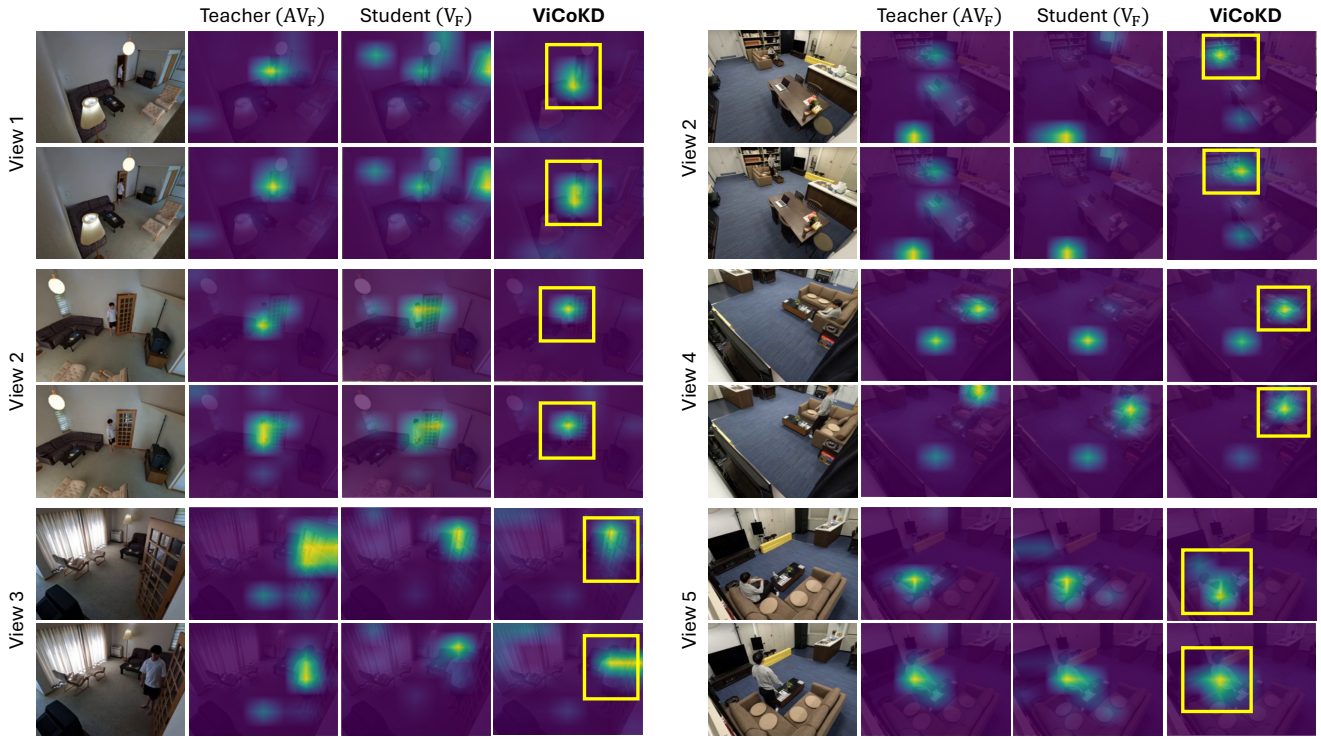
Table 4. Ablation study on the proposed ViCoKD method. Results are reported in mAP [%] using the MultiASL [23] backbone. Numbers in parentheses show the difference from ViCoKD.

Method	Home-1		Home-2	
	$AV_F \rightarrow V_F$	$AV_F \rightarrow V_S$	$AV_F \rightarrow V_F$	$AV_F \rightarrow V_S$
ViCoKD	76.83	65.23	91.80	89.89
View-aware Consistency Module				
w/o ConfW	73.15 (−3.68)	62.27 (−2.96)	90.51 (−1.29)	88.90 (−0.99)
w/o Mask	72.33 (−4.50)	63.34 (−1.89)	90.08 (−1.72)	89.19 (−0.70)
JS → KL	75.50 (−1.33)	64.87 (−0.36)	90.11 (−1.69)	89.62 (−0.27)
Knowledge Distillation Strategies				
w/o FD	76.52 (−0.31)	64.85 (−0.38)	90.44 (−1.36)	88.90 (−0.99)
w/o LD	76.23 (−0.60)	61.51 (−3.72)	92.02 (+0.22)	88.76 (−1.13)

while both FD and LD contributed to the overall improvement, LD was particularly important in scenarios with only sequence-level supervision.

4.4. Qualitative Results

Figure 4 compares the attention maps of the teacher, the baseline non-distilled student, and the proposed ViCoKD for two representative actions across multiple sensor views. In the Home-1 environment example ("enter the room"), the baseline student often exhibited diffuse or background-focused attention, missing the key human regions. In contrast, ViCoKD consistently produced sharp, human-centric activation that tightly aligned with the actor's location across all views, even when the viewpoint changed substantially. Similarly, in the Home-2 environment example ("stand up"), the attention of the student was scattered and shifted toward irrelevant areas such as furnitures. In contrast, ViCoKD focused precisely on the subject's body, capturing the motion cues critical for action recognition. In several views, ViCoKD attended more precisely to human regions where the action occured, indicating that the proposed method transfered and refined the teacher's learned features, enabling the model to focus on discriminative regions despite partial view overlaps and reduced modalities.



(a) Home-1 environment: Action *"enter the room"* on Views 1, 2, and 3.

(b) Home-2 environment: Action *"stand up"* on Views 2, 4, and 5.

Figure 4. Qualitative comparison of attention maps for the teacher, baseline student, and the proposed ViCoKD method using the MultiASL [23] backbone on the MultiSensor-Home dataset [25]. Each row corresponds to a different sensor view. ViCoKD produces more precise and human-centric attentions (yellow boxes) compared to the baseline student.

4.5. Limitations and Future Work

While the proposed ViCoKD method demonstrated strong performance in partially overlapping multi-view action recognition, it has several limitations. First, the framework relies on pre-trained human detectors to generate human-detection masks, which may introduce errors under heavy occlusion, poor lighting, or unconventional poses. These errors can affect the reliability of View-aware Consistency supervision. Second, the method depends on a strong multi-modal teacher, which limits applicability in domains where high-quality multi-modal data are scarce or unavailable. Finally, the current experiments are limited by available multi-view action recognition datasets, which typically provide only audio and visual modalities and lack other informative sources such as depth or skeleton signals.

Future work will explore end-to-end learned visibility estimation for improved robustness, as well as extending ViCoKD to additional modalities and cross-domain generalization.

5. Conclusion

We addressed the challenge of multi-view action recognition in partially overlapping sensor setups under modality and annotation-limited conditions. We introduced **Vi**ew-aware **Cro**ss-modal **K**nowledge **D**istillation (ViCoKD), a novel knowledge distillation method that transfers supervision from a fully supervised multi-modal teacher to a constrained student through cross-modal attention and a view-aware consistency mechanism. The proposed human-detection masks with confidence-weighted Jensen–Shannon divergence [22] ensure that distillation focuses on frames with reliable, view-consistent evidence. Extensive experiments on the MultiSensor-Home [25] dataset demonstrated that ViCoKD achieved substantial gains over non-distilled students and competitive baselines, with consistent improvements across backbones and environments. These results highlight the importance of explicitly modeling View-aware Consistency in partially overlapping scenarios and provide a foundation for more robust multi-view action recognition in real-world deployments.

Acknowledgment

This work was partly supported by JSPS KAKENHI JP21H03519 and JP24H00733. The computation was carried out using the General Projects on supercomputer "Flow" at IT Center, Nagoya University.

References

[1] Alaa Eldin Abdelaal, Apeksha Avinash, Megha Kalia, Gregory D Hager, and Septimiu E Salcudean. A multi-camera, multi-view system for training and skill assessment for robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery*, 15(8):1369–1377, 2020. 1

[2] Abien Fred Agarap. Deep learning using Rectified Linear Units (ReLU). *Computing Research Repository arXiv Preprints,* arXiv:1803.08375, pages 1–7, 2018. 5

[3] Yanbei Chen, Yongqin Xian, A Sophia Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7016–7025, 2021. 2

[4] Srijan Das and Michael S Ryoo. ViewCLR: Learning self-supervised video representation for unseen viewpoints. In *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5573–5583, 2023. 1, 2

[5] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota Smarthome: Real-world activities of daily living. In *Proceedings of the 17th IEEE/CVF International Conference on Computer Vision*, pages 833–842, 2019. 2

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the 2021 International Conference on Learning Representations*, pages 1–22, 2021. 4

[7] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the 15th European Conference on Computer Vision, Part VIII*, pages 103–118, 2018. 5

[8] Nuno Cruz Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. Distillation multiple choice learning for multimodal action recognition. In *Proceedings of the 2021 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2755–2764, 2021. 2

[9] Yuan Gong, Yu-An Chung, and James R Glass. AST: Audio Spectrogram Transformer. In *Proceedings of the 22nd International Conference on Speech Communication and Technology*, pages 571–575, 2021. 4

[10] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2827–2836, 2016. 2

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Proceedings of the 2015 NIPS Deep Learning and Representation Learning Workshop*, pages 1–9, 2015. 1, 2, 5

[12] Vijay John and Yasutomo Kawanishi. Frame-level latent embedding using weak labels for multi-view action recognition. In *Proceedings of the 7th IEEE International Conference on Multimedia Information Processing and Retrieval*, pages 235–238, 2024. 1, 2

[13] Vijay John and Yasutomo Kawanishi. Multimodal cascaded framework with multimodal latent loss functions robust to missing modalities. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(9): 1–21, 2025. 1

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 2015 International Conference on Learning Representations*, pages 1–15, 2015. 5

[15] Takumi Kobayashi. Two-way multi-label loss. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7476–7485, 2023. 4, 5

[16] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22 (1):79–86, 1951. 4

[17] Qun Li, Rui Yang, Fu Xiao, Bir Bhanu, and Feng Zhang. Attention-based anomaly detection in multi-view surveillance videos. *Knowledge-Based Systems*, 252(109348):1–11, 2022. 1

[18] 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*, 42(10):2684–2701, 2019. 1, 2

[19] Wenxuan Liu, Xian Zhong, Zhuo Zhou, Kui Jiang, Zheng Wang, and Chia-Wen Lin. Dual-recommendation disentanglement network for view fuzz in action recognition. *IEEE Transactions on Image Processing*, 32:2719–2733, 2023. 1, 2

[20] Yang Liu, Keze Wang, Guanbin Li, and Liang Lin. Semantics-aware adaptive knowledge distillation for sensor-to-vision action recognition. *IEEE Transactions on Image Processing*, 30:5573–5588, 2021. 5

[21] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with warm Restarts. *Computing Research Repository arXiv Preprints,* arXiv:1608.03983, pages 1–16, 2016. 5

[22] María Luisa Menéndez, Julio Angel Pardo, Leandro Pardo, and María del C Pardo. The Jensen–Shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318, 1997. 2, 5, 8

[23] Trung Thanh Nguyen, Yasutomo Kawanishi, Takahiro Komamizu, and Ichiro Ide. Action selection learning for multilabel multiview action recognition. In *Proceedings of the 2024 ACM Multimedia Asia Conference*, pages 1–7, 2024. 1, 2, 5, 6, 7, 8

[24] Trung Thanh Nguyen, Yasutomo Kawanishi, Vijay John, Takahiro Komamizu, and Ichiro Ide. Action selection learning for weakly labeled multi-modal multi-view action recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, pages 1–22, 2025. 1

[25] Trung Thanh Nguyen, Yasutomo Kawanishi, John Vijay, Takahiro Komamizu, and Ichiro Ide. MultiSensor-Home: A wide-area multi-modal multi-view dataset for action recognition and Transformer-based sensor fusion. In *Proceedings of*

the 19th IEEE International Conference on Automatic Face and Gesture Recognition, pages 1–10, 2025. 1, 2, 5, 6, 7, 8

[26] Adeshina Sirajdin Olagoke, Haidi Ibrahim, and Soo Siang Teoh. Literature survey on multi-camera system and its application. *IEEE Access*, 8:172892–172922, 2020. 1

[27] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019. 2

[28] Andres Perez, Valentina Sanguineti, Pietro Morerio, and Vittorio Murino. Audio-visual model distillation using acoustic images. In *Proceedings of the 2020 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2854–2863, 2020. 2

[29] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings of the 19th IEEE/CVF International Conference on Computer Vision*, pages 5213–5224, 2023. 2, 5

[30] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *Proceedings of the 2015 International Conference on Learning Representations*, pages 1–13, 2015. 2

[31] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158, 2011. 5

[32] Ketul Shah, Anshul Shah, Chun Pong Lau, Celso M de Melo, and Rama Chellappa. Multi-view action recognition using contrastive learning. In *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3381–3391, 2023. 1, 2

[33] 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 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016. 1, 2

[34] Fida Mohammad Thoker and Juergen Gall. Cross-modal knowledge distillation for action recognition. In *Proceedings of the 2019 IEEE International Conference on Image Processing*, pages 6–10, 2019. 5

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:6000–6010, 2017. 4

[36] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. YOLOv10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024. 4, 5

[37] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning and recognition. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2649–2656, 2014. 1, 2

[38] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10941–10950, 2020. 4

[39] Yingyuan Yang, Guoyuan Liang, Can Wang, and Xiaojun Wu. Trunk–branch contrastive network with multi-view deformable aggregation for multi-view action recognition. *Pattern Recognition*, 169(111923):1–10, 2026. 1

[40] Masahiro Yasuda, Yasunori Ohishi, Shoichiro Saito, and Noboru Harada. Multi-view and multi-modal event detection utilizing Transformer-based multi-sensor fusion. In *Proceedings of the 47th IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4638–4642, 2022. 1, 2, 5, 6, 7

[41] Masahiro Yasuda, Noboru Harada, Yasunori Ohishi, Shoichiro Saito, Akira Nakayama, and Nobutaka Ono. Guided masked self-distillation modeling for distributed multimedia sensor event analysis. *Computing Research Repository arXiv Preprints,* arXiv:2404.08264, pages 1–13, 2024. 1, 2

[42] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10502–10511, 2019. 4

[43] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of the 2017 International Conference on Learning Representations*, pages 1–13, 2017. 2

[44] Leying Zhang, Zhengyang Chen, and Yanmin Qian. Knowledge distillation from multi-modality to single-modality for person verification. *Proceedings of the 22nd International Conference on Speech Communication and Technology*, pages 1897–1901, 2021. 2