

Received 6 June 2024, accepted 10 July 2024, date of publication 19 July 2024, date of current version 17 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3431227

RESEARCH ARTICLE

Privacy-Safe Action Recognition via Cross-Modality Distillation

YUHYUN KIM¹, (Member, IEEE), JINWOOK JUNG¹, (Member, IEEE),
HYEONCHEOL NOH¹, (Student Member, IEEE), BYUNGTAEE AHN²,
JUNGHYE KWON³, AND DONG-GEOL CHOI¹, (Member, IEEE)

¹Department of Information and Communication Engineering, Hanbat National University, Daejeon 34158, Republic of Korea

²Korea Institute of Machinery and Materials, Daejeon 34103, Republic of Korea

³Division of Hematology and Oncology, Department of Internal Medicine, College of Medicine, Chungnam National University, Daejeon 34134, Republic of Korea

Corresponding author: Dong-Geol Choi (dgchoi@hanbat.ac.kr)

This work was supported in part by Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport under Grant RS-2023-00238018; in part by the NRF grant funded by Korean Government Ministry of Science and ICT (MIST) under Grant 2023R1A2C1007428; and in part by the Research Fund of Hanbat National University, in 2022, under Grant 202204870001.

ABSTRACT Human action recognition systems enhance public safety by detecting abnormal behavior autonomously. RGB sensors commonly used in such systems capture personal information of subjects and, as a result, run the risk of potential privacy leakage. On the other hand, privacy-safe alternatives, such as depth or thermal sensors, exhibit poorer performance because they lack the semantic context provided by RGB sensors. Moreover, the data availability of privacy-safe alternatives is significantly lower than RGB sensors. To address these problems, we explore effective cross-modality distillation methods in this paper, aiming to distill the knowledge of context-rich large-scale pre-trained RGB-based models into privacy-safe depth-based models. Based on extensive experiments on multiple architectures and benchmark datasets, we propose an effective method for training privacy-safe depth-based action recognition models via cross-modality distillation: cross-modality mixing distillation. This approach improves both the performance and efficiency by enabling interaction between depth and RGB modalities through a linear combination of their features. By utilizing the proposed cross-modal mixing distillation approach, we achieve state-of-the-art accuracy in two depth-based action recognition benchmarks. The code and the pre-trained models will be available upon publication.

INDEX TERMS Action recognition, knowledge distillation, cross-modality distillation, deep learning, multi modal, privacy-safe.

I. INTRODUCTION

Action recognition is an important task in computer vision. Actions can only be differentiated using video modality, in which sequences of image frames provide a temporal dimension to the data. Contextual information is extracted from the spatial dimension and dynamics information is extracted from the temporal dimension. Associated methods include simple action classification based on RGB-trimmed videos [1], [2], temporal action localization in untrimmed videos [3], [4], and dense video captioning [5], [6]. In recent

years, the availability of large-scale benchmark datasets has improved action recognition performance, motivating its implementation in real-world applications, such as surveillance systems.

Real-world action recognition applications, however, raise concerns about the privacy of subjects captured in videos. The General Data Protection Regulation [7], published by the European Union, reflects the rising privacy concerns. The root cause of such concerns is that RGB sensors record more information, including details about the subjects' appearance, than is required for action recognition. Thus, RGB sensor-based systems inherently run the risk of privacy leakage owing to unauthorized access to sensor

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues¹.

data. As an alternative, privacy-safe depth-based systems have been tested in real-world environments [8], [9]. The widespread adoption of privacy-safe sensors can prevent private information from entering the automated system, thereby avoiding potential privacy leakage.

Utilizing privacy-safe modalities, such as depth/thermal sensors, exhibits certain advantages and disadvantages. As observed empirically in the NTU RGB+D 120 benchmark [10], depth-based models exhibit significantly worse performance compared to RGB-based models. However, ensembling multiple modalities improves performance quality. On the other hand, they are advantageous in scenarios involving shape-related object interactions, such as distinguishing between a box and a paper. This is in contrast to RGB sensors, which excel in appearance-related recognition, e.g., distinguishing when a subject wearing a jacket or a backpack. Appearance information is crucial in action recognition, but it can also be considered to introduce bias [11] as appearance does not comprise temporally associated information (i.e., single-frame classification is possible). In addition to the limitations of each modality, data availability is a major challenge for non-RGB modalities. Multi-modal approaches to action recognition have garnered increasing attention, yet exhibit performance standards far below the scale of RGB-based datasets. The availability of large-scale datasets is crucial for pre-training models (e.g., a foundation model) for downstream tasks. Thus, data scarcity for non-RGB modalities may limit downstream task models.

Cross-modality distillation methods are often used to address the limitations of a single modality. This enables the rich knowledge of RGB-based models to be distilled into privacy-safe models. Cross-modality distillation has been studied in the context of various computer vision tasks [14], [15], [16]. Taking motivation from previous works, we consider unidirectional cross-modality distillation in this paper, with the sole focus on the test-time performance of the privacy-safe model.

To ensure thorough benchmarking of cross-modality distillation for privacy-safe action recognition, we implement various distillation methods based on simple baselines to state-of-the-art methods—a simple distillation [12] from RGB to depth models and pre-training depth-based models using model-inferenced depth [18] in large-scale datasets. We propose a new method for cross-modality distillation: cross-modality mixing distillation. Owing to the limited range of commercially available depth sensors [19], their implementation is usually limited to indoor environments. Among the various available RGB-D benchmarks, we consider Toyota Smarthome [20] and NTU RGB+D 120 [10] as our main target downstream tasks. Optionally, we use Kinetics-400 [21] when a large-scale RGB dataset is required.

The primary contributions of this paper are 3-fold:

- Various cross-modality distillation methods are benchmarked with various action recognition methods for human action recognition.

- Various distillation methods are adapted to cross-modality distillation in action recognition methods.
- A novel architecture-agnostic cross-modality distillation method, which can be viewed in Figure 1, is proposed that significantly outperforms baseline methods.

II. RELATED WORKS

A. ACTION RECOGNITION

Significant advances have been made in action recognition since the emergence of large-scale video datasets [4], [22], [23]. Besides the simplest classification of trimmed video clips [53], significant research has been conducted on untrimmed video recognition [52], temporal action localization [54], video captioning [41], [42], ground video situation recognition [55], and video question-answering [56], [57], [58]. In terms of model architecture, training using large-scale datasets has been adopted, yielding 3D convolution-based networks [23], [24], [25], [26] with strong recognition capabilities. This is enabled by training the models to aggregate spatial and temporal information simultaneously. Recently proposed action recognition models exploit powerful attention-based transformer architectures, setting new state-of-the-art standards [47], [48], [49], [50], [51].

B. MULTI-MODAL ACTION RECOGNITION

Utilizing complementary information obtained from multiple modalities has been demonstrated to yield better performance and generalizability compared to unimodal approaches on various computer vision benchmarks [27], [28]. The most widely used modalities for action recognition include RGB frames, optical flow [35], depth [44], and human body pose [43]. However, multimodal recognition does not allow combinations of input modalities to be varied during the test time [29]. In this context, cross-modal knowledge distillation [30], which combines multimodal training with unimodal testing, is considered to be the appropriate solution for privacy-preserving applications.

C. KNOWLEDGE DISTILLATION

Initially, knowledge distillation comprised the transfer of knowledge from larger networks (teachers) to smaller networks (students) via joint training of pairs [12]. Subsequently, this idea has been expanded to semi-supervised learning, in which the size of the student network is equal to or larger than that of the teacher network [13]. In this study, we focus on knowledge distillation across different modalities [30], aiming to transfer the knowledge of one modality to another and improve the test time performance of the final unimodal model. Even though cross-modal knowledge distillation has not yet been investigated in detail in the context of action recognition, we adopt knowledge distillation as the baseline [12], [31], followed by our proposed method of cross-modal mixing distillation.

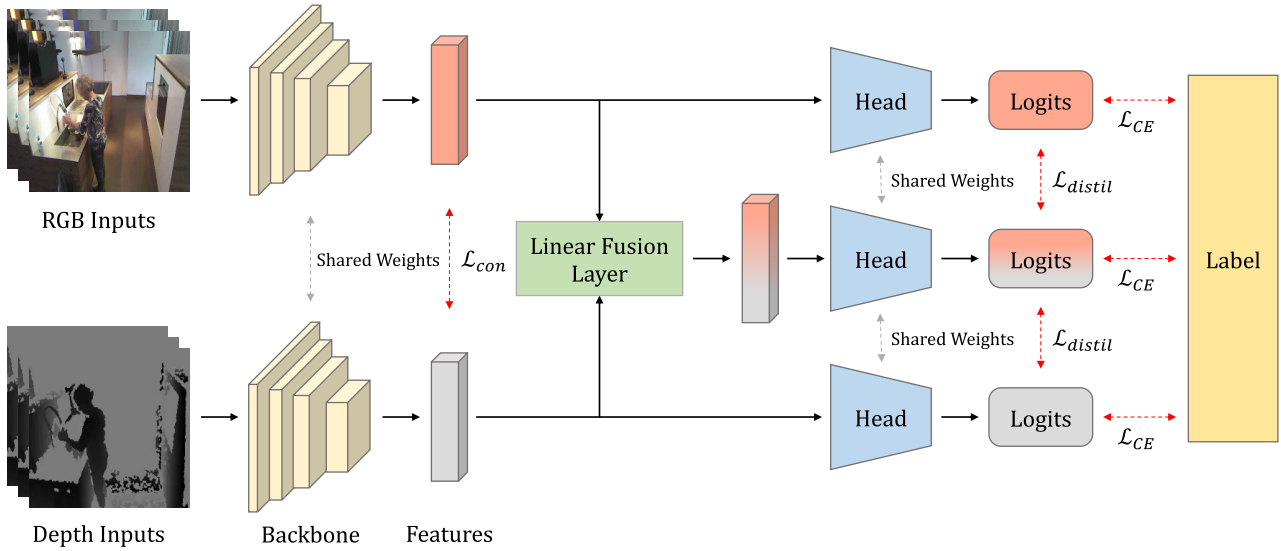


FIGURE 1. The proposed cross-modal mixing distillation framework. The proposed method extracts features for each modality via the shared backbone network, linearly combines them to produce an additional integrated feature, and transmits them through a shared head to obtain three logits (RGB logit, depth logit, and modality-mixed logit). The consistency loss between modality features and the distillation loss among the 3 logits facilitate the exchange of complementary information between modalities.

III. METHODS

In this section, we provide an overview of various distillation methods. Subsequently, we introduce a cross-modal mixed distillation strategy that enhances the efficiency of inference processes while also emphasizing privacy considerations and their practical applicability in real-world scenarios.

Our research methodology distinctly differs from previous studies. While most prior research focused on methods requiring skeleton annotations, our approach can be utilized in environments with limited out skeleton annotations. Specifically, this study emphasizes improving performance using only RGB and Depth modalities without the need for skeleton annotations. Additionally, by applying the distillation technique, we demonstrated the ability to maintain robust and stable performance across flexible inputs, accommodating both uni-modal and multi-modal types. These aspects are the key elements that distinguish our methodology from existing methods.

A. PSEUDO-DEPTH

In deep learning, fine-tuning with pre-trained weights [37], [38] has recently gained widespread recognition for its efficacy. Particularly in action recognition, generation and fine-tuning of pre-trained weights based on supervised learning using the Kinetics [21], [32], [33] RGB modality dataset is common. However, with respect to depth modality, there is a noticeable dearth of large-scale labeled data akin to Kinetics, which results in the absence of pre-trained weights trained with a substantial dataset via supervised learning. To address this issue, a pseudo-depth map is generated by implementing the Zoe depth [34] model on a large-scale RGB dataset, Kinetics-400. The model on these

data to generate pre-trained weights via supervised learning, ensuring sufficient depth knowledge. These pre-trained weights are subsequently used to fine-tune the target dataset.

B. NAIVE DISTILLATION

We also consider some conventional distillation methods. The objective of this study is to examine the effects of leveraging information from a large-scale RGB dataset on the training quality of a small-scale RGB-depth paired dataset. To this end, knowledge distillation is performed with identical teacher and student model architectures. The knowledge distillation process is subdivided in two phases. In the first stage, a large-scale RGB dataset, Kinetics 400 [21], is employed to supervise the training of the teacher network. This enables us to generate supervised learning-based pre-trained weights that capture ample RGB knowledge. In the second stage, the teacher model trained in the first stage is frozen, and RGB images are input into it and corresponding depth images are input into the student model from a dataset comprising RGB-Depth pairs. Subsequently, we compute the Mean Squared Error (MSE) loss based on the final layer's features to obtain the transferred knowledge. The primary objective of this methodology is to discover potent strategies for the integration of information from larger datasets, including RGB rich spatial information, using conventional methods for smaller datasets, which are paired combinations of RGB and other sensors with lesser information. By learning from extensive RGB datasets, it is possible to transfer RGB rich spatial information to smaller datasets of different modalities, thereby augmenting learning outcomes.

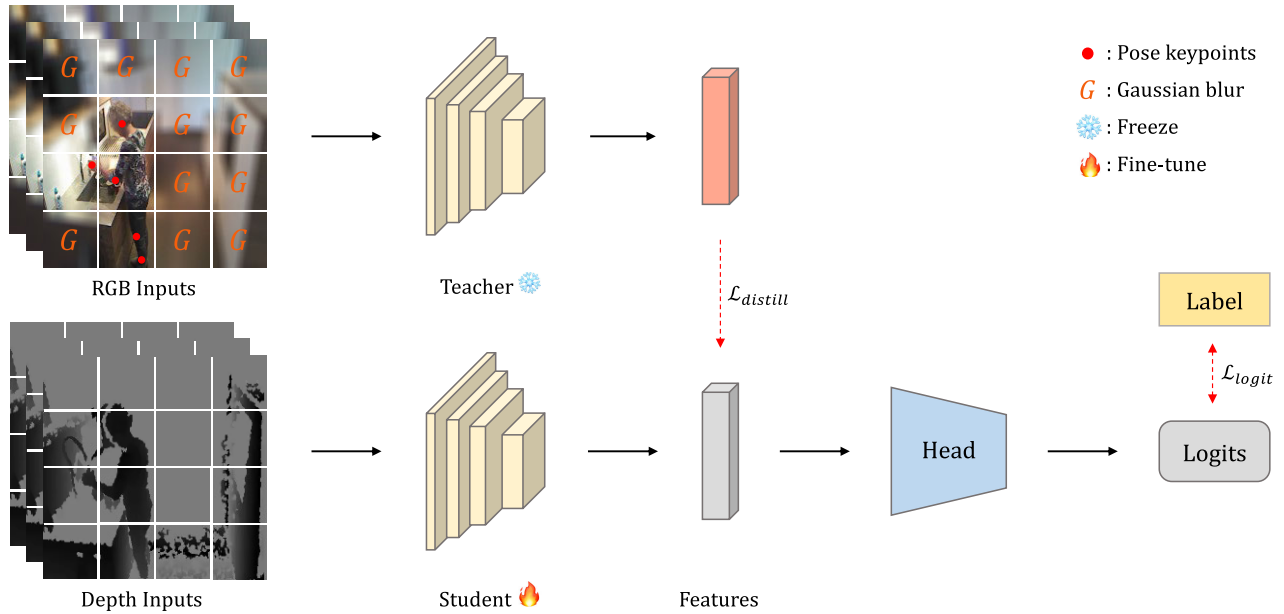


FIGURE 2. The body pose-aware distillation framework. The RGB-depth paired frames are divided into patches of equal size, which are in turn categorized as foreground or background using keypoint labels. Gaussian blur is applied to the background patches and fed into the teacher model, with the final layer yielding an output feature. This feature is then used for knowledge distillation to the student model, which accepts depth images as input. During this process, the teacher model remains frozen, and only the student model is trained via distillation.

C. BODY POSE-AWARE DISTILLATION

Conventional action recognition methods [35], [36] employ feature distillation without weighting and transfer the teacher model's entire knowledge to the student model without modification. Consequently, the student model learns not only the significant foreground information, but also the comparatively less important background information. To address this issue and enable the student model to focus more on the foreground, we implement body pose-aware feature distillation, inspired by the pose-aware attention block [17] utilized in conventional action recognition that leverages body pose information.

The implemented method, termed body pose-aware distillation, is designed to help the model place greater focus on the object's motion. It distills knowledge by applying Gaussian blur to the background of the input images obtained from the teacher model. Its framework is illustrated in Figure 2. In the first step, supervised learning is performed using the large-scale action recognition RGB dataset, Kinetics 400 [21], which is subsequently used as the teacher model. Next, RGB images from a smaller RGB-Depth paired dataset are subdivided into patches of equal size. Using keypoint labels, each patch is classified as either foreground or background. Gaussian blur is applied to the categorized background patches, which are then input into the teacher model. The features obtained from the final layer of the teacher model are distilled into the student model, and the depth images are also input into it. Throughout this process, the teacher model is frozen and only the student model is trained. As both foreground and background information are

required during the model's learning process, the application of Gaussian blur helps make the model focus more on the foreground, while retaining a basic level of background information.

$$x_{(p,t)} = \begin{cases} x_{(p,t)} & \text{if } (t,k,x,y) \in \mathcal{K} \\ \mathcal{G} \cdot x_{(p,t)} & \text{otherwise} \end{cases} \quad (1)$$

The input frames are denoted by $x \in \mathbb{R}^{C \times T \times H \times W}$ and consist of T frames with a size of $C \times H \times W$, where C denotes the number of channels, H denotes the height of the frame, and W denotes the width of the frame. To categorize the foreground and background, we subdivide each frame into N 2D patches, $N = HW/P^2$, each with a size of $P \times P$. We tensorize each patch $x_{(p,t)} \in \mathbb{R}^{(C \times P^2)}$, where $p = 1, \dots, N$ denotes the spatial locations, and $t = 1, \dots, T$ denotes an index for the frames. The keypoints are denoted by $\mathcal{K} = (k, t, x, y) : 1 \leq k \leq K, 1 \leq t \leq T$, where K denotes the number of keypoints. Gaussian blur was applied to all teacher input patches not containing any keypoint ($\mathcal{G}$).

Distillation loss uses mean-squared error (MSE) loss:

$$\mathcal{L}_{distil} = \mathcal{L}_{MSE}(F_S, F_T) \quad (2)$$

F_S denotes the final feature of the student model and F_T denotes the final feature of the teacher model.

Logit loss uses cross entropy (CE) loss:

$$\mathcal{L}_{logit} = \mathcal{L}_{CE}(\text{Logits}, \text{Label}) \quad (3)$$

Logits comprise the final output logits of the student model, and the label is taken to be the correct answer label.

The total training objective of body pose-aware feature distillation is taken to be the weighted sum of the aforementioned losses:

$$\mathcal{L}_{total} = \lambda \mathcal{L}_{distill} + \mathcal{L}_{logit} \quad (4)$$

where λ denotes a hyperparameter with a default value of 10.

D. CROSS-MODAL MIXING DISTILLATION

By design, the proposed cross-modal mixing distillation methodology is based on interaction between the depth and RGB modalities. By linearly combining the RGB and Depth features, this distillation method enables more effective cross-modality interaction and achieves improved performance and efficient inferencing. Importantly, being a distillation approach, it allows for flexible inputs, accommodating both uni-modal and multi-modal types, thereby increasing its versatility and applicability. The framework of this process is depicted in Figure 1.

1) MULTI-MODAL ALIGNMENT

Image data captured using different sensors differ in terms of various aspects, such as field-of-view, resolution, and location. These differences arise from various factors, including sensor characteristics, shooting environments, and shooting methods; as a result, the same scene can be perceived differently depending on the sensor used. These differences may degrade the performance of multimodal or cross-modal models. To resolve this problem, we propose an alignment method at the feature level. All modalities share the same backbone network, and use it to generate its features. These generated features are mapped to similar locations using consistency loss, indicated by $\mathcal{L}_{con}$ in Figure 1. Consistency loss, which measures the similarity between features and calculates the difference, is defined as follows:

$$\mathcal{L}_{con} = \frac{1}{N} \sum_{i=1}^N \|x_i - y_i\|_2^2 \quad (5)$$

We denote the batch number by N , an RGB modality feature by x , and a depth modality feature by y , and induce each modality to generate features sharing similar patterns. The proposed alignment allows them to indicate similar locations. Thus, ensuring the consistency of features across data extracted from diverse sensors enables integrated processing, thereby fostering effective learning.

2) MIXING DISTILLATION

We propose a mixing distillation method to determine RGB and depth modalities complementarily. First, we extract the features of each modality using a shared backbone network. Subsequently, we consider a linear combination of the extracted modality features, yielding three features: one mixed feature, and two unimodal features. These are transmitted through a fully connected (FC) layer to obtain three logits. The obtained logits are optimized using cross-entropy loss

for ground-truth labels. This corresponds to $\mathcal{L}_{CE}$ in Figure 1, where an average of different cross-entropy losses is used.

$$\begin{aligned} \mathcal{L}_{logit} = & \text{mean}(\mathcal{L}_{CE}(\text{Logits}_x, \text{Label}) \\ & + \mathcal{L}_{CE}(\text{Logits}_y, \text{Label}) \\ & + \mathcal{L}_{CE}(\text{Logits}_m, \text{Label})) \end{aligned} \quad (6)$$

Logits_x and Logits_y denote logits obtained by processing each modality using a shared backbone network and shared FC layer. Logits_m are obtained by mixing the output features obtained by transmitting each modality first through the shared backbone network and then through the shared FC layer.

In addition, knowledge distillation is performed to enable the knowledge of different modalities to be exchanged and complemented. As the difference between different modalities is considerably large, direct knowledge distillation is avoided. Instead, we perform knowledge distillation based on the logits obtained using the integrated feature. A previous study [39] demonstrated that using MSE loss instead of Kullback-Leibler divergence (KLD) loss in the knowledge distillation process yields more effective results. Therefore, MSE loss is adopted in this paper to foster harmony between features of different modalities. This is denoted by $\mathcal{L}_{distil}$ in Figure 1.

$$\begin{aligned} \mathcal{L}_{distil} = & \mathcal{L}_{MSE}(\text{Logits}_x, \text{Logits}_m) \\ & + \mathcal{L}_{MSE}(\text{Logits}_y, \text{Logits}_m) \end{aligned} \quad (7)$$

This approach enables more effective information transfer than using each modality's information independently, and increases the accuracy of action recognition. Moreover, the proposed model is not constrained by a specific constituent modality, and functions effectively even when there is only one modality. This demonstrates that the model can respond flexibly to various input conditions, which aids the effective management of maintenance costs. Consequently, the proposed model is expected to exhibit significantly enhanced applicability in real-world environments.

The final loss is expressed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{logit} + \mathcal{L}_{distil} + \mathcal{L}_{con} \quad (8)$$

IV. EXPERIMENTS

A. DATASETS

Toyota Smarthome (Toyota) dataset [20] is a real-world dataset corresponding to activities of daily living. (e.g., making coffee, watching TV) It consists of 16,115 RGB+D video clips of 31 activity classes performed by 18 subjects. For the Toyota Smarthome dataset, we followed cross-subject (CS) and cross-view (CV1, CV2) evaluation protocols [20].

NTU RGB+D 120(NTU) dataset [10] is a real-world dataset corresponding to daily, mutual, and health-related actions. (e.g., drinking water, blowing one's nose, kicking) It consists of 114,480 RGB+D video clips arranged into 120 classes performed by 106 subjects. For the NTU RGB+D 120 dataset, we follow the cross-subject (x-sub) and cross-setup (x-set) evaluation protocols [10].

TABLE 1. Performances of I3D-based video classification models trained using various methodologies on the Toyota Smarthome dataset.

Model	Method	Modality		Toyota		
		RGB	Depth	CS	CV1	CV2
I3D [23]	RGB fine-tune	✓		76.76	48.52	62.73
	Depth scratch		✓	71.04	38.11	51.43
	Depth fine-tune		✓	73.46	46.36	56.43
	Distillation method					
	Distillation during pre-training with pseudo-depth		✓	66.78	30.90	46.46
	Naive distillation		✓	72.80	41.13	52.43
	Body pose-aware distillation		✓	73.76	46.73	59.14
	Cross-modal mixing distillation (Ours)		✓	74.43	53.46	61.09

TABLE 2. Performances of various video classification models trained using various methodologies on the Toyota Smarthome dataset and the NTU RGB+D 120 dataset. Existing state-of-the-art (SOTA) methods achieve higher performance by utilizing skeleton information for training, which makes direct comparison with our method challenging. However, our method demonstrates respectable performance using only depth modality for inference, despite the relatively insufficient information it provides. †: Utilize the skeleton modality for training.

Model	Method	Modality		Toyota			NTU	
		RGB	Depth	CS	CV1	CV2	x-sub	x-set
STAR-Transformer [†] [60]	RGB-Skeleton fine-tune	✓		-	-	-	90.3	92.7
3Mformer [†] [61]	RGB-Skeleton fine-tune	✓		-	-	-	92.0	93.8
TimeSformer [41]	RGB fine-tune	✓		77.37	56.50	66.12	70.48	71.82
	Depth scratch		✓	68.75	17.64	44.02	49.48	11.84
	Cross-modal mixing distillation (ours)		✓	74.08	50.91	60.77	62.94	79.54
Video Swin Transformer [46]	RGB fine-tune	✓		75.87	53.22	62.78	87.40	87.23
	Depth scratch		✓	52.57	30.81	32.84	49.99	39.47
	Cross-modal mixing distillation (ours)		✓	69.69	50.43	57.20	83.71	84.10

TABLE 3. Ablation study on the impact of various loss components on model performance for the Toyota dataset.

Model	Variants			Toyota		
I3D [23]	$\mathcal{L}_{logit}$	$\mathcal{L}_{con}$	$\mathcal{L}_{distil}$	CS	CV1	CV2
				72.80	41.13	52.43
	✓			73.02	42.35	56.12
	✓	✓		70.67	48.86	57.79
	✓		✓	74.05	50.03	60.17
	✓	✓	✓	74.43	53.46	61.09

B. IMPLEMENTATION DETAILS

Pre-trained weights obtained by performing supervised learning on Kinetics-400 are used, except in methods involving Scratch. The sizes of the frames constituting the clips of all datasets are set to 224×224 . During the implementation of cross-modal mixing distillation, the mixing ratio of depth-to-RGB is set to 7:3 across all datasets.

1) I3D

The standard I3D [23] configuration is adopted, with input clips of the model comprising 64 frames sampled in two steps in the case of the Toyota Smarthome dataset [20] and

64 frames sampled in two steps in the case of NTU RGB+D 120 [10].

2) VIDEO SWIN TRANSFORMER

The standard configuration is utilized for a video swine transformer [45] using the base model. The model's input clips consist of 64 frames sampled in two steps for the Toyota Smarthome dataset [20] and 32 frames sampled in two steps for the NTU RGB+D 120 [10] dataset.

3) TIMESFORMER

The standard configuration of TimeSformer [40] is adopted using the base model. Input clips of the model are composed of eight frames sampled in 32 steps for the Toyota Smarthome dataset [20] and 16 frames sampled in four steps for the NTU RGB+D 120 dataset [10].

V. MAIN RESULTS AND ANALYSIS

In this section, we present the quantitative results obtained from our experiments and an analysis of the action recognition performance of the proposed method. We propose a cross-modal mixing distillation method that uses cross-modal learning during training, but can operate effectively with both uni-modal or cross-modal inputs during inference. Compared

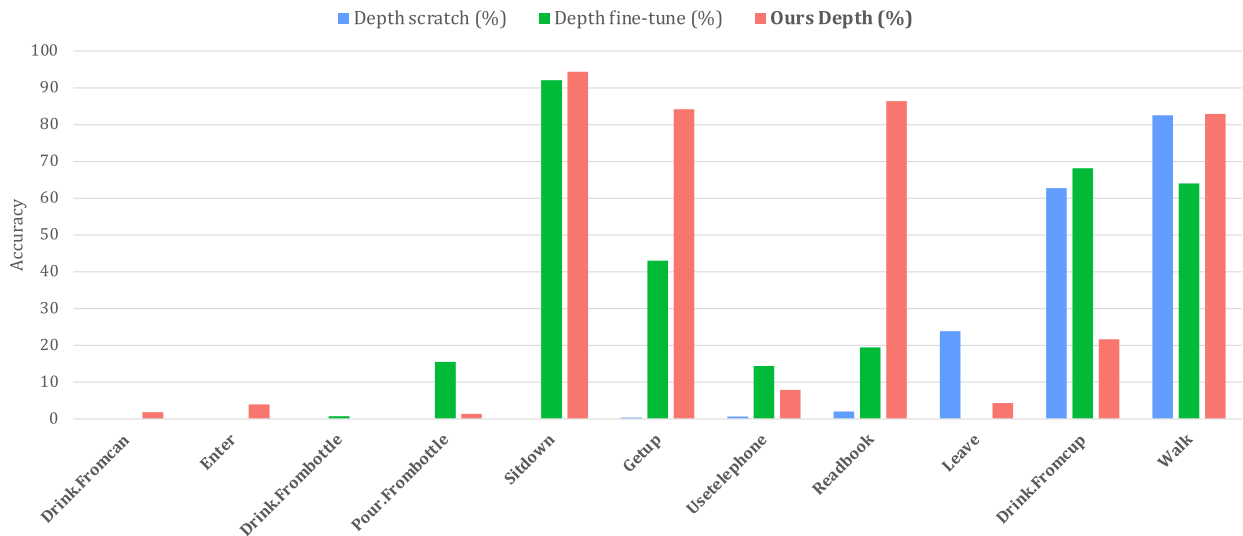


FIGURE 3. Performance comparison of I3D-based models on the Toyota Smarthome dataset (CV1 protocol). ‘Ours Depth’ refers to the depth-based unimodal inference of cross-modal mixing distillation. The 2 classes absent from the CV1 validation dataset (i.e., ‘Cutbread,’ ‘Eat.Attable’) and 6 classes with no performance change (i.e., ‘Usetablet,’ ‘Drink.Fromglass,’ ‘Eat.Snack,’ ‘Pour.Fromcan,’ ‘Takepills,’ and ‘Uuselaptop’) have been removed from the graph to enhance clarity of interpretation.

to conventional methodologies, our model achieves superior performance. Notably, in specific experiments, our model outperforms the conventional RGB fine-tune methods even when inferring with depth modality alone. In addition, during inference, it can work effectively with both unimodal and cross-modal inputs. The effectiveness of the proposed method is experimentally validated in comparison with naive and body pose-aware distillation methods.

A. COMPARISON OF DISTILLATION METHODS

Table 1 summarizes the results of the performance evaluation of the I3D-based models paired with various learning methodologies. All experiments in this table are conducted on the I3D model, and pre-trained weights obtained by performing supervised learning on Kinetics-400 are used in all cases except the depth scratch experiment. In Table 1, cross-modal mixing distillation refers to the learning of depth and RGB images as pairs, with only the depth images used for inference. These experiments are conducted using the Toyota Smarthome dataset [20], and the model performance is measured following the CS, CV1, and CV2 evaluation protocols. The various distillation methods are evaluated using depth scratch, RGB fine-tuning, and depth fine-tuning as performance metrics. In most cases, the performance is observed to either decline or exhibit marginal improvement. However, as is evident from the Table 1, the proposed cross-modal mixing distillation approach exhibits superior performance enhancement compared to other cases, achieving the best performance in all evaluation protocols. In particular, it induces a noticeable performance improvement in CV1, where smaller and more imbalanced training data are included, thereby verifying the generalizability of the model. Furthermore, our model outperforms the conventional RGB fine-tune by 4.94%. This suggests that the proposed method

effectively extracts features based on interactions between various modalities, thereby enabling more accurate learning.

B. COMPARISON WITH OTHER BASELINES

Further, additional experiments are performed using the Toyota Smarthome [20] and NTU RGB+D 120 datasets [10]. As indicated by Table 2, our proposed method is valid and effective across various models and datasets, demonstrating excellent performance improvements.

Although existing state-of-the-art (SOTA) methods achieve higher performance by utilizing the skeleton modality for training, directly comparing these methods with our proposed method is challenging. Nevertheless, our method demonstrates respectable performance using only the depth modality for inference, despite the relatively insufficient information it provides. This not only showcases the robustness and efficiency of our approach but also highlights its potential for practical applications where only limited data is available. By leveraging just the depth modality, our method still manages to perform competitively, underscoring its effectiveness even with relatively sparse information.

Notably, the NTU RGB+D 120 [10] x-set of TimeSformer corresponds to a more improved performance than the conventional RGB fine-tune by 7.72%, and the learning instability issue of depth scratch in the x-set is also resolved. These results demonstrate that the proposed methodology contributes significantly to performance improvement by complementing various modalities and enhancing learning stability.

C. ABLATION STUDY OF LOSS COMPONENTS

To validate our proposed loss functions, we conducted an ablation study on the Toyota Smarthome dataset [20] using the I3D model as the baseline. The results, presented

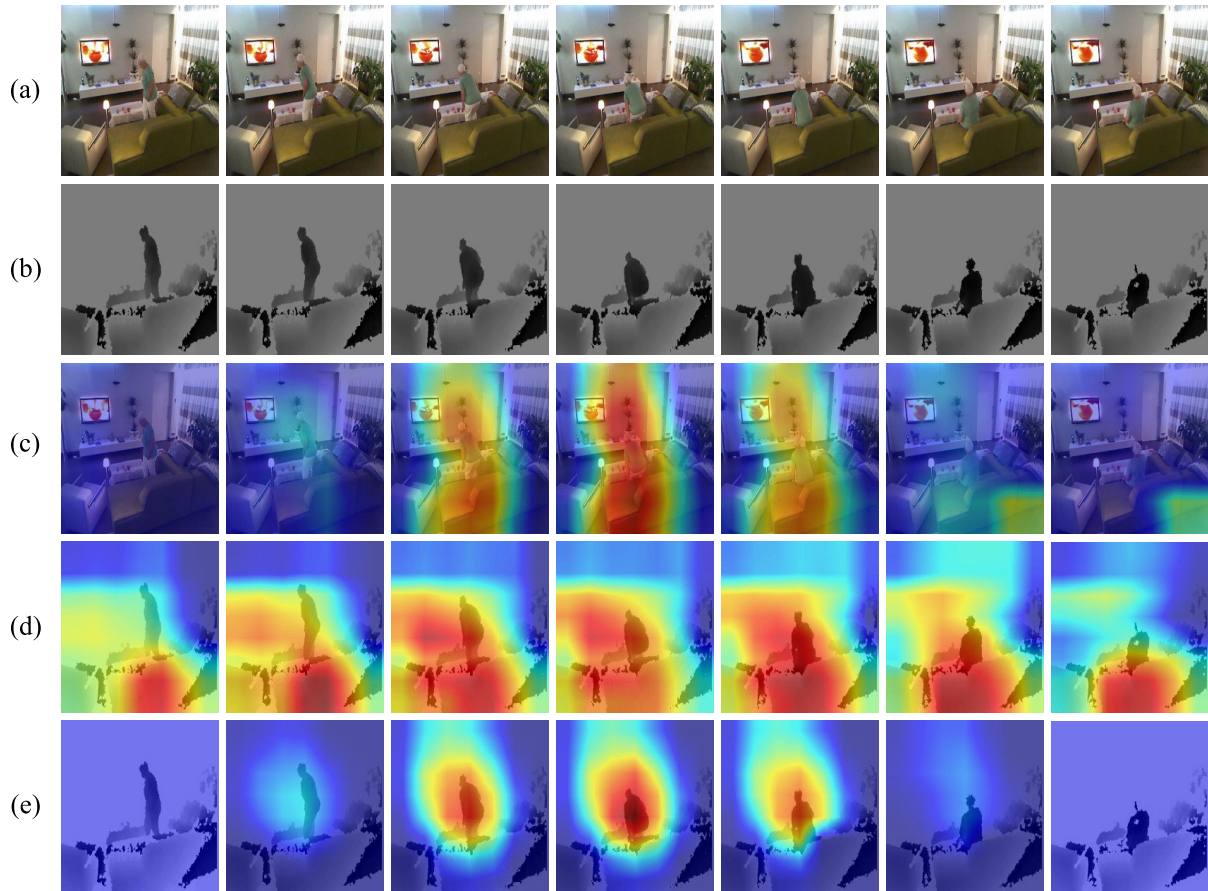


FIGURE 4. Visualization of the last feature map from the I3D backbone. From top to bottom: (a) RGB input, (b) Depth input, (c) RGB fine-tuned model, (d) Depth scratch model, (e) Depth-based unimodal inference of cross-modal mixing distillation model. In (d), the depth scratch model incorrectly predicts the given sample as ‘Walk,’ while in (e), the proposed cross-modal mixing distillation correctly predicts it as ‘SitDown.’

in Table 3, show the impact of each loss component on the model’s performance. The baseline “Naive distillation” method used none of the proposed loss components.

Our study demonstrates that the three loss components effectively improve model performance. The best results were achieved when all three components were used together, underscoring their contribution to the model’s performance enhancements. Notably, the substantial improvement in CV1 performance, which is designed to evaluate the generalization capability of the model, indicates that our proposed loss functions significantly enhance the model’s generalization performance.

D. PER-CLASS SUBGROUP ANALYSIS

Figure 3 graphically represents the performance of each class corresponding to the methods presented in Table 1. This is obtained using the Toyota Smarthome dataset [20] and the CV1 evaluation protocol. However, the two classes not included in the CV1 validation dataset (i.e., ‘Cutbread’ and ‘Eat.Attable’) and 6 classes with no performance change (i.e., ‘Usetablet’, ‘Drink.Fromglass’, ‘Eat.Snack’, ‘Pour.Fromcan’, ‘Takepills’, and ‘Uselaptop’) are excluded from the graph for clearer interpretation. Application of the proposed cross-modality mixing distillation improves the

performance of each class. This is particularly noticeable in the classes ‘Sitdown’ and ‘Getup,’ which are similar to each other, but correspond to completely different meanings depending on the sequence of frame playback. When depth fine-tune is used, the top three error rates for the ‘Getup’ class are 23.16% with respect to ‘Sitdown,’ 20.22% with respect to ‘Drink.Fromcup,’ and 8.82% with respect to ‘Walk.’ This demonstrates the difficulty in accurately analyzing and understanding these behaviors. However, when the proposed method is applied, the error rates are observed to decrease significantly. The error rate for ‘Sitdown’ is reduced to 4.78%, that for ‘Drink.Fromcup’ to 0.74%, and that for ‘Walk’ to 1.10%. Thus, the proposed method aids the accurate analysis and understanding of complex behavior patterns significantly, contributing to an overall improvement in class performance.

E. VISUALIZATION

Figure 4 depicts the visualization results of the final feature map of the I3D model backbone using Grad-CAM [46] on the Toyota Smarthome dataset [20]. In the figure, (a) RGB image, (b) depth image, (c) RGB fine-tune, (d) depth scratch, and (e) depth unimodal inference of cross-modal mixing distillation are depicted. While the baseline (I3D depth

scratch) predicts the given sample incorrectly as ‘Walk,’ the proposed cross-modal mixing distillation predicts it correctly as ‘Sitdown.’ RGB fine-tuning is observed to promote an overall focus on the object. However, Depth scratch promotes less focus on the object, and as a result, exhibits poorer characterization of objects and their behavior, resulting in poorer overall performance. However, the proposed method maintains a high level of attention to foreground objects and their behavior, even in low-information depth images. This indicates that the object-centric action recognition behavior of the RGB-based model has been successfully distilled into a depth-based model using the proposed distillation method.

VI. CONCLUSION

Recently, the importance of privacy protection in computer vision has been highlighted. In the past, solutions to this problem have primarily involved changing the sensor type used, such as depth sensors or infrared sensors. However, these exhibit poor performance owing to insufficient information. Alternatively, the use of multimodal data has also been considered as a solution, but this makes the learning and inference processes computationally expensive because of the need to process each modality independently. In addition, multimodal methods also suffer from high sensor costs because all sensors need to be available at all times. In this context, we propose a new method that uses various modalities during learning but operates conventionally with both unimodal and multimodal data during inference. It integrates the information obtained from each sensor efficiently, maintains the consistency of the patterns, and exchanges and complements information obtained from different sensors. It is experimentally confirmed to yield effective results while simultaneously achieving privacy protection, enabling flexible responses reducing computational complexity, and saving sensor costs. In turn, this improves the practicality and applicability of the method. On the Toyota Smarthome dataset [20] and the NTU RGB+D 120 dataset [10], this method is observed to exhibit superior performance compared to existing alternatives. In addition, it promotes greater focus on foreground objects. Our results demonstrate that the proposed method can be effectively applied for action recognition, and proves that effective action recognition is possible even when modalities with insufficient information are used, while simultaneously enabling privacy protection. However, the proposed model is still not perfect. In future research on this topic, models capable of responding to more diverse sensor environments and situations should be developed, and the effective implementation of these resources in the learning process of the model should be studied. We expect the research presented in this paper to serve as an important contribution to research in this field.

ACKNOWLEDGMENT

(Yuhyun Kim and Jinwook Jung are co-first authors.)

REFERENCES

- [1] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, “HMDB: A large video database for human motion recognition,” in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.
- [2] K. Soomro, A. Roshan Zamir, and M. Shah, “UCF101: A dataset of 101 human actions classes from videos in the wild,” 2012, *arXiv:1212.0402*.
- [3] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Ghorban, I. Laptev, R. Sukthankar, and M. Shah, “The THUMOS challenge on action recognition for videos ‘in the wild,’” *Comput. Vis. Image Understand.*, vol. 155, pp. 1–23, Feb. 2017.
- [4] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, “ActivityNet: A large-scale video benchmark for human activity understanding,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 961–970.
- [5] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, “Dense-captioning events in videos,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 706–715.
- [6] L. Zhou, Y. Kalantidis, X. Chen, J. J. Corso, and M. Rohrbach, “Grounded video description,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6571–6580.
- [7] *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*, Council Eur. Union, Off. J. Eur. Union, Apr. 2016. [Online]. Available: <https://eur-lex.europa.eu/eli/reg/2016/679/oj>
- [8] S. Yeung, A. Alahi, A. Haque, B. Peng, Z. Luo, A. Singh, T. Platchek, A. Milstein, and L. Fei-Fei, “Vision-based hand hygiene monitoring in hospitals,” in *Proc. AMIA*, 2016, pp. 1–11.
- [9] A. Haque, A. Milstein, and L. Fei-Fei, “Illuminating the dark spaces of healthcare with ambient intelligence,” *Nature*, vol. 585, no. 7824, pp. 193–202, Sep. 2020.
- [10] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, “NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.
- [11] F. Ilic, T. Pock, and R. P. Wildes, “Is appearance free action recognition possible?” in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 156–173.
- [12] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” in *Proc. NIPS*, 2015, pp. 1–9.
- [13] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le, “Self-training with noisy student improves ImageNet classification,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10684–10695.
- [14] F. M. Hafner, A. Bhuyian, J. F. P. Kooij, and E. Granger, “Cross-modal distillation for RGB-depth person re-identification,” *Comput. Vis. Image Understand.*, vol. 216, Feb. 2022, Art. no. 103352.
- [15] J. Hoffman, S. Gupta, and T. Darrell, “Learning with side information through modality hallucination,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 826–834.
- [16] G. Radevski, D. Grujicic, M. Blaschko, M.-F. Moens, and T. Tuytelaars, “Multimodal distillation for egocentric action recognition,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 5213–5224.
- [17] D. Reilly, A. Chadha, and S. Das, “Seeing the pose in the pixels: Learning pose-aware representations in vision transformers,” 2023, *arXiv:2306.09331*.
- [18] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, “MARS: Motion-augmented RGB stream for action recognition,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7874–7883.
- [19] G. Kurillo, E. Hemingway, M.-L. Cheng, and L. Cheng, “Evaluating the accuracy of the Azure Kinect and Kinect v2,” *Sensors*, vol. 22, no. 7, p. 2469, Mar. 2022.
- [20] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, “Toyota smarthome: Real-world activities of daily living,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 833–842.
- [21] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, “The kinetics human action video dataset,” 2017, *arXiv:1705.06950*.

- [22] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6047–6056.
- [23] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.
- [24] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.
- [25] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, and B. Gong, "MoViNets: Mobile video networks for efficient video recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16015–16025.
- [26] J. Wang, Z. Sun, Y. Qian, D. Gong, X. Sun, M. Lin, M. Pagnucco, and Y. Song, "Maximizing spatio-temporal entropy of deep 3D CNNs for efficient video recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [27] Z. Lu, "A theory of multimodal learning," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2024.
- [28] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, "KAIST multi-spectral day/night data set for autonomous and assisted driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 934–948, Mar. 2018.
- [29] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal learning with missing modality via shared-specific feature modelling," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 15878–15887.
- [30] F. M. Thoker and J. Gall, "Cross-modal knowledge distillation for action recognition," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 6–10.
- [31] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, vol. 3, no. 2, p. 896.
- [32] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," 2018, *arXiv:1808.01340*.
- [33] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman, "A short note on the Kinetics-700–2020 human action dataset," 2020, *arXiv:2010.10864*.
- [34] S. F. Bhat, R. Birkel, D. Wofk, P. Wonka, and M. Müller, "ZoeDepth: Zero-shot transfer by combining relative and metric depth," 2023, *arXiv:2302.12288*.
- [35] P. Lee, T. Kim, M. Shim, D. Wee, and H. Byun, "Decomposed cross-modal distillation for RGB-based temporal action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, p. 2373.
- [36] N. C. Garcia and P. M. V. Murino, "Modality distillation with multiple stream networks for action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 103–118.
- [37] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2019, pp. 2712–2721.
- [38] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," 2020, *arXiv:2004.10964*.
- [39] T. Kim, J. Oh, N. Y. Kim, S. Cho, and S.-Y. Yun, "Comparing Kullback–Leibler divergence and mean squared error loss in knowledge distillation," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 2628–2635.
- [40] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proc. Int. Conf. Mach. Learn.*, 2021, p. 4.
- [41] W. Choi, J. Chen, and J. Yoon, "PWS-DVC: Enhancing weakly supervised dense video captioning with pretraining approach," *IEEE Access*, vol. 11, pp. 128162–128174, 2023.
- [42] R. S. Bhooshan, "A multimodal framework for video caption generation," *IEEE Access*, vol. 10, pp. 92166–92176, 2022.
- [43] H. Chen, Y. Jiang, and H. Ko, "Pose-guided graph convolutional networks for skeleton-based action recognition," *IEEE Access*, vol. 10, pp. 111725–111731, 2022.
- [44] X. Weiya, W. Muqing, Z. Min, L. Yifeng, L. Bo, and X. Ting, "Human action recognition using multilevel depth motion maps," *IEEE Access*, vol. 7, pp. 41811–41822, 2019.
- [45] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video Swin transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 3192–3201.
- [46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-Cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.
- [47] A. Piergiovanni, W. Kuo, and A. Angelova, "Rethinking video ViTs: Sparse video tubes for joint image and video learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2214–2224.
- [48] Y. Wang, K. Li, Y. Li, Y. He, B. Huang, Z. Zhao, H. Zhang, J. Xu, Y. Liu, Z. Wang, S. Xing, G. Chen, J. Pan, J. Yu, Y. Wang, L. Wang, and Y. Qiao, "InternVideo: General video foundation models via generative and discriminative learning," 2022, *arXiv:2212.03191*.
- [49] S. Srivastava and G. Sharma, "OmniVec: Learning robust representations with cross modal sharing," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 1236–1248.
- [50] K. Li, Y. Wang, Y. Li, Y. Wang, Y. He, L. Wang, and Y. Qiao, "Unmasked teacher: Towards training-efficient video foundation models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19948–19960.
- [51] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao, "UniformerV2: Spatiotemporal learning by arming image ViTs with video uniformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1632–1643.
- [52] M. Seo, D. Cho, S. Lee, J. Park, D. Kim, J. Lee, J. Ju, H. Noh, and D.-G. Choi, "A self-supervised sampler for efficient action recognition: Real-world applications in surveillance systems," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 1752–1759, Apr. 2022.
- [53] K. Li, Y. Wang, P. Gao, G. Song, Y. Liu, H. Li, and Y. Qiao, "Uniformer: Unified transformer for efficient spatiotemporal representation learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 14733–14743.
- [54] S. Nag, X. Zhu, Y. Song, and T. Xiang, "Proposal-free temporal action detection via global segmentation mask learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 645–662.
- [55] Z. Khan, C. V. Jawahar, and M. Tapaswi, "Grounded video situation recognition," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2022, pp. 8199–8210.
- [56] Y. Zhong, W. Ji, J. Xiao, Y. Li, W. Deng, and T.-S. Chua, "Video question answering: Datasets, algorithms and challenges," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2022, pp. 6439–6455.
- [57] M. Maaz, H. Rasheed, S. Khan, and F. Shahbaz Khan, "Video-ChatGPT: Towards detailed video understanding via large vision and language models," 2023, *arXiv:2306.05424*.
- [58] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-LLaVA: Learning united visual representation by alignment before projection," 2023, *arXiv:2311.10122*.
- [59] D. Ahn, S. Kim, H. Hong, and B. Chul Ko, "STAR-transformer: A spatio-temporal cross attention transformer for human action recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3319–3328.
- [60] L. Wang and P. Koniusz, "3Mformer: Multi-order multi-mode transformer for skeletal action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5620–5631.



YUHYUN KIM (Member, IEEE) received the B.S. degree in information and communication engineering from Hanbat National University, South Korea, in 2023. He is currently pursuing the M.S. degree in information and communication engineering. His research interests include deep learning, computer vision, multimodal recognition, long tailed learning, and active learning.



JINWOOK JUNG (Member, IEEE) received the B.S. degree in mechanical engineering from Hanbat National University, South Korea, in 2024. He is currently pursuing the M.S. degree in information and communication engineering. His research interests include deep learning, multi-modal recognition, and self-supervision.



HYEONCHEOL NOH (Student Member, IEEE) received the B.S. degree in information and communication engineering from Hanbat National University, South Korea, in 2022. He is currently pursuing the M.S. degree in information and communication engineering. His research interests include deep learning, self-supervised learning, and change detection.



BYUNGTAE AHN received the B.S. degree in electronic engineering from Kumoh National Institute of Technology, South Korea, in 2007, and the M.S. degree in biomechanics from Sungkyunkwan University, South Korea, in 2011, and the Ph.D. degree in robotics program from Korea Advanced Institute of Science and Technology (KAIST), in 2018. He is currently a Senior Researcher with the Department of AI Machinery, Korea Institute of Machinery and Materials. His research interests include computer vision, robot vision, time series analysis, and artificial intelligence. He was awarded the Qualcomm Innovation Award, in 2013 and has been listed in Marquis Who's Who in the World, since 2016.



JUNGHYE KWON is currently a Medical Oncologist and a Professor with the College of Medicine, Chungnam National University who began her career, in 2004. She upholds the belief that cancer treatment requires equal emphasis on curing the disease and providing care for patients. Her foray into palliative care began during her visit to the MD Anderson Cancer Center, in 2011, marking a significant pivot in her professional journey towards specialization in this field. She has been prolific in her research efforts, contributing numerous articles to palliative care literature. Her research interests include centered on symptom management, behaviors of cancer patients, and end-of-life care, with a burgeoning interest in the integration of digital health technologies in these areas. She is leading the Division of Hemato-Oncology, Department of Internal Medicine, Chungnam National University Sejong Hospital, where she continues to make significant contributions to the field.



DONG-GEOL CHOI (Member, IEEE) received the B.S. and M.S. degrees in electric engineering and computer science from Hanyang University, in 2005 and 2007, respectively, and the Ph.D. degree in robotics program from Korea Advanced Institute of Science and Technology (KAIST), in 2016. After working as a Postdoctoral Researcher and a Research Assistant Professor with KAIST, until August 2018. He is currently an Associate Professor with the Department of Information and Communication Engineering, Hanbat National University. His research interests include sensor fusion, autonomous robotics, and artificial intelligence. He was a member of 'Team KAIST,' which won the first place in the DARPA Robotics Challenge Finals, in 2015. He received a fellowship from the Qualcomm Korea Research and Development Center, in 2013.

...