Distill, Suppress, and Fuse: Cross-Modal Knowledge Integration for Optical Flow-Free Temporal Action Segmentation

Seungjin Han [*1] **Gyeong-hyeon Kim** [*2] **Eunwoo Kim** [1 2]

Abstract

Cross-modal knowledge distillation (CMKD) enables efficient inference by transferring knowledge from a teacher model trained on a computationally heavy modality (i.e., optical flow) to a student model operating on a lightweight modality (i.e., RGB). However, we find that most current CMKD methods are hindered by a key limitation when applied in temporal action segmentation: motion cues transferred from optical flow often lead the student to produce frame representations that are misaligned with the underlying action structure. To address this, we propose RELATE, an optical flow-free framework that selectively integrates transferred cues while suppressing misaligned cues. We further introduce a prediction refinement strategy to resolve ambiguous segments using multiple predictions. Experiments on three benchmarks with multiple segmenters show that RELATE consistently outperforms RGB-only baselines, approaches two-stream performance, and achieves up to 175× faster inference.

1. Introduction

Temporal action segmentation (TAS) aims to assign action labels to every frame in long procedural videos, enabling applications such as AR/VR and human-robot interaction (Ding et al., 2023). In recent years, several studies have shown that two-stream methods that use both RGB and optical flow achieve strong performance (Gan et al., 2024; Lu & Elhamifar, 2024), as they provide complementary appearance and motion cues (Carreira & Zisserman, 2017). However, they suffer from substantial latency due to the high computational cost of optical flow extraction. For example, processing optical flow for a 37-second video requires 16.6 minutes (Sec. 4.4.2), limiting practical deployment.

Cross-modal knowledge distillation (CMKD) (Gupta et al., 2016) enables efficient inference without optical flow by transferring knowledge from a computationally heavy modality (i.e., optical flow) to a student operating on a lightweight modality (i.e., RGB). Leveraging this capability, CMKD has been widely adopted in short video tasks, where models operate on video clips around 10 seconds (Crasto et al., 2019) or focus on aggregating local temporal cues within such short temporal windows (Lee et al., 2023). In particular, prior works in action recognition and localization have shown that transferring motion knowledge leads to strong efficiency and performance gains.

However, most CMKD methods are directly adapted from image classification (Gupta et al., 2016), often overlooking the requirements of long video tasks with sequences exceeding 10 minutes (Stein & McKenna, 2013). Notably, existing approaches distill individual feature maps or logits (Huo et al., 2024; Lee et al., 2023), but do not explicitly distill which frames the teacher model focuses on, nor how frame features align with action semantics. Fig. 1 shows that this leads to an imbalance between classification and segmentation abilities. Specifically, segmentation performance should improve when more frames are correctly recognized via CMKD, provided they offer appropriate temporal context. However, prior CMKD methods that indiscriminately leverage transferred cues (e.g., MARS (Crasto et al., 2019)) degrade segmentation performance despite improving frame-wise accuracy, suggesting that the learned representations are often misaligned with the underlying action structure and less informative for temporal reasoning.

These observations inspire us to minimize the effect of transferred representations that are less informative for temporal reasoning. We propose **R**GB-based action s**E**gmentation with a**L**ignment g**ATE**d fusion (**RELATE**), a framework that selectively integrates transferred motion cues with RGB cues by suppressing the uninformative ones. As illustrated in Fig. 2 (Left), we begin by learning modality-specific features from RGB inputs alone, using two independent segmenters for RGB and optical flow representations. We then

[*]Equal contribution [1]Department of Artificial Intelligence, Chung-Ang University, Seoul, Republic of Korea [2]School of Computer Science and Engineering, Chung-Ang University, Seoul, Republic of Korea. Correspondence to: Eunwoo Kim <eunwoo@cau.ac.kr>.

ICML 2026 Workshop on Resource-Adaptive Foundation Model Inference (AdaptFM), Seoul, South Korea.

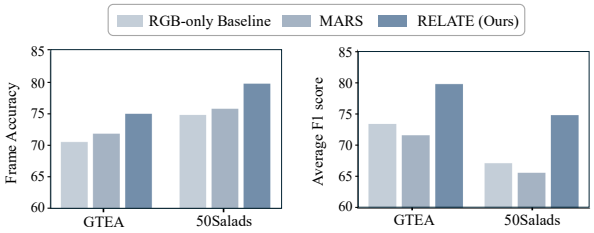

Figure 1. Comparison of three methods that use only RGB inputs at inference, on the GTEA and 50Salads datasets. The left plot shows frame accuracy, while the right plot reports the average segmental F1 score at IoU thresholds of 10%, 25%, and 50%. Relative to the RGB-only Baseline without CMKD, both MARS and RELATE (Ours) distill knowledge from an optical flow teacher to an RGB student, yet exhibit different behaviors. In all datasets, MARS improves accuracy but degrades the F1 score. RELATE mitigates this imbalance and improves both metrics by selectively integrating transferred motion cues with RGB cues. All comparisons use MS-TCN (Farha & Gall, 2019) as the backbone model.

introduce an alignment fusion transformer that incorporates these modality-specific features by identifying the modality that is more aligned with the temporal context (i.e., action anchors). In addition, we introduce a training-free refinement step that mitigates ambiguous outputs by selecting the most reliable prediction among modalities.

Extensive experiments on three benchmark datasets (Fathi et al., 2011; Stein & McKenna, 2013; Kuehne et al., 2014) with three action segmentation models (Farha & Gall, 2019; Yi et al., 2021; Lu & Elhamifar, 2024) demonstrate that our method surpasses RGB-only models and achieves performance comparable to the two-stream models. RELATE achieves these improvements while being significantly more efficient, with approximately 175× faster inference.

2. Related work

2.1. Temporal Action Segmentation

In TAS, existing works can be broadly categorized into two directions: temporal modeling and temporal matching. The former involves modeling temporal dependencies across frame sequences to aggregate contextual information over time, typically through convolutional (Farha & Gall, 2019; Zhong et al., 2024) or attention-based architectures (Yi et al., 2021; Bahrami et al., 2023). The latter explicitly matches frames or temporal segments with action queries, focusing on learning frame-to-action associations while preserving temporal structure (Gan et al., 2024; Lu & Elhamifar, 2024). Notably, most existing approaches are based on two-stream inputs. Among them, optical flow extraction incurs high latency that limits their applicability in real-world setups. In this paper, we propose a novel RGB-based action segmentation framework that eliminates the need for optical flow during inference by effectively leveraging transferred cues.

2.2. Cross-modal Knowledge Distillation

Cross-modal knowledge distillation (CMKD) aims to transfer knowledge from one modality to another, which enables a student model to mimic the teacher model during inference (Gupta et al., 2016; Crasto et al., 2019; Huo et al., 2024). Several works have adopted CMKD in tasks such as action recognition (Crasto et al., 2019) and localization (Lee et al., 2023). These approaches employ a teacher model trained on optical flow to transfer motion cues to an RGB student model, focusing on task-specific objectives. For instance, they distill action features (Crasto et al., 2019; Ni et al., 2022), or boundary regression and completeness cues (Dai et al., 2021; Lee et al., 2023). In addition, some works distill temporal relations encoded in the channel covariance (Dai et al., 2021) for action detection, as it can transfer the temporal context. However, not all transferred information leads to improved segmentation-level performance due to the misalignment between actions and frames. In this paper, we propose a CMKD framework that selectively incorporates informative distilled cues with the RGB modality during multi-modal fusion.

3. The Proposed Method

3.1. Dual-Branch Pipeline

To extract both modality-specific representations using only RGB frames, the proposed method adopts a dual-branch design that independently learns each modality. As illustrated in Fig. 2 (Left), each branch learns the modality-specific features of optical flow and RGB, using two action segmenters that predict frame-wise actions.

Specifically, the RGB branch directly encodes appearance features $\mathbf{Z}_{\text{RGB}}$, while the flow branch learns motion-aware features $\mathbf{Z}_{\text{OF}}$ via CMKD. For knowledge distillation, the teacher segmenter trained on optical flow modality supervises the flow segmenter to mimic its logits, enabling it to learn motion-aware features of optical flow. We note that $\mathbf{Z}_{\text{RGB}}$ and $\mathbf{Z}_{\text{OF}}$ are obtained from the last layer of each segmenter. To facilitate CMKD, we additionally adopt two fully connected (FC) layers that bridge the discrepancy between modalities (Huo et al., 2024) (i.e., a distributional gap between I3D features (Carreira & Zisserman, 2017)).

3.2. Alignment Fusion Transformer

To first determine which modality positively contributes to segmentation, we focus on the temporal modeling mechanism in multi-modal scenarios. Particularly, most two-stream methods adopt early fusion (i.e., concatenation) (Gan et al., 2024), allowing frame features predictive of surrounding actions to be abstracted across modalities. To capture this trait, we introduce the alignment fusion transformer that employs an action-gated attention to estimate the action-

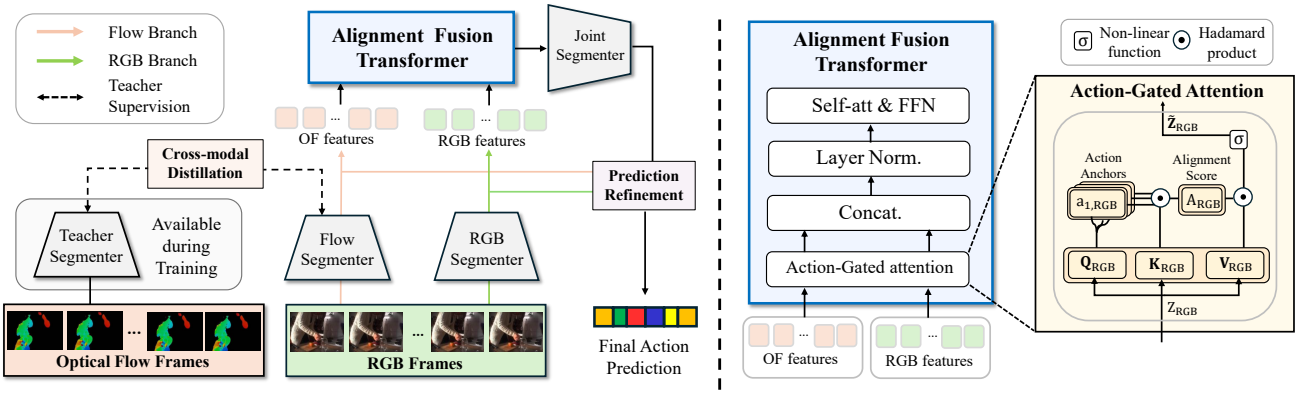

Figure 2. (Left) The proposed framework consists of RGB and flow branches to learn modalities, a teacher segmenter that distills optical flow features to the flow branch, an alignment fusion transformer to incorporate modality features, and a joint segmenter producing the final prediction after prediction refinement. Note that the teacher segmenter as well as optical flow are only available during training. (Right) The alignment fusion transformer consists of an action-gated attention and subsequent layers. In the action-gated attention module, an alignment score is computed for each modality by measuring the relevance between frame features and action anchors. We then modulate modality features based on this alignment score, after which subsequent layers learn cross-modal importance for fusion.

frame alignment and adaptively modulate the modalities. We then learn cross-modal importance based on this modulation and fuse the modalities accordingly.

To estimate the action-frame alignment, a straightforward approach would be to apply self-attention (Vaswani et al., 2017). However, we observe two drawbacks: (i) it often fails to capture action-level semantics because it relies on frame-wise interactions (Ahn & Lee, 2021), and (ii) it struggles to model fine-grained alignment where detailed frame features correspond to actions to varying extents. Together, these limitations hinder fine-grained action-frame alignment. To address the first issue, we define an action anchor to estimate the semantic alignment between frames and actions. For the second issue, we design action-gated attention using the Hadamard product to capture fine-grained channel-wise alignment.

As illustrated in Fig. 2 (Right), we first generate query, key, and value embeddings $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ by applying linear projections to the modality-specific features $\mathbf{Z} \in \mathbb{R}^{T\times d}$. Here, T denotes the number of frames, and d denotes the channel dimension. Note that we apply action-gated attention to both modalities to avoid biased learning toward a specific modality, but omit modality-specific notation for clarity. To obtain the action anchors that represent the action semantics, $\mathbf{a}_i \in \mathbb{R}^{1\times d}$, we average the query embeddings over the temporal range of the i^{th} predicted segment where $i \in \{1,\dots,s\}$. Thereby, we have $\mathbf{a}_i = \frac{1}{|e_i|}\sum_{t=l_i}^{l_i+e_i} \mathbf{q}_t$, where $\mathbf{q}_t \in \mathbb{R}^{1\times d}$ is the query embedding at time $t \in \{1,\dots,T\}$, l_i and e_i denote the start time and duration of the corresponding segment, respectively. Note that we derive anchors from predicted segments rather than ground truth to align training with inference, where early predictions provide diverse alignment signals that facilitate suppressing misaligned cues.

Then, we aggregate the channel-wise similarities between $\mathbf{K}$ and all action anchors to obtain the alignment score $\mathbf{A} \in \mathbb{R}^{T\times d}$. Accordingly, $\mathbf{A}$ measures the relevance of frame representations to the action semantics at each channel and time step, assigning lower values to irrelevant regions and higher values to relevant ones. We gate the value embeddings $\mathbf{V}$ with $\mathbf{A}$ via the Hadamard product, yielding the modulated feature $\tilde{\mathbf{Z}}$ as

$$\mathbf{A} = \sigma\left(\sum_{i=1}^{s} \frac{\mathbf{K}\odot \mathbf{a}_i}{\sqrt{d}}\right), \quad \tilde{\mathbf{Z}} = \mathbf{A}\odot\mathbf{V} \tag{1}$$

where $\odot$ denotes the Hadamard product, and $\sigma(\cdot)$ represents the sigmoid function. The normalization factor $\sqrt{d}$ is introduced to stabilize training and prevent vanishing gradients.

To regulate the fusion process based on modality informativeness, we propagate the modulated feature derived from action-gated attention in the subsequent layers. Specifically, we concatenate the modulated features along the channel dimension as $\tilde{\mathbf{Z}}_{\text{J}} = \text{concat}(\tilde{\mathbf{Z}}_{\text{RGB}}, \tilde{\mathbf{Z}}_{\text{OF}}) \in \mathbb{R}^{T\times 2d}$, and apply layer normalization to balance the feature statistics across modalities. We then apply self-attention to $\tilde{\mathbf{Z}}_{\text{J}}$, since the model can learn cross-modal importance based on the modulated cues (Yi et al., 2021). Finally, the resulting $\hat{\mathbf{Z}}_{\text{J}} = \text{self-attention}(\tilde{\mathbf{Q}}_{\text{J}}, \tilde{\mathbf{K}}_{\text{J}}, \tilde{\mathbf{V}}_{\text{J}})$ is then passed through a feed-forward network (FFN) and fed into the joint segmenter to refine features based on multi-modal context, producing a joint prediction $\mathbf{p}_{\text{J}} \in \mathbb{R}^{T\times n}$, where n denotes the number of action classes.

3.3. Training Objective

To distill the knowledge of optical flow into a flow segmenter, we adopt logit-level distillation (Gupta et al., 2016),

which minimizes the discrepancy between probability distributions. While relational-level (Dai et al., 2021; Ni et al., 2022) distillation that transfers temporal relations (e.g., pairwise distance of frame features) or feature-level distillation that matches intermediate representations (Crasto et al., 2019) are also beneficial, logit-level distillation yields better empirical performance and thus becomes the main training objective. The cross-modal knowledge distillation loss is denoted as $\mathcal{L}_{CMKD}$. We also adopt two fully connected (FC) layers to bridge the input-domain discrepancy, using a mean squared error loss between RGB I3D features and optical flow I3D features. In addition, we adopt the training objectives from prior works (Yi et al., 2021; Lu & Elhamifar, 2024) for the segmenters. For clarity, we denote the losses for the RGB, optical flow, and joint segmenter as $\mathcal{L}_{RGB}$, $\mathcal{L}_{OF}$, and $\mathcal{L}_{J}$, respectively. Finally, the entire framework is trained using a weighted combination of these losses, $\mathcal{L} = \mathcal{L}_{RGB} + \mathcal{L}_{OF} + \mathcal{L}_{J} + \gamma \cdot \mathcal{L}_{CMKD}$, where γ controls the contribution of knowledge distillation.

3.4. Prediction Refinement

While the alignment fusion transformer effectively incorporates modalities based on action-frame alignment, prediction uncertainty can still arise in regions where the alignment remains weak. In such cases, a unimodal prediction or the previous prediction can be more reliable (Zhong et al., 2024; Zhang et al., 2024) than the current joint prediction. To mitigate prediction uncertainty, we introduce a training-free prediction refinement step that leverages two unimodal predictions to improve the reliability of the joint prediction.

Following (Zhong et al., 2024), we define the confidence score for modality $m \in \{\text{RGB}, \text{OF}, \text{J}\}$ at time t as $o_{m,t} = \max_c p_{m,t,c}$, where $p_{m,t,c} \in [0, 1]$ denotes the predicted probability of class c. If the joint prediction $\mathbf{p}_{\text{J},t}$ has the lowest confidence and falls below the threshold $\delta \in [0, 1]$, we replace it with the unimodal prediction that had the highest confidence at the previous time, *i.e.,* $\mathbf{p}_{\text{J},t} = \mathbf{p}_{m^*,t-1}$ where $m^* = \arg\max_m(o_{m,t-1})$. Otherwise, if the confidence score of the joint prediction $\mathbf{p}_{\text{J},t}$ is below the threshold but is not the lowest among the predictions, we retain the previous joint prediction $\mathbf{p}_{\text{J},t} = \mathbf{p}_{\text{J},t-1}$, since no unimodal alternative is clearly superior.

4. Experiments

4.1. Experimental Settings

4.1.1. Datasets and Evaluation Metrics

Following prior works (Lu & Elhamifar, 2024), we evaluate our method on three benchmark datasets. GTEA (Fathi et al., 2011) has 28 videos and 11 action classes, with an average duration of 37 seconds. 50Salads (Stein & McKenna, 2013) has 50 videos and 17 action classes, and is the longest benchmark with an average duration of 6.4 minutes. Breakfast (Kuehne et al., 2014) is the largest benchmark, containing 1,712 videos and 48 action classes, with video lengths ranging from about 30 seconds to 5 minutes. We report segmental F1 scores at IoU thresholds $\{10, 25, 50\}$, along with Edit and frame-wise accuracy (Acc).

4.1.2. Implementation Details

For the modality inputs, we used I3D features (Carreira & Zisserman, 2017) with 2,048 channel dimensions. The first 1,024 dimensions were used for the RGB inputs and the others for the optical flow inputs. We employed three action segmenters in our experiments: MS-TCN (Farha & Gall, 2019) and ASFormer (Yi et al., 2021) for temporal modeling methods, and FACT (Lu & Elhamifar, 2024) for the matching method. For fair capacity comparison, we matched the parameter budget by reducing the layers in our segmenters, while keeping all baselines unchanged. The hyperparameters γ and δ were set to 0.5 and 0.7, respectively. All experiments were conducted on RTX 2080Ti GPUs.

4.2. Performance Comparison

To evaluate the effectiveness of our method, we compare it against RGB-only and two-stream (RGB+OF) methods, as shown in Tab. 1. First, across all three datasets and three segmenters, the RGB-only models underperform their two-stream counterparts due to the absence of optical flow. Compared with the RGB-only baselines, our method consistently improves performance across all datasets and segmenters, highlighting its effectiveness and generalizability.

Among the three datasets, the largest gains are observed on 50Salads, which has the longest average video duration. The proposed method improves F1@50 by 11.3, 7.2, and 4.4 over the RGB-only baselines of MS-TCN, ASFormer, and FACT, respectively. Notably, on this dataset, our method achieves performance comparable to the two-stream model across all segmenters and even surpasses it in some cases.

For the Breakfast and GTEA datasets, the proposed method consistently improves F1@50 over RGB-only baselines. Specifically, it achieves gains of 4.2, 3.6, and 4.2 on Breakfast, and 7.6, 6.4, and 5.5 on GTEA for MS-TCN, ASFormer, and FACT, respectively. Notably, the proposed method with FACT achieves comparable gains to the other TAS methods despite frame-action matching, indicating that our method complements the temporal matching approach by gating uninformative cues that are misaligned with action anchors.

While the proposed method achieves strong performance without prediction refinement (PR), incorporating PR provides additional improvements. For instance, it enabled our method with ASFormer to match the F1 scores of the two-stream model on 50Salads. However, its effect on FACT is

Table 1. Performance comparison on three datasets using three segmenters. All methods use only RGB inputs at inference, except two-stream methods that require both RGB and optical flow. We bold the state-of-the-art RGB method and italicize reproduced results.

Methods	Modality	GTEA					50Salads					Breakfast				
		F1@{10,25,50}			Edit	Acc	F1@{10,25,50}			Edit	Acc	F1@{10,25,50}			Edit	Acc
MS-TCN (Farha & Gall, 2019)	RGB+OF	85.8	83.4	69.8	79.0	76.3	76.3	74.0	64.5	67.9	80.7	52.6	48.1	37.9	61.7	66.3
MS-TCN	RGB	80.0	76.5	63.4	75.3	70.6	72.2	69.8	59.5	65.3	74.9	56.8	50.1	36.1	58.9	45.9
RELATE (w/o PR)	RGB	**85.8**	82.3	70.1	**84.1**	**74.1**	78.9	76.7	68.4	71.6	79.8	60.2	52.9	39.0	61.6	52.9
RELATE	RGB	85.6	**82.4**	**71.0**	83.3	73.8	**81.0**	**78.0**	**70.8**	**74.2**	**80.9**	**61.7**	**54.2**	**40.3**	**61.8**	**53.5**
ASFormer (Yi et al., 2021)	RGB+OF	90.1	88.8	79.2	84.6	79.7	85.1	83.4	76.0	79.6	85.6	76.0	70.6	57.4	75.0	73.5
ASFormer	RGB	84.5	82.5	69.7	82.3	73.4	80.7	78.0	69.7	74.1	81.5	67.9	61.1	46.9	68.0	62.4
RELATE (w/o PR)	RGB	86.3	85.5	74.5	83.9	75.9	84.8	83.5	75.1	77.7	84.0	69.5	63.2	49.4	68.7	**65.3**
RELATE	RGB	**87.4**	**85.6**	**76.1**	**84.4**	**76.6**	**85.6**	**83.7**	**76.9**	**79.1**	**84.2**	**70.8**	**64.7**	**50.5**	**69.6**	**65.3**
FACT (Lu & Elhamifar, 2024)	RGB+OF	*91.4*	*88.7*	*79.4*	*89.4*	*83.9*	*84.7*	*83.6*	*76.9*	*80.4*	*85.8*	81.4	76.5	66.2	79.7	76.2
FACT	RGB	86.4	84.6	68.0	85.8	76.4	79.9	78.1	70.4	74.1	79.1	69.0	63.1	51.2	68.9	62.4
RELATE (w/o PR)	RGB	**89.3**	**87.0**	**73.5**	**87.7**	79.6	**83.1**	81.0	74.8	**76.5**	81.8	**73.0**	**67.6**	**55.4**	**72.2**	**66.2**
RELATE	RGB	89.2	86.9	**73.5**	87.6	**79.7**	83.0	**81.2**	**76.0**	76.2	**82.0**	72.8	67.3	**55.4**	72.1	**66.2**

Table 2. An ablation study of distillation, dual-branch design and action-gated attention.

CMKD	DB	AGA		F1@{10,25,50}	Edit	Acc	Avg.
		RGB	OF				
✗	✗	✗	✗	72.2 / 69.8 / 59.5	65.3	74.9	68.3
✓	✗	✗	✗	67.4 / 64.9 / 57.1	62.7	75.9	65.6
✓	✓	✗	✗	69.1 / 68.8 / 58.9	62.3	76.6	67.1
✓	✓	✓	✗	69.9 / 65.6 / 57.1	61.2	76.3	66.0
✓	✓	✗	✓	71.7 / 67.4 / 59.3	63.2	77.4	67.8
✓	✓	✓	✓	**78.9 / 76.7 / 68.4**	**71.6**	**79.8**	**75.1**

mixed, as FACT already reduces prediction ambiguity via unique token-to-segment matching (Lu & Elhamifar, 2024), leaving limited room for further improvement.

4.3. Ablation Study

In Tab. 2, we provide an ablation study of each component in RELATE on 50Salads, using MS-TCN. When action-gated attention (AGA) is not used, the features are passed directly to the concatenation operation. Introducing CMKD without dual-branch (DB) in the second row, or employing CMKD with DB in the third row, fails to improve the RGB-only baseline in the first row. Instead, these variants improve frame-wise accuracy but degrade segment-level metrics, consistent with our preliminary study in Fig. 1. These results indicate that simply increasing architectural complexity (e.g., dual-branch design) or applying CMKD is insufficient to mitigate the misaligned representations.

Applying AGA to only a single modality yields marginal gains, as shown in the fourth and fifth rows. This is because single-modality gating allows only one branch to be learnable for AGA, which may lead to a biased reliance on the learnable branch while neglecting the other branch. Meanwhile, adopting AGA on the flow branch rather than the RGB branch yields clearer improvements, as it can benefit more from AGA to address the action-frame misalignment. When all components are enabled, the proposed method outperforms the RGB-only baseline by an average of 6.8, demonstrating that our method effectively leverages transferred modality cues for improved temporal reasoning.

Table 3. Comparison of different fusion mechanisms. The methods are evaluated on GTEA using ASFormer as the segmenters.

Fusion	F1@{10,25,50}	Edit	Acc	Average
Naïve	83.8 / 82.1 / 67.4	79.5	75.3	77.6 (±0.0)
Self-attention	83.8 / 82.4 / 72.3	81.1	75.1	78.9 (+1.3)
Cross-attention	84.7 / 82.5 / 71.9	81.3	74.8	79.0 (+1.4)
Local-attentive	85.3 / 83.8 / 71.6	82.1	74.4	79.4 (+1.8)
Ours	**86.3 / 85.5 / 74.5**	**83.9**	**75.9**	**81.2** (+3.6)

4.4. Analysis

4.4.1. Effectiveness of the Alignment Fusion Transformer

To validate the effectiveness of our alignment fusion transformer (AFT), we compared it with various fusion methods under the CMKD setting in Tab. 3. In particular, we evaluate a naïve fusion method (*i.e.,* concatenation) and three attentive fusion methods; *i.e.,* self-attention, cross-attention (Vaswani et al., 2017), and local-attentive fusion (Lee et al., 2023). Compared with attentive fusion methods, the naïve fusion scheme achieves reasonable accuracy owing to CMKD, but performs poorly on segment-level metrics, as it fails to capture and incorporate informative temporal cues for effective fusion.

While attentive fusion methods improve segmental evaluation metrics, they still lag behind AFT by up to 3.1 in F1 scores. This performance gap can be attributed to their limited ability to handle the action-frame misalignment, due to their frame-level (Vaswani et al., 2017) or video-level modeling nature (Lee et al., 2023). In contrast, AFT effectively addresses this misalignment by estimating the alignment between actions and frames in a fine-grained manner, thereby providing informative cues for fusion and achieving the best overall performance. This result suggests that effective fu-

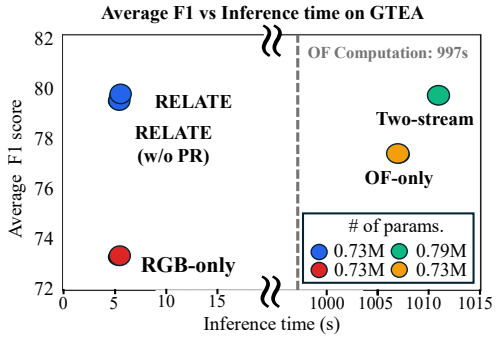

Figure 3. Average F1 score against inference time on GTEA compared with various modality setups, using the MS-TCN segmenter.

sion is particularly critical under CMKD, where transferred cues may be misaligned with the action semantics.

4.4.2. Computational Cost

Fig. 3 illustrates the average F1 score against the inference time per video on GTEA, where the average video length is 37 seconds. The inference time includes optical flow extraction [1], I3D pre-processing, and the model forward pass. The unimodal models exhibit a clear trade-off between performance and efficiency. The F1 score in the RGB-only setup decreases by 4.0 points compared to the optical flow-only setup (OF-only), while maintaining a high computational efficiency. The two-stream model leverages both modalities and achieves the highest F1 score of 79.7, but incurs a latency of 16.8 minutes due to the optical flow extraction step. As indicated by the gray dashed line, optical flow extraction accounts for 98% of the total inference time (16.6 min).

To mitigate this overhead, our method avoids optical flow extraction at inference while selectively leveraging informative cross-modal cues. Thus, RELATE matches the two-stream model with an average F1 score of 79.7. This is achieved with an inference time of 5.7 seconds, which is approximately 175× shorter, demonstrating improved efficiency.

4.4.3. Effectiveness of RELATE on various distillation methods

Tab. 4 presents our method extended to various distillation approaches, including one instance-level (Gupta et al., 2016) (first and second rows) and two relational-level distillation methods (Ni et al., 2022; Dai et al., 2021) that distill temporal relational cues (third through sixth rows). Compared with the instance-level method, relational-level methods result in lower performance, both with and without our method. This is because relational objectives must satisfy multiple temporal constraints rather than mimicking a single target, making them more vulnerable to misalignment.

[1]Following prior works (Farha & Gall, 2019), optical flow was extracted using the TV-L1 algorithm (Wedel et al., 2009) at a spatial resolution of 224 × 224 with a window size of 21 frames.

Table 4. Comparison of different CMKD methods. We use pairwise L2 distance between frame features for relational knowledge distillation (RKD) and channel covariance between frame features for global contextual relation distillation (GCR) following (Ni et al., 2022; Dai et al., 2021). Methods are evaluated on Breakfast using MS-TCN as segmenters.

CMKD	DB	F1@{10,25,50}	Edit	Acc	Average
KD (Gupta et al., 2016)	✗	57.1 / 50.4 / 36.1	59.2	49.2	50.4
RELATE + KD	✓	60.2 / 52.9 / 39.0	61.6	52.9	53.3 (+2.9)
RKD (Park et al., 2019)	✗	56.5 / 49.3 / 36.1	58.7	48.5	49.8
RELATE + RKD	✓	58.0 / 50.9 / 37.5	60.1	50.2	51.3 (+1.5)
GCR (Dai et al., 2021)	✗	55.5 / 48.7 / 36.1	58.0	47.9	49.2
RELATE + GCR	✓	56.9 / 50.2 / 37.4	59.7	49.9	50.8 (+1.6)

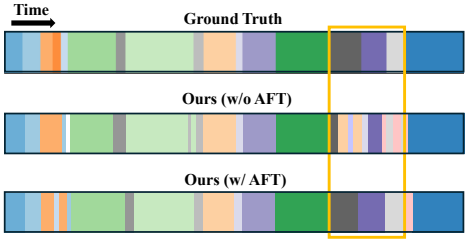

Figure 4. Visualizations of the predictions for the sequence "*rgb-23-1*" from 50Salads using MS-TCN segmenter for all predictions.

Nevertheless, applying RELATE to leverage the transferred knowledge improves all distillation approaches across all metrics, highlighting its compatibility and effectiveness.

4.4.4. Qualitative Results

In Fig. 4, we present a qualitative comparison between models with and without AFT. In the absence of AFT, naïve concatenation is applied for fusion. As shown in the yellow boxes, fragmented and repetitive predictions are observed without AFT, indicating a disrupted understanding of the temporal structure. Such repeated predictions are not observed when using AFT. This improvement is attributed to the integration of highly informative cues from each modality. Consequently, the model reasons more effectively over meaningful temporal cues, resulting in predicted segments that align more closely with the ground truth.

5. Conclusion

In this paper, we demonstrate that selective integration of transferred knowledge is an essential part of effective CMKD in TAS. To this end, we present RELATE, an RGB-based TAS framework that effectively leverages transferred motion cues without requiring a computationally heavy modality (i.e., optical flow) during inference. Through the proposed alignment fusion transformer and the prediction refinement step, RELATE selectively incorporates informative modality cues. As a result, RELATE consistently outperforms RGB-only baselines and even approaches two-stream performance while significantly improving computational efficiency.

Acknowledgements

This research was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

References

Ahn, H. and Lee, D. Refining action segmentation with hierarchical video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Bahrami, E., Francesca, G., and Gall, J. How much temporal long-term context is needed for action segmentation? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13820–13830, 2023.

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Crasto, N., Weinzaepfel, P., Harchaoui, Z., and Schmid, C. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Dai, R., Das, S., and Bremond, F. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Ding, G., Sener, F., and Yao, A. Temporal action segmentation: An analysis of modern techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (2):1011–1030, 2023.

Farha, Y. A. and Gall, J. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Fathi, A., Ren, X., and Rehg, J. M. Learning to recognize objects in egocentric activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3281–3288, 2011.

Gan, Z. et al. Asquery: A query-based model for action segmentation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. i–vi. IEEE, 2024. doi: 10.1109/ICME57554.2024.10687535.

Gupta, S., Hoffman, J., and Malik, J. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2827–2836, 2016.

Huo, F. et al. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Kuehne, H., Arslan, A., and Serre, T. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 780–787, 2014.

Lee, P. et al. Decomposed cross-modal distillation for rgb-based temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Lu, Z. and Elhamifar, E. Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ni, J., Wang, Y., Tang, H., and Cui, Z. Cross-modal knowledge distillation for vision-to-sensor action recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4613–4617. IEEE, 2022.

Park, W., Kim, D., Lu, Y., and Cho, M. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3967–3976, 2019.

Stein, S. and McKenna, S. J. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 729–738, 2013.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wedel, A., Pock, T., Zach, C., Bischof, H., and Cremers, D. An improved algorithm for tv-l1 optical flow. In *Statistical and Geometrical Approaches to Visual Motion Analysis*, pp. 23–45. Springer, 2009.

Yi, F., Wen, H., and Jiang, T. Asformer: Transformer for action segmentation. In *British Machine Vision Conference (BMVC)*, 2021.

Zhang, Y., Latham, P. E., and Saxe, A. M. Understanding unimodal bias in multimodal deep linear networks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Zhong, Q., Ding, G., and Yao, A. Onlinetas: An online baseline for temporal action segmentation. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.