Multimodal Distillation for Egocentric Action Recognition

Gorjan Radevski* Dusan Grujicic* Marie-Francine Moens Tinne Tuytelaars
KU Leuven University, Belgium
{firstname}.{lastname}@kuleuven.be

Abstract

The focal point of egocentric video understanding is modelling hand-object interactions. Standard models, e.g. CNNs or Vision Transformers, which receive RGB frames as input perform well, however, their performance improves further by employing additional input modalities (e.g. object detections, optical flow, audio, etc.) which provide cues complementary to the RGB modality. The added complexity of the modality-specific modules, on the other hand, makes these models impractical for deployment. The goal of this work is to retain the performance of such a multimodal approach, while using only the RGB frames as input at inference time. We demonstrate that for egocentric action recognition on the Epic-Kitchens and the Something-Something datasets, students which are taught by multimodal teachers tend to be more accurate and better calibrated than architecturally equivalent models trained on ground truth labels in a unimodal or multimodal fashion. We further adopt a principled multimodal knowledge distillation framework, allowing us to deal with issues which occur when applying multimodal knowledge distillation in a naïve manner. Lastly, we demonstrate the achieved reduction in computational complexity, and show that our approach maintains higher performance with the reduction of the number of input views. We release our code at: https://github.com/gorjanradevski/multimodal-distillation

1. Introduction

The purpose of egocentric vision is enabling machines to interpret real-world data taken from a human’s perspective. Its applications are numerous, ranging from recognizing [63] or anticipating [15] actions, to more complex tasks such as recognizing egocentric object-state changes, localizing action instances of a particular video moment [22], etc. The focal point of egocentric vision is hand-object interactions. Usually, these hand-object interactions take place in cluttered environments, where the object of interest is often occluded, or occurs only during a short time period. Furthermore, egocentric vision often suffers from motion blur – due to the movement of the scene objects or the camera itself – and thus, understanding video content from RGB frames alone may be challenging.

To cope with these challenges, various egocentric action recognition methods [25, 29, 33, 43, 56, 59, 61] demonstrate that explicitly modelling hand-object interactions (usually represented via bounding boxes & object categories) significantly improves the action recognition performance, most notably in a compositional generalization setup [43]. Similarly, other works show that leveraging multiple modalities (optical flow, audio, etc.) at inference time yields improved performance [16, 28, 36, 57]. The assumptions these methods make are (i) that all modalities used during training are also available at inference time, and (ii) the compute budget at inference time would be sufficient to obtain and process the additional modalities. Such assumptions make these methods cumbersome or even impossible to use in practice, e.g. on a limited compute budget such as in the case of embedded devices. Namely, using dedicated models for each additional modality (e.g. object detector, object tracker and a transformer when using bounding boxes & object categories as input [43]), increases both the memory footprint as well as the inference time. Ideally, video understanding...
models would leverage additional modalities during training, while the resulting model would use only RGB frames at inference time, i.e. when deployed in practice.

One way to achieve the aforementioned goal is training Omnivorous models, i.e. models trained jointly on multiple modalities, which have been shown to generalize better [20] than unimodal counterparts. In this work, we take a different route and transfer multimodal knowledge to models subsequendy used in a unimodal setting. Namely, we distill the knowledge from a multimodal ensemble – exhibiting superior performance, but unviable for deployment – to a standard RGB-based action recognition model [5] (see Fig. 1).

Contributions. We employ state-of-the-art knowledge distillation practices [6] and 1 show that a student [31] taught by a multimodal teacher, is both more accurate and better calibrated than the same model trained from scratch or in an omnivorous fashion (§4.1); 2 We provide motivation and establish a simple but reliable multimodal distillation approach, which overcomes the issue of potentially suboptimal modality-specific teachers (§4.2); 3 We demonstrate that the distilled student performs on par with significantly larger models, and maintains performance in computationally cheaper inference setups (§4.3).

2. Related Work

Knowledge distillation. Originally introduced by Hinton et al. [26], knowledge distillation is used to transfer the knowledge from one model, i.e. a teacher, to another model, i.e. a student, by training the student to match the teacher’s (intermediate) outputs on a certain dataset. Shown to be useful in a variety of contexts, the primary goal is model compression – transferring the knowledge from a larger, cumbersome teacher model, or from a teacher exhibiting a different inductive bias [9, 52], to a typically lightweight student model [10, 35, 55]. Another line of work [2, 47] proposes to distill from ensembles of large teacher models to lightweight student models, obtaining promising results. Compared to these works, we focus on knowledge distillation from a multimodal teacher ensemble, i.e. a set of models where each is trained on a distinct modality.

Multimodal knowledge distillation has been previously used mainly in a cross-modal fashion, where the teacher and the student receive different modalities as input. In some works, the teacher receives RGB images while the student receives depth or optical flow images [24], while in others, the teacher receives RGB images as input, while the student receives audio as input [3]. In contrast, other works explore a multimodal knowledge expansion scenario, where a multimodal student learns from pseudo-labels of a unimodal teacher [58]. We, on the other hand, focus on scenarios where obtaining additional modalities (optical flow, object detections, audio, etc.) during inference is prohibitive due to a limited compute budget, and therefore multimodal data is only used during training time. Multimodal knowledge distillation for action recognition has been previously explored in the works of Gracia et al. [17, 18]. They propose to train a model on aligned data from two modalities, where a model which receives data from modality A is trained to imitate the intermediate features of a model which receives training data from modality B. This approach is shown to yield performance improvements compared to an RGB baseline when using RGB and depth data. Moreover, the Mars [11] and D3d [49] methods, similarly to us, leverage RGB frames and optical flow during training to improve test-time performance of a model that performs inference using RGB frames alone. This is achieved by matching the corresponding features or probabilistic outputs of the modality-specific models during training. Compared to these works, we consider a more general knowledge distillation setting [6, 26], with a multimodal teacher ensemble used to provide a better approximation of the true posterior. Lastly, we consider a more diverse and broader set of modalities (RGB frames, optical flow, audio and object detections) compared to the aforementioned works.

Multimodal (egocentric) video understanding. In the context of (egocentric) video understanding, several works have shown that using additional modalities at inference time significantly improves performance [25, 29, 33, 36, 43, 50, 56, 61]. The hypothesis is intuitive – certain actions are more easily understood from specific modalities, e.g. to recognize that a person is “pushing something from left to right,” the bounding boxes alone are sufficient [33, 43]. Nevertheless, the assumption these works make is that all modalities used during training are available during inference, and that the compute budget allows for processing additional modalities other than the RGB frames. To that end, multiple works [25, 29, 33, 43, 56, 61] effectively use a Faster R-CNN [44], multiple object tracker (MOT), and object detection-specific models at inference time. In contrast, we posit that for egocentric video understanding, computing additional modalities on the fly may be prohibitive. Therefore, we propose a distillation approach which uses multimodal data only during training, while the resulting model is dependent on RGB frames alone during inference.

Models robust to missing modalities during inference. A parallel route to our goal – retaining the performance of multimodal approaches, while using only the RGB frames at inference time – is to explicitly train models to be robust to missing modalities during inference [37, 39, 62]; or more recently, to process different modalities altogether interchangeably – Omnivorous models [12, 19, 20]. These models have been shown to generalize better than models trained on unimodal data. In this work, we train an Omnivorous model using the same architecture as our student, and show that the student distilled from a multimodal teacher generalizes better than its Omnivorous variant.
3. Methodology

3.1. Egocentric Action Recognition

We assume we are given an input \( \mathbf{x} \in \mathbb{R}^{T \times D_1 \times D_2 \times \cdots \times D_k} \), which describes an egocentric action sequence, where \( T \) is the number of time-steps, while \( D_1 \times D_2 \times \cdots \times D_L \) represent other dimensions of the input data, e.g. the height, width and the number of channels of a video frame. The goal of the model \( f \) is to produce a discrete probability distribution over a predefined set of \( C \) classes, i.e. \( \hat{y} = \sigma ( f(\mathbf{x}) ) \in \mathbb{R}_+^C \), where \( \sigma \) is the softmax operator. The classes represent the actions, or alternatively, the nouns and verbs constituting the actions (e.g. the active video object and the activity).

Given a dataset \( \mathcal{D} = \{ (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N) \} \) of \( N \) egocentric action sequences \( \mathbf{x}_i \) paired with labels \( y_i \in \mathbb{R}_+^C \), the model is trained by minimizing the standard cross-entropy objective \( \mathcal{L}_{CE} = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot \log \sigma ( f(\mathbf{x}_i) ) \), where \( \cdot \) represents the scalar product. In the case of actions characterized by separate nouns and verbs, and accompanied by their respective labels \( y_i^n \) and \( y_i^v \), separate prediction heads produce \( f^n(\mathbf{x}_i) \) and \( f^v(\mathbf{x}_i) \). Finally, the model is trained by minimizing the sum of the loss terms corresponding to the nouns and verbs, \( \mathcal{L}_{CE}^n \) and \( \mathcal{L}_{CE}^v \) respectively.

3.2. Multimodal Knowledge Distillation

In egocentric vision, there often exist more than one input modality that characterize the same actions. The action recognition task may thus be performed via the use of multiple modalities, leveraged only during training [20, 42], or both during training and inference [16, 28, 36, 57] by ensembling [57] models, multimodal-fusion [28, 43], etc. However, in the case of the latter, processing multimodal data may be computationally prohibitive at inference time (e.g. due to a limited compute budget).

The fundamental concept our method builds on is knowledge distillation [26], featuring a teacher (usually a larger model, exhibiting strong performance, but cumbersome to use in practice) and a student (typically a smaller model, trained to mimic the teacher [6]). Focusing on the most accessible data modality – RGB video frames (e.g. obtained using a single monocular video camera) – we opt for distilling the knowledge of a multimodal ensemble to a single model that relies on RGB inputs alone. We make a modification to the standard knowledge distillation approach, by altering the teacher such that (i) it is not a single model, but rather an ensemble of models, and (ii) the constituting models receive different modalities as input.

**Teacher ensemble.** Given \( M \) datasets \( \mathcal{D}_m = \{ (\mathbf{x}_1^m, y_1), \ldots, (\mathbf{x}_N^m, y_N) \} \) of different modalities, we train a separate model \( f_m \) by minimizing the learning objective \( \mathcal{L}_{CE}^m = -\frac{1}{N} \sum_{i=1}^{N} y_i \cdot \log \sigma ( f_m(\mathbf{x}_i^m) ) \) for each modality. Finally, the output of the ensemble can be obtained by averaging the outputs of the individual teachers:

\[
\hat{y}_i = \sigma \left( \frac{1}{M} \sum_{m=1}^{M} f_m(\mathbf{x}_i^m) \right). \tag{1}
\]

Intuitively, under-performing modality-specific models could negatively affect the performance of the ensemble. We thus consider assigning different weights to the output logits of each model in the ensemble, before aggregating their predictions. Ideally, we would want to perform a Bayesian prediction: \( p(\mathbf{y}|\mathbf{x}, \mathcal{D}) = \int_{\mathcal{D}} p(\mathbf{y}|x, f)p(f|\mathcal{D})df. \) For a given finite ensemble of \( M \) diverse predictors, we replace the integral with a sum over the individual models: \( p(\mathbf{y}|\mathbf{x}, \mathcal{D}) \approx \sum_{m=1}^{M} p(\mathbf{y}|x, f_m)p(f_m|\mathcal{D}). \) We further approximate \( p(f_m|\mathcal{D}) \) via its proportionality to the data likelihood \( p(\mathcal{D}|f_m) \) under the Bayes rule: \( p(f_m|\mathcal{D}) \propto p(\mathcal{D}|f_m); \) which itself can be expressed in terms of the cross-entropy that the model \( f_m \) exhibits on the dataset \( \mathcal{D} \).

The cross-entropy \( e_m \) of each modality-specific model in the ensemble can be estimated on a holdout set:

\[
e_m = -\frac{1}{Z} \sum_{i=1}^{Z} \mathbf{y}_i \cdot \log \sigma ( f_m(\mathbf{x}_i^m) ), \tag{2}
\]

where \( Z \) is the number of held-out samples used to estimate the weights. Then, we can obtain the weights for the modality-specific models via softmax normalization of the negative cross-entropy terms:

\[
w^m \propto \exp(-e_m / \gamma), \tag{3}
\]

where \( \gamma \) is a temperature term which controls the entropy of the model weights, e.g. if \( \gamma \to \infty \), equal weights would be given to each teacher – resulting in an arithmetic mean.

We finally compute the weighted average of the predictions of \( M \) modality-specific models as the teacher output:

\[
\hat{y}_i = \sigma \left( \frac{1}{M} \sum_{m=1}^{M} w^m f_m(\mathbf{x}_i^m) \right). \tag{4}
\]

Figure 2 presents a high-level overview of our approach. In summary, our student is taught by a multimodal teacher which is itself an ensemble of multiple modality-specific models, trained separately on each modality.

**Training objective.** During training, we perform multimodal knowledge distillation, as originally proposed by Hinton et al. [26]. Specifically, we minimize the KL-divergence \( \mathcal{L}_{KL} \) between the class probabilities predicted by the teacher \( \hat{y}_i = [\hat{y}_{i,1}, \ldots, \hat{y}_{i,C}] \in \mathbb{R}_+^C \) and the class probabilities of the student \( \hat{y}_i^s = [\hat{y}_{i,1}^s, \ldots, \hat{y}_{i,C}^s] \in \mathbb{R}_+^C \) as \( \mathcal{L}_{KL} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{-} \hat{y}_i^s \cdot \log \hat{y}_i^s + \hat{y}_i^t \cdot \log \hat{y}_i^t). \)

Additionally, we use a temperature parameter \( \tau \) to control the entropy of the predicted probability scores while
preserving their ranking, i.e. \( \hat{y}^j \propto \exp(y^j/\tau) \). As per standard practice [26], we use the temperature parameter \( \tau \) to also rescale the \( KL \)-divergence loss, i.e. \( L_{KL} = L_{KL} \cdot \tau^2 \). We further use the standard cross-entropy action recognition loss \( L_{CE} \), and compute the final loss:

\[
L = \lambda \cdot L_{KL} + (1 - \lambda) \cdot L_{CE},
\]

where \( \lambda \) balances the distillation loss \( L_{KL} \) and the action recognition loss \( L_{CE} \). For example, with \( \lambda = 0.0 \) we would effectively be training the modality-specific RGB model, while with \( \lambda = 1.0 \), we would perform solely multimodal knowledge distillation.

**Optimality of a multimodal ensemble teacher.** The work of Menon et al. [34] further demonstrates that in the context of distillation, a teacher that predicts the Bayes class-probability distribution over the labels \( p(x) = \mathbb{P}(y|x)_{y \in [C]} \) exhibits the lowest possible variance of the student objective for any convex loss function.

The student trained to minimize the \( KL \)-divergence between its output and the Bayes class-probabilities would thus generalize the best. In a preliminary experiment shown in Table 1, we demonstrate that the multimodal ensemble (both for \( \gamma = 1 \) and \( \gamma = 30.0 \)) achieves a significantly higher accuracy and lower calibration error, and thus represents a better approximation of the Bayes class probabilities than a single modality-specific model. This may lead to a lower-variance objective for the student and improved generalization during knowledge distillation [34].

### 4. Experiments & Discussion

**Datasets.** We use the Something-Something (V2) [21] and the Epic-Kitchens (100) [14] datasets. Something-Something contains videos of people performing 174 (object agnostic) unique actions with their hands, e.g. “pushing [something] left”, “taking [something] out of [something]”, etc. Epic-Kitchens’ videos take place in kitchen environments, where the actions are noun-verb compositions. The 300 unique nouns indicate the active object in the video, while 97 unique verbs indicate the activity, e.g. “cutting carrot”, “washing pan”, etc. During evaluation, the action is considered correct if both the noun and the verb are correctly predicted. Additionally, we use the Something-Else [33] dataset and the Epic-Kitchens Unseen split to measure the compositional generalization ability of the models w.r.t. unseen objects and environments.

**Modalities.** In addition to the RGB frames (the modality of interest at inference time), in our experiments we consider the following modalities:

1. (ii) **Object detections (OBJ):** Shown to significantly improve the performance of standard (RGB) egocentric video understanding models across datasets [25, 33, 43, 56, 61]. As per [46], we use the object detector trained on object-agnostic annotations, i.e. with “hands” and “objects” object labels.

2. (iii) **Audio (A):** For certain datasets [13,14], several works [28,57] have shown that using audio improves action recognition performance. The audio is obtained directly from the recorded video.

Moreover, if available, one may include additional modalities, e.g. depth estimates, heat maps, etc. For Something-Something and Something-Else there is no audio available, while we use the object detections provided by [25, 33], and optical flow from [54]. For Epic-Kitchens we use the modalities available and released with the dataset itself [14]: optical flow and audio.

**Models.** We use a Swin-Tiny (Swin-T) Transformer [31] to encode RGB frames, optical flow, and audio. Each optical flow frame is represented as a \( 224 \times 224 \times 2 \) tensor, where the two values at each spatial location represent the \( x \) and \( y \) velocity components. In the case of audio, we extract 1.116 second-audio segments (0.558s before its corresponding time-step and 0.558s after it) for each frame. We compute the mel-spectrogram of the audio segment (see details in Supp. B), which is subsequently resized to desired width and height. We thus treat each modality as a sequence of...
224 × 224 multi-channel images, which we provide as input to the vision transformer, as per the common practice in recent vision models [19, 20]. To encode the object detections, i.e. bounding boxes and object categories of the scene objects, we use a state-of-the-art model – STLT [43]. In STLT, a spatial and temporal transformer separately encode the spatial and temporal arrangement of the objects occurring in the video. The multimodal teacher we use during knowledge distillation is an ensemble of individual, modality-specific models. Unless noted otherwise, the student is a Swin-T model which receives RGB frames as input, both during training (distillation) and inference. In addition to Swin-T, in §4.3, we also consider ResNet3D [27] (18 and 50 layers deep) light-weight student models which receive video frames of size 112 × 112 as input.

**Metrics.** Besides accuracy, we measure Expected Calibration Error (ECE) [23]. As per [23, 45], we sort the predictions based on the per-class confidence scores and group them into K bins B_k, each associated with a confidence interval I_{B_k} = (\frac{k-1}{K-1}, \frac{k}{K}). For a prediction f(x_i) \in I_{B_k} \subset [0, 1] \subset \mathbb{R}, we define the model’s confidence in this prediction as conf(f(x_i)). The Expected Calibration Error is thus 

\[ECE = \frac{1}{N} \sum_{k=1}^{K} \frac{|B_k|}{N} |\text{acc}(B_k) - \text{conf}(f(B_k))|,\]

where N is the number of evaluation samples.

**Implementation details.** We train all models for 60 epochs using AdamW [32], with a peak learning rate of 1e−4, linearly increased for the first 5% of the training and decreased to 0.0 by the end of the specified 60 epochs. We use weight decay with a regularization coefficient of 5e−2, and clip the gradients when their norm exceeds 5.0. For Epic-Kitchens, we sample 32 frames with a fixed stride of 2, and for Something-Something and Something-Else we evenly sample 16 frames to cover the whole video. We use a single spatial and temporal crop, unless stated otherwise. During training, we chose a random start frame, while during inference, we select the start frame such that the sequence covers the central portion of the video. If we use multiple temporal crops as test-time augmentation, we chose the start frames such that the video is covered uniformly. During training we apply standard data augmentations – random spatial video crops, color jittering, and horizontal flips (for Epic-Kitchens only). The temperature parameter τ is fixed to 10.0 for both the student and the teacher during multimodal knowledge distillation. In §4.2 we ablate the impact of the loss balancing term λ and the Ensemble Teacher Weighting temperature term γ.

During training we follow the consistent teaching paradigm [6] where the student and teacher strictly receive the same views of the data – we ensure for spatial and temporal consistency, i.e. the models receive the same frame indices, same random crops, and horizontal flips.

4.1. Multimodal Distillation for Egocentric Vision

We first verify the overall effectiveness of multimodal knowledge distillation on the task of egocentric action recognition for both object-agnostic actions (Something-Something) and actions represented as noun-verb compositions (Epic-Kitchens). Across all experiments, we fix λ to 1.0, i.e. we train solely with multimodal knowledge distillation. Similarly, we set γ to a large value (γ = 30.0), where effectively each teacher equally contributes to the ensemble output. Note that in this setting, models trained on modalities such as optical flow and audio may underperform, and thus adversely affect the ensemble teacher performance of recognizing active objects, i.e nouns.

**4.1.1 Recognizing Egocentric Actions**

In Table 2, we report performance on Something-Something V2 [21] and Epic-Kitchens 100 [14]. In line with the previous findings reported in the literature [33, 43, 57], we find that employing additional modalities at inference time significantly improves the performance compared to the RGB baseline model, for both the Epic-Kitchens and the Something-Something datasets.

1 As the datasets’ test sets either do not exist [33], or have restricted access, we report results using the model after the final training epoch.

---

Table 2: Egocentric action recognition. RGB = Video frames; OF = Optical flow; A = Audio; OBJ = Object detections. Multimodal distillation with λ = 1.0 and γ = 30.0. Improvement over RGB frames baseline [31] in red.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Modalities</th>
<th>Inference Modalities</th>
<th>Noun@1</th>
<th>Verb@1</th>
<th>Action@1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>RGB</td>
<td>RGB</td>
<td>52.0</td>
<td>61.7</td>
<td>38.3</td>
</tr>
<tr>
<td>Modality-specific Teacher</td>
<td>OF</td>
<td>OF</td>
<td>34.1</td>
<td>59.0</td>
<td>25.9</td>
</tr>
<tr>
<td>Student</td>
<td>RGB &amp; OF</td>
<td>RGB &amp; OF</td>
<td>51.9</td>
<td>65.3</td>
<td>39.5</td>
</tr>
<tr>
<td>Modality-specific Teacher</td>
<td>A</td>
<td>A</td>
<td>22.3</td>
<td>46.5</td>
<td>15.1</td>
</tr>
<tr>
<td>Student</td>
<td>RGB &amp; A</td>
<td>RGB &amp; A</td>
<td>52.7</td>
<td>64.4</td>
<td>39.8</td>
</tr>
<tr>
<td>Teacher</td>
<td>RGB &amp; OF &amp; A</td>
<td>RGB &amp; OF &amp; A</td>
<td>52.3</td>
<td>66.8</td>
<td>40.5</td>
</tr>
</tbody>
</table>

(a) Epic-Kitchens [14]: Recognition of the active video object (noun) and the activity (verb).

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Modalities</th>
<th>Inference Modalities</th>
<th>Action@1</th>
<th>Action@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>RGB</td>
<td>RGB</td>
<td>60.3</td>
<td>86.4</td>
</tr>
<tr>
<td>Modality-specific Teacher</td>
<td>OF</td>
<td>OF</td>
<td>49.3</td>
<td>79.0</td>
</tr>
<tr>
<td>Student</td>
<td>RGB &amp; OF</td>
<td>RGB &amp; OF</td>
<td>64.3</td>
<td>88.9</td>
</tr>
<tr>
<td>Modality-specific Teacher</td>
<td>OBJ</td>
<td>OBJ</td>
<td>47.9</td>
<td>76.2</td>
</tr>
<tr>
<td>Student</td>
<td>RGB &amp; OBJ</td>
<td>RGB &amp; OBJ</td>
<td>65.3</td>
<td>89.5</td>
</tr>
<tr>
<td>Teacher</td>
<td>RGB &amp; OF &amp; OBJ</td>
<td>RGB &amp; OF &amp; OBJ</td>
<td>66.6</td>
<td>90.5</td>
</tr>
</tbody>
</table>

(b) Something-Something [21]: Object-agnostic action recognition.
A novel observation by our work is that multimodal knowledge distillation performs well in the context of egocentric video understanding, with student models often approaching the performance of the multimodal teacher ensemble. For Epic-Kitchens (Table 2a), we observe that when the student is distilled from an RGB & OF, or RGB & OF & A teacher, it is superior to the baseline model, as well as all modality-specific models, for recognizing actions. On the other hand, distilling from an RGB & A teacher yields performance lower than the baseline, due to the low performance of the model trained only on audio data. Specifically, we observe that the audio-specific teacher lowers the noun (active object) recognition performance. In §4.2, we propose a solution for this issue. For Something-Something (Table 2b), the student model is superior to the baseline for each combination of modalities. When distilling from all available modalities (RGB & OF & OBJ), the resulting model outperforms the baseline by 3.7% in terms of the top-1 accuracy. In terms of the top-5 accuracy on Something-Something, the students achieve performance that is nearly on par with the multimodal teacher ensemble.

### 4.1.2 Generalizing to Unseen Environments & Objects

We investigate to what extent our findings translate to the compositional generalization setting 2, in which the performance of standard video models deteriorates significantly.

#### Table 4: Comparison with Omnivorous models [20].

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Modalities</th>
<th>Inference Modalities</th>
<th>Noun@1</th>
<th>Verb@1</th>
<th>Action@1</th>
<th>Noun@5</th>
<th>Verb@5</th>
<th>Action@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Omnivore [20]</td>
<td>RGB</td>
<td>RGB</td>
<td>38.3</td>
<td>51.7</td>
<td>25.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student</td>
<td>RGB &amp; OF</td>
<td>RGB</td>
<td>28.0</td>
<td>53.2</td>
<td>21.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>OF</td>
<td>RGB</td>
<td>41.0</td>
<td>54.9</td>
<td>28.4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>RGB &amp; OF</td>
<td>RGB</td>
<td>42.5±0.2</td>
<td>55.9±0.4</td>
<td>30.2±0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>A</td>
<td>A</td>
<td>15.0±0.2</td>
<td>41.5±0.1</td>
<td>9.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>RGB &amp; A</td>
<td>RGB</td>
<td>41.9±0.2</td>
<td>55.3±0.1</td>
<td>28.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>RGB &amp; OF &amp; A</td>
<td>RGB</td>
<td>42.9±0.2</td>
<td>58.0±0.2</td>
<td>36.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>RGB &amp; OF &amp; OBJ</td>
<td>RGB</td>
<td>43.7±0.2</td>
<td>54.3±0.2</td>
<td>29.6±0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>RGB</td>
<td>RGB</td>
<td>47.8±0.2</td>
<td>62.8±0.2</td>
<td>35.9</td>
<td>58.4±0.2</td>
<td>86.2±0.2</td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>RGB &amp; OF</td>
<td>RGB</td>
<td>51.7±0.2</td>
<td>65.4±0.2</td>
<td>39.3</td>
<td>63.0±0.2</td>
<td>88.9±0.2</td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>RGB &amp; OF &amp; OBJ</td>
<td>RGB</td>
<td>58.5±0.2</td>
<td>64.5±0.2</td>
<td>27.9</td>
<td>59.3±0.2</td>
<td>84.9±0.2</td>
<td></td>
</tr>
<tr>
<td>Teacher</td>
<td>RGB</td>
<td>RGB</td>
<td>43.7±0.2</td>
<td>54.1±0.2</td>
<td>29.6</td>
<td>59.1±0.2</td>
<td>86.1±0.2</td>
<td></td>
</tr>
</tbody>
</table>

2Compositional generalization measures to what extent the model can generalize to novel combinations of concepts observed during training.

#### Table 3: Egocentric action recognition with unseen environments and objects. RGB = Video frames; OF = Optical flow; A = Audio; OBJ = Object detections. Multimodal distillation with λ = 1.0 and γ = 30.0. Improvement over RGB frames baseline [31] in red.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training Modalities</th>
<th>Inference Modalities</th>
<th>Noun@1</th>
<th>Verb@1</th>
<th>Action@1</th>
<th>Noun@5</th>
<th>Verb@5</th>
<th>Action@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>RGB</td>
<td>RGB</td>
<td>51.8</td>
<td>79.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modality-specific Teacher</td>
<td>OF</td>
<td>OF</td>
<td>49.0</td>
<td>77.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modality-specific Teacher</td>
<td>RGB &amp; OF</td>
<td>RGB &amp; OF</td>
<td>61.0</td>
<td>86.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modality-specific Teacher</td>
<td>RGB &amp; OBJ</td>
<td>RGB &amp; OBJ</td>
<td>58.2±0.2</td>
<td>85.1±0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modality-specific Teacher</td>
<td>OBJ</td>
<td>OBJ</td>
<td>41.4</td>
<td>67.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modality-specific Teacher</td>
<td>RGB &amp; OBJ</td>
<td>RGB &amp; OBJ</td>
<td>59.4</td>
<td>84.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Modality-specific Teacher</td>
<td>RGB &amp; OF &amp; OBJ</td>
<td>RGB &amp; OF &amp; OBJ</td>
<td>63.6±0.2</td>
<td>87.7±0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher Student</td>
<td>RGB &amp; OF &amp; OBJ</td>
<td>RGB &amp; OF &amp; OBJ</td>
<td>58.0±0.2</td>
<td>85.1±0.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Table 3a: Egocentric action recognition with unseen environments and objects. RGB = Video frames; OF = Optical flow; A = Audio; OBJ = Object detections. Multimodal distillation with λ = 1.0 and γ = 30.0. Improvement over RGB frames baseline [31] in red.

#### Table 3b: Something-Something [33]: Object-agnostic action recognition featuring objects unseen during training.

#### Figure 3: Expected Calibration Error across datasets. The student is Swin-T trained with λ = 1.0 and γ = 30.0.
that using homogeneous batches yields better performance. ECE for an RGB model trained using the ground truth labels on the Something-Something and Something-Else datasets. We report the results of Table 5: Ablation study on the Epic-Kitchens dataset. $\lambda$: Distillation and Cross-Entropy loss balancing term; $\gamma$: Temperature of the Ensemble Teacher Weighting.

4.1.5 Per-Class Performance Breakdown

In Fig 4, we present the relative change in action recognition accuracy of the student model obtained via multimodal distillation w.r.t. the architecturally equivalent baseline RGB model trained on ground truth labels, computed on the top-20 most frequent classes (actions) on Epic-Kitchens and Something-Something. Overall, we observe that multimodal distillation generally improves performance across different actions, and particularly so on Something-Something, where we achieve improvements in terms of all of the top-20 most frequent action classes.

4.2 Ensemble Teacher Weighting

In §4.1 we observe that distilling from multimodal teachers yields students which are superior to models trained on the ground truth labels. Nevertheless, if a weak teacher is added to the ensemble, it negatively affects the knowledge distillation and yields a student which is inferior than using the ground truth labels, e.g. adding the audio-specific teacher in the ensemble for Epic-Kitchens. To cope with this, we weigh the logits of the teacher ensemble as discussed in §3.2. Namely, we use the two hyperparameters ($\lambda$ and $\gamma$) for (i) balancing between the ground truth and the distillation loss: $\lambda$ (Equation 5), and (ii) controlling how the predictions of modality-specific models are combined in the ensemble: $\gamma$ (Equation 3). We report results for Epic-Kitchens in Table 5, including the values for $\lambda$ and $\gamma$ as well as the objective we effectively minimize. We observe that the model trained only on the ground truth labels ($\lambda = 0.0$), is inferior to all other models. Using a large $\gamma$ (e.g. 30.0) – effectively assigning equal weights to all models in the ensemble – we observe to perform well despite the simple setup (we use this model in the experiments in §4.1). Additionally, training using the task loss in addition to the distillation loss ($\lambda = 0.8$) further improves the performance. The
Figure 5: Top-1 action accuracy of Teacher, Student & Baseline models and their associated computational cost in giga-FLOPs ($10^9$) required to update the input and predict the action. Note: Top-left corner is optimal (i.e. faster and most accurate models).

Figure 6: Performance degradation when reducing the number of inference clips/crops (Epic-Kitchens/Something-Something).

best performing model uses both the task and the distillation loss ($\lambda = 0.8$), and assigns weights to each teacher in the ensemble based on its performance on $Z = 1000$ randomly sampled videos held-out from the training dataset, with a normalization temperature $\gamma = 1.0$. Moreover, we find that lowering the normalization temperature $\gamma$ further ($\gamma = 0.33$) – giving higher weight to the best performing model in the ensemble – yields lower performance. Similarly, increasing the normalization temperature to $\gamma = 3.0$, thus equalizing the model weights, also negatively affects performance.

4.3. Efficiency Analysis

Despite the strong performance multimodal action recognition models exhibit, we argue that the high computational complexity makes them cumbersome for deployment, particularly compared to a model which uses only RGB video during inference. The teacher ensemble used for Epic-Kitchens uses three modality-specific Swin-T models, where each has 28.22M parameters, and requires 140.33 GFLOPs for processing single-view 32 frame/spectrogram video. Assuming such an ensemble is deployed, the optical flow would have to be computed on-the-fly\(^4\). Using RAFT, we measure a total added computation of 163.37 GFLOPs for such a model. We consider the computation required to obtain the spectrograms of 1.116s audio segments to be negligible in comparison. Therefore, when using all three modalities, updating the input sequences for each newly observed frame and performing action recognition would require 584.36 GFLOPs. In contrast, the distilled student is a single RGB model, and in the case of Swin-T requires 140.33 GFLOPs – which represents a reduction of 75.98%.

We report results on the Epic-Kitchens dataset in Fig. 5 for: (i) All variations of teacher models (RGB & OF, RGB & A, and RGB & OF & A); (ii) The Swin-T student model, distilled with $\lambda = 0.8$ and $\gamma = 1.0$; (iii) The Swin-T baseline model, trained with $\lambda = 0.0$; (iv) Two new ResNet3D models [27] (with depth of 50 and 18 layers), exhibiting less parameters and GFLOPs compared to the Swin-T model. The resolution size of the ResNet3D models is $112 \times 112$. For each ResNet3D, we report action recognition performance of the baseline with $\lambda = 0.0$, and performance of the distilled students with $\lambda = 0.8$ and $\gamma = 1.0$.

We observe a consistently higher performance of the student models compared to the same model architecture trained on ground truth labels alone. Notably, our best performing student achieves comparable performance to the significantly more expensive RGB & OF & A teacher. Furthermore, the R3D-18 and R3D-50 students outperform

\(^4\)The Duality-based TV-L1 [40, 60] can be efficiently computed on a GPU (5-10 FPS) [4]. Deep Learning-based approaches, e.g. RAFT [51], require 163.37 GFLOPs, however, achieve higher FPS of 21.10, measured with 10 refinement iterations and resolution of $256 \times 456$.
Figure 7: Qualitative example for the Epic Kitchens (Left) & the Something-Something (Right) datasets.

their counterparts trained on class labels. Finally, we observe that the R3D-18 student matches the performance of the larger, and computationally more expensive R3D-50 trained on ground truth labels.

4.3.1 Effect of Test-Time Augmentation

Lastly, we inspect the relationship between action recognition performance and the number of temporal clips (on Epic-Kitchens) and spatial crops (on Something-Something) sampled from the video during inference. Note that our goal is to reduce the computational complexity while maintaining the performance. Since standard video models [1, 7, 30, 31] use multiple temporal clips and spatial crops as test-time augmentation to obtain further performance improvements, we explore the extent to which the distilled model is dependent on their availability during inference. We report results in Fig. 6, where we observe that the distilled model ($\lambda = 1.0$) is much less adversely affected by the reduction of both sampled temporal clips and spatial crops during inference, compared to the same model trained on the ground truth labels ($\lambda = 0.0$).

4.4. Qualitative Examples

In Fig. 7, we showcase the classes corresponding to the highest scores predicted by the student and the individual modality-specific models in the teacher ensemble, as well as the ground truth label (on Epic-Kitchens and Something-Something datasets). For both examples, we observe that the student picks up on relevant cues from each modality and accurately predicts the action of interest (see Supp. H for additional qualitative examples).

5. Conclusion

We demonstrated a simple, yet effective distillation-based approach for leveraging multimodal data only during training in order to improve a model that uses solely RGB frames during inference. Our experiments indicate clear performance improvements over models trained on ground truth labels. We further showed an advantageous trade-off between the high performance of a cumbersome multimodal ensemble, and low computational complexity of unimodal approaches. Moreover, our approach relies less on expensive test-time augmentations, otherwise widely used in the literature to improve the egocentric action recognition models’ performance.

Limitations & Future work. Notably, in this work we considered only the task of action recognition, while multimodal distillation can be readily applied to other egocentric tasks [22]. Future work may also feature additional modalities such as depth, hand poses, motion captured by inertial sensors (IMU), etc., available in recent large-scale egocentric datasets [22].

Acknowledgement

We acknowledge the funding from the Flemish Government under the Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen programme.
References


[52] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training


