Multimodal Distillation for Egocentric Action Recognition

Gorjan Radevski* Dusan Grujicic* Matthew Blaschko
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

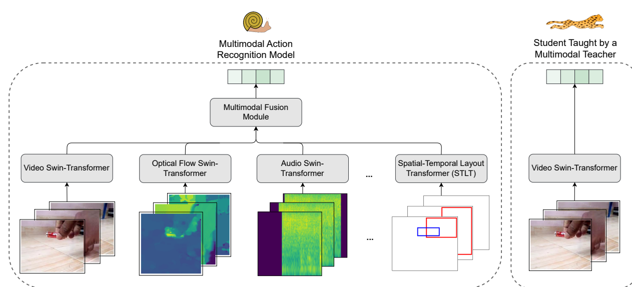

Figure 1: **Motivation:** The multimodal action recognition model is powerful, but too slow to be used in practice. The distilled student is significantly faster yet achieves competitive performance.

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*

[*] Authors contributed equally.

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 subsequently 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 $\mathbf{x} \in \mathbb{R}^{T \times D_1 \times D_2 \ldots \times D_L}$, which describes an egocentric action sequence, where T is the number of time-steps, while $D_1 \times D_2 \ldots \times D_L$ 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. $\hat{\mathbf{y}} = \sigma\left(f\left(\mathbf{x}\right)\right) \in \mathbb{R}_+^C$, where σ 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 $\mathbb{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_N, \mathbf{y}_N)\}$ of N egocentric action sequences $\mathbf{x}_i$ paired with labels $\mathbf{y}_i \in \mathbb{R}_+^C$, the model is trained by minimizing the standard cross-entropy objective $\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i \cdot \log \sigma\left(f\left(\mathbf{x}_i\right)\right)$, where $\cdot$ represents the scalar product. In the case of actions characterized by separate nouns and verbs, and accompanied by their respective labels $\mathbf{y}_i^n$ and $\mathbf{y}_i^v$, 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, $\mathcal{L}_{\text{CE}}^n$ and $\mathcal{L}_{\text{CE}}^v$ 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 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 $\mathbb{D}^m = \{(\mathbf{x}_1^m, \mathbf{y}_1), \ldots, (\mathbf{x}_N^m, \mathbf{y}_N)\}$ of different modalities, we train a separate model f^m by minimizing the learning objective $\mathcal{L}_{\text{CE}}^m = -\frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i \cdot \log \sigma\left(f^m\left(\mathbf{x}_i^m\right)\right)$ for each modality. Finally, the output of the ensemble can be obtained by averaging the outputs of the individual teachers:

$$\hat{\mathbf{y}}_i^t = \sigma\left(\frac{1}{M}\sum_{m=1}^{M} f^m(\mathbf{x}_i^m)\right). \tag{1}$$

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: $p(\mathbf{y}|\mathbf{x}, \mathbb{D}) = \int_{f \in F} p(\mathbf{y}|\mathbf{x}, f) p(f|\mathbb{D}) df$. For a given finite ensemble of M diverse predictors, we replace the integral with a sum over the individual models: $p(\mathbf{y}|\mathbf{x}, \mathbb{D}) \approx \sum_{m=1}^{M} p(\mathbf{y}|\mathbf{x}, f^m) p(f^m|\mathbb{D})$. We further approximate $p(f^m|\mathbb{D})$ via its proportionality to the data likelihood $p(\mathbb{D}|f^m)$ under the Bayes rule: $p(f^m|\mathbb{D}) \propto p(\mathbb{D}|f^m)$; which itself can be expressed in terms of the cross-entropy that the model f^m exhibits on the dataset $\mathbb{D}$.

The cross-entropy e^m of each modality-specific model in the ensemble can be estimated on a holdout set:

$$e^m = -\frac{1}{Z}\sum_{i=1}^{Z} \mathbf{y}_i \cdot \log \sigma\left(f^m\left(\mathbf{x}_i^m\right)\right), \tag{2}$$

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 γ 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:

$$\hat{\mathbf{y}}_i^t = \sigma\left(\sum_{m=1}^{M} w^m f^m(\mathbf{x}_i^m)\right). \tag{4}$$

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 $\mathcal{L}_{\text{KL}}$ between the class probabilities predicted by the teacher $\hat{\mathbf{y}}_i^t = [\hat{y}_{i,1}^t, \ldots, \hat{y}_{i,C}^t] \in \mathbb{R}_+^C$ and the class probabilities of the student $\hat{\mathbf{y}}_i^s = [\hat{y}_{i,1}^s, \ldots, \hat{y}_{i,C}^s] \in \mathbb{R}_+^C$ as $\mathcal{L}_{\text{KL}} = \frac{1}{N}\sum_{i=1}^{N}(-\hat{\mathbf{y}}_i^t \cdot \log \hat{\mathbf{y}}_i^s + \hat{\mathbf{y}}_i^t \cdot \log \hat{\mathbf{y}}_i^t)$.

Additionally, we use a temperature parameter τ to control the entropy of the predicted probability scores while

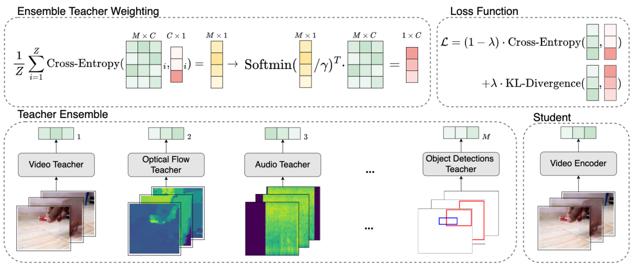

Figure 2: Overview of the main components of our approach.

	Epic-Kitchens Regular				Something-Something	
Method	Noun Acc.	Noun ECE	Verb Acc.	Verb ECE	Action Acc.	Action ECE
RGB Baseline	52.0	19.9	61.7	17.1	59.3	21.2
Teacher ($\gamma = 30.0$)	52.3	2.04	66.8	3.4	66.6	8.1
Teacher ($\gamma = 1.0$)	53.9	8.9	66.9	4.2	66.7	9.0
	Epic-Kitchens Unseen Environments				Something-Else	
Method	Noun Acc.	Noun ECE	Verb Acc.	Verb ECE	Action Acc.	Action ECE
RGB Baseline	38.3	26.9	51.7	24.0	51.8	21.4
Teacher ($\gamma = 30.0$)	42.9	6.02	58.0	9.9	63.3	6.5
Teacher ($\gamma = 1.0$)	43.8	9.5	57.7	9.7	63.5	8.1

Table 1: Accuracy & Expected Calibration Error (ECE) of the RGB baseline and Teacher ensemble on Epic-Kitchens and Something-Something (regular & unseen splits). All available modalities for the respective dataset used in the ensemble.

preserving their ranking, i.e. $\hat{y}^s_j \propto \exp(\hat{y}^s_j/\tau)$. As per standard practice [26], we use the temperature parameter τ to also rescale the KL-divergence loss, i.e. $\mathcal{L}_{\text{KL}} = \mathcal{L}_{\text{KL}} \cdot \tau^2$. We further use the standard cross-entropy action recognition loss $\mathcal{L}_{\text{CE}}$, and compute the final loss:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\text{KL}} + (1-\lambda) \cdot \mathcal{L}_{\text{CE}}, \tag{5}$$

where λ balances the distillation loss $\mathcal{L}_{\text{KL}}$ and the action recognition loss $\mathcal{L}_{\text{CE}}$. 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 $\mathbf{p}^*(\mathbf{x}) = [\mathbb{P}(y|\mathbf{x})]_{y\in[C]}$ 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:

(i) *Optical flow* (OF): First used by I3D [8] for action recognition, it is also used by recent state-of-the-art (egocentric) video understanding methods [57]. We use optical flow obtained using TV-L_1 [60].

(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.

(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

Method	Training Modalities	Inference Modalities	Noun@1	Verb@1	Action@1
Baseline	RGB	RGB	52.0	61.7	38.3
Modality-specific	OF	OF	34.1	59.0	25.9
Teacher	—	RGB & OF	51.9	65.3	39.5
Student	RGB & OF	RGB	$52.2_{+0.2}$	$65.6_{+3.9}$	$39.9_{+1.6}$
Modality-specific	A	A	22.3	46.5	15.1
Teacher	—	RGB & A	52.7	64.4	39.8
Student	RGB & A	RGB	$51.5_{-0.5}$	$62.4_{+0.7}$	$37.9_{-0.4}$
Teacher	—	RGB & OF & A	52.3	66.8	40.5
Student	RGB & OF & A	RGB	$51.7_{-0.3}$	$65.4_{+3.7}$	$39.3_{+1.0}$

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

Method	Training Modalities	Inference Modalities	Action@1	Action@5
Baseline	RGB	RGB	60.3	86.4
Modality-specific	OF	OF	49.3	79.0
Teacher	—	RGB & OF	64.3	88.9
Student	RGB & OF	RGB	$62.8_{+2.5}$	$88.9_{+2.5}$
Modality-specific	OBJ	OBJ	47.9	76.2
Teacher	—	RGB & OBJ	65.3	89.5
Student	RGB & OBJ	RGB	$63.2_{+2.9}$	$88.7_{+2.3}$
Teacher	—	RGB & OF & OBJ	66.6	90.5
Student	RGB & OF & OBJ	RGB	$63.0_{+2.7}$	$88.9_{+2.5}$

(b) Something-Something [21]: Object-agnostic action recognition.

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

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 $I_{B_k} = (\frac{k-1}{K}, \frac{k}{K})$, where $K = 15$. ECE represents the discrepancy between the average accuracy $acc(B_k)$ and the average confidence $conf(B_k)$ in each bin B_k:

$$\text{ECE} = \sum_{k=1}^{K} \frac{|B_k|}{N} |acc(B_k) - conf(B_k)|, \tag{6}$$

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[1].

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 ($\gamma = 30.0$), 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.

Method	Training Modalities	Inference Modalities	Noun@1	Verb@1	Action@1
Baseline	RGB	RGB	38.3	51.7	25.4
Modality-specific	OF	OF	28.0	53.2	21.6
Teacher	—	RGB & OF	41.0	54.9	28.4
Student	RGB & OF	RGB	42.5 +4.2	55.9 +4.2	30.2 +4.8
Modality-specific	A	A	15.0	41.5	9.1
Teacher	—	RGB & A	41.9	55.3	28.5
Student	RGB & A	RGB	41.8 +3.5	51.8 +0.1	27.5 +2.1
Teacher	—	RGB & OF & A	42.9	58.0	30.3
Student	RGB & OF & A	RGB	43.7 +5.4	54.1 +3.4	29.6 +4.2

(a) Epic-Kitchens Unseen Environments [14]: Recognition of the active object (noun) and the activity (verb) on participants unseen during training.

Method	Training Modalities	Inference Modalities	Action@1	Action@5
Baseline	RGB	RGB	51.8	79.5
Modality-specific	OF	OF	49.0	77.4
Teacher	—	RGB & OF	61.0	86.4
Student	RGB & OF	RGB	58.2 +6.4	85.1 +5.6
Modality-specific	OBJ	OBJ	41.4	67.3
Teacher	—	RGB & OBJ	59.4	84.5
Student	RGB & OBJ	RGB	57.5 +5.7	84.1 +4.6
Teacher	—	RGB & OF & OBJ	63.6	87.7
Student	RGB & OF & OBJ	RGB	59.1 +7.3	86.1 +6.6

(b) Something-Else [33]: Object-agnostic action recognition featuring objects unseen 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 $\lambda = 1.0$ and $\gamma = 30.0$. Improvement over RGB frames baseline [31] in **red.**

	Epic-Kitchens Regular split			Something-Something	
Method	Noun@1	Verb@1	Action@1	Action@1	Action@5
Omnivore [20]	47.8	62.8	35.9	58.4	86.2
Student	**51.7**	**65.4**	**39.3**	**63.0**	**88.9**
	Epic-Kitchens Unseen Participants			Something-Else	
Omnivore [20]	38.5	**54.5**	27.9	58.3	84.9
Student	**43.7**	54.1	**29.6**	**59.1**	**86.1**

Table 4: Comparison with Omnivorous models [20]. All models are Swin-T and perform inference using only RGB frames. Multimodal distillation using all modalities with $\lambda = 1.0$ and $\gamma = 30.0$.

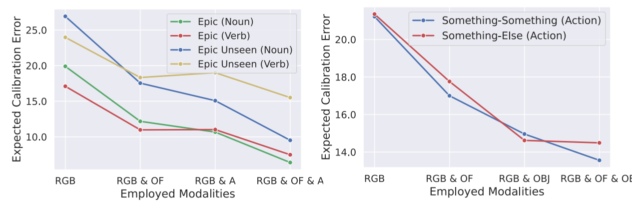

Figure 3: Expected Calibration Error across datasets. The student is Swin-T trained with $\lambda = 1.0$ and $\gamma = 30.0$.

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 [33]. In the Something-Something dataset, the objects the people interact with might overlap between training and testing, potentially allowing models to pick up undesirable biases w.r.t. object appearance to discriminate between the actions. Therefore, Something-Else proposes a compositional generalization data split, on which the objects at training and testing time do not overlap, such that the models encounter strictly novel objects during testing. Similarly, the rich visual environments featured in the Epic-Kitchen training data may also provide the model with spurious cues for predicting the actions and lead to significant performance drop on new, unobserved environments. The Epic-Kitchens Unseen split consists of a subset of videos in the validation set from participants whose videos were not present in the training dataset, giving insight into how our approach generalizes to new visual environments. We report the performance of our approach Epic-Kitchens Unseen split and the Something-Else [33] split in Table 3. The general observation across datasets and modalities is that the distilled students significantly outperform the RGB baseline model, and are sometimes even competitive with their respective teacher. Notably, on Something-Else, the student distilled from an RGB & OF & OBJ teacher outperforms the baseline by **7.3%** in terms of top-1 accuracy.

4.1.3 Comparison with Omnivorous Models

We explicitly compare multimodal knowledge distillation against Omnivorous models [19, 20], which are trained

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

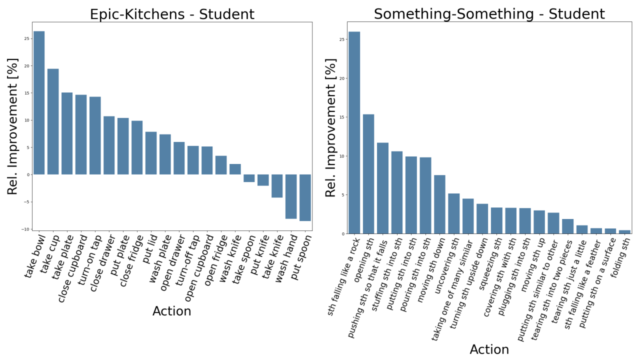

Figure 4: Per-class improvement of the student over the RGB baseline on the top-20 most frequent actions across datasets.

jointly on all modalities featuring non-aligned data, e.g. with RGB frames obtained from one dataset and depth maps from another [20]. These models have been proven to generalize better compared to their unimodal counterparts. We train the Omnivore model using batches comprised of data from the same modality, where we randomly sample the modality for each training batch[3]. Additionally, we train the Omnivore models for $M\times$ more epochs (M is the number of modalities), to account for the random sampling of modalities during training. We represent the object detections modality as bounding boxes with a line thickness of 2px pasted on a white canvas where each category is colored uniquely, e.g. hand in blue and object in red.

We report results in Table 4. Even though on the Epic-Kitchens Unseen split and the Something-Else dataset the Omnivore model exhibits strong performance ($\sim$5.5% improvement on top of the baseline on Something-Else), the multimodal distillation approach achieves even higher performance. Nevertheless, we conclude that both approaches represent viable options for training multimodal models which leverage unimodal inputs during inference.

4.1.4 Effect on Model Calibration

Next to the improvements in action recognition performance we observe, we also investigate how the distilled student fares against the baseline in terms of model calibration. That is, whether the probability score of the predicted class reflects the accuracy of predicting the said class [23, 41]. More importantly, we verify whether distilling from additional modalities yields better calibrated students. In Fig. 3, we measure the Expected Calibration error [23] (ECE) on the Epic-Kitchens (regular & unseen), and the Something-Something and Something-Else datasets. We report the ECE for an RGB model trained using the ground truth labels, as well as all distilled students reported in Table 2 and Table 3. The general observation is that *across datasets, distillation improves the models' calibration*. Furthermore, we find that increasing the number of modalities used in the ensemble improves the model calibration.

[3]Girdhar *et al*. [20] show that there is no performance difference if the batches contain mixed data from different modalities, however, we found that using homogeneous batches yields better performance.

Objective	λ	γ	Noun@1	Verb@1	Action@1
$\mathcal{L}_{\text{CE}}$	0.0	—	52.0	61.7	38.3
$\mathcal{L}_{\text{KL}}$	1.0	30.0	51.7	65.4	39.3
$\mathcal{L}_{\text{CE}} \wedge \mathcal{L}_{\text{KL}}$	0.8	30.0	52.6	65.1	40.4
$\mathcal{L}_{\text{CE}} \wedge \mathcal{L}_{\text{KL}}$	0.8	3.0	53.0	66.9	41.0
$\mathcal{L}_{\text{KL}}$	1.0	1.0	53.1	65.5	40.5
$\mathcal{L}_{\text{CE}} \wedge \mathcal{L}_{\text{KL}}$	0.8	1.0	53.5	65.4	41.2
$\mathcal{L}_{\text{CE}} \wedge \mathcal{L}_{\text{KL}}$	0.8	0.33	53.6	64.7	40.5

Table 5: Ablation study on the Epic-Kitchens dataset. λ: Distillation and Cross-Entropy loss balancing term; γ: 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 in 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 (λ and γ) for (i) balancing between the ground truth and the distillation loss: λ (Equation 5), and (ii) controlling how the predictions of modality-specific models are combined in the ensemble: γ (Equation 3). We report results for Epic-Kitchens in Table 5, including the values for λ and γ 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 γ (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

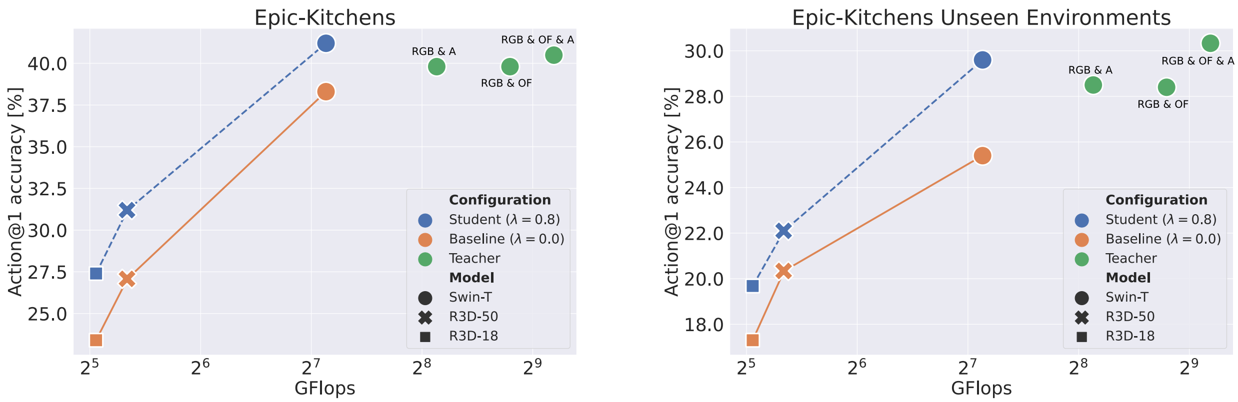

Figure 5: Top-1 action accuracy of Teacher, Student & Baseline models and their associated computational cost in giga-FLOPs (10^9) required to update the input and predict the action. Note: Top-left corner is optimal (i.e. faster and most accurate models).

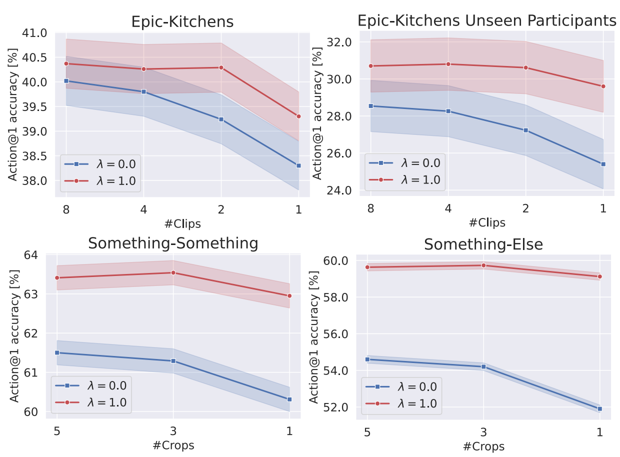

Figure 6: Performance degradation when reducing the number of inference clips/crops (Epic-Kitchens/Something-Something).

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 lowerig the normalization temperature γ 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 112×112. 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 256×456.

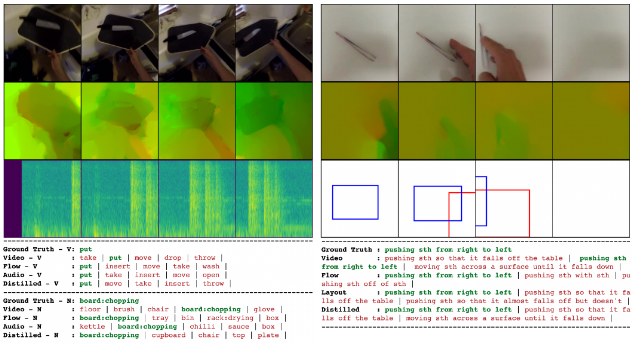

Figure 7: Qualitative example for the Epic Kitchens (Left) & the Something-Something (Right) datasets.

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 multi-modal ensemble, and low computational complexity of uni-modal 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 multi-modal 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

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021. 9

[2] Umar Asif, Jianbin Tang, and Stefan Harrer. Ensemble knowledge distillation for learning improved and efficient networks. *arXiv preprint arXiv:1909.08097*, 2019. 2

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 2

[4] Linchao Bao, Hailin Jin, Byungmoon Kim, and Qingxiong Yang. A comparison of tv-l1 optical flow solvers on gpu. *GTC Posters*, 6, 2014. 8

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 2

[6] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10925–10934, 2022. 2, 3, 5

[7] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. *Advances in Neural Information Processing Systems*, 34:19594–19607, 2021. 9

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 4, 13

[9] Xianing Chen, Qiong Cao, Yujie Zhong, Jing Zhang, Shenghua Gao, and Dacheng Tao. Dearkd: Data-efficient early knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12052–12062, 2022. 2

[10] Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019. 2

[11] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7882–7891, 2019. 2

[12] Yong Dai, Duyu Tang, Liangxin Liu, Minghuan Tan, Cong Zhou, Jingquan Wang, Zhangyin Feng, Fan Zhang, Xueyu Hu, and Shuming Shi. One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code. *arXiv preprint arXiv:2205.06126*, 2022. 2

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 4, 13

[14] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020. 4, 5, 6, 13

[15] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6252–6261, 2019. 1

[16] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *European Conference on Computer Vision*, pages 214–229. Springer, 2020. 1, 3

[17] Nuno C Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. Dmcl: Distillation multiple choice learning for multimodal action recognition. *arXiv preprint arXiv:1912.10982*, 2019. 2, 15

[18] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 2, 15

[19] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. *arXiv preprint arXiv:2206.08356*, 2022. 2, 5, 6

[20] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 2, 3, 5, 6, 7, 15

[21] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 4, 5, 13

[22] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 9

[23] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 5, 7

[24] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016. 2

[25] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition, pages 3148–3159, 2022. 1, 2, 4, 13, 14

[26] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2, 3, 4

[27] Hirokatsu Kataoka, Tenga Wakamiya, Kensho Hara, and Yutaka Satoh. Would mega-scale datasets further enhance spatiotemporal 3d cnns? *arXiv preprint arXiv:2004.04968*, 2020. 5, 8, 14

[28] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. *arXiv preprint arXiv:2111.01024*, 2021. 1, 3, 4

[29] Tae Soo Kim, Jonathan Jones, and Gregory D Hager. Motion guided attention fusion to recognize interactions from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13076–13086, 2021. 1, 2

[30] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 9

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 4, 5, 6, 9, 14

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[33] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1049–1059, 2020. 1, 2, 4, 5, 6, 13, 14

[34] Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *International Conference on Machine Learning*, pages 7632–7642. PMLR, 2021. 4

[35] Asit Mishra and Debbie Marr. Apprentice: Using knowledge distillation techniques to improve low-precision network accuracy. *arXiv preprint arXiv:1711.05852*, 2017. 2

[36] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 1, 2, 3

[37] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015. 2, 15

[38] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*, 2019. 14

[39] Srinivas Parthasarathy and Shiva Sundaram. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 400–404, 2020. 2

[40] Javier Sánchez Pérez, Enric Meinhardt-Llopis, and Gabriele Facciolo. Tv-l1 optical flow estimation. *Image Processing On Line*, 2013:137–150, 2013. 8

[41] Teodora Popordanoska, Raphael Sayer, and Matthew B Blaschko. A consistent and differentiable l_p canonical calibration error estimator. In *Advances in Neural Information Processing Systems*, 2022. 7

[42] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Students taught by multimodal teachers are superior action recognizers. *arXiv preprint arXiv:2210.04331*, 2022. 3

[43] Gorjan Radevski, Marie-Francine Moens, and Tinne Tuytelaars. Revisiting spatio-temporal layouts for compositional action recognition. *arXiv preprint arXiv:2111.01936*, 2021. 1, 2, 3, 4, 5, 13, 14

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2, 13, 14

[45] Axel-Jan Rousseau, Thijs Becker, Jeroen Bertels, Matthew B Blaschko, and Dirk Valkenborg. Post training uncertainty calibration of deep networks for medical image segmentation. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1052–1056. IEEE, 2021. 5

[46] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 4, 14

[47] Zhiqiang Shen and Marios Savvides. Meal v2: Boosting vanilla resnet-50 to 80%+ top-1 accuracy on imagenet without tricks. *arXiv preprint arXiv:2009.08453*, 2020. 2

[48] Alexandros Stergiou and Dima Damen. Play it back: Iterative attention for audio recognition. *arXiv preprint arXiv:2210.11328*, 2022. 14

[49] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3d: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 625–634, 2020. 2

[50] Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. Egodistill: Egocentric head motion distillation for efficient video understanding. *arXiv preprint arXiv:2301.02217*, 2023. 2

[51] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 8

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

data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 2

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 14

[54] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 4

[55] Lin Wang and Kuk-Jin Yoon. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[56] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*, pages 399–417, 2018. 1, 2, 4

[57] Xuehan Xiong, Anurag Arnab, Arsha Nagrani, and Cordelia Schmid. M&m mix: A multimodal multiview transformer ensemble. *arXiv preprint arXiv:2206.09852*, 2022. 1, 3, 4, 5

[58] Zihui Xue, Sucheng Ren, Zhengqi Gao, and Hang Zhao. Multimodal knowledge expansion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 854–863, 2021. 2

[59] Rui Yan, Lingxi Xie, Xiangbo Shu, and Jinhui Tang. Interactive fusion of multi-level features for compositional activity recognition. *arXiv preprint arXiv:2012.05689*, 2020. 1

[60] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 4, 8

[61] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Is an object-centric video representation beneficial for transfer? *arXiv preprint arXiv:2207.10075*, 2022. 1, 2, 4

[62] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2608–2618, 2021. 2

[63] Yipin Zhou and Tamara L Berg. Temporal perception and prediction in ego-centric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4498–4506, 2015. 1