Edinburgh Research Explorer

An Action Is Worth Multiple Words: Handling Ambiguity in Action Recognition

Link:
[Link to publication record in Edinburgh Research Explorer](#)

Document Version:
Peer reviewed version

An Action Is Worth Multiple Words: Handling Ambiguity in Action Recognition

Kiyoon Kim
kiyoon.kim@ed.ac.uk

Davide Moltisanti
davide.moltisanti@ed.ac.uk

Oisin Mac Aodha
oisin.macaodha@ed.ac.uk

Laura Sevilla-Lara
lsevilla@ed.ac.uk

School of Informatics
University of Edinburgh
Edinburgh, UK

Abstract

Precisely naming the action depicted in a video can be a challenging and oftentimes ambiguous task. In contrast to object instances represented as nouns (e.g. dog, cat, chair, etc.), in the case of actions, human annotators typically lack a consensus as to what constitutes a specific action (e.g. jogging versus running). In practice, a given video can contain multiple valid positive annotations for the same action. As a result, video datasets often contain significant levels of label noise and overlap between the atomic action classes. In this work, we address the challenge of training multi-label action recognition models from only single positive training labels. We propose two approaches that are based on generating pseudo training examples sampled from similar instances within the train set. Unlike other approaches that use model-derived pseudo-labels, our pseudo-labels come from human annotations and are selected based on feature similarity. To validate our approaches, we create a new evaluation benchmark by manually annotating a subset of EPIC-Kitchens-100's validation set with multiple verb labels. We present results on this new test set along with additional results on a new version of HMDB-51, called Confusing-HMDB-102, where we outperform existing methods in both cases. Data and code are available at https://github.com/kiyoon/verb_ambiguity

1 Introduction

Actions are often fundamentally ambiguous. Barker and Wright [2] explained that the hierarchical nature of an action is described differently depending on the viewpoint. In their example of "children going to school", one could consider the atomic body movement (*e.g.* "going down the stairs") to the goal and intention of the action (*e.g.* "walking to the school", "getting an education", *etc.*). This fundamental ambiguity and complexity in actions seems to influence the way humans learn verbs. For many languages, it has been observed that infants tend to learn nouns before verbs [58]. Furthermore, according to WordNet [21], the number of unique nouns in the English vocabulary is more than ten times that of verbs, allowing for a more precise categorization. Similarly, researchers also face difficulties in the computational setting when naming, categorising and learning actions in videos [39].
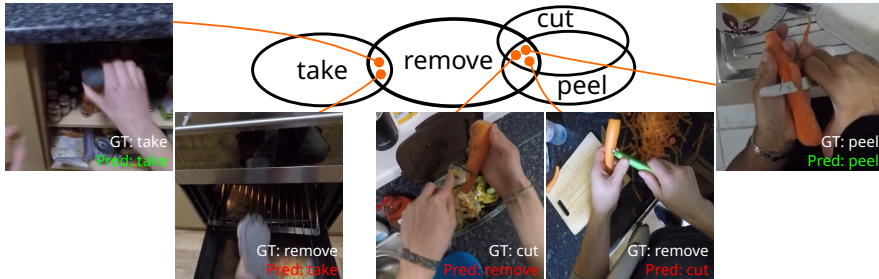
Figure 1: Ambiguity caused by single-verb labels. The same action can be described in different ways, *e.g.* the act of peeling a carrot can be annotated as "cut", "peel" or "remove" (skin). Large action recognition datasets gather several instances with different but equivalent verb labels. Here we show some examples from the EPIC-Kitchens dataset [5]. Recognition models can be confused by this semantic ambiguity and are usually not evaluated taking this verb overlap into account. Reasonable and semantically correct predictions can be counted as incorrect (*e.g.* red text, middle frames) by standard evaluation protocols as the predicted label does not match the noisy ground truth.

Designing an action label space for a given dataset where each video is assigned a single unique label is an extremely difficult task. For example, consider the verbs "stir" and "mix" as labels for a cooking video. You can stir to mix the ingredients, but stirring is not the only way to mix them (*e.g.* you can also "shake"). Since each verb can often have multiple meanings, the label space can partially overlap. Specifically, one label can represent multiple types of actions, and many labels can represent a single type of action, as in Figure 1. This is particularly difficult when datasets are large, and so is the label space, making it a combinatorial problem of finding overlapping classes for every instance. Labelling datasets in a thorough multi-label fashion can be prohibitively expensive. As a result, we typically only get partially-labelled video datasets, which negatively impacts both training and testing.

To address this, we propose a new training procedure to handle the problem of missing and ambiguous labels in action recognition. In particular, we generate pseudo-labels by searching neighbouring video instances in the feature space of the training set and analyse how similar videos are labelled differently. We take the commonly-agreed labels to be pseudo-labels (*i.e.* labels that are highly likely to be positive and thus relevant for the query video) and train action recognition models to either not penalise the potentially-positive classes or boost the learning signal for them. In order to evaluate our proposed methods we manually annotated a large subset of the validation set of the EPIC-Kitchens-100 dataset [5] with multiple verbs. We also evaluate our methods on a modified version of the popular HMDB-51 dataset [16], where we added synthetic ambiguity. Overall, our methods improve multi-label accuracy up to 18% over several recent baselines. To summarise:

- We show that the ambiguity in current action recognition labelling leads to issues in training and testing models. We show "failure cases" on challenging datasets where the model predicts reasonable labels but the accuracy is penalised because of label ambiguity.
- To mitigate the issue, we introduce two novel approaches for training multi-label action recognition models when only single positive labels are available. Our approaches attempt to disambiguate the label space by generating pseudo-labels from similar instances that are labelled differently.

- We introduce a multi-label subset of the EPIC-Kitchens-100 validation set for the purpose of multi-label evaluation. On this dataset, and another with synthetically generated label ambiguity, we report results that outperform existing methods.

2 Related Work

Action recognition. Action recognition refers to a video classification problem where the goal is to predict the action depicted in a video. Many different approaches have been proposed. For example, 3D convolutions [3, 53] have been widely used to capture spatio-temporal information, variants of which include using relation networks [56] and capturing slow and fast motion in separate networks [9]. A recent trend is to use transformer-based architectures [20]. For efficiency, 2D networks have been a popular choice, and this includes averaging the predictions from 2D image classifiers [55], using relation information [45] or temporal shift modules [18] to share different frame features from the 2D backbone.

Noisy labels and images. [10] provides a summary of methods that learn from noisy labels. Label smoothing [51] is one popular and simple method to prevent models from being overly confident by converting hard labels to soft ones. Progressive Self Label Correction [37] is a label correction technique that mixes model predictions with labels. This method assigns a larger weight to the prediction of a given class if the model still outputs high confidence for a class after some initial training. Other methods represent uncertainty with multiple hypotheses [27], use contrastive learning [24], correct the loss by estimating probabilities of one class being flipped to another [25], or optimise network parameters and labels alternatively [32]. These methods are commonly evaluated on image datasets with synthetic noise or on noisy image datasets collected from the web. In the latter case, noise derives from incorrect annotations or imperfect data collection. In our setting, noise stems from semantic ambiguity, *i.e.* multiple equivalent labels can apply to the same action, which involves arguably more complicated dynamics compared to noise caused by labelling omissions.

Noisy labels and videos. Video datasets are generally noisy. [41] proposed a Q-learning method to mitigate the problem. [28] introduced a video dataset where noise in the labels comes from inaccurate parsing of web tags. [12] explored pre-training methods using large-scale video datasets with label noise (*i.e.* missing or incorrect labels) and temporal segmentation noise. [22] studied temporal ambiguity in action labelling showing that the perception of temporal bounds is highly subjective, which in turns leads to recognition robustness issues. Multiple instance learning has also been used to tackle noisy video datasets, where noise comes from less reliable annotations or sparse labelling [1, 17]. However, none of these methods specifically address the noise generated by the ambiguous nature of verbs. This type of noise exhibits different distributions compared to noise stemming from the other problems discussed above. This is because the meaning of a verb varies depending on the video context, and as a consequence there is only a partial overlap across different labels, *i.e.* labels are not exactly interchangeable across all videos.

Visual and semantic ambiguity. [40] proposed a method to learn actions from visually similar instances with different semantic labels. Their framework builds a graph where nodes are connected if they are visually similar or semantically related. A Markov walk through this graph estimates the action label of a new video. This work is very close in spirit to ours, however experiments showed that it requires fine-grained verb meaning labels in order to achieve good results. On the contrary, we do not require further annotations and only use the available single-label ground truth. Furthermore, we release multi-label annotations for

better testing of our methods. [39] identified the issues arising from single-verb labels due to semantic ambiguity. They proposed training with multi-verb representations, which were obtained by asking annotators to label one representative sample per action with all verbs they deemed relevant. Verbs collected for a given action sample were then attached to the other instances of the action. This work is also closely related to ours, however we stress again that we only require single-label ground truth for training. We also show in our experiments that instance-based multi-verb representations better alleviate semantic ambiguity compared to global dataset-based representations. Visual Sense Disambiguation [11] aims to identify the context of an action given an image and a verb. Understanding visual and semantic relationships in the label space has also been shown to improve multi-label image classification [7]. [44] use language priors in the form of word embeddings to find relationships between verbs in human-object interaction data. However, our preliminary experiments with word embeddings failed to produce sensible similarity estimates for the finer-grained datasets we use in this work. [14] address the problem of missing label imputation for scene graph generation proposing an iterative pseudo label approach.

Single positive multi-label learning. There has been recent interest in the problem of multi-label learning from only single positive labels. The problem is outlined in [4], where the authors propose a loss that encourages models to predict both known and unlabelled positives. The loss also adds a regularisation term to match the expected number of positives per image. [46] learns from the annotated positive labels using an entropy maximisation approach and decreases the loss penalty for unknown labels. Their method also treats samples with the lowest confidence as pseudo negatives. [34] outlined an image-specific spatial consistency loss that measures consistency in model predictions over multiple random crops from the input image. A similar problem of multi-label learning from missing labels has been tackled by leveraging the estimated distribution of missing and known labels [8, 43].

Semi-supervised learning. Semi-supervised learning approaches tackle the problem of having partially annotated labels, which translates to having noisy label estimates. This problem considers partially annotated instances in single-label classification tasks and does not entail multi-label classification. In the context of image datasets, [13] showed that self-supervised contrastive learning can result in representations that are more robust to label noise. Contrastive learning has also been explored for action recognition by using only a small subset of the labels [29]. Recently, an uncertainty-aware pseudo-label selection framework was proposed for both image and video [26]. Pseudo-labelling was also used to group potentially similar classes in [23]. The major benefit of pseudo-labelling over contrastive learning is that it does not rely on domain-specific data augmentation.

3 Problem and Methodology

As outlined above, in contrast to nouns, verb annotations are prone to semantic ambiguity. Such ambiguity stems from two issues: i) verbs have different meanings, thus a single verb can be attached to visually different actions and ii) individuals can often use different verbs to annotate the same action. Figure 1 shows an example of semantic ambiguity, where the act of peeling a carrot in EPIC Kitchens [5] is labelled as "peel", "cut", and "remove" (skin). This is a common issue in large datasets where multiple annotators label videos and thus it is hard to enforce consistent classes. While noun labels alleviate semantic ambiguity, they add to the annotation burden and make learning actions a fragmented process. In fact, as noted in [39, 40], nouns are mostly used to facilitate learning, however they unnaturally split the same

action into different classes. For example, "open-door" and "open-drawer" are the same action, but belong to separate classes when categories are indicated by both verbs and nouns. Annotating actions with multiple verbs [39] also mitigates ambiguity, however it comes with the cost of choosing multiple labels for each video. Models are typically not designed, nor evaluated, taking this semantic ambiguity into account. We address the semantic ambiguity introduced by single-verb labels with two methods which generate multi-verb pseudo-labels from single-verb annotations during training. We also evaluate our methods across several metrics better suited for datasets with semantic ambiguity. We next formalise the problem and review common baselines, before introducing our methods.

Formalisation. We consider our problem to be related to that of single positive multi-label learning (SPML) [4, 46]. Here, we have a training and validation set where instances have only one positive label. Other categories are unknown and could be either positive or negative (*i.e.* present or absent). For evaluation, instances in the test set have multiple labels. Formally, let $\Gamma = \{1, \ldots, C\}$ be the set of classes in a given dataset, where C is the number of classes. Each training/validation video is annotated with a single positive label $y_i \in \Gamma$. Remaining classes in $\Gamma \setminus \{y_i\}$ are not annotated and cannot be assumed to be negative. A test video is annotated with a positive label set $\mathcal{Y}_i \subset \Gamma$ which can contain multiple classes, and remaining the classes $\Gamma \setminus \mathcal{Y}_i$ are treated as confirmed negatives.

SPML. One of the most common approaches in SPML is the "assume negative" loss (AN) [4]. This treats all unknown labels as negative ones and uses the binary cross-entropy (BCE) loss for multi-label learning. Given a training video-label pair $(\mathbf{x}_i, y_i)$:

$$\mathcal{L}_{\text{AN}}(\mathbf{x}_i, y_i) = -\frac{1}{C} \sum_{c=1}^{C} \left[\mathbb{1}_{[y_i=c]} \log\left(f^{(c)}(\mathbf{x}_i)\right) + \mathbb{1}_{[y_i \neq c]} \log\left(1 - f^{(c)}(\mathbf{x}_i)\right) \right] \tag{1}$$

where $\mathbb{1}_{[.]}$ is the indicator function and $f^{(c)}(\mathbf{x}_i)$ is the model prediction for class c. To address the imbalance between the known positive labels and the unknown assumed-negative labels, an alternative version of AN is proposed: the "weak assume negative" loss (WAN) [4]. This improves over AN in that it gives equal weight for all assumed negatives and the single positive, however all unknown labels are still treated as negatives. Label smoothing (LS) [31] is another popular approach to prevent models from being overly confident. A variant of LS only for assumed negative labels (N-LS) was also proposed in [46]. This improves over LS in that it only alters unknown (assumed negative) labels. The focal loss [19], originally designed for object detection, has been used for SPML as well. This works well in the presence of imbalanced labels, however it requires additional parameter tuning to achieve optimal performance. Finally, the entropy maximisation (EM) loss [46] allows the unknown labels to be unknown rather than assuming them to be negatives. EM learns mainly from annotated labels and attains state-of-the-art results in SPML on *image* datasets. We provide details for each of these losses in the supplementary material and compare our methods against these losses in Section 4.1.

3.1 Mask and Pseudo+Single-label BCE

The above methods were mostly designed for partial labelled setting, where instances potentially contain multiple classes but *classes do not semantically overlap*. In this setting there
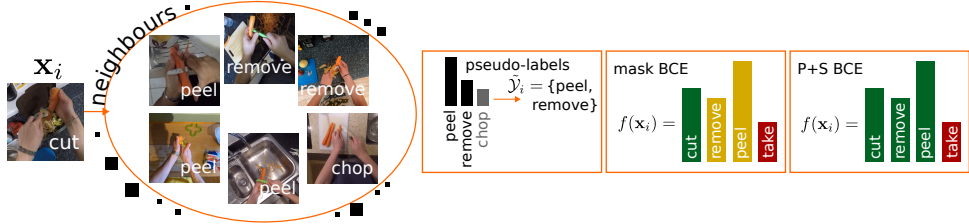
Figure 2: Summary of our approach. Given a training video $\mathbf{x}_i$ we first retrieve its nearest neighbours in the feature space. Looking at the frequency of the neighbours' classes we then obtain the pseudo-labels set $\tilde{\mathcal{Y}}_i$. Pseudo-labels are used to optimise the model given its predictions $f(\mathbf{x}_i)$. The model is forced to predict the ground truth class (green bar) and penalised when predicting negative classes with high confidence (red bar). When using our Mask BCE loss, the model predictions for the pseudo-labels (yellow bars) are masked out during back-propagation, thus the model is neither penalised nor rewarded for guessing reasonable classes. When using our P+S BCE loss, pseudo-labels are assumed to be positive classes, thus the model is forced to predict both the ground truth and the pseudo-labels.

are multiple "things" present in an instance and the task is to label all of them, using only the available single-label annotations at training time. In our case, however, instances can belong to multiple classes *because classes can semantically overlap*. We thus treat the unknown labels such that the model does not get penalised if it predicts an unknown label with high likelihood. Specifically, we pick classes that are likely to be positive based on feature similarity. We treat these as pseudo-labels. The model learns from single-verb ground truth annotations, however it is not penalised if it predicts a pseudo-label with confidence. Classes that are neither ground truth nor pseudo-labels are assumed negative.

To obtain pseudo-labels we extract features for each sample in the training set using a pre-trained action recognition backbone. Given a training pair $(\mathbf{x}_i, y_i)$ we take K neighbouring videos in the feature space. These videos are likely to share the same label as $\mathbf{x}_i$, however we also expect these videos to belong to different classes given the semantically ambiguous setting we work in. We build pseudo-labels for $\mathbf{x}_i$ by looking at the label frequency of its K neighbours. Formally, let Ω be the sequence of labels found among the K neighbours, and let Υ be the set of unique labels in Ω. For each $y_j \in \Upsilon$ we count its frequency in Ω, *i.e.* $\omega(y_j) = |(\eta \in \Omega : \eta = y_j)|/K$. The pseudo-label set $\tilde{\mathcal{Y}}_i$ is finally obtained from Υ by applying a threshold τ to $\omega(y_j)$, *i.e.* $\tilde{\mathcal{Y}}_i = \{y_j \in \Upsilon : \omega(y_j) > \tau, y_j \neq y_i\}$. Note that we do not add the original label to the pseudo-labels. Unlike other methods [37, 46] which generate pseudo-labels from model predictions, we use actual human annotations to produce pseudo-labels. With this approach, the ambiguous label space is realistically well-represented. Clearly we rely on selecting informative neighbours, but we find empirically that the backbone we use is sufficient for this task.

With the above defined, we now propose our **Mask BCE** loss. This loss treats the single-label ground truth as the only positive, pseudo-labels as unknown classes, and assumes the remaining classes are negative:

$$\mathcal{L}_{\text{mask}}(\mathbf{x}_i, y_i) = -\frac{1}{C - |\tilde{\mathcal{Y}}_i|} \sum_{\substack{c=1 \\ c \notin \tilde{\mathcal{Y}}_i}}^{C} \left[\mathbb{1}_{[y_i=c]} \log\left(f^{(c)}(\mathbf{x}_i)\right) + \mathbb{1}_{[y_i \neq c]} \log\left(1 - f^{(c)}(\mathbf{x}_i)\right) \right] \quad (2)$$

Figure 3: Samples from our EPIC-Kitchens-100-SPMV dataset. The yellow text indicates the original label from EPIC-Kitchens-100 [6], which is also included in our annotations. The white text shows the additional labels we gathered from multiple annotators.

Here, the output of the model $f^{(c)}(\mathbf{x}_i)$ for classes $c \in \tilde{\mathcal{Y}}_i$ is detached from the gradient computation and back-propagation of the loss, *i.e.* pseudo-label classes are frozen and do not provide any learning signal. This does not penalise the model if it produces reasonable predictions for relevant classes, *i.e.* if it outputs a high likelihood for a class that is potentially an equivalent label for the video. Importantly, this loss does not send incorrect positive learning signals when the pseudo-labels are not accurate. Instead, when pseudo-labels can be assumed to be correct, treating pseudo-labels as positive would be a better strategy.

With this in mind, we propose an alternative loss named **Pseudo+Single-label BCE**:

$$
\begin{aligned}
\mathcal{L}_{\text{P+S}}(\mathbf{x}_i, y_i) = -\frac{1}{C} \sum_{c=1}^{C} \Big[& \mathbb{1}_{[y_i=c]} \log\left(f^{(c)}(\mathbf{x}_i)\right) + \mathbb{1}_{[c\in\tilde{\mathcal{Y}}_i]} \log\left(f^{(c)}(\mathbf{x}_i)\right) \\
& + \mathbb{1}_{[y_i\neq c]} \mathbb{1}_{[c\notin\tilde{\mathcal{Y}}_i]} \log\left(1 - f^{(c)}(\mathbf{x}_i)\right) \Big]
\end{aligned}
\tag{3}
$$

Intuitively, $\mathcal{L}_{\text{P+S}}$ will work better than $\mathcal{L}_{\text{mask}}$ when the pseudo-labels are accurate, but it will hurt performance when they are not. Figure 2 summarises our pseudo-label generation and the two losses described above.

4 Experiments

Datasets. We manually annotated 40% of the EPIC-Kitchens-100 [6] validation set with multiple verbs, using the procedure described in the supplementary material. This subset is our test set and consists of 3,493 multi-verb videos with 2.4 labels per video on average. The remaining videos in the original validation set form our own validation set. The training
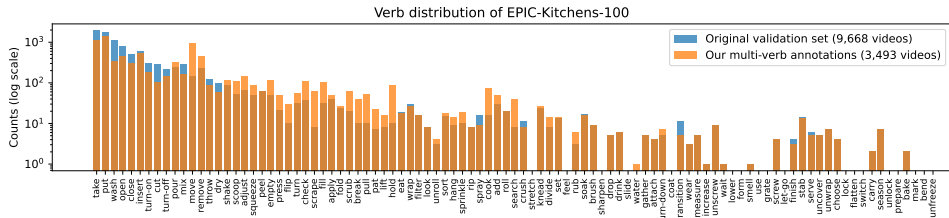
Figure 4: EPIC-Kitchens-100 verb distribution from our multi-label annotations (orange) compared to the original labels (blue). Although we only annotated 40% of the validation dataset, we obtained similar or even more labels per class. Note that some generic verbs (*e.g.* move, hold, search) gained a significant number of annotations compared to the original.

set is the same as the original one. Figure 3 illustrates a few examples from our dataset, while Figure 4 shows the verb distribution from our multi-label annotations and the original labels. Our annotations share a very similar distribution compared to the original ones, thus no bias towards specific classes is introduced with our labels. We call this dataset *EPIC-Kitchens-100-SPMV* (Single Positive Multi-Verb learning). We also constructed a dataset from the popular HMDB-51 dataset [16] with synthetic labels in order to simulate ambiguous annotations. We doubled the original 51 classes to 102 classes, where each class c is split into c^0 and c^1. Given a training video $\mathbf{x}_i^{train}$ originally belonging to c, we randomly assign it to *either* c^0 or c^1. Given a test video $\mathbf{x}_i^{test}$ originally belonging to c, we assign it to *both* c^0 and c^1. This simulates verb ambiguity, where actions are represented with multiple verbs. We name this dataset *Confusing-HMDB-102*.

Metrics. We evaluate *Top-Set Multi-label Accuracy* following [39]. This measures how many top predictions match the ground truth: $A_{\text{Top-set\_ML}} = \frac{1}{N} \sum_{i=1}^{N} |\mathcal{Y}_i \cap \hat{\mathcal{Y}}_i| / |\mathcal{Y}_i|$, where N is the total number of videos, $\mathcal{Y}_i$ is the multi-label ground truth, and $\hat{\mathcal{Y}}_i$ is the top-K predicted classes, where K is the number of ground-truth labels. We also report another metric called *Top-1 Multi-label Accuracy*: $A_{\text{Top1\_ML}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{[\hat{y}_i \in \mathcal{Y}_i]}$, where $\hat{y}_i$ is the top-1 predicted class. This metric is more relaxed and does not penalise the model when it predicts a relevant verb, regardless of the specific choice. For example, when the ground-truth is ("stir", "mix"), the model is free to choose either "stir" or "mix", and not predicting the other one will not lower this accuracy. This is different from Top-Set Multi-label Accuracy, where predicting only one label in this case would give 50% accuracy, and the model would have to predict both "stir" and "mix" to obtain 100%. We also evaluate *IOU Accuracy* [30], which is the intersection over union between the ground truth and predicted labels, and F_1-*Measure*, which is the harmonic mean of precision and recall. We set the confidence threshold to 0.5, so any predictions over 0.5 will be treated as positive labels. Finally, we report mean average precision *mAP*. This is a popular metric for detection tasks. However, mAP is not very suitable for long-tail datasets such as EPIC-Kitchens, since class scores are averaged with equal weight. We provide a formal definition of IOU Accuracy and F_1 in the supplementary material. For EPIC-Kitchens-100-SPMV, we choose the best model by looking at the standard top-1 accuracy on our single-label validation set, reporting results obtained on five different runs with different seeds. For Confusing-HMDB-102 we take the best score per metric and take the average over the official three splits.

Implementation details. We pre-extracted RGB and optical flow features using TSM [18] pre-trained on each dataset's training set using a standard cross entropy loss with the single

positive labels. TSM shares information across frames and has been shown to perform well on multiple action-orientated datasets. The concatenated RGB and optical flow features form the input to our model, which is a multi-layer perceptron (MLP) with three layers and hidden dimension of 1024. We train the MLP using the Adam optimiser [15] with a batch size of 64 and learning rate of 5e-6. We stopped training whenever the validation accuracy did not improve for 20 epochs. For EPIC-Kitchens-100-SPMV all experiments took less than 130 epochs to saturate. For Confusing-HMDB-102 it took longer, less than 390 epochs, given the much smaller size. Pseudo-labels are generated only once before training. We set the number of neighbours $K = 15$ and use a threshold of $\tau = 0.1$ to obtain pseudo-labels.

4.1 Results

Table 1 compares our methods to multiple recent SPML baselines introduced in Section 3. We observe that our methods outperform all baselines across all metrics except mAP on EPIC-Kitchens-100-SPMV. As mentioned above, this metric is not well suited for imbalanced datasets. We note that in most cases the P+S loss outperforms the Mask loss. Given that the P+S loss treats pseudo-labels as positives, this demonstrates that our method for obtaining pseudo-labels is able to find good matches. We observe a clear boost especially in IOU and F_1-measure, which are well suited metrics for our multi-label setting. This indicates that the SPML baselines suffer from being penalised for correct (*i.e.* related) classes. This also possibly suggests that SPML methods designed for partial-label settings struggle to work well when the source of the multi-label noise originates from semantic ambiguity.

Interestingly, the EM [46] loss, despite working well on static images, shows the worst IOU accuracy and F_1-measure. We hypothesise this is because the loss does not treat unknown labels as negative. This results in having no negative signal, which in turns gives overly confident predictions for most classes. Looking deeper into the model's output on EPIC-Kitchens-100-SPMV, the EM loss predicts 51.5 positive labels on average per video, whereas our mask method predicts on average 1.2 and P+S method predicts on average 1.8, with a confidence threshold of 0.5. Considering that our test videos have 2.4 positive labels on average, it is evident that the EM loss predicts *too many* false positives. Our methods instead effectively push confidence up only for the relevant labels, keeping confidence for irrelevant labels low.

Table 2 (bottom) reports an ablation study on Confusing-HMDB-102 with an ideal label search, *i.e.* the retrieved pseudo-labels are intentionally chosen as the correct classes. Results show that it is theoretically possible to improve performance with perfect pseudo-labels, however we do not observe a dramatic improvement compared to the labels obtained through our visual neighbour search. Table 2 (top) illustrates another ablation experiment, where we obtain pseudo-labels from class co-occurrences retrieved on our EPIC-Kitchens-100-SPMV test set. In this case, pseudo-labels are formed globally by looking at the co-occurrence of our multi-verb annotations across the test set. Given a verb v_i we count how many times it was annotated together with any other verb v_j. If v_i and v_j occur at least half of the time together v_j will form the pseudo-label set for instances labelled with v_i. The large performance drop observed when using these co-occurrence-based global pseudo-labels suggests that the semantic ambiguity in our setting requires *instance-level* disambiguation rather than *class-level* pseudo-labels. This is because verbs can have subtly different meaning which can change according to the context. Clustering verbs using co-occurrences ignores this signal, and thus hurts performance.

Dataset	Loss	Top-set ML	Top-1 ML	IOU Acc.	F_1	mAP
EPIC-Kitchens-100-SPMV	AN	43.7 ± 0.5	51.0 ± 0.4	11.2 ± 0.4	15.0 ± 0.6	22.8 ± 1.8
	WAN	44.8 ± 0.3	52.5 ± 0.6	15.2 ± 5.0	24.6 ± 6.6	**26.3 ± 1.3**
	LS	43.7 ± 0.9	51.8 ± 0.7	9.6 ± 0.4	12.9 ± 0.6	24.4 ± 1.5
	N-LS	43.4 ± 0.6	50.7 ± 0.4	11.0 ± 0.4	14.8 ± 0.5	22.1 ± 0.8
	Focal	43.3 ± 0.6	51.5 ± 0.3	5.7 ± 0.2	7.8 ± 0.2	23.9 ± 0.6
	EM	44.7 ± 0.4	52.9 ± 0.3	4.1 ± 0.0	7.8 ± 0.0	25.1 ± 1.0
	Mask (ours)	<u>46.6 ± 0.2</u>	<u>55.2 ± 0.4</u>	<u>27.8 ± 0.4</u>	<u>36.9 ± 0.4</u>	<u>25.9 ± 0.8</u>
	P+S (ours)	**46.9 ± 0.1**	**56.0 ± 0.6**	**33.5 ± 0.2**	**44.9 ± 0.3**	25.8 ± 0.8
Confusing-HMDB-102	AN	32.0 ± 1.6	38.5 ± 2.0	18.9 ± 1.6	24.8 ± 1.6	20.6 ± 3.8
	WAN	36.8 ± 0.7	40.7 ± 0.6	4.1 ± 0.1	7.9 ± 0.1	32.0 ± 1.4
	LS	32.4 ± 2.3	38.8 ± 1.8	19.3 ± 2.3	24.9 ± 2.3	19.7 ± 3.8
	N-LS	32.2 ± 1.3	38.8 ± 1.7	19.3 ± 1.7	25.3 ± 1.8	20.1 ± 3.7
	Focal	31.6 ± 1.9	38.0 ± 2.0	13.0 ± 0.8	17.6 ± 0.7	14.8 ± 3.8
	EM	31.9 ± 0.6	37.4 ± 0.9	3.2 ± 0.0	6.2 ± 0.1	18.6 ± 3.1
	Mask (ours)	<u>41.8 ± 1.1</u>	<u>43.3 ± 0.9</u>	**30.8 ± 2.5**	**36.3 ± 2.7**	<u>40.3 ± 2.2</u>
	P+S (ours)	**41.9 ± 0.9**	**43.4 ± 0.5**	<u>29.9 ± 2.2</u>	<u>35.9 ± 2.0</u>	**40.7 ± 2.0**

Table 1: Results with ± standard deviation calculated over five runs for EPIC-Kitchens-100-SPMV and three splits for Confusing-HMDB-102. For EPIC-Kitchens-100-SPMV the mAP metric shows unstable results due to the very long-tailed distribution of the verbs. Best and second-best results are bolded and underlined respectively.

Dataset	Loss	Top-set ML	Top-1 ML	IOU Acc.	F_1	mAP
EPIC-Kitchens-100-SPMV	Mask	<u>46.6 ± 0.2</u>	<u>55.2 ± 0.4</u>	<u>27.8 ± 0.4</u>	<u>36.9 ± 0.4</u>	**25.9 ± 0.8**
	Mask[†]	37.8 ± 1.3	48.4 ± 1.0	18.1 ± 0.4	23.7 ± 0.4	21.7 ± 0.9
	P+S	**46.9 ± 0.1**	**56.0 ± 0.6**	**33.5 ± 0.2**	**44.9 ± 0.3**	<u>25.8 ± 0.8</u>
	P+S[†]	23.0 ± 0.4	28.0 ± 0.5	12.4 ± 0.4	18.0 ± 0.6	20.8 ± 1.3
Confusing-HMDB-102	Mask	41.8 ± 1.1	43.3 ± 0.9	<u>30.8 ± 2.5</u>	**36.3 ± 2.7**	<u>40.3 ± 2.2</u>
	Mask*	<u>42.6 ± 2.2</u>	<u>43.8 ± 2.3</u>	30.4 ± 2.3	34.2 ± 2.4	37.4 ± 3.6
	P+S	41.9 ± 0.9	43.4 ± 0.5	29.9 ± 2.2	<u>35.9 ± 2.0</u>	**40.7 ± 2.0**
	P+S*	**43.2 ± 1.8**	**44.3 ± 2.1**	**31.4 ± 2.5**	35.9 ± 2.1	39.1 ± 3.1

Table 2: Ablation experiments using class-level pseudo labelling (†) or ideal pseudo-label search (∗). We report results with ± standard deviation calculated over five runs for EPIC-Kitchens-100-SPMV and three splits for Confusing-HMDB-102. Best and second-best results are bolded and underlined.

5   Conclusion

We addressed the problem of action label ambiguity at training time in video action recognition. Here, videos can contain multiple equivalent labels due to the semantic ambiguity of verbs. This is a distinct problem compared to most partial-label settings, where multiple actions are present but not all are annotated. We showed that state-of-the-art SPML approaches struggle in the presence of this semantic ambiguity. To address this, we proposed two pseudo-labelled-based losses: *Mask BCE*, which treats pseudo-labels neutrally and *Pseudo+Single-label BCE*, which trusts pseudo-labels and uses them as positive labels. Through results on one dataset with synthetically generated ambiguity and another consisting of a new set of annotations for EPIC-Kitchens, we show that our methods outperform current SPML methods. One interesting direction for future research is addressing the additional ambiguity resulting from temporally overlapping actions that are often present in dense video datasets.

Supplementary Material

A EPIC-Kitchens-100-SPMV annotations

Here we describe the annotation protocol we followed to compile our EPIC-Kitchens-100-SPMV dataset. Figure 5 shows the co-occurrence frequency of head classes. Figure 6 depicts our annotation interface.

Annotating difficult samples. We first trained a TSM [13] RGB model on EPIC-Kitchens-100 [6]. This model achieved 60.9% verb accuracy on the original validation set. We then annotated failure cases where the top-1 prediction was incorrect, which amounted to 39.1% of EPIC-Kitchens-100's validation set (3,782 videos).

Verb candidates. Annotators were presented with a video containing an action and were asked to choose any verb they saw fit to describe the action. Annotators could also choose no labels in case the video was noisy and did not show any action. Annotators could not see the original ground truth and could only choose from a limited list of verbs. We exploited the TSM model predictions to compile a list of verb candidates for each video. Specifically, we observed that in many cases the model was predicting reasonable verbs within its top predictions. To facilitate labelling we thus show annotators only the 10 verbs corresponding to the top 10 predictions of the model for a given video. We also included the ground truth single-verb label among the verb list (without telling annotators that this was the ground truth). This was to analyse the performance of the annotators: if they did not choose the ground truth verb in most cases then they likely did not pay enough attention to the video. Regardless of the chosen verbs we always included the original ground truth in the final multi-labels.

Multiple annotators. For robustness each video was annotated by at least three different annotators. In some cases annotators did not meet our quality requirements, *e.g.* they chose too few verbs on average per video or failed to choose the ground truth most of the time. In these cases we asked other annotators to label the same videos. In total we employed 26 annotators. People were students mostly from our computer science department.

Filtering noisy samples. For each video annotators were asked whether they were confident about their verb selection or not. We discarded samples where more than half annotators did not feel confident. Many of the discarded videos include actions happening outside the viewpoint, segments with bad temporal boundaries or where no action was happening. Altogether, we filtered out 289 videos from the initial 3,782 samples, compiling a set of 3,493 clean annotations.

Finalising labels. For a given video we keep labels that were chosen by at least half of the annotators. For example, for a video annotated by three people a verb would have to be chosen at least twice to be kept in the video multi-verb annotations. We also included the original single-label ground truth regardless of the annotators' choice.

Held-out validation set. Our 3,493 multi-verb annotations constitute our test set. The remaining $9,668 - 3,493 = 6,175$ videos from EPIC-Kitchens-100's validation set constitute our own validation set, where videos are annotated with only one verb.

Fleiss' kappa agreement score. We computed Fleiss' kappa to measure agreement. We report an average κ of 0.52, which is typically translated as moderate agreement. Labelling actions is highly subjective, thus we believe annotators showed a satisfactory agreement.
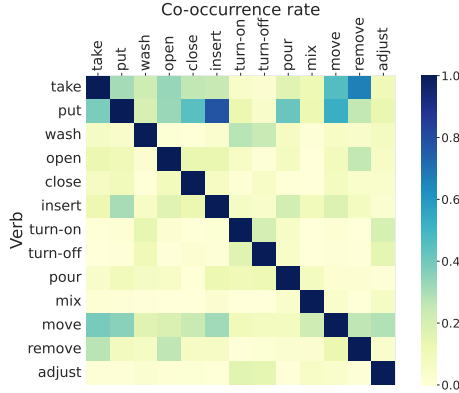
Figure 5: Co-occurrence ratio between verbs in our multi-label annotations. Rows are normalised by the number of classes, *e.g.* 1 indicates two verbs always co-occur (diagonal), 0.5 means they occur half of the time together, etc. "Take" and "remove" frequently co-occur, and so do "put" and "insert". Only head classes are visualised.

B SPML Losses

In this section we formally define all the SMPL losses introduced in the paper. For convenience, $f_i^{(c)} = f^{(c)}(\mathbf{x}_i)$ denotes the model prediction confidence for class c. Firstly, Weak Assume Negative (WAN) loss [4] has a negative balancing weight $\frac{1}{C}$ from the Assume Negative (AN) loss:

$$\mathcal{L}_{\text{WAN}}(\mathbf{x}_i, y_i) = -\frac{1}{C} \sum_{c=1}^{C} \left[\mathbb{1}_{[y_i=c]} \log\left(f^{(c)}(\mathbf{x}_i)\right) + \mathbb{1}_{[y_i \neq c]} \frac{1}{C-1} \log\left(1 - f^{(c)}(\mathbf{x}_i)\right) \right] \quad (4)$$

Label Smoothing (LS) loss [51] is defined as:

$$\mathcal{L}_{\text{LS}}(\mathbf{x}_i, y_i) = -\frac{1}{C} \sum_{c=1}^{C} \left[\mathbb{1}_{[y_i=c]}^{\frac{\varepsilon}{2}} \log\left(f^{(c)}(\mathbf{x}_i)\right) + \mathbb{1}_{[y_i \neq c]}^{\frac{\varepsilon}{2}} \log\left(1 - f^{(c)}(\mathbf{x}_i)\right) \right] \quad (5)$$

where $\mathbb{1}_{[Q]}^{\alpha} = (1-\alpha)\mathbb{1}_{[Q]} + \alpha\mathbb{1}_{[\neg Q]}$ for any logical preposition Q. The label smoothing only for assumed negatives (N-LS) [46] would be the same only for the assumed negatives:

$$\mathcal{L}_{\text{N-LS}}(\mathbf{x}_i, y_i) = -\frac{1}{C} \sum_{c=1}^{C} \Bigg[\mathbb{1}_{[y_i=c]} \log\left(f^{(c)}(\mathbf{x}_i)\right)$$
$$+ \mathbb{1}_{[y_i \neq c]} \left[(1-\varepsilon) \log\left(1 - f^{(c)}(\mathbf{x}_i)\right) + \varepsilon \log\left(f^{(c)}(\mathbf{x}_i)\right) \right] \Bigg] \quad (6)$$

We used $\varepsilon = 0.1$ throughout the paper. Focal loss [19] has a balancing parameter α and a focussing parameter γ.

$$\mathcal{L}_{\text{Focal}}(\mathbf{x}_i, y_i) = -\frac{1}{C} \sum_{c=1}^{C} \Bigg[\mathbb{1}_{[y_i=c]} \left[\alpha(1 - f^{(c)}(\mathbf{x}_i))^{\gamma} \log\left(f^{(c)}(\mathbf{x}_i)\right) \right]$$
$$+ \mathbb{1}_{[y_i \neq c]} \left[(1-\alpha)(f^{(c)}(\mathbf{x}_i))^{\gamma} \log\left(1 - f^{(c)}(\mathbf{x}_i)\right) \right] \Bigg] \quad (7)$$
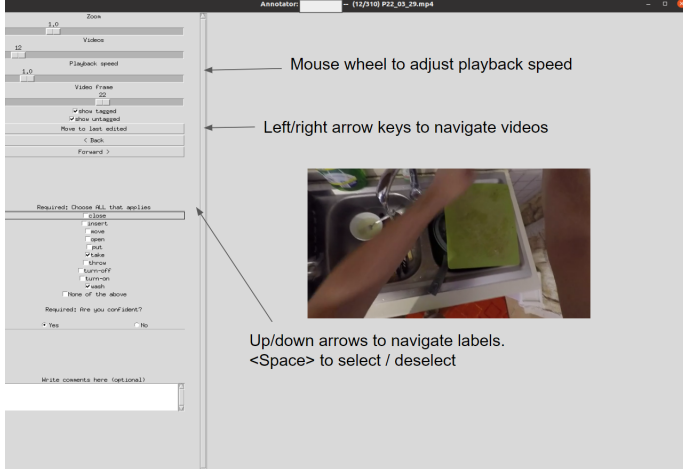
Figure 6: Our annotation interface.

We set $\alpha = 0.25$ and $\gamma = 2$. Finally, the Entropy Maximisation loss [46] is defined as follows:

$$\mathcal{L}_{\text{EM}}(\mathbf{x}_i, y_i) = -\frac{1}{C} \sum_{c=1}^{C} \left[\mathbb{1}_{[y_i=c]} \log \left(f^{(c)}(\mathbf{x}_i) \right) + \mathbb{1}_{[y_i \neq c]} \alpha H(f^{(c)}(\mathbf{x}_i)) \right] \tag{8}$$

$$H(f^{(c)}(\mathbf{x}_i)) = -\left[f^{(c)}(\mathbf{x}_i) \log \left(f^{(c)}(\mathbf{x}_i) \right) + (1 - f^{(c)}(\mathbf{x}_i)) \log \left(1 - f^{(c)}(\mathbf{x}_i) \right) \right] \tag{9}$$

C   Definitions of IOU Accuracy and F_1

Let N be the number of videos in the test set, $\mathcal{Y}_i$ be the set of ground truth, and $\hat{\mathcal{Y}}_i$ be the set of predicted classes with over 50% prediction confidence. The IOU accuracy and F_1-Measure are defined as follow:

IOU accuracy:

$$A_{\text{IOU}} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathcal{Y}_i \cap \hat{\mathcal{Y}}_i|}{|\mathcal{Y}_i \cup \hat{\mathcal{Y}}_i|} \tag{10}$$

F_1-Measure:

$$F_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{2|\mathcal{Y}_i \cap \hat{\mathcal{Y}}_i|}{|\mathcal{Y}_i| + |\hat{\mathcal{Y}}_i|} \tag{11}$$

D   Impact of K and τ

There are two hyperparameters involved in our methods: the number of neighbours (K) and the pseudo-label threshold (τ). Table 3 shows the impact of the different choices of K and τ. Results improve with a larger number of neighbours and a smaller τ, which entail a larger set of pseudo-labels. This shows the benefits of using more pseudo-labels to better alleviate semantic ambiguity.

(a) Mask BCE loss						(b) P+S BCE loss				
τ/K	3	5	10	15	20	τ/K	3	5	10	15	20
0.1	27.2	30.5	31.0	36.4	36.2	0.1	37.9	42.7	41.6	44.5	43.7
0.2	27.2	23.8	26.3	27.5	28.6	0.2	37.9	30.5	34.3	34.7	37.6
0.3	27.2	23.8	23.9	25.8	25.1	0.3	37.9	30.5	29.7	33.5	31.9

Table 3: F_1-Measure for our Mask and P+S BCE loss on EPIC-Kitchens-100-SPMV with different values of τ and K.

Loss	Top-set ML	Top-1 ML	IOU Acc.	F_1	mAP
AN	46.7± 1.5	49.5± 1.7	11.5± 0.8	13.7± 1.1	47.8± 1.2
WAN	38.2± 2.4	40.9± 2.7	13.7± 0.9	23.0± 1.3	37.4± 2.2
LS	47.8± 0.8	51.0± 0.5	9.5± 1.2	11.4± 1.5	49.8± 0.8
N-LS	48.6± 1.3	51.1± 1.2	12.0± 2.2	14.2± 2.8	50.0± 2.1
EM	47.7± 1.7	48.5± 1.7	3.4± 0.1	6.5± 0.1	49.6± 2.8
Mask	50.4± 0.1	**53.2± 0.2**	25.6± 1.8	27.9± 2.0	51.9± 1.2
P+S	**50.5± 0.3**	51.8± 0.3	**30.9± 2.3**	**33.4± 2.6**	**52.4± 1.3**
Mask*	53.4± 0.9	54.7± 1.5	32.5± 0.8	34.3± 0.9	55.4± 1.1
P+S*	57.2± 0.6	57.3± 0.8	41.7± 2.1	42.7± 1.9	59.9± 1.3

Table 4: Results obtained training the model end-to-end with optical flow images. Results in percent (%) ± standard deviation on the three splits of Confusing-HMDB-102. Losses marked with ∗ assume the case when the pseudo label search is ideal, and gets to use actual ground truth labels during training. Mask and P+S (without ∗) are trained with $K = 10$ and $\tau = 0.2$.

E   Qualitative results

Figure 7 illustrates a few qualitative examples where we compare the predictions obtained with our methods and the other SPML baselines. Verbs shown here were selected by thresholding the model confidence (sigmoid) at 50%. A "-" indicates no predictions, which happens when all classes have low confidence. We note that WAN and EM tend to over-predict numerous verbs, whereas AN, LS and N-LS tend to predict only one verb. Our methods instead correctly output relevant verbs without over-predicting a large number of labels.

F   End-to-end experiments

We additionally performed end-to-end training on the Confusing-HMDB-102 dataset, using optical flow images. Optical flow was extracted using the TV-L1 [42] algorithm. In this case the training set features and the pseudo labels were updated every 5 epochs. Table 4 reports results obtained with this setting. Performance with end-to-end training naturally improves compared to results reported in the main paper, where pre-extracted features were used throughout training. We observe that our methods consistently outperform all baselines.

References

[1] Anurag Arnab, Chen Sun, Arsha Nagrani, and Cordelia Schmid. Uncertainty-aware weakly supervised action detection from untrimmed videos. In *European Conference on Computer Vision*, pages 751–768. Springer, 2020.

Figure 7: Qualitative examples from our methods and the other SPML baselines. White indicates the ground truth, yellow denotes a partial match, while red and green denote incorrect and successful total matches, respectively. We cap the model predictions shown here to five. A dash "–" denotes the case where the are no predictions due to the model having low confidence for all classes.

[2] Roger G Barker and Herbert F Wright. Midwest and its children: The psychological ecology of an american town. 1955.

[3] 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] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2021.

[5] 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.

[6] 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. *International Journal of Computer Vision*, 2021.

[7] Xiang Deng, Songhe Feng, Gengyu Lyu, Tao Wang, and Congyan Lang. Beyond word embeddings: Heterogeneous prior knowledge driven multi-label image classification. *IEEE Transactions on Multimedia*, 2022.

[8] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 647–657, 2019.

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[10] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

[11] Spandana Gella, Mirella Lapata, and Frank Keller. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[12] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12046–12055, 2019.

[13] Aritra Ghosh and Andrew Lan. Contrastive learning improves model robustness under label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2703–2708, 2021.

[14] Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Not all relations are equal: Mining informative labels for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15596–15606, 2022.

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[16] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[17] Thomas Leung, Yang Song, and John Zhang. Handling label noise in video classification via multiple instance learning. In *2011 International Conference on Computer Vision*, pages 2056–2063. IEEE, 2011.

[18] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[20] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[21] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

[22] Davide Moltisanti, Michael Wray, Walterio Mayol-Cuevas, and Dima Damen. Trespassing the boundaries: Labeling temporal bounds for object interactions in egocentric video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[23] Islam Nassar, Samitha Herath, Ehsan Abbasnejad, Wray Buntine, and Gholamreza Haffari. All labels are not created equal: Enhancing semi-supervision via label grouping and co-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7241–7250, 2021.

[24] Diego Ortego, Eric Arazo, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6606–6615, 2021.

[25] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952, 2017.

[26] Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. In *International Conference on Learning Representations*, 2021.

[27] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE international conference on computer vision*, pages 3591–3600, 2017.

[28] Mohit Sharma, Raj Aaryaman Patra, Harshal Desai, Shruti Vyas, Yogesh Rawat, and Rajiv Ratn Shah. Noisyactions2m: A multimedia dataset for video understanding from noisy labels. In *ACM Multimedia Asia*, pages 1–5. 2021.

[29] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10399, 2021.

[30] Mohammad S Sorower. A literature survey on algorithms for multi-label learning. *Oregon State University, Corvallis*, 18:1–25, 2010.

[31] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[32] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5552–5560, 2018.

[33] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[34] Thomas Verelst, Paul K Rubenstein, Marcin Eichner, Tinne Tuytelaars, and Maxim Berman. Spatial consistency loss for training multi-label classifiers from single-label annotations. *arXiv preprint arXiv:2203.06127*, 2022.

[35] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018.

[36] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[37] Xinshao Wang, Yang Hua, Elyor Kodirov, David A Clifton, and Neil M Robertson. Proselflc: Progressive self label correction for training robust deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 752–761, 2021.

[38] Sandra Waxman, Xiaolan Fu, Sudha Arunachalam, Erin Leddon, Kathleen Geraghty, and Hyun-joo Song. Are nouns learned before verbs? infants provide insight into a long-standing debate. *Child development perspectives*, 7(3):155–159, 2013.

[39] Michael Wray and Dima Damen. Learning visual actions using multiple verb-only labels. *British Machine Vision Conference*, 2019.

[40] Michael Wray, Davide Moltisanti, Walterio Mayol-Cuevas, and Dima Damen. Sembed: Semantic embedding of egocentric action videos. In *European Conference on Computer Vision*, pages 532–545. Springer, 2016.

[41] Serena Yeung, Vignesh Ramanathan, Olga Russakovsky, Liyue Shen, Greg Mori, and Li Fei-Fei. Learning to learn from noisy web videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5154–5162, 2017.

[42] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime TV-L 1 optical flow. *Pattern Recognition*, pages 214–223, 2007.

[43] Youcai Zhang, Yuhao Cheng, Xinyu Huang, Fei Wen, Rui Feng, Yaqian Li, and Yandong Guo. Simple and robust loss design for multi-label learning with missing labels. *arXiv preprint arXiv:2112.07368*, 2021.

[44] Xubin Zhong, Changxing Ding, Xian Qu, and Dacheng Tao. Polysemy deciphering network for robust human–object interaction detection. *International Journal of Computer Vision*, 129(6):1910–1929, 2021.

[45] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018.

[46] Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. Acknowledging the unknown for multi-label learning with single positive labels. In *ECCV*, 2022.