ASPnet: Action Segmentation with Shared-Private Representation of Multiple Data Sources

Beatrice van Amsterdam[1,2], Abdolrahim Kadkhodamohammadi[2], Imanol Luengo[2], Danail Stoyanov[1,2]
[1] Wellcome/EPSRC Centre for Interventional and Surgical Sciences,
[2] Medtronic plc

Abstract

Most state-of-the-art methods for action segmentation are based on single input modalities or naïve fusion of multiple data sources. However, effective fusion of complementary information can potentially strengthen segmentation models and make them more robust to sensor noise and more accurate with smaller training datasets. In order to improve multimodal representation learning for action segmentation, we propose to disentangle hidden features of a multi-stream segmentation model into modality-shared components, containing common information across data sources, and private components; we then use an attention bottleneck to capture long-range temporal dependencies in the data while preserving disentanglement in consecutive processing layers. Evaluation on 50salads, Breakfast and RARP45 datasets shows that our multimodal approach outperforms different data fusion baselines on both multiview and multimodal data sources, obtaining competitive or better results compared with the state-of-the-art. Our model is also more robust to additive sensor noise and can achieve performance on par with strong video baselines even with less training data.

1. Introduction

Action segmentation is the task of predicting which action is occurring at each frame in untrimmed videos of complex and semantically structured human activities [18, 32]. While conventional methods for human action understanding focus on classification of short video clips [6, 27, 34], action segmentation models have to learn the semantics of all action classes as well as their temporal boundaries and contextual relations, which is challenging and requires the design of efficient strategies to capture long range temporal information and inter-action correlations.

Recent methods for action segmentation input precomputed low-dimensional visual features [6, 11] into different long-range temporal processing units, such as temporal convolutions [11, 19], temporal self-attention [38, 45] or graph neural networks [46]. While these methods utilize

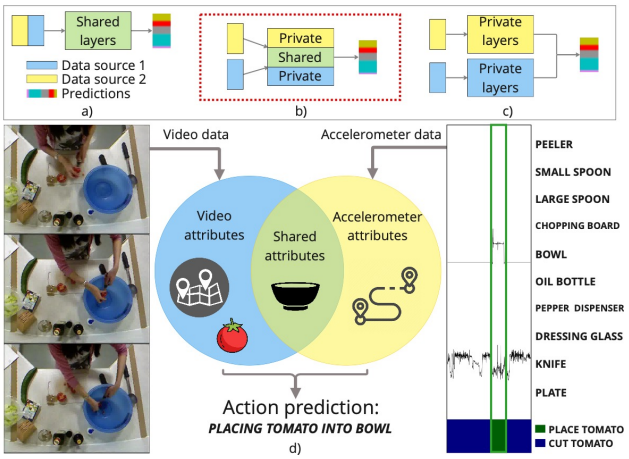

Figure 1. Different paradigms for multi-source data fusion via (a) early fusion, (b) disentanglement of modality-shared and modality-specific representations (our model) and (c) late fusion; (d) Example from 50salads highlighting shared and private information that can be extracted from video and accelerometer data. While both modalities can detect the activation of relevant tools and common motion cues, RGB videos additionally capture fundamental details about objects without acceleration sensors and their state (*e.g.* chopped tomatoes), the overall spatial configuration and the localization of motion in the scene. Accelerometer signals, on the other hand, contain explicit and complementary information about 3D fine motion patterns of activated objects and their co-occurrence. In the presence of noise (*e.g.* video occlusions) or other variability factors, some shared attributes could become part of the private space of the uncorrupted modality.

This research was funded in part by the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [203145/Z/16/Z]; the Engineering and Physical Sciences Research Council (EPSRC) [EP/P012841/1]; and the Royal Academy of Engineering Chair in Emerging Technologies Scheme. For the purpose of open access, the author has applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

only video data, recent computer vision datasets have increasing availability of multiple synchronized data sources [9,24,32], some of which could be collected readily in real-case scenarios (*e.g.* audio recordings [9], teleoperated robot kinematics [35]). Effective fusion of different data modalities or different ' views' of the same modality (here we use the term 'view' to denote any different representation of the same data source) is not trivial and still a very active area of research, as potential advantages include higher recognition performance, improved robustness to sensor noise and mitigating the need for large training datasets [3].

Action segmentation with multiple data sources has not been investigated as extensively as similar tasks like action classification. It has generally been addressed via naïve fusion strategies such as multimodal feature concatenation [5, 45] and prediction fusion [37], or limited to the feature encoding stage [22]. However, sensor fusion can also benefit from long-range temporal modelling performed in later stages. Inspired by work on multimodal representation learning [5, 20], we approach the problem implementing a multi-stream action segmentation model, one stream for each available data source, and disentangling their latent space into modality-shared versus modality-specific representations (Fig. 1b and 1d), aiming at learning more discriminative features and more robust action recognition. We assume that creating a shared feature space across data sources produces more abstract action representations and reduces over-fitting to modality-specific nuances and noise, while private features could retain useful complementary information for the downstream task. Instead of relying on adversarial mechanisms [41], autoencoders [5, 20] or generative approaches [20], we learn shared feature spaces with minimal model modification by minimizing Maximum Mean Discrepancy (MMD) on partitions of the latent spaces to reduce the distance between their distributions. In order to capture long-range temporal dependencies in the data while preserving feature disentanglement in consecutive processing layers, an attention bottleneck [26] is then integrated into the segmentation model and initialized with learned modality-shared features, allowing independent processing of all private features. We called the model ASPnet (Action Shared-Private network).

Evaluation results of our model on three challenging benchmark datasets show improvement over unimodal baselines and different fusion strategies using both multimodal (*e.g.* video and accelerometer) and multiview (*e.g.* RGB and optical flow) inputs, leading to competitive or better results than the state-of-the-art. In addition, results suggest that ASPnet could generalize well to multiple data sources, improving its performance with growing number of inputs. Despite requiring synchronized recordings of multiple sensors, we demonstrated that our model is also more robust to additive input noise and can match the performance of strong video baselines with less data. In summary, our contributions are the following:

- We present ASPnet, a new multi-source activity recognition model to effectively exploit shared and complementary information contained in multiple data sources for robust action segmentation. ASPnet partitions the latent representation of each modality and exploits a bottleneck mechanism to allow feature interaction at multiple levels of abstraction while preserving disentanglement. Additionally, modality fusion is influenced by long-range temporal dynamics captured at different scales.
- We show the advantage of feature disentanglement to fuse not only multimodal data, but also multiple representations of the same modality.
- We perform extensive ablation studies to evaluate ASPnet against strong baselines, different levels of noise and less training data.
- We evaluate ASPnet on three challenging benchmark datasets and achieved competitive or better results than state-of-the-art models.

2. Related Work

Action segmentation: Many studies on action segmentation classify video frames using temporal convolutional networks [19], that capture multi-scale temporal dependencies in the data through temporal pooling layers [19, 21, 47] and/or dilated convolutions [11]. While performing well in frame-wise accuracy, over-segmentation errors are very common among models designed to predict one action class for each frame. Different strategies were devised to alleviate this issue, from auxiliary smoothing losses [11, 28] to self-supervised domain adaptation techniques [7, 8], prediction refinement modules [1, 15, 16, 28, 31, 36, 40], and post-processing strategies [4, 23]. In contrast, graph-based models attempt to directly regularize model predictions by explicitly modelling contextual relations between successive actions [46].

Recent studies [10, 38, 45] have shown the potential of the attention mechanism in capturing long-range temporal dependencies in long video sequences. ASFormer [45], for example, uses sliding-window attention to reduce complexity of transformers and integrates it with temporal convolutions. Predictions can be refined with different types of attention-based decoders [4, 38]. In this paper, we focus on the idea of efficiently fusing multiple data sources in long-range action segmentation models and use ASFormer as our backbone model, as in related work [22, 38]. However, the proposed methodology can be integrated with arbitrary decoders [4,38], refinement modules [1,16] or post-processing strategies [23] to improve final predictions.

Multimodal action recognition: While several ac-

tion segmentation studies used features from multiple data sources [11, 35, 38, 45, 46], many relied on naïve merging strategies such as early concatenation of RGB and flow features. Alternatively, multimodal fusion was elaborated at the video encoding level [22], failing to capture common longer-range temporal dependencies among data sources.

Modality fusion has been explored more extensively for action classification, the task of labelling trimmed action clips. While CNNs have dominated the scene for several years [6, 30], they rely on strong architectural priors that are often modality-specific, offering limited flexibility to data fusion and leading to a variety of customized schemes that need to be re-adapted to each application and dataset [3]. More flexible fusion strategies relied on weighted blending of supervision signals in multistream systems [39].

The research focus has recently shifted towards transformers [43], representing flexible perceptual models able to handle a wide range of data sources with minimal changes to the model structure [13, 17]. The self-attention operation in transformers represents a straight-forward solution to combine different signals [42, 44], but it does not account for information redundancy and it does not scale well to longer temporal sequences. To mitigate these issues, [26] introduced 'attention bottleneck' to restrict the attention flow between tokens from different data sources and force each modality to share only what is necessary with the other modalities. In our model, we explicitly separate modality-shared and private feature representations in a long-range action segmentation model and exploit the bottleneck mechanism to preserve feature disentanglement in subsequent temporal processing layers.

Shared-private representation learning: Shared and private feature disentanglement was explored in Domain Separation Network (DNS) for unsupervised domain adaptation [5]. DNS uses a shared-weight encoder to capture domain-shared features for a given input sample, and a private encoder for each domain to capture domain–specific attributes. To generate such disentangled representations and avoid trivial solutions, auxiliary losses are employed to bring shared representations close, while pushing them apart from the private features; a shared decoder is also employed to reconstruct input samples from their partitioned representations. The shared representation of the source domain is finally used to train the network on the task of interest. We differ from this work under multiple aspects: first, we use multimodal data rather than multi-domain images. This implies that modality-specific feature representations could contain useful information for our downstream task and are considered also for prediction. Second, feature disentanglement is obtained by partitioning the latent space of each unimodal encoder, rather than building separate private and shared encoders, using an auxiliary similarity loss and a bottleneck mechanism, but no additional layers or auxiliary tasks. Our solution allows information exchange at multiple abstraction levels while preserving disentanglement.

Few other studies have used similar decompositions for different multimodal tasks, such as representation learning [20] and cross-modal retrieval [41]. These models are however based on probabilistic frameworks [20] or generative adversarial networks [41], which are not trivial to train.

3. Methods

This section presents a detailed description of ASPnet. The model is illustrated in Fig. 2 in the case of two input modalities. Pre-extracted frame-wise features from multiple data sources are disentangled into modality-shared and private spaces (Section 3.1) and then refined via temporal processing with a shared attention bottleneck (Section 3.2) to generate frame-wise action predictions.

3.1. Shared-private feature disentanglement

The core hypothesis in this paper is that we can partition latent representations of multimodal or multiview networks into a shared space, containing common information across all sources, and a private space for each modality, and that such disentangled representations are more robust and facilitate action prediction. This is because shared knowledge can help abstraction from modality-specific details, while private spaces retain useful complementary knowledge.

The goal of the first stage of ASPnet is to obtain well-separated shared and private representations of the input signals. Given N synchronized input sequences X_i of length T and size D_i, $i = 1 : N$, we project them into low dimensional features of size F via independent fully connected layers FC_i followed by normalization layers [2]. We then partition the latent space of each modality into private and shared spaces (P_i, S_i) of size $F/2$, as in [20]. To effectively make all S_i features contain shared information across data sources, the Maximum Mean Discrepancy (MMD) [14] between all S_i pairs is minimized during training:

$$L_{mmd} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} MMD(S_i, S_j). \quad (1)$$

3.2. Temporal attention bottleneck

The goal of the second stage of ASPnet is to process effectively the disentangled feature sequences from the first stage and generate frame-wise action predictions. Our solution consists of a multi-stream segmentation model, one stream for each data source. As our focus is to optimize information fusion, we chose the encoder of a popular segmentation model, ASFormer [45], as the backbone of all ASPnet streams.

To preserve feature disentanglement in consecutive processing layers, we integrated a temporal attention bottle-

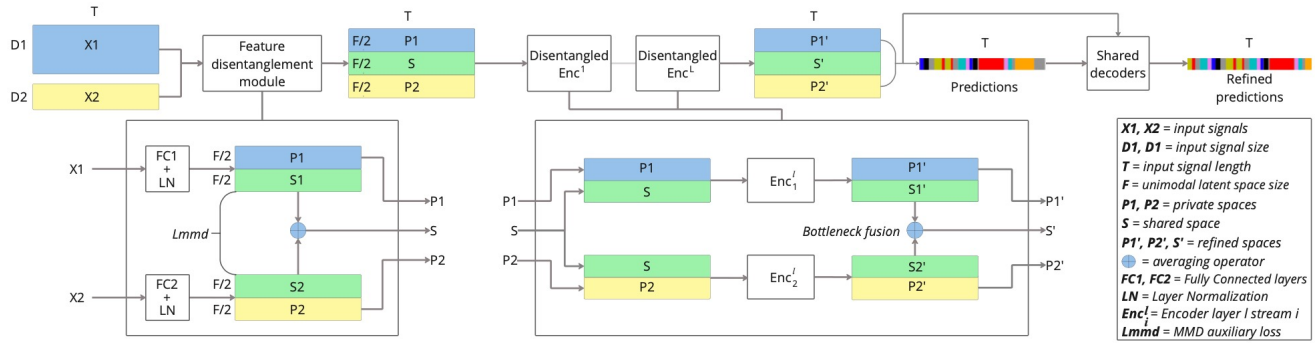

Figure 2. ASPnet schematic. The modules introduced in Sections 3.1 and 3.2 are illustrated in separate boxes. Frame-wise features from multiple data sources are first disentangled into modality-shared and private features. They then go through a sequence of L encoder layers with temporal attention bottleneck to generate frame-wise action predictions, later refined through multiple decoders.

neck into our multi-stream architecture. The bottleneck is shared among all modalities, similar to [26], and the first layer is initialized by the average S of the shared spaces S_i generated in the first ASPnet stage. At each layer $Enc^l, l \in \{1, ..., L\}$, shared (S) and private (P_i) features are processed independently for each modality i, and all the refined shared spaces S_i' are then averaged again:

$$P_i', S_i' = Enc_i^l(P_i, S), \tag{2}$$

$$S' = \frac{1}{N}\sum_{i=1}^{N} S_i', \tag{3}$$

where S' is the refined bottleneck and P_i' is the refined private space of modality i.

The concatenation of the shared bottleneck and all private spaces in the last layer is used to generate action predictions, later refined with multiple ASFormer decoders and moving-average post-processing.

3.3. Loss function

As for ASFormer, the loss function is a combination of cross-entropy classification loss (L_{ce}) and smooth loss (L_{sm}) [11] computed at the encoder ($p = 0$) and decoder ($p = 1 : D$) prediction stages. In addition, we introduce the auxiliary MMD loss (L_{mmd}) for feature disentanglement:

$$L = \gamma L_{mmd} + \sum_{p=0}^{D}(L_{ce}(p) + \lambda L_{sm}(p)), \tag{4}$$

D is the number of decoder stages, λ and γ are loss weights.

3.4. Implementation and training details

As our goal was to optimize data fusion, we did not tune ASPnet on 50salads and Breakfast, but used the same set of model and training hyperparameters as ASFormer [45]. The final model consists of N encoder streams, where N is the number of available data sources, and one common 3-stage decoder. Each encoder stream and decoder stage contains 10 attention layers with feature size = 64 (shared feature size = private feature size = 32). Smooth loss weight λ is 0.25. For MMD we used multiscale kernels with bandwidth range [0.2, 0.5, 0.9, 1.3]. We set γ to 1 without tuning. On 50 salads, we trained the model for 100 epochs using Adam optimizer and learning rate 0.0005. On Breakfast, we trained it for 100 epochs with learning rate 0.0001. Predictions were post-processed with a moving average filter of 7 seconds in 50salads and 2 seconds in Breakfast, with grid search performed on a range from 1 to 10 seconds.

On RARP45, we tuned ASPnet on a separate validation set, and then re-trained the model on the full train set with the chosen hyperparameters. We optimized the number of layers (set to 8, with search in [10, 9, 8, 7]), initial learning rate (set to 0.0005, with search in [0.0005, 0.0001], number of training epochs (set to 50) and smooting window size (set to 3 seconds, with grid search in the range 1 to 5 seconds). The other parameters remained the same.

We implemented ASPnet in PyTorch and trained it on NVIDIA Tesla V100. Optical flow features, if not already available, were extracted from RAFT [33] flow frames using I3D [6] pre-trained on Kinetics with window size = 9.

4. Experiments and Results

We benchmarked ASPnet on challenging action segmentation datasets and performed extensive ablation studies.

4.1. Datasets

50Salads [32] dataset contains 50 top-view videos of salad preparation activities performed by 25 different users in the same kitchen and is annotated with 19 action classes. It also contains 3-axis accelerometer signals of devices attached to the cooking tools and synchronization parameters for temporal alignment with the videos. In line with related

Table 1. Comparison of multimodal ASPnet and different unimodal and fusion baselines on 50salads. MA = moving average.

	$F1_{50}$	$F1_{25}$	$F1_{10}$	Acc	$Edit$	#Param(M)
ASFormer accel	58.4	66.7	69.1	66.0	66.5	1.01
ASFormer video	76.3	83.1	84.8	86.2	80.0	1.13
Late fusion	71.9	79.0	80.7	85.4	73.2	2.14
Early fusion	78.9	84.6	86.0	88.4	79.7	1.14
Mid fusion	78.8	84.3	86.1	87.0	79.6	1.39
ASPnet - Gaus	79.1	85.3	86.5	87.4	80.6	1.33
ASPnet - Eucl	80.0	85.8	87.0	86.0	81.5	1.39
ASPnet	84.7	88.2	89.2	**89.8**	83.8	1.39
ASPnet + MA	**85.6**	**89.5**	**90.4**	**89.8**	**85.6**	1.39

Table 2. Results with different input sources on 50salads. R = RGB, F = optical flow, A = accelerometer. + = concatenation.

	$F1_{50}$	$F1_{25}$	$F1_{10}$	Acc	$Edit$	#Param(M)
ASFormer (R+F)	79.7	86.6	87.8	86.2	82.2	1.13
ASPnet (R, F)	80.9	86.8	88.6	87.2	82.7	1.39
ASPnet (R+F, A)	85.6	89.5	90.4	89.8	85.6	1.39
ASPnet (R, F, A)	**86.4**	**90.4**	**91.3**	**90.3**	**85.8**	1.64

Table 3. Testing the contribution of learned shared-private features towards prediction performance via feature masking.

	$F1_{50}$	$F1_{25}$	$F1_{10}$	Acc	$Edit$
Mask private	74.6	81.8	83.0	82.2	76.3
Mask shared	79.2	84.9	86.5	85.4	80.6
No mask	**85.6**	**89.5**	**90.4**	**89.8**	**85.6**

work, evaluation on 50salads is performed at 15Hz via 5-fold cross-validation and the average results are reported.

Breakfast [18] is a much larger dataset containing 1712 videos of breakfast preparation activities recorded from multiple points of view in 18 different kitchens and annotated with 48 action classes. To compare with ASFormer, we reported average results over 4 cross-validation folds.

RARP45 [35] is a recent action segmentation dataset containing surgical activities extracted from 45 robot-assisted radical prostatectomies performed by 8 surgeons with different expertise, and it is annotated with 8 action classes. The data consist of synchronized endoscopic videos and kinematic trajectories recorded from the robotic platform, but only the videos are publicly available. This dataset is challenging not only because it contains real-life activities in uncontrolled environment, but also because images are noisy (due to occlusions and specularities) and motion is analyzed at finer granularity, so that action segmentation models must learn to discriminate subtle motion cues rather than the identity of the objects in use. Results on RARP45 are reported as average scores over the test videos.

For our experiments on 50salads and Breakfast we used the I3D features extracted from RGB and flow frames by [11], unless stated otherwise. For RARP45, we extracted the same type of features ourselves.

4.2. Evaluation metrics

We analyzed segmentation performance using accuracy, edit distance and segmental F1-scores [19]. Accuracy evaluates predictions in a frame-wise manner, but it is not able to assess temporal properties. The other scores measure the ability of a network to understand the structure of complex activities. While the edit distance only evaluates action ordering, the F1-scores additionally measure the temporal overlap between predicted and ground truth segments at different thresholds: 10%, 25%, 50%.

4.3. Result on multimodal data

We tested the ability of ASPnet to fuse multimodal data using the video and accelerometer signals contained in 50salads (Table 1). Our model outperforms all unimodal baselines as well as three different modality fusion strategies: early fusion corresponds to the original ASFormer model, where multimodal features are concatenated; late fusion corresponds to parallel unimodal ASFormer streams with average output logits; middle fusion corresponds to parallel ASFormer encoders and a common ASFormer decoder, which is equivalent to ASPnet with zero-sized bottleneck. We also trained ASPnet-Gaus, a variant of our model where the attention bottleneck is initiliazed with a Gaussian with zero mean and standard deviation of 0.02, as in [26]. We observed that modality-shared features provide a more effective initialization (improvement ranging from +2.4% to +5.6% on different scores). Naïve moving-average post-processing further increases the final segmental scores and is used in the rest of the experiments. Qualitative results are shown in Fig. 3.

We finally experimented training ASPnet with a simpler similarity loss minimizing the Euclidean distance between modality-shared features (ASPnet-Eucl), but it degraded performance compared with MMD alignment; this could be partially due to the fact that the auxiliary loss is applied at a very early processing stage, where features cannot yet reach abstraction from their sensor-specific characteristics. We also tried introducing a second auxiliary loss to push shared and private features apart [41], but it was not helpful; feature separation occurred spontaneously while training one disentangled stream per modality, as opposed to separate shared-private streams [5], and the separation loss was minimized just after a few training iterations.

Feature disentanglement was sanity checked against trivial solutions by testing ASPnet with a mask on either the bottleneck or the private features. In both cases we observed a moderate drop in prediction performance (Table 3), indicating that both representations contain useful information and are needed for action segmentation.

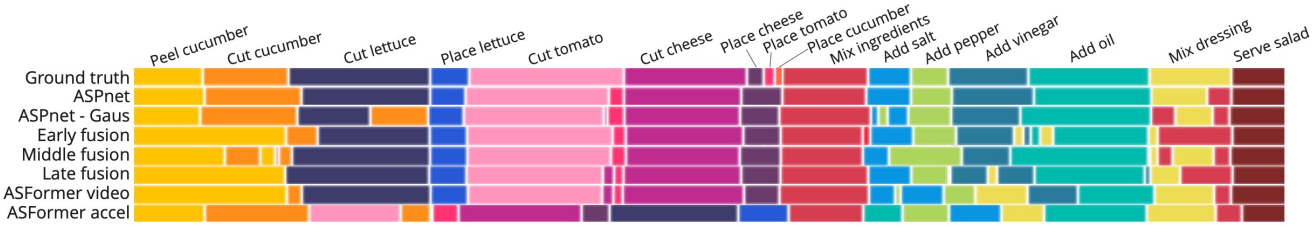

Figure 3. Qualitative results on 50salads. ASPnet exploits multimodal information better than all fusion baselines and improves upon video-based predictions despite the low predictive power of accelerometers alone.

4.4. Results on multiview data

The 2048-dimensional video features used in the previous experiment were extracted by [11] using both RGB and optical flow video frames. This gave us the opportunity to test the ability of our model to effectively fuse multiview data by separating those features. As reported in Table 2, ASPnet achieved slighlty better performance compared with the original ASFormer, that uses a simple concatenation of the same data views. However, improvement is less significant than the multimodal case; different sensors generally contain more complementary information than multiple views of the same source, explaining why modality-shared and modality-specific feature disentanglement can be more beneficial in multimodal case.

4.5. Scalability to multiple data streams

ASPnet has a flexible design that supports in principle an arbitrary number of input modalities. On 50salads, three-stream ASPnet using I3D RGB, I3D optical flow and accelerometer features outperformed all two-stream solutions (Table 2). This suggests that ASPnet could generalize readily to multiple data sources, increasing its accuracy as the number of input sources grows.

In practice, however, two modifications are necessary to make ASPnet scalable to a large number of input sensors: first and foremost, all encoder streams can share weights, so that the network size becomes independent from the number of inputs. In our experiments, this resulted in only a moderate drop in performance (Table 4), showing that 3-stream ASPnet with shared weights can be used with reduced computational resources while still offering competitive performance. We also observe that ASPnet size is comparable to ASFormer (1.13M) and smaller than other less competitive models (ranging from 0.8M to 19.2M [28]).

The second modification is on the auxiliary loss. Instead of computing MMD between all pairs of modality-shared spaces, we can compute MMD between each shared space and the corresponding average bottleneck ($\hat{L}_{mmd}$). This is not convenient with two modalities (the number of losses grows from 1 to 2), and it is irrelevant with three modalities (the number of losses is 3 in both cases), but it is efficient with more than three modalities. In our experiments with 3 data streams, recognition performance was not significantly affected by this design choice (Table 4).

Table 4. Computational scalability with multiple data streams. We assess ASPnet performance when all encoders have shared weights and using a scalable variation of the MMD loss ($\hat{L}_{mmd}$).

	$F1_{50}$	$F1_{25}$	$F1_{10}$	Acc	$Edit$	#Param(M)
Shared-$\hat{L}_{mmd}$	85.5	89.4	90.9	89.6	85.7	1.15
Shared	85.6	89.6	91.0	89.6	85.7	1.15
Non-shared	**86.4**	**90.4**	**91.3**	**90.3**	**85.8**	1.64

4.6. Robustness to noise

Improved prediction performance is not the only potential advantage of multimodal data fusion. Multiple data sources generally contain complementary information and are subject to different types of noise, and when one modality is corrupted or insufficient to discriminate a certain action, the other modalities could compensate and rectify the model predictions.

Analysis of model robustness to different levels of additive zero-mean Gaussian noise (standard deviation $s \in [0.5, 1, 1.5, 2]$, corresponding to about [10, 20, 30, 40]% of the input feature range) is presented in Fig. 4a. Results are reported as average crossvalidation scores over 5 testing runs using different instances of the same random noise. Compared with video ASFormer, 3-stream ASPnet shows significantly reduced sensitivity to data corruption, whether on the video or both modalities. Moreover, ASPnet with corrupted accelerometer signals still outperforms ASFormer on original uncorrupted videos. Accelerometer signals in 50salads are very compact and easy to process, but insufficient to discriminate all action classes (as shown in Table 1 and Fig. 3). ASPnet thus strongly relies on the visual features and is more sensitive to video noise than accelerometer noise; however, the complementary motion information from the accelerometers is exploited effectively to improve prediction performance and make ASPnet remarkably more robust to video noise than the corresponding video baseline (the performance drop of ASPnet from uncorrupted inputs is about 50% smaller than ASFormer when

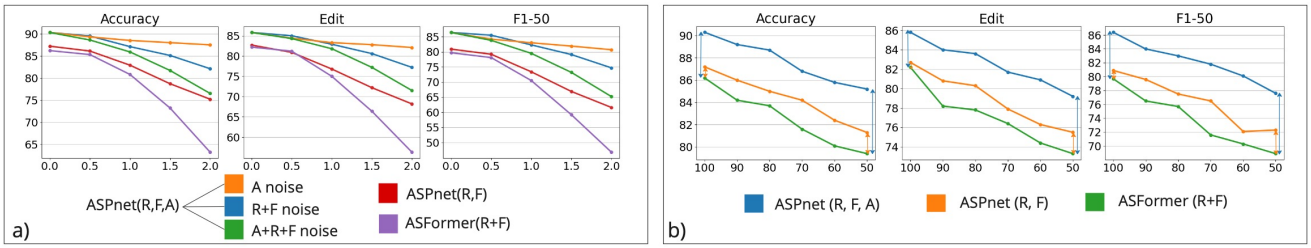

Figure 4. (a) ASPnet robustness to noise. Multimodal ASPnet and multiview video ASPnet are compared with video ASFormer under different levels of additive zero-mean Gaussian noise (x-axis denotes noise standard deviation). Multimodal ASPnet is tested with acceleration noise, video noise and both. (b) Impact of reducing the training set size on different models. The x-axis shows the amount of training data in percentage. Vertical double arrows highlight performance gap between our models and ASFormer, which tends to increase as the number of training sequences decreases. F1-25 and F1-10 scores follow a trend similar to other evaluation scores in both experiments. R = RGB, F = flow, A = accelerometer. + = feature concatenation.

$s = 1$, and this gap increases with stronger noise).

Although in multiview scenarios we cannot take advantage of a clean input when the other one is corrupted because all inputs derive from the same source, we observed that video ASFormer is notably more sensitive to noise than multiview video ASPnet based on the same input features, highlighting another advantage of the proposed feature-disentanglement strategy.

4.7. Sensitivity to training set size

The second potential gain when using richer data representations such as multiple sensors or views could be the ability to reach competitive performance with a reduced number of training videos, and therefore reduced annotation effort and costs. Fig. 4b shows prediction scores of multimodal ASPnet, multiview ASPnet and video ASFormer trained with decreasing amounts of data. Both multimodal and multiview ASPnets show smaller performance drops than ASFormer. In addition, multiview ASPnet trained with 50% of the videos outperforms video ASFormer trained with 70% of the videos and the same input features. Similar but amplified trend is observed for multimodal ASPnet, matching the performance of video ASFormer using only 70% of the training data, and outperforming video ASFormer trained with 90% of the videos using only 50%.

4.8. Comparison with SOTA

On 50salads, I3D features have been recently replaced with stronger video representations (Br-Prompt) [22] aimed at improving ASFormer results (Br-Prompt+ASFormer). Following [22], we used Br-Prompt RGB features for comparison with SOTA methods on this dataset, together with I3D optical flow [33] features and accelerometer data. As reported in Table 5, our multiview video ASPnet outperforms the state-of-the-art in accuracy and gets close to or matches the top segmental scores. When adding the accelerometer signals, ASPnet outperforms the state-of-the-art in all the metrics but the edit score. The top ranking method is a graph-based model, which is well suited to learn sequences and avoid over-segmentation errors, thus achieving large edit scores.

Table 5. Results on 50salads. R=RGB, F=flow, A=accelerometer.

	$F1_{50}$	$F1_{25}$	$F1_{10}$	Acc	$Edit$
ASFormer [45]	76.0	83.4	85.1	85.6	79.6
ASFormer+ASRF [45]	79.3	85.4	85.1	85.6	81.7
CETnet [38]	80.1	86.5	87.6	86.9	81.7
DPRN [28]	79.4	86.3	87.8	87.2	82.0
UVAST [4]	81.7	87.6	89.1	87.4	83.9
Br-Prompt+ASFormer [22]	81.3	87.8	89.2	88.1	83.8
Semantic2Graph [46]	87.3	90.2	91.5	88.6	**89.1**
Br-Prompt+ASPnet (R,F)	86.0	90.3	91.2	90.4	86.0
Br-Prompt+ASPnet (R,A)	87.3	91.3	92.2	90.9	87.5
Br-Prompt+ASPnet (R,F,A)	**88.5**	**91.6**	**92.7**	**91.4**	87.5

We then tested multiview RGB-flow ASPnet on Breakfast (Table 6), which does not contain multimodal data, but is larger and more complex than 50salads. ASPnet proved again to be superior to video ASFormer in all the evaluation metrics, using the same input features and sharing most of the network structure. It also demonstrated to be competitive with the state-of-the-art. We note that CETnet differs from ASFormer only in the decoder stage, which is much larger (100 layers in CETnet as opposed to 30 in ASFormer), so ASPnet could be readily integrated into CETnet to potentially improve the prediction scores on larger datasets such as Breakfast. We also note that there could be room for improvement with systematic hyperparameter search, smoothing losses [11, 28] or refinement stages [16,23,28], that we did not explore in this study.

We finally tested multiview RGB-flow ASPnet on RARP45 (Table 7), investigating the ability of our model to work in a different data domain. While using only video-derived information, we outperformed the state-of-the-art method fusing video and robot kinematics [35]. This result shows that our model could be used in a wide range of

Table 6. Results on Breakfast.

	$F1_{50}$	$F1_{25}$	$F1_{10}$	Acc	$Edit$
ASFormer [45]	57.4	70.6	76.0	73.5	75.0
EUT [10]	59.8	71.8	76.2	75.0	74.6
C2F-TCN [31]	57.6	68.7	72.2	**76.0**	69.6
CETnet [38]	**61.9**	**74.3**	**79.3**	74.9	**77.8**
ASPnet	60.8	72.9	78.1	75.9	76.3

Table 7. Results on RARP45.

	$F1_{50}$	$F1_{25}$	$F1_{10}$	Acc	$Edit$
MA-TCN [35]	-	-	83.7	80.9	79.6
ASPnet	**74.8**	**84.0**	**86.7**	**82.7**	**79.8**

surgical procedures where robot kinematics is not available, such as traditional laparoscopy and endoscopy. Kinematic information could potentially be replaced also with compact surgical tool representations automatically extracted from surgical videos using pre-trained object and key-point detection models [12, 25].

4.9. Other experiments

We tested a few variations of ASPnet architecture. For example, we tried extending the attention bottleneck to the decoder stage, in all or part of its layers, but we did not observe any significant improvement. The role of the decoder is to refine the encoder predictions, improving specially the segmental scores [45]. If such refinement is performed separately for each stream, the model could overfit individual modalities rather than achieve effective data fusion.

We also experimented with different types of attention layers. While attention bottlenecks were originally applied on spatio-temporal tokens extracted from short video clips [26], full spatio-temporal self-attention is not tractable with long video sequences. We can however decompose the computation into a temporal dimension, that is the dimension regarded in ASPnet, and a spatial dimension to capture complementary relations among all features at the same timestamp. We thus tried introducing into ASPnet additional spatial-attention layers with attention bottleneck. We experimented integrating them in the encoder or in the full model, sequentially or interlaced with the temporal-attention modules, but we didn't obtain any relevant gain in performance. We also observed that, replacing all temporal-attention layers with spatial-attention, the model reaches about 80% accuracy, but significantly lower segmental scores (*e.g.* less than 5% edit score) on split 1 of 50salads. This indicates that useful information can be captured via spatial self-attention, but it is challenging to integrate it optimally into long-range temporal models where temporal regularity is fundamental. Spatial attention could also give us interesting insight on which features and modalities the model focuses on at each timestamp. We will investigate the problem more extensively in future work.

5. Conclusion

In this paper, we tackled the problem of automatic action segmentation using multiple data sources. We presented ASPnet, a flexible model to fuse multiple inputs while simultaneously capturing long-range temporal dynamics in sequential data. Despite requiring synchronized recordings from multiple sensors, which might not always be possible, or time-consuming data processing to generate multiple input views, ASPnet has important advantages over strong baselines, including higher recognition performance, reduced sensitivity to input noise and smaller training sets. The latter could have a large impact in reducing annotation efforts and costs, data storage requirements (when the other modalities are low-dimensional such as accelerometers, lidars, robot kinematics, etc), as well as training time, mitigating the model environmental impact.

Limitations and future work: while showing similar advantages, multiview ASPnet achieves only a marginal performance gain compared with multimodal ASPnet. In the case of optical flow, the amount of information that is complementary to RGB features is more limited than, for example, 3D acceleration trajectories of multiple objects. We could therefore speculate that RGB-flow fusion will benefit less from the disentanglement of modality-shared and private feature representations. Future work will be aimed at evaluating ASPnet on alternative views of the video frames, such as human skeletons automatically identified in the scene.

Improvement in prediction performance could also be achieved by tuning the relative size of shared and private latent spaces for every combination of inputs; we will investigate learning the optimal partition.

We finally noted that large-scale action detection datasets such as as Epic-Kitchens-100 [9] and EGTEA [24] include multiple synchronized data sources such as RGB, accelerometer, audio and gaze signals, and constitute ideal benchmarks to compare data fusion strategies. We plan to expand our approach to the action detection task and to test fusion efficacy with such diverse signals.

Societal impact: action segmentation represents a core step in a wide range of applications, including delicate tasks such as monitoring of surgical procedures. In this context, adversarial attacks could put patients' health at risk, especially with non-visual data sources such as robot kinematics, which are harder to inspect. While appropriate defense strategies should always be implemented [29], we believe that effective modality fusion is by itself a defense mechanism, making models more robust to input corruption.

References

[1] Hyemin Ahn and Dongheui Lee. Refining action segmentation with hierarchical video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16302–16310, 2021. 2

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *Advances in Neural Information Processing Systems (NIPS)*, 2016. 3

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 2, 3

[4] Nadine Behrmann, S Alireza Golestaneh, Zico Kolter, Juergen Gall, and Mehdi Noroozi. Unified fully and timestamp supervised temporal action segmentation via sequence to sequence translation. In *European Conference on Computer Vision*, pages 52–68. Springer, 2022. 2, 7

[5] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016. 2, 3, 5

[6] 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. 1, 3, 4

[7] Min-Hung Chen, Baopu Li, Yingze Bao, and Ghassan AlRegib. Action segmentation with mixed temporal domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 605–614, 2020. 2

[8] Min-Hung Chen, Baopu Li, Yingze Bao, Ghassan AlRegib, and Zsolt Kira. Action segmentation with joint self-supervised temporal domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9454–9463, 2020. 2

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 2, 8

[10] Dazhao Du, Bing Su, Yu Li, Zhongang Qi, Lingyu Si, and Ying Shan. Efficient u-transformer with boundary-aware loss for action segmentation. *arXiv preprint arXiv:2205.13425*, 2022. 2, 8

[11] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019. 1, 2, 3, 4, 5, 6, 7

[12] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 8

[13] 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. 3

[14] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 3

[15] Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14024–14034, 2020. 2

[16] Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2322–2331, 2021. 2, 7

[17] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 3

[18] H. Kuehne, A. B. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014. 1, 5

[19] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 5

[20] Mihee Lee and Vladimir Pavlovic. Private-shared disentangled multimodal vae for learning of latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1700, 2021. 2, 3

[21] Peng Lei and Sinisa Todorovic. Temporal deformable residual networks for action segmentation in videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6742–6751, 2018. 2

[22] Muheng Li, Lei Chen, Yueqi Duan, Zhilan Hu, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Bridge-prompt: Towards ordinal action understanding in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19880–19889, 2022. 2, 3, 7

[23] Yunheng Li, Zhuben Dong, Kaiyuan Liu, Lin Feng, Lianyu Hu, Jie Zhu, Li Xu, Shenglan Liu, et al. Efficient two-step networks for temporal action segmentation. *Neurocomputing*, 454:373–381, 2021. 2, 7

[24] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision (ECCV)*, pages 619–635, 2018. 2, 8

[25] William McNally, Kanav Vats, Alexander Wong, and John McPhee. Rethinking keypoint representations: Modeling keypoints and poses as objects for multi-person human pose estimation. *European Conference on Computer Vision (ECCV)*, 2021. 8

[26] 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. 2, 3, 4, 5, 8

[27] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3163–3172, October 2021. 1

[28] Junyong Park, Daekyum Kim, Sejoon Huh, and Sungho Jo. Maximization and restoration: Action segmentation through dilation passing and temporal reconstruction. *Pattern Recognition*, 129:108764, 2022. 2, 6, 7

[29] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020. 8

[30] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 3

[31] Dipika Singhania, Rahul Rahaman, and Angela Yao. Coarse to fine multi-resolution temporal convolutional network. *arXiv preprint arXiv:2105.10859*, 2021. 2, 8

[32] Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738, 2013. 1, 2, 4

[33] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 4, 7

[34] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 1

[35] Beatrice Van Amsterdam, Isabel Funke, Eddie Edwards, Stefanie Speidel, Justin Collins, Ashwin Sridhar, John Kelly, Matthew J Clarkson, and Danail Stoyanov. Gesture recognition in robotic surgery with multimodal attention. *IEEE Transactions on Medical Imaging*, 2022. 2, 3, 5, 7, 8

[36] Dong Wang, Di Hu, Xingjian Li, and Dejing Dou. Temporal relational modeling with self-supervision for action segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2729–2737, 2021. 2

[37] Huogen Wang, Pichao Wang, Zhanjie Song, and Wanqing Li. Large-scale multimodal gesture segmentation and recognition based on convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3138–3146, 2017. 2

[38] Jiahui Wang, Zhenyou Wang, Shanna Zhuang, and Hui Wang. Cross-enhancement transformer for action segmentation. *arXiv preprint arXiv:2205.09445*, 2022. 1, 2, 3, 7, 8

[39] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 3

[40] Zhenzhi Wang, Ziteng Gao, Limin Wang, Zhifeng Li, and Gangshan Wu. Boundary-aware cascade networks for temporal action segmentation. In *European Conference on Computer Vision*, pages 34–51. Springer, 2020. 2

[41] Fei Wu, Xiao-Yuan Jing, Zhiyong Wu, Yimu Ji, Xiwei Dong, Xiaokai Luo, Qinghua Huang, and Ruchuan Wang. Modality-specific and shared generative adversarial network for cross-modal retrieval. *Pattern Recognition*, 104:107335, 2020. 2, 3, 5

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

[43] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: a survey. *arXiv preprint arXiv:2206.06488*, 2022. 3

[44] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3333–3343, 2022. 3

[45] Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. *British Machine Vision Conference (BMVC)*, 2021. 1, 2, 3, 4, 7, 8

[46] Junbin Zhang, Pei-Hsuan Tsai, and Meng-Hsun Tsai. Semantic2graph: Graph-based multi-modal feature for action segmentation in videos. *arXiv preprint arXiv:2209.05653*, 2022. 1, 2, 3, 7

[47] Yan Zhang, Siyu Tang, Krikamol Muandet, Christian Jarvers, and Heiko Neumann. Local temporal bilinear pooling for fine-grained action parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12005–12015, 2019. 2