

# ModSelect: Automatic Modality Selection for Synthetic-to-Real Domain Generalization

Zdravko Marinov, Alina Roitberg,  
David Schneider, and Rainer Stiefelhagen

Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology  
firstname.lastname@kit.edu

**Abstract.** Modality selection is an important step when designing multimodal systems, especially in the case of cross-domain activity recognition as certain modalities are more robust to domain shift than others. However, selecting only the modalities which have a positive contribution requires a systematic approach. We tackle this problem by proposing an unsupervised modality selection method (ModSelect), which does not require any ground-truth labels. We determine the correlation between the predictions of multiple unimodal classifiers and the domain discrepancy between their embeddings. Then, we systematically compute modality selection thresholds, which select only modalities with a high correlation and low domain discrepancy. We show in our experiments that our method ModSelect chooses only modalities with positive contributions and consistently improves the performance on a SYNTHETIC→REAL domain adaptation benchmark, narrowing the domain gap.

**Keywords:** Modality selection, domain generalization, action recognition, robust vision, synthetic-to-real, cross-domain

## 1 Introduction

Human activity analysis is vital for intuitive human-machine interaction, with applications ranging from driver assistance [55] to smart homes and assistive robotics [71]. Domain shifts, such as appearance changes, constitute a significant bottleneck for deploying such models in real-life. For example, while simulations are an excellent way of economical data collection, a SYNTHETIC→REAL domain shift leads to > **60%** drop in accuracy when recognizing daily living activities [70]. *Multimodality* is a way of mitigating this effect, since different types of data, such as RGB videos, optical flow and body poses, exhibit individual strengths and weaknesses. For example, models operating on body poses are less affected by appearance changes, as the relations between different joints are more stable given a good skeleton detector [23,22]. RGB videos, in contrast, are more sensitive to domain shifts [72,57] but also convenient since they cover the complete scene and video is the most ubiquitous modality [14,26,73,75].

Given the complementary nature of different data types (see Figure 2), we believe, that multimodality has a strong potential for improving domain generalization of activity recognition models, but *which modalities to select* and *how*

to fuse the information become important questions. Despite its high relevance for applications, the question of modality selection has been often overlooked in this field. The main goal of our work is to develop a systematic framework for studying the contribution of individual modalities in cross-domain human activity recognition. We specifically focus on the SYNTHETIC→REAL distributional shift [70], which opens new doors for economical data acquisition but comes with an especially large domain gap. We study five different modalities and examine how the prediction outcomes of multiple unimodal classifiers correlate as well as the domain discrepancy between their embeddings. We hope that our study will provide guidance for a better modality selection process in the future.

**Contributions and Summary.** We aim to make a step towards effective use of multimodality in the context of cross-domain activity recognition, which has been studied mostly for RGB videos in the past [14,26,73,75]. This work develops the modality selection framework ModSelect for quantifying the importance of individual data streams and can be summarized in two major contributions. (1) We propose a metric for quantifying the contribution of each modality for the specific task by calculating how the performance changes when the modality is included in the late fusion. Our new metric can be used by future research to justify decisions in modality selection. However, to estimate these performance changes, we use the ground-truth labels from the test data. (2) To detach ourselves from supervised labels, we propose to study the domain discrepancy between the embeddings and the correlation between the predictions of the unimodal classifiers of each modality. We use the discrepancy and the correlation to compute modality selection thresholds and show that these thresholds can be used to select only modalities with positive contributions w.r.t. our proposed metric in (1). Our unsupervised modality selection ModSelect can be applied in settings where no labels are present, e.g., in a multi-sensor setup deployed in unseen environments, where ModSelect would identify which sensors to trust.

## 2 Related Work

### 2.1 Multimodal Action Recognition

The usage of multimodal data represents a common technique in the field of action recognition, and is applied for both: increasing performance in supervised learning as well as unsupervised representation learning. Multimodal methods for action recognition include approaches which make use of video and audio [50,4,64,62,3], optical flow [39,64], text [52] or pose information [23,22,65]. Such methods can be divided into lower level *early / feature* fusion which is based on merging latent space information from multiple modality streams [43,94,56,1,2,63,34,84] and *late / score fusion* which combines the predictions of individual classifiers or representation encoders either with learned fusion modules [77,76,1,93,47,61] or with rule-based algorithms.

For this work, we focus on the latter, since the variety of early fusion techniques and learned late fusion impedes a systematic comparison, while rule-based

late fusion builds on few basic but successful techniques such as averaging single-modal scores [89,42,7,5,24,13,27,12,28], the max rule [45,5,27,67], product rule [42,78,45,48,25,91,81,27,67] or median rule. Ranking based solutions [30,31,66,20] like Borda count are less commonly used for action recognition but recognised in other fields of computer vision.

## 2.2 Modality Contribution Quantification

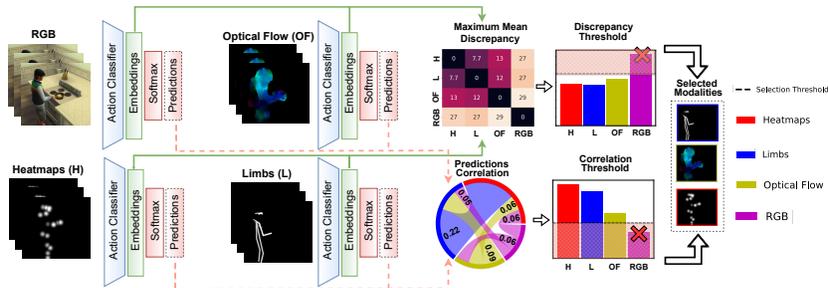
While the performance contribution of modalities has been analyzed in multiple previous works, e.g., by measuring the signal-to-noise ratio between modalities [83], determining class-wise modality contribution by learning an optimal linear modality combination [46,6] or extracting modality relations with threshold-based rules [80], in-depth analysis of modality contributions in the field of action recognition remains sparse and mostly limited to small ablation studies. Metrics to measure data distribution distances like Maximum Mean Discrepancy (MMD) or Mean Pairwise Distance (MPD) have been applied in fields like domain adaptation. MMD is commonly used to estimate and to reduce domain shift [60,54,35,79] and can be adapted to be robust against class bias, e.g. in the form of weighted MMD [87], Mean Pairwise Distance (MPD) was applied to analyze semantic similarities of word embeddings, e.g., in [29]. In this work, we introduce a systematic approach for analyzing modality contributions in the context of cross-domain activity recognition, which, to the best of our knowledge, has not been addressed in the past.

## 2.3 Domain Generalization and Adaptation

Both domain generalization and domain adaptation present strategies to learn knowledge from a source domain which is transferable to a given target domain. While domain adaptation allows access to data from the target domain to fulfill this task, either paired with labels [69,16,59] or in the form of unsupervised domain adaptation [10,15,69,16,59,19,18,74], domain generalization assumes an unknown target domain and builds upon methods which condition a neural network to make use of features which are found to be more generalizable [88], apply heavy augmentations to increase robustness [90] or explore different methods of leveraging temporal data [90].

## 3 Approach

Our approach consists of three main steps. (1) We extract multiple modalities and train a unimodal action recognition classifier on each modality. Afterwards, we evaluate all possible combinations of the modalities with different late fusion methods. We define the action recognition task in Section 3.1, the datasets we use in Section 3.2, and the modality extraction and training in Section 3.3. (2) In Section 3.4, we determine which modalities lead to a performance gain based on our evaluation results from (1). This establishes a baseline for the (3) third



**Fig. 1.** ModSelect: out approach for unsupervised modality selection which uses predictions correlations and domain discrepancy.

step (Section 3.5), where we show how to systematically select these beneficial modalities in an unsupervised way with our framework ModSelect - without the need of labels nor evaluation results. We offer an optional notation table in the Supplementary for a better understanding of all of our equations.

We intentionally do not make use of learned late fusion techniques, such as [77,76,1,93,47], since such methods do not allow for comparing the contribution of individual modalities. Instead, a specific learned late fusion architecture could be better suited to some modalities in contrast to others, overshadowing a neutral evaluation. However, our work can be used to select modalities upon which such learned late fusion techniques can be designed.

### 3.1 Action Recognition Task

Our goal is to produce a systematic method for unsupervised modality selection in multimodal action recognition. More specifically, we focus on SYNTHETIC→REAL domain generalization to show the need for a modality selection approach when a large domain gap is present. In this scenario an action classifier is trained only on samples from a SYNTHETIC source domain  $x_s \in X_s$  with action labels  $y \in Y$ . In domain generalization, the goal is to generalize to an unseen target domain  $X_t$ , without using any samples  $x_t \in X_t$  from it during training. In our case, the target domain  $X_t$  consists of REAL data and the source and target data originate from distinct probability distributions  $x_s \sim p_{synthetic}$  and  $x_t \sim p_{real}$ . The goal is to classify each instance  $x_t$  from the REAL target test domain  $X_t$ , which has a shared action label set  $Y$  with the training set. To achieve this, we use the synthetic Sims4Action dataset [70] for training and the real Toyota Smarthome (Toyota) [21] and ETRI-Activity3D-LivingLab (ETRI) [44] as two separate target test sets. We also evaluate our models on the Sims4Action official test split [70] in our additional SYNTHETIC→SYNTHETIC experiments.

### 3.2 Datasets

We focus on SYNTHETIC→REAL domain generalization between the synthetic Sims4Action [70] as a training dataset and the real Toyota Smarthome [21] and

ETRI [44] as test datasets. Sims4Action consists of ten hours of video material recorded from the computer game Sims 4, covering 10 activities of daily living which have direct correspondences in the two REAL datasets. Toyota Smarthome [21] contains videos of 18 subjects performing 31 different everyday actions within a single apartment, and ETRI [44] consists of 50 subjects performing 55 actions recorded from perspectives of home service robots in various residential spaces. However, we use only the 10 action correspondences to Sims4Action from the REAL datasets for our evaluation.

### 3.3 Modality Extraction and Training

We leverage the multimodal nature of actions to extract additional modalities for our training data, such as body pose, movement dynamics, and object detections. To this end, we utilize the RGB videos from the synthetic Sims4Action [70] to produce four new modalities - heatmaps, limbs, optical flow, and object detections. An overview of all modalities can be seen in Figure 2.

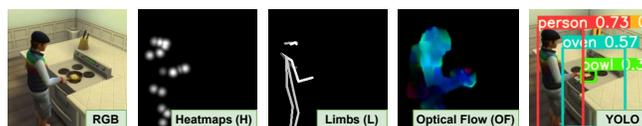
**Heatmaps and Limbs.** The heatmaps and limbs (H and L) are extracted via AlphaPose [32,51,86], which infers 17 joint locations of the human body. The heatmaps modality  $h(x, y)$  at pixel  $(x, y)$  is obtained by stacking 2D Gaussian maps, which are centered at each joint location  $(x_i, y_i)$  and each map is weighted by its detection confidence  $c_i$  as shown in Equation 1, where  $\sigma = 6$ .

$$h(x, y) := \exp\left(\frac{-((x - x_i)^2 + (y - y_i)^2)}{2\sigma^2}\right) \cdot c_i \quad (1)$$

The limbs modality is produced by connecting the joints with white lines and weighting each line by the smaller confidence of its endpoints. We weight both modalities by the detection confidences  $c_i$  so that uncertain and occluded body parts are dimmer and have a smaller contribution.

**Optical Flow.** The optical flow modality (OF) is estimated via the Gunnar-Farneback method [33]. The optical flow  $of(x, y)$  at pixel  $(x, y)$  encodes the *magnitude* and *angle* of the pixel intensity changes between two frames in the *value* and *hue* components of the HSV color space. The *saturation* is used to adjust the visibility and we set it to its maximum value. The heatmaps, limbs, and optical flow are all *image-based* and are used as an input to models which usually utilize RGB images.

**Object Detections.** Our last modality (YOLO) consists of object detections obtained by YOLOv3 [68], which detects 80 different objects. Unlike the other



**Fig. 2.** Examples of all extracted modalities. Note: the YOLO modality is represented as a vector  $\mathbf{v}$ , which encodes distances to the person’s detection (see Section 3.3).

modalities, we represent the detections as a vector, instead of an image. We show that such a simple representation achieves good domain generalization in our experiments. The YOLO modality for an image sample consists of a  $k$ -dimensional vector  $\mathbf{v}$ , where  $\mathbf{v}[i]$  corresponds to the reciprocal Euclidean distance between the person’s and the  $i^{\text{th}}$  object’s bounding box centers, and  $k = 80$  is the number of detection classes. This way, objects closer to the person have a larger weight in  $\mathbf{v}$  than ones which are further away. After computing the distances,  $\mathbf{v}$  is normalized by its norm:  $\mathbf{v} \leftarrow \mathbf{v}/\|\mathbf{v}\|$ . We denote the set of all modalities as  $\mathcal{M} := \{\text{H, L, OF, RGB, YOLO}\}$  and use the term  $\mathcal{M}$  in our equations.

**Training.** We train unimodal classifiers on each modality and evaluate all possible modality combinations with different late fusion methods. We utilize 3D-CNN models with the S3D backbone [85] for each one of the RGB, H, L, and OF modalities. The YOLO modality utilizes an MLP model as it is not image-based. We train all 5 action recognition models end-to-end on Sims4Action [70].

**Evaluation.** For our late fusion experiments, we combine the predictions of all unimodal classifiers at the class score level and obtain results for all  $\sum_{i=1}^5 \binom{i}{5} = 31$  modality combinations. We investigate 6 late fusion strategies - Sum, Squared Sum, Product, Maximum, Median, and Borda Count [40,8], which all operate on the class probability scores. Borda Count also uses the ranking of the class scores. For brevity, we refer to a late fusion of unimodal classifiers as a *multimodal classifier* and present our late fusion results in Section 4.1.

### 3.4 Quantification Study: Modality Contributions

In this section, we propose how to quantify the contributions of each modality based on the performance of the models on the target test sets. To this end, we propose a with-without metric, which computes the average difference  $f(m)$  of the performance of a multimodal classifier *with* a modality  $m$  to the performance *without* it. Formally the contribution  $f(m)$  of a modality  $m$  is defined as:

$$f(m) := \mathbb{E}_{C \in \mathcal{C}} [acc(C \cup \{m\}) - acc(C)] \quad (2)$$

where  $\mathcal{C}$  is the set of all modality combinations,  $acc(C)$  is the test accuracy of the multimodal classifier with the modality combination  $C$ , and  $m \in \mathcal{M}$ . We compute  $f(m)$  for all modalities based on the late fusion results listed in Section 4.1. The contribution of each modality can be used to determine the modalities, which positively influence the performance on the test dataset  $\mathcal{M}^+ \subseteq \mathcal{M}$ .

### 3.5 ModSelect: Unsupervised Modality Selection

In this section, we introduce our method ModSelect for unsupervised modality selection. In this setting, we assume that we do not have any labels in the target test domain  $X_t$ . This is exactly the case for SYNTHETIC→REAL domain generalization, where a model trained on simulated data is deployed in real-world conditions. In this case, the contribution of each modality cannot be estimated with Equation 2 as  $acc(C)$  cannot be computed without ground-truth labels.

Note that we do have labels in our test sets but we only use them in our quantification study and ignore them for our unsupervised experiments.

We propose ModSelect - a method for unsupervised modality selection based on the consensus of two metrics: (1) the correlation between the unimodal classifiers' predictions  $\rho$  and (2) the Maximum Mean Discrepancy (MMD) [36] between the classifiers' embeddings. We compute both metrics with our unimodal classifiers and propose how to systematically estimate modality selection thresholds. We show that the thresholds select the same modalities with positive contributions  $\mathcal{M}^+$  as our quantification study in Section 3.4.

**Correlation Metric.** We define the predictions correlation vector  $\rho_{mn}$  between modalities  $m$  and  $n$  as:

$$\rho_{mn} := \frac{\mathbb{E}[(\mathbf{z}_m - \boldsymbol{\mu}_m) \odot (\mathbf{z}_n - \boldsymbol{\mu}_n)]}{\boldsymbol{\sigma}_m \odot \boldsymbol{\sigma}_n} \quad (3)$$

where  $\mathbf{z}_m, \mathbf{z}_n$  are the softmax class scores of the action classifiers trained on modalities  $m$  and  $n$  respectively,  $(\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m), (\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n)$  are the mean and standard deviation vectors of  $\mathbf{z}_m, \mathbf{z}_n$ , and  $\odot$  is the element-wise multiplication operator. We define the predictions correlation  $\rho(m, n)$  between two modalities  $m, n$  as:

$$\rho(m, n) := \frac{1}{N} \sum_{i=1}^N \rho_{mn}[i] \quad (4)$$

where  $N = 10$  is the number of action classes.

**MMD Metric.** We show that the distance between the distributions of the embeddings of two unimodal classifiers can also be used to compute a modality selection threshold. The MMD metric [36] between two distributions  $P$  and  $Q$  over a set  $\mathcal{X}$  is formally defined as:

$$MMD(P, Q) := \|\mathbb{E}_{X \sim P}[\varphi(X)] - \mathbb{E}_{Y \sim Q}[\varphi(Y)]\|_{\mathcal{H}} \quad (5)$$

where  $\varphi : \mathcal{X} \mapsto \mathcal{H}$  is a feature map, and  $\mathcal{H}$  is a reproducing kernel Hilbert space (RKHS) [37,9,36]. For our empirical calculation of MMD between the embeddings  $\mathbf{h}_m, \mathbf{h}_n$  of two modalities  $m, n$  we set  $\mathcal{X} = \mathcal{H} = \mathbb{R}^d$  and  $\varphi(x) = x$ :

$$MMD(m, n) := \|\mathbb{E}[\mathbf{h}_m] - \mathbb{E}[\mathbf{h}_n]\| \quad (6)$$

where  $\mathbf{h}_m, \mathbf{h}_n$  are the embeddings from the second-to-last linear layer of the action classifiers for modalities  $m$  and  $n$  respectively, and  $d$  is the embedding size. Note that using a linear feature map  $\varphi(x) = x$  lets us determine only the discrepancy between the distributions' means. A linear mapping is sufficient to produce a good modality selection threshold, but one could also consider more complex alternatives, such as  $\varphi(x) = (x, x^2)$  or a Gaussian kernel [38].

We make the following observations regarding both metrics for modality selection. Firstly, a high correlation between correct predictions is statistically more likely than a high correlation between wrong predictions, since there is only 1 correct class and  $N - 1$  possibilities for error. We believe that a stronger correlation between the predictions results in a higher performance. Secondly, unimodal

classifiers should have a high agreement on easy samples and a disagreement on difficult cases [40]. A higher domain discrepancy between the classifiers’ embeddings has been shown to indicate a lower agreement on their predictions [53,92], and hence, a decline in performance when fused. We therefore believe that good modalities are characterized by a low discrepancy and high correlation.

**Modality Selection Thresholds.** After computing  $\rho(m, n)$  and  $MMD(m, n)$  for all pairs  $(m, n) \in \mathcal{M}^2$ , we systematically calculate modality selection thresholds for each metric  $\rho$  and MMD. We consider two types of thresholds: (1) an aggregated threshold  $\delta_{\text{agg}}$ , which selects a set of *individual modalities*  $\mathcal{M}_{\text{agg}} \subseteq \mathcal{M}$ , and (2) a pairs-threshold  $\delta_{\text{pair}}$ , which selects a set of *modality pairs*  $\mathcal{C}_{\text{pair}} \subseteq \mathcal{M}^2$ .

**Aggregated Threshold  $\delta_{\text{agg}}$ .** For the first threshold, we aggregate the  $\rho$  and  $MMD$  values for a modality  $m$  by averaging over all of its pairs:

$$\rho(m) := \frac{1}{|\mathcal{M}|} \sum_{n \in \mathcal{M}} \rho(m, n) \quad MMD(m) := \frac{1}{|\mathcal{M}|} \sum_{n \in \mathcal{M}} MMD(m, n) \quad (7)$$

Thus, we produce the sets  $A_\rho := \{\rho(m) | m \in \mathcal{M}\}$  and  $A_{MMD} := \{MMD(m) | m \in \mathcal{M}\}$ . A simple approach would be to use the mean or median as a threshold for  $A_\rho$  and  $A_{MMD}$ . However, such thresholds are sensitive to outliers (mean) or do not use all the information from the values (median). Additionally, one cannot tune the threshold with prior knowledge. To mitigate these issues, we propose to use the Winsorized Mean [41,82]  $\mu_\lambda(A)$  for both sets, which is defined as:

$$\mu_\lambda(A) := \lambda a_\lambda + (1 - 2\lambda)\bar{a}_\lambda + \lambda a_{1-\lambda} \quad (8)$$

where  $a_\lambda$  is the  $\lambda$ -percentile of  $A$ ,  $\bar{a}_\lambda$  is the  $\lambda$ -trimmed mean of  $A$ , and  $\lambda \in [0, 0.5]$  is a “trust” hyperparameter. A higher  $\lambda$  results in a lower contribution of edge values in  $A$  and a bigger trust in values near the center. Therefore, we set  $\lambda = 0.2$  as we have 5 modalities and expect to trust at least 3. We compute two separate thresholds  $\delta_{\text{agg}}^\rho := \mu_{0.2}(A_\rho)$  and  $\delta_{\text{agg}}^{MMD} := \mu_{0.2}(A_{MMD})$  and select the modalities  $\mathcal{M}_{\text{agg}}$  as a consensus between the two metrics as:

$$\mathcal{M}_{\text{agg}} := \{m \in \mathcal{M} | \rho(m) \geq \delta_{\text{agg}}^\rho \vee MMD(m) \leq \delta_{\text{agg}}^{MMD}\} \quad (9)$$

**Pairs-Threshold  $\delta_{\text{pair}}$ .** The second type of selection threshold skips the aggregation step of  $\delta_{\text{agg}}$  and directly computes the Winsorized Means  $\mu_{0.2}(\cdot)$  over the sets of all  $\rho(m, n)$  and  $MMD(m, n)$  values to obtain  $\delta_{\text{pair}}^\rho$  and  $\delta_{\text{pair}}^{MMD}$  respectively. This results in a selection of *modality pairs*, rather than individual modalities as in Equation 9. In other words, the  $\delta_{\text{pair}}$  thresholds are suitable when one is searching for the *best pairs* of modalities, and  $\delta_{\text{agg}}$  for the *best individual* modalities. The selected modality pairs  $\mathcal{C}_{\text{pair}}$  with this method are:

$$\mathcal{C}_{\text{pair}} := \{(m, n) \in \mathcal{M}^2 | \rho(m, n) \geq \delta_{\text{pair}}^\rho \vee MMD(m, n) \leq \delta_{\text{pair}}^{MMD}\} \quad (10)$$

**Summary of our Approach.** A summary of our unsupervised modality selection method ModSelect is illustrated in Figure 1. We use the embeddings of unimodal action classifiers to compute the Maximum Mean Discrepancy (MMD) between all pairs of modalities. We also compute the correlation  $\rho$  between the

predictions of all pairs of classifiers. We systematically estimate thresholds for MMD and  $\rho$  which discard certain modalities and select only modalities on which both metrics have a consensus. In the following Experiments 4 we show that the selected modalities  $m \in \mathcal{M}$  with our method ModSelect are exactly the modalities with a positive contribution  $f(m) > 0$  according to Equation 2, although our unsupervised selection does not utilize any ground-truth labels.

## 4 Experiments

### 4.1 Late Fusion: Results

We evaluate our late fusion multimodal classifiers following the cross-subject protocol from [21] for Toyota, the inter-dataset protocol from [49] for ETRI, and the official test split for Sims4Action from [70]. We follow the original SIMS4ACTION $\rightarrow$ TOYOTA evaluation protocol of [70] and utilize the mean-per-class accuracy (mPCA) as the number of samples per class are imbalanced in the REAL test sets. The mPCA metric avoids bias towards overrepresented classes and is often used in unbalanced activity recognition datasets [21,11,55,70].

The results from our evaluation are displayed in Table 1. The domain gap of transferring to REAL data is apparent in the drastically lower performance, especially on the ETRI dataset. Combinations including the H, L, or RGB modalities exhibit the best performance on the Sims4Action dataset, whereas OF and YOLO are weaker. However, combinations including the RGB modality seem to have an overall lower performance on the REAL datasets, perhaps due to the large appearance change. Combinations with the YOLO modality show the best performance for both REAL test sets. Inspecting the results in Table 1 is tedious and prone to misinterpretation or confirmation bias [58]. It is also possible to overlook important tendencies. Hence, we show in Section 4.2 how our quantification study tackles these problems by systematically disentangling the modalities with a positive contribution  $\mathcal{M}^+$  from the rest  $\mathcal{M}^-$ .

### 4.2 Quantification Study: Results

We use the results from Table 1 for the  $acc(\cdot)$  term in Equation 2 and compute the contribution of each modality  $f(m)$ . We do this for all late fusion strategies and all three test datasets and plot the results in Figure 3. The SYNTHETIC $\rightarrow$ REAL domain gap is clearly seen in the substantial difference in the height of the bars in the test split of Sims4Action [70] compared to the REAL test datasets. The limbs and RGB modalities have the largest contribution on Sims4Action [70], followed by the heatmaps. The only modalities with negative contributions are the optical flow and YOLO, where YOLO reaches a drastic drop of over ( $-10\%$ ) for the Squared Sum and Maximum late fusion methods. We conclude that YOLO and optical flow have a negative contribution on Sims4Action [70].

The results on the REAL test datasets Toyota [21] and ETRI [44] show different tendencies. The contributions are smaller due to the domain shift, especially

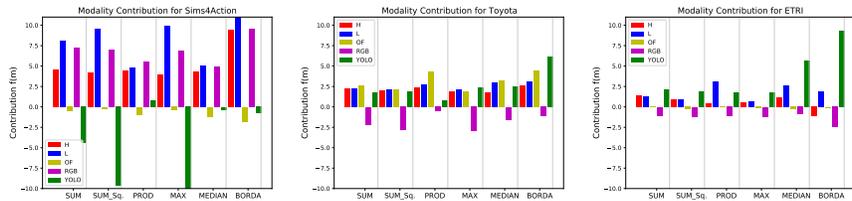
**Table 1.** Results for the action classifiers trained on Sims4Action [70] in the mPCA metric. The late fusion results are averaged over the 6 fusion strategies discussed in Section 3.3. H: Heatmaps, L: Limbs, OF: Optical Flow.

Test Set	mPCA [%]						
	SYNTHETIC			REAL			
	Sims4Action [70]	Toyota [21]	ETRI [44]	Test Set	Sims4Action [70]	Toyota [21]	ETRI [44]
Modalities				Modalities			
H	71.38	20.23	12.12	H L OF RGB	96.95	26.00	16.22
L	75.09	22.00	12.88	H L OF YOLO	90.05	28.98	20.27
OF	44.50	21.34	9.28	H L RGB YOLO	92.88	25.36	18.55
RGB	61.79	13.74	9.37	H OF RGB YOLO	91.14	26.10	17.20
YOLO	50.54	26.08	38.25	L OF RGB YOLO	92.56	26.12	18.56
H L	91.48	23.44	16.95	H L OF	94.03	26.70	17.26
H OF	89.86	25.97	15.10	H L RGB	95.72	23.25	15.78
H RGB	94.44	21.38	13.91	H OF RGB	95.41	23.73	13.79
L OF	90.84	25.60	15.90	L OF RGB	95.61	24.09	15.49
L RGB	94.42	20.58	15.46	H L YOLO	86.57	25.80	20.46
OF RGB	86.15	18.50	12.85	H OF YOLO	82.97	29.48	19.91
H YOLO	74.86	25.51	20.55	H RGB YOLO	88.79	23.27	17.77
L YOLO	80.25	24.77	19.91	L OF YOLO	86.31	28.51	20.68
OF YOLO	62.21	27.56	20.34	L RGB YOLO	90.59	23.29	19.09
RGB YOLO	76.23	17.29	19.41	OF RGB YOLO	82.37	21.15	16.92
H L OF RGB YOLO	94.27	27.52	18.53				

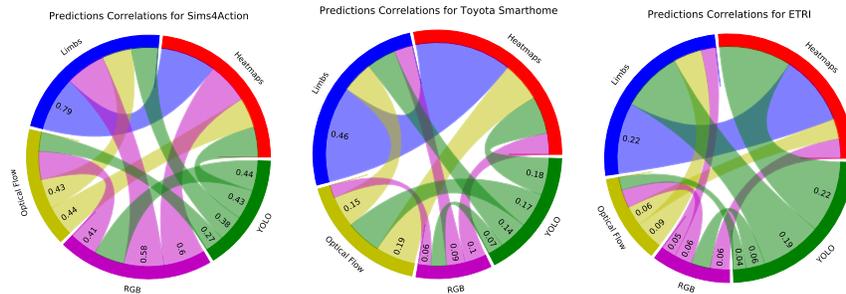
on the ETRI dataset. The RGB modality has explicitly negative contributions on both REAL datasets. We hypothesize that this is due to the appearance changes when transitioning from synthetic to real data. Apart from RGB, optical flow also has a consistently negative contribution on the ETRI dataset. An interesting observation is that the domain gap is much larger on ETRI than on Toyota. A reason for this might be that Roitberg et al. [70] design Sims4Action specifically as a SYNTHETIC→REAL domain adaptation benchmark to Toyota Smarthome [21], e.g., in Sims4Action the rooms are furnished the same way as in Toyota Smarthome. Our results indicate that optical flow has a negative contribution in ETRI, whereas RGB is negative in both REAL datasets. The average contribution of each modality over the 6 fusion methods is in Table 2(a).

### 4.3 Results from ModSelect: Unsupervised Modality Selection

Table 2(a) shows which modalities have a negative contribution on each target test dataset. However, to estimate these values we needed the performance, and



**Fig. 3.** Quantification Study: Quantification of the contribution of each modality for 6 late fusion methods and on three different test sets. The height of each bar corresponds to the contribution value  $f(m)$  which is computed with Equation 2.



**Fig. 4.** Chord plots of the prediction correlations  $\rho(m, n)$  for all modality pairs  $(m, n)$ . The thickness of each arch corresponds to the correlation  $\rho(m, n)$  between its two endpoint modalities  $m$  and  $n$ . Each value is computed according to Equation 4.

hence, the labels for the target test sets. In this section, we show how to select only the modalities with a positive contribution without using any labels.

**Predictions Correlation.** We utilize the predictions correlation metric  $\rho(m, n)$  and compute it for all modality pairs using Equation 4. The results for all datasets are illustrated in the chord diagrams in Figure 4. The chord diagrams allow us to identify the same tendencies, which we observed in our quantification study. Each arch connects two modalities  $(m, n)$  and its thickness corresponds to the value  $\rho(m, n)$ . The YOLO modality has the weakest correlations on Sims4Action [70], depicted in the thinner green arches. Optical flow also exhibits weaker correlations compared to the heatmaps, limbs, and RGB. We see significantly thinner arches for the RGB modality in both REAL test datasets, and for optical flow in ETRI, which matches our results in Table 2(a).

However, simply inspecting the chord plots is not a systematic method for modality selection. Hence, we first compute the aggregated correlations  $\rho(m)$  with Equation 7, which constitute the set  $A_\rho$ . Then, we compute the aggregated threshold  $\delta_{\text{agg}}^\rho := \mu_{0.2}(A_\rho)$  using the 0.2-Winsorized Mean [82] from Equation 8. We compute these terms for each test dataset and obtain three thresholds. The aggregated correlations  $\rho(m)$  and their thresholds  $\delta_{\text{agg}}^\rho$  for each test dataset are

**Table 2.** (a) Average contribution  $f(m)$  over the 6 late fusion methods of each modality  $m$ . Negative contributions are colored in red. (b) Aggregated prediction correlation  $\rho(m)$  values for each modality  $m$  on the three test datasets and the aggregated thresholds  $\delta_{\text{agg}}^\rho$ . Values below the threshold are colored in red. (c) Aggregated  $MMD(m)$  values for each modality  $m$  on the three test datasets and the aggregated thresholds  $\delta_{\text{agg}}^{MMD}$ . Values above the threshold are colored in red.

Test Dataset	(a) Contribution $f(m)$					(b) Aggregated $\rho(m)$					$\delta_{\text{agg}}^\rho$	(c) Aggregated $MMD(m)$				$\delta_{\text{agg}}^{MMD}$
	H	L	OF	RGB	YOLO	H	L	OF	RGB	YOLO		H	L	OF	RGB	
Sims4Action [70]	4.37	8.86	-0.74	6.98	-2.57	0.57	0.55	0.38	0.50	0.37	0.40	9.49	8.12	13.07	9.92	10.15
Toyota [21]	2.14	2.46	2.90	-1.86	2.13	0.23	0.21	0.14	0.08	0.14	0.10	11.93	11.47	13.34	20.79	14.38
ETRI [44]	0.76	1.60	-0.13	-1.17	2.02	0.14	0.14	0.06	0.05	0.13	0.08	17.84	17.76	22.04	24.91	20.64

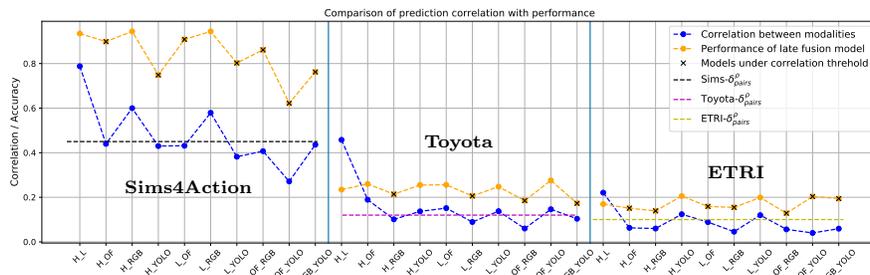
illustrated in Table 2(b). The modalities underneath the thresholds are exactly the ones with negative contributions from Table 2(a).

We also show that applying the pairs-threshold  $\delta_{\text{pair}}^\rho$  leads to the same results. We skip the aggregation step of  $\delta_{\text{agg}}^\rho$  and compose the  $A_\rho$  set out of the  $\rho(m, n)$  values, i.e. we focus on *modality pairs* instead of *individual modalities*. We compute the threshold  $\delta_{\text{pair}}^\rho := \mu_{0.2}(A_\rho)$  again with the 0.2-Winsorized Mean [82] from Equation 8. The correlation values  $\rho(m, n)$  as well as the accuracies of all bimodal action classifiers from Table 1 are shown in Figure 5. The pairs-thresholds  $\delta_{\text{pair}}^\rho$  for each test set are drawn as dashed lines and divide the modality pairs into two groups. The modality pairs below the thresholds are marked with an  $\times$  so that it is possible to identify which pairs are selected.

For Sims4Action [70], optical flow (OF) and RGB show a negative  $f(m)$  in Table 2(a) and are also discarded by  $\delta_{\text{agg}}^\rho$ . Figure 5 shows that all modality pairs containing either OF or RGB are below the threshold, i.e.  $\delta_{\text{pair}}^\rho$  discards the same modalities as  $\delta_{\text{agg}}^\rho$  for Sims4Action. The same is true for the REAL test datasets, where all models containing RGB are discarded in Toyota and ETRI, as well as OF for ETRI. Another observation is that the majority of "peaks" in the yellow *accuracy* lines coincide with the peaks in the blue *correlation* lines. This result is in agreement with our theory that a high correlation between correct predictions is statistically more likely than a high correlation between wrong predictions, since there is one only correct class and multiple incorrect ones. Moreover, while we do achieve the same results with both thresholds, we recommend using  $\delta_{\text{agg}}^\rho$  when discarding an entire input modality, e.g. a faulty sensor in a multi-sensor setup, and  $\delta_{\text{pair}}^\rho$  when searching for the *best synergies* from all modality combinations.

**Domain Discrepancy.** The second metric we use to discern the contributing modalities from the rest is the Maximum Mean Discrepancy (MMD) [36] between the embeddings of the action classifiers. Note that the YOLO modality is not included in this experiment as its MLP model’s embedding size is different than the other 4 *image-based* modalities. While MMD is widely used as a loss term for minimizing the domain gap between source and target domains [79,87,17], a large MMD is also associated with a decline in performance in fusion methods [53,92]. To utilize the MMD metrics to separate the modalities, we first compute  $MMD(m, n)$  for all modality pairs and the aggregated discrepancies  $MMD(m)$  with Equations 6 and 7. We then compute the pairs- $\delta_{\text{pair}}^{MMD}$  and aggregated  $\delta_{\text{agg}}^{MMD}$  thresholds with the 0.2-Winsorized Mean [82] from Equation 8 the same way as we did for the predictions’ correlations  $\rho$ .

The  $MMD(m, n)$  values for all modality pairs in the three test datasets are illustrated in Figure 6. The higher discrepancy values are clearly apparent by their bright colors and contrast to the rest of the values. Optical flow has the largest values on Sims4Action, and RGB has the highest discrepancy on Toyota and ETRI. Optical flow also exhibits a high discrepancy on ETRI. Once again, we can see that the domain gap to the ETRI dataset is much larger, which is manifested in drastically higher  $MMD(m, n)$  values. The pairs-thresholds  $\delta_{\text{pair}}^{MMD}$  for the three datasets are {13.68, 19.18, 27.50} and separate exactly the same modal-

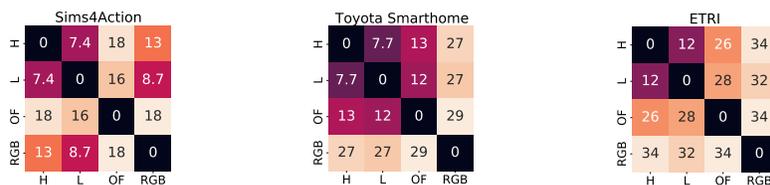


**Fig. 5.** Prediction correlations  $\rho(m, n)$  between all modality pairs and the late fusion accuracy of the bi-modal action classifiers. The pairs-thresholds  $\delta_{\text{pair}}^{\rho}$  are depicted as dashed lines. Pairs under the thresholds are crossed out with an  $\times$  in the yellow line.

ity pairs as our quantification study and the  $\delta_{\text{pair}}^{\rho}$  thresholds, with the exception of the (H,OF) pair in ETRI. The aggregated discrepancies  $MMD(m)$  for each modality and the aggregated thresholds are listed in Table 2(c), where the values above the thresholds are colored in red. The red values coincide exactly with the negative contributions  $f(m)$  from our quantification study in Table 2(a).

**ModSelect: Unsupervised Modality Selection.** Finally, we select the modalities with either the aggregated  $\delta_{\text{agg}}$  or the pairs-thresholds  $\delta_{\text{pair}}$ , by constructing the consensus between our two metrics  $\rho$  and  $MMD$  (see Equations 9 and 10). The selected modalities  $\mathcal{M}_{\text{agg}}$  with our aggregated thresholds  $\delta_{\text{agg}}$  and selected modality pairs  $\mathcal{C}_{\text{pair}}$  with our pairs-thresholds  $\delta_{\text{pair}}$  are listed in Table 3. The selected modalities  $\mathcal{M}_{\text{agg}}$  from our aggregated thresholds  $\delta_{\text{agg}}$  are exactly the ones with a positive contribution  $\mathcal{M}^+$  in Table 2(a) from our quantification study in Section 4.2. The pairs-thresholds  $\delta_{\text{pair}}$  have selected only modality pairs  $\mathcal{C}_{\text{pair}}$  which are constituted out of modalities from  $\mathcal{M}^+$ , which means that  $\mathcal{C}_{\text{pair}}$  contains only pairs of modalities  $(m, n)$  with positive contributions, i.e.  $f(m), f(n) > 0$ . In other words, our proposed unsupervised modality selection is able to select only the modalities with positive contributions by utilizing the predictions correlation and MMD between the embeddings of the unimodal action classifiers, without the need of any ground-truth labels on the test datasets.

**Impact on the multimodal accuracy.** The impact of the unsupervised modality selection on the mean multimodal accuracy can be seen in Table 3. Selecting the modalities with our proposed thresholds leads to an average improve-



**Fig. 6.** Maximum Mean Discrepancy values  $MMD(m, n)$  computed with Equation 6 for all modality pairs  $(m, n)$ . Warmer colors correspond to a higher discrepancy.

**Table 3.** Results from ModSelect: Selected modalities  $\mathcal{M}_{\text{agg}}$  with  $\delta_{\text{agg}}$  and selected modality pairs  $\mathcal{C}_{\text{pair}}$  with  $\delta_{\text{pair}}$  computed with Equations 9 and 10.

Test Dataset	$\mathcal{M}_{\text{agg}}$	$\mathcal{C}_{\text{pair}}$	Average multimodal accuracy	
			All Modalities $\mathcal{M}$	Ours: $\mathcal{M}_{\text{agg}}$
Sims4Action [70]	{H, L, RGB}		85.7%	90.9% (+5.2%)
Toyota [21]	{H, L, OF, YOLO}	$\{(m, n) \in \mathcal{M}^2   m \in \mathcal{M}^+ \wedge n \in \mathcal{M}^+\}$	22.9%	26.5% (+3.6%)
ETRI [44]	{H, L, YOLO}		17.7%	22.0% (+4.3%)

ment of 5.2%, 3.6%, and 4.3% for Sims4Action [70], Toyota [21], and ETRI [44] respectively. This is a substantial improvement, given the low accuracies on the REAL test datasets due to the synthetic-to-real domain gap. These results confirm that our modality selection approach is able to discern between good and bad sources of information, even in the case of a large distributional shift.

## 5 Limitations and Conclusion

**Limitations.** A limitation of our work is that the contributions quantification metric relies on the evaluation results on the test datasets, i.e., the ground-truth labels are needed to calculate the metric. Moreover, the with-without metric  $f(m)$  in Equation 2 requires  $\mathcal{O}(M * 2^{M-1})$  computations, where  $M$  is the number of modalities. However, one should note that in practice  $M$  is not too large, e.g.,  $M \leq 10$ . Additionally, since our method is novel, it has only been tested on the task of cross-domain action recognition. To safely apply our method to other multimodal tasks, e.g., object recognition, future investigations are needed. Moreover, the overall accuracy is still relatively low for cross-domain activity recognition (< 30%) and more research is needed for deployment-ready systems.

**Conclusion.** This is the first systematic study of modality selection in the context of cross-domain activity recognition, aimed at providing guidance for future work in multimodal domain generalization. Our experiments validate our assumption, that cross-domain activity recognition clearly benefits from multimodality, but not all modalities improve the recognition and a systematic modality selection is vital for achieving good results. We proposed a way to measure the contribution of each modality when it is included in a late fusion workflow. The contribution can be used to quantify the importance of each modality and to justify which sources of information are included in a multimodal framework. Our experiments indicate that the correlation between the predictions of unimodal classifiers and the Maximum Mean Discrepancy between their embeddings are both suitable metrics for unsupervised modality selection. The metrics allow to compute thresholds which select only modalities with positive contributions, which opens the possibility to automatically discard bad or uncertain sources of information and to improve the performance on unseen domains. We hope that our findings will provide guidance for a better modality selection process in the future, which is based on more structured and justified decisions.

**Acknowledgements.** This work was supported by the JuBot project sponsored by the Carl Zeiss Stiftung and Competence Center Karlsruhe for AI Systems Engineering (CC-KING) sponsored by the Ministry of Economic Affairs, Labour and Housing Baden-Württemberg.

## References

1. Ahmad, Z., Khan, N.: Human action recognition using deep multilevel multimodal ( $M^2$ ) fusion of depth and inertial sensors. *IEEE Sensors Journal* **20**(3), 1445–1455 (2019)
2. Ahmad, Z., Khan, N.: Cnn-based multistage gated average fusion (mgaf) for human action recognition using depth and inertial sensors. *IEEE Sensors Journal* **21**(3), 3623–3634 (2020)
3. Alayrac, J.B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., Zisserman, A.: Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems* **33**, 25–37 (2020)
4. Alwassel, H., Mahajan, D., Korbar, B., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems* **33** (2020)
5. Ardianto, S., Hang, H.M.: Multi-view and multi-modal action recognition with learned fusion. In: 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1601–1604. IEEE (2018)
6. Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. *Multimedia systems* **16**(6), 345–379 (2010)
7. Baradel, F., Wolf, C., Mille, J.: Human action recognition: Pose-based attention draws focus to hands. In: *IEEE International Conference on Computer Vision Workshops*. pp. 604–613 (2017)
8. Black, D., et al.: *The theory of committees and elections* (1958)
9. Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.P., Schölkopf, B., Smola, A.J.: Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics* **22**(14), e49–e57 (2006)
10. Busto, P.P., Iqbal, A., Gall, J.: Open set domain adaptation for image and action recognition. *IEEE transactions on pattern analysis and machine intelligence* **42**(2), 413–429 (2018)
11. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 961–970 (2015)
12. Cai, J., Jiang, N., Han, X., Jia, K., Lu, J.: Jolo-gcn: mining joint-centered lightweight information for skeleton-based action recognition. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 2735–2744 (2021)
13. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 6299–6308 (2017)
14. Chaquet, J.M., Carmona, E.J., Fernández-Caballero, A.: A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding* **117**(6), 633–659 (2013)
15. Chen, M.H., Kira, Z., AlRegib, G., Yoo, J., Chen, R., Zheng, J.: Temporal attentive alignment for large-scale video domain adaptation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6321–6330 (2019)
16. Chen, M.H., Li, B., Bao, Y., AlRegib, G.: Action segmentation with mixed temporal domain adaptation. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 605–614 (2020)

17. Chen, Y., Song, S., Li, S., Wu, C.: A graph embedding framework for maximum mean discrepancy-based domain adaptation algorithms. *IEEE Transactions on Image Processing* **29**, 199–213 (2019)
18. Choi, J., Sharma, G., Chandraker, M., Huang, J.B.: Unsupervised and semi-supervised domain adaptation for action recognition from drones. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 1717–1726 (2020)
19. Choi, J., Sharma, G., Schuler, S., Huang, J.B.: Shuffle and attend: Video domain adaptation. In: *European Conference on Computer Vision*. pp. 678–695. Springer (2020)
20. Cormack, G.V., Clarke, C.L., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: *International ACM SIGIR conference on Research and development in information retrieval*. pp. 758–759 (2009)
21. Das, S., Dai, R., Koperski, M., Minciullo, L., Garattoni, L., Bremond, F., Francesca, G.: Toyota smarthome: Real-world activities of daily living. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 833–842 (2019)
22. Das, S., Dai, R., Yang, D., Bremond, F.: Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021)
23. Das, S., Sharma, S., Dai, R., Bremond, F., Thonnat, M.: Vpn: Learning video-pose embedding for activities of daily living. In: *European Conference on Computer Vision*. pp. 72–90. Springer (2020)
24. Dawar, N., Kehtarnavaz, N.: A convolutional neural network-based sensor fusion system for monitoring transition movements in healthcare applications. In: *2018 IEEE 14th International Conference on Control and Automation (ICCA)*. pp. 482–485. IEEE (2018)
25. Dawar, N., Ostadabbas, S., Kehtarnavaz, N.: Data augmentation in deep learning-based fusion of depth and inertial sensing for action recognition. *IEEE Sensors Letters* **3**(1), 1–4 (2018)
26. Delaitre, V., Laptev, I., Sivic, J.: Recognizing human actions in still images: a study of bag-of-features and part-based representations. In: *BMVC 2010-21st British Machine Vision Conference* (2010)
27. Dhiman, C., Vishwakarma, D.K.: View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics. *IEEE Transactions on Image Processing* **29**, 3835–3844 (2020)
28. Duan, H., Zhao, Y., Chen, K., Shao, D., Lin, D., Dai, B.: Revisiting skeleton-based action recognition. *arXiv preprint arXiv:2104.13586* (2021)
29. Elekes, Á., Schäler, M., Böhm, K.: On the various semantics of similarity in word embedding models. In: *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*. pp. 1–10. IEEE (2017)
30. Emerson, P.: The original borda count and partial voting. *Social Choice and Welfare* **40**(2), 353–358 (2013)
31. van Erp, M., Vuurpijl, L., Schomaker, L.: An overview and comparison of voting methods for pattern recognition. In: *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*. pp. 195–200 (2002). <https://doi.org/10.1109/IWFHR.2002.1030908>
32. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: Regional multi-person pose estimation. In: *ICCV* (2017)

33. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In: Scandinavian conference on Image analysis. pp. 363–370. Springer (2003)
34. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10457–10467 (2020)
35. Ghifary, M., Kleijn, W.B., Zhang, M.: Domain adaptive neural networks for object recognition. In: Pacific Rim international conference on artificial intelligence. pp. 898–904. Springer (2014)
36. Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A kernel method for the two-sample-problem. *Advances in neural information processing systems* **19** (2006)
37. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. *The Journal of Machine Learning Research* **13**(1), 723–773 (2012)
38. Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., Sriperumbudur, B.K.: Optimal kernel choice for large-scale two-sample tests. *Advances in neural information processing systems* **25** (2012)
39. Han, T., Xie, W., Zisserman, A.: Self-supervised Co-training for Video Representation Learning (NeurIPS) (2020), <http://arxiv.org/abs/2010.09709>
40. Ho, T.K., Hull, J.J., Srihari, S.N.: Decision combination in multiple classifier systems. *IEEE transactions on pattern analysis and machine intelligence* **16**(1), 66–75 (1994)
41. Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in statistics, pp. 492–518. Springer (1992)
42. Imran, J., Kumar, P.: Human action recognition using rgb-d sensor and deep convolutional neural networks. In: 2016 international conference on advances in computing, communications and informatics (ICACCI). pp. 144–148. IEEE (2016)
43. Imran, J., Raman, B.: Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition. *Journal of Ambient Intelligence and Humanized Computing* **11**(1), 189–208 (2020)
44. Jang, J., Kim, D., Park, C., Jang, M., Lee, J., Kim, J.: Etri-activity3d: A large-scale rgb-d dataset for robots to recognize daily activities of the elderly. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10990–10997. IEEE (2020)
45. Kamel, A., Sheng, B., Yang, P., Li, P., Shen, R., Feng, D.D.: Deep convolutional neural networks for human action recognition using depth maps and postures. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **49**(9), 1806–1819 (2018)
46. Kampman, O., Barezi, E.J., Bertero, D., Fung, P.: Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction. arXiv preprint arXiv:1805.00705 (2018)
47. Kazakos, E., Nagrani, A., Zisserman, A., Damen, D.: Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5492–5501 (2019)
48. Khaire, P., Imran, J., Kumar, P.: Human activity recognition by fusion of rgb, depth, and skeletal data. In: Chaudhuri, B.B., Kankanhalli, M.S., Raman, B. (eds.) Proceedings of 2nd International Conference on Computer Vision & Image Processing. pp. 409–421. Springer Singapore, Singapore (2018)
49. Kim, D., Lee, I., Kim, D., Lee, S.: Action recognition using close-up of maximum activation and etri-activity3d livinglab dataset. *Sensors* **21**(20), 6774 (2021)

50. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 31*, pp. 7763–7774. Curran Associates, Inc. (2018)
51. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324* (2018)
52. Li, T., Wang, L.: Learning spatiotemporal features via video and text pair discrimination (2020)
53. Liang, T., Lin, G., Feng, L., Zhang, Y., Lv, F.: Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 8148–8156 (2021)
54. Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S.: Transfer feature learning with joint distribution adaptation. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2200–2207 (2013)
55. Martin, M., Roitberg, A., Haurilet, M., Horne, M., Reiß, S., Voit, M., Stiefelhagen, R.: Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2801–2810 (2019)
56. Memmesheimer, R., Theisen, N., Paulus, D.: Gimme signals: Discriminative signal encoding for multimodal activity recognition. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 10394–10401. IEEE (2020)
57. Munro, J., Damen, D.: Multi-modal domain adaptation for fine-grained action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 122–132 (2020)
58. Nickerson, R.S.: Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* **2**(2), 175–220 (1998)
59. Pan, B., Cao, Z., Adeli, E., Niebles, J.C.: Adversarial cross-domain action recognition with co-attention. In: *AAAI*. vol. 34, pp. 11815–11822 (2020)
60. Pan, S.J., Tsang, I.W., Kwok, J.T., Yang, Q.: Domain adaptation via transfer component analysis. *IEEE transactions on neural networks* **22**(2), 199–210 (2010)
61. Panda, R., Chen, C.F.R., Fan, Q., Sun, X., Saenko, K., Oliva, A., Feris, R.: AdaMML: Adaptive Multi-Modal Learning for Efficient Video Recognition. pp. 7576–7585 (2021), [https://openaccess.thecvf.com/content/ICCV2021/html/Panda\\_AdaMML\\_Adaptive\\_Multi-Modal\\_Learning\\_for\\_Efficient\\_Video\\_Recognition\\_ICCV\\_2021\\_paper.html](https://openaccess.thecvf.com/content/ICCV2021/html/Panda_AdaMML_Adaptive_Multi-Modal_Learning_for_Efficient_Video_Recognition_ICCV_2021_paper.html)
62. Patrick, M., Asano, Y., Fong, R., Henriques, J.F., Zweig, G., Vedaldi, A.: Multi-modal self-supervision from generalized data transformations. *ArXiv abs/2003.04298* (2020)
63. Pham, C., Nguyen, L., Nguyen, A., Nguyen, N., Nguyen, V.T.: Combining skeleton and accelerometer data for human fine-grained activity recognition and abnormal behaviour detection with deep temporal convolutional networks. *Multimedia Tools and Applications* **80**(19), 28919–28940 (2021)
64. Piergiovanni, A., Angelova, A., Ryoo, M.S.: Evolving losses for unsupervised video representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 133–142 (2020)
65. Rai, N., Adeli, E., Lee, K.H., Gaidon, A., Niebles, J.C.: Cocon: Cooperative-contrastive learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 3384–3393 (2021)

66. Ramanathan, M., Kochanowicz, J., Thalmann, N.M.: Combining pose-invariant kinematic features and object context features for rgb-d action recognition. *International Journal of Machine Learning and Computing* **9**(1), 44–50 (2019)
67. Rani, S.S., Naidu, G.A., Shree, V.U.: Kinematic joint descriptor and depth motion descriptor with convolutional neural networks for human action recognition. *Materials Today: Proceedings* (2021)
68. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* (2018)
69. Reiß, S., Roitberg, A., Haurilet, M., Stiefelhagen, R.: Deep classification-driven domain adaptation for cross-modal driver behavior recognition. In: *2020 IEEE Intelligent Vehicles Symposium (IV)*. pp. 1042–1047. IEEE (2020)
70. Roitberg, A., Schneider, D., Djamal, A., Seibold, C., Reiß, S., Stiefelhagen, R.: Let’s play for action: Recognizing activities of daily living by learning from life simulation video games. In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 8563–8569. IEEE (2021)
71. Roitberg, A., Somani, N., Perzylo, A., Rickert, M., Knoll, A.: Multimodal human activity recognition for industrial manufacturing processes in robotic workcells. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. pp. 259–266 (2015)
72. Sankaranarayanan, S., Balaji, Y., Jain, A., Lim, S.N., Chellappa, R.: Learning from synthetic data: Addressing domain shift for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3752–3761 (2018)
73. Sharma, G., Jurie, F., Schmid, C.: Discriminative spatial saliency for image classification. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3506–3513. IEEE (2012)
74. Song, X., Zhao, S., Yang, J., Yue, H., Xu, P., Hu, R., Chai, H.: Spatio-temporal contrastive domain adaptation for action recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9787–9795 (2021)
75. Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., Liu, J.: Human action recognition from various data modalities: A review. *arXiv preprint arXiv:2012.11866* (2020)
76. Wang, C., Yang, H., Meinel, C.: Exploring multimodal video representation for action recognition. In: *International Joint Conference on Neural Networks*. pp. 1924–1931. IEEE (2016)
77. Wang, L., Ding, Z., Tao, Z., Liu, Y., Fu, Y.: Generative multi-view human action recognition. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 6212–6221 (2019)
78. Wang, P., Li, W., Gao, Z., Zhang, Y., Tang, C., Ogunbona, P.: Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In: *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 595–604 (2017)
79. Wang, W., Li, H., Ding, Z., Wang, Z.: Rethink maximum mean discrepancy for domain adaptation. *arXiv preprint arXiv:2007.00689* (2020)
80. Wang, X., He, J., Jin, Z., Yang, M., Wang, Y., Qu, H.: M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics* **28**(1), 802–812 (2021)
81. Wei, H., Jafari, R., Kehtarnavaz, N.: Fusion of video and inertial sensing for deep learning-based human action recognition. *Sensors* **19**(17), 3680 (2019)

82. Wilcox, R.R., Keselman, H.: Modern robust data analysis methods: measures of central tendency. *Psychological methods* **8**(3), 254 (2003)
83. Wu, P., Liu, H., Li, X., Fan, T., Zhang, X.: A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion. *IEEE Transactions on Multimedia* **18**(3), 326–338 (2016)
84. Xiao, F., Lee, Y.J., Grauman, K., Malik, J., Feichtenhofer, C.: Audiovisual slowfast networks for video recognition. arXiv preprint arXiv:2001.08740 (2020)
85. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: *ECCV*. pp. 305–321 (2018)
86. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose Flow: Efficient online pose tracking. In: *BMVC* (2018)
87. Yan, H., Ding, Y., Li, P., Wang, Q., Xu, Y., Zuo, W.: Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2272–2281 (2017)
88. Yao, Z., Wang, Y., Wang, J., Yu, P., Long, M.: Videodg: Generalizing temporal relations in videos to novel domains. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021)
89. Ye, J., Li, K., Qi, G.J., Hua, K.A.: Temporal order-preserving dynamic quantization for human action recognition from multimodal sensor streams. In: *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. pp. 99–106 (2015)
90. Yi, C., Yang, S., Li, H., Tan, Y.p., Kot, A.: Benchmarking the robustness of spatial-temporal models against corruptions. arXiv preprint arXiv:2110.06513 (2021)
91. Zhao, C., Chen, M., Zhao, J., Wang, Q., Shen, Y.: 3d behavior recognition based on multi-modal deep space-time learning. *Applied Sciences* **9**(4), 716 (2019)
92. Zheng, Y.: Methodologies for cross-domain data fusion: An overview. *IEEE transactions on big data* **1**(1), 16–34 (2015)
93. Zou, H., Yang, J., Prasanna Das, H., Liu, H., Zhou, Y., Spanos, C.J.: Wifi and vision multimodal learning for accurate and robust device-free human activity recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. pp. 0–0 (2019)
94. Zou, Q., Wang, Y., Wang, Q., Zhao, Y., Li, Q.: Deep learning-based gait recognition using smartphones in the wild. *IEEE Transactions on Information Forensics and Security* **15**, 3197–3212 (2020)