Interact before Align: Leveraging Cross-Modal Knowledge for Domain Adaptive Action Recognition

Lijin Yang, Yifei Huang*, Yusuke Sugano, Yoichi Sato
Institute of Industrial Science, The University of Tokyo
{yang-lj,hyf,sugano,ysato}@iis.u-tokyo.ac.jp

Abstract

*Unsupervised domain adaptive video action recognition aims to recognize actions of a target domain using a model trained with only out-of-domain (source) annotations. The inherent complexity of videos makes this task challenging but also provides ground for leveraging multi-modal inputs (*e.g.*, RGB, Flow, Audio). Most previous works utilize the multi-modal information by either aligning each modality individually or learning representation via cross-modal self-supervision. Different from previous works, we find that the cross-domain alignment can be more effectively done by using cross-modal interaction first. Cross-modal knowledge interaction allows other modalities to supplement missing transferable information because of the* ***cross-modal complementarity****. Also, the most transferable aspects of data can be highlighted using* ***cross-modal consensus****.*

In this work, we present a novel model that jointly considers these two characteristics for domain adaptive action recognition. We achieve this by implementing two modules, where the first module exchanges complementary transferable information across modalities through the semantic space, and the second module finds the most transferable spatial region based on the consensus of all modalities. Extensive experiments validate that our proposed method can significantly outperform state-of-the-art methods on multiple benchmark datasets, including the complex fine-grained dataset EPIC-Kitchens-100.

1. Introduction

Unsupervised domain adaptation (UDA) models aim at learning features on the source dataset that can also be used on the target dataset. Due to its potential in reducing the necessity of large-scale labeling, UDA has been extensively explored for tasks such as image recognition [33,49,52,58], semantic segmentation [3,64] and object detection [5,8].

With one additional temporal dimension, video data is more complex than image data, and the domain gap not only resides in the appearance difference of environments but also in the motion variance when different people perform the same action. This prevents the direct application of image-based domain adaptation methods on the domain adaptive action recognition task [6, 20]. One direction to address this complexity is to use additional modality information (*e.g.* optical flow, audio). Other than directly combining multi-modal inputs [38], recent works add self-supervised modality alignment to implicitly learn properties of source and target data [24,36,47]. However, since the objectives of cross-modal alignment and cross-domain alignment are not perfectly consistent, simultaneously aligning *modality* and aligning *domain* can distract the learning target, *i.e.*, minimizing the *domain* discrepancy.

Due to different characteristics, the transferability (*i.e.*, invariance of feature across domains) of each modality lies in different and complementary perspectives. For example, for an action "wash cup" on the target domain, since the sound of water is similar across domains, the audio modality is more transferable to determine the verb "wash" of the action. Meanwhile, although RGB cannot perform as good as audio when recognizing the verb on the target domain, it can well recognize the noun "cup" on the target domain based on its domain-transferable appearance knowledge. If

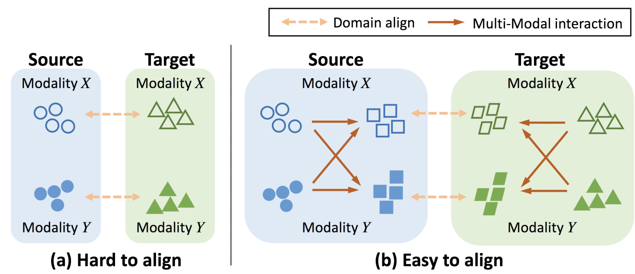

Figure 1. Different from existing UDA works that directly align the multi-modal inputs (a), we find that it is more effective to first enhance the transferability of each modality by cross-modal interaction, and then perform cross-domain alignment (b).

*Corresponding author.

these two modalities can interact with each other and exchange their unique domain-transferable knowledge, both of them can enhance their transferability and finally determine the action "wash cup" accurately. Based on this observation, we leverage this **cross-modal complementarity** and propose a Mutual Complementarity (MC) module that allows each modality to refine its feature by absorbing the transferable knowledge from other modalities, thus *the transferability of all modalities can be enhanced*.

Another aspect brought by multiple modalities is the **cross-modal consensus**. Since domain shift is often accompanied by changes of the scenario background, finding and focusing on more transferable foreground objects is critical. Rather than applying spatial attention like previous works [27, 55] which introduce additional parameters that also suffer from domain gap, we instead use a parameter-free correlation-based spatial consensus operation. Leveraging multi-modal features, we find and emphasize the transferable regions which share consensus among different modalities by developing a cross-modal Spatial Consensus (SC) module. Compared with spatial attention, our proposed consensus operation is proved in the experiments to be more suitable for domain adaptation.

We conduct experiments on the standard UCF-HMDB dataset and EPIC-Kitchens-55 dataset. Our experiments demonstrate that with cross-modal knowledge interaction, our proposed method can outperform state-of-the-art methods significantly. We also show that our method can bring remarkable enhancement on the EPIC-Kitchens-100 dataset that contains challenging fine-grained actions.
Our contributions can be summarized as follows:

- We propose a novel model to enhance multi-modality features for domain adaptive action recognition. To our best knowledge, this is the first work to consider cross-modal interaction for increasing the feature transferability across domains.
- We propose to use correlation-based operation to evaluate the transferability of spatial locations, which is proved to be simple and effective compared with spatial attention in the context of domain adaptation.
- Our proposed model achieves state-of-the-art performance on multiple datasets, including the challenging EPIC-Kitchens-100 dataset with fine-grained actions.

2. Related work

Unsupervised domain adaptation (UDA) other than action recognition. For solving the domain gap problem which exists widely in various applications such as object recognition [13, 41, 49], image segmentation [3, 14, 17, 60, 64], and natural language understanding [39, 44, 51], domain adaptation has been extensively studied especially in recent years. The goal of domain adaptation is to improve the performance on the target domain with a model trained on the source domain. Some works try to mitigate the domain gap from the input level by modifying the source input to become similar to the target domain via approaches like image-to-image translation [2, 37]. Another direction addresses this task from the representation level with Maximum Mean Discrepancy (MMD) [34] or adversarial training [52]. Very recently, self-supervised training becomes a new direction for domain adaptation [3, 22, 50]. Kang *et al*. [22] proposed to build the pixel-level cycle association between source and target pixel pairs for the task of domain adaptive semantic segmentation. Incorporating multiple modalities for UDA has been recently investigated for the task of emotion recognition and image retrieval [40]. They used single and multi-modal discriminators with cross-modality attention, showing that using multiple modalities can be more robust to domain shift compared with a single modality.

Action recognition and its UDA. Action recognition enjoyed a huge advance with the help of deep learning [4, 12, 19, 28, 35, 57]. Recent methods use multiple modalities such as RGB frames, optical flow and audio as input, and demonstrate the advantage of each modality [23]. Besides the rapid development in action recognition, domain adaptive action recognition also got a considerable amount of research attention. Most research works focus on the cross-viewpoint domain adaptation [25]. These works aim to adapt to the geometric transformations of a camera in the same environment, with optional additional information like the human pose [31] and temporal correspondence [45].

Another line of research focuses on the unsupervised domain adaptation for action recognition in different environments. These include methods that align source and target domain using hand-crafted features [11, 63], and recent works based on deep neural networks [1, 7, 32], using the RGB modality. Recently, several works [36, 38, 46, 59] explored the use of multiple modalities (RGB and flow) for domain adaptive action recognition. In [38], the authors use temporal alignment independently on each modality and only fuse modalities during inference. In [24, 36, 47], self-supervised alignment of modality is adopted. However, self-supervised modality alignment has a different learning target with domain adaptation, and simultaneously learning a model with two targets distracts the model from the primary task – minimizing the domain discrepancy.

In this work, we allow cross-modal interaction to increase the transferability by re-evaluating semantic transferability based on information from other modalities, and using the cross-modal spatial consensus to find the most transferable regions. Different from previous methods [24, 36, 47], our cross-modal interaction does not add self-supervision loss, so that the interaction can be optimized to solely improve the domain transferability.

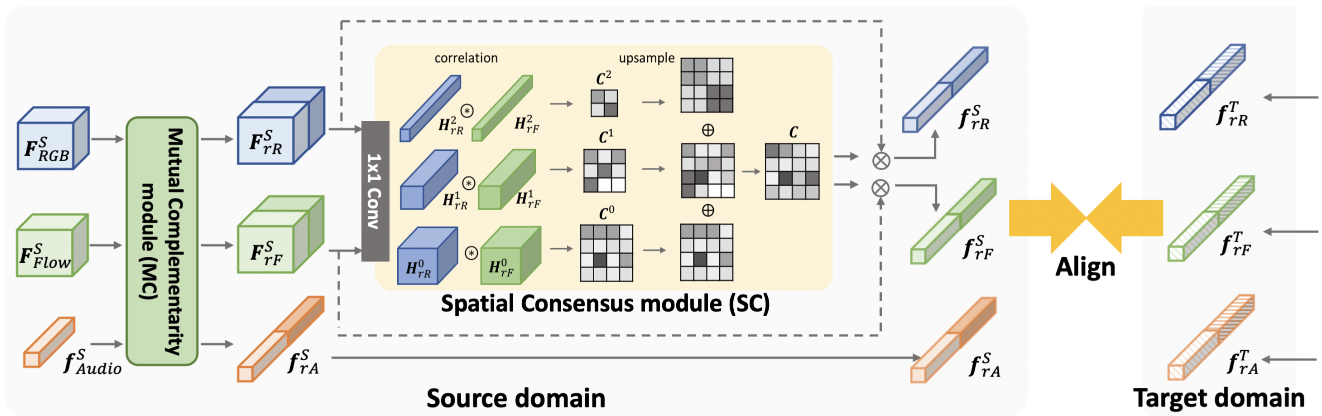

Figure 2. Overview of the proposed CIA model. We showcase three modalities RGB, Flow and Audio as input but it can be easily extended to add other modalities such as depth. In the figure, $\oplus$ denotes element-wise summation, $\otimes$ is element-wise multiplication, and $\circledast$ means the correlation operation that calculates the Pearson correlation coefficient on each spatial position.

3. Method

To effectively leverage the cross-modal complementarity and cross-modal consensus for domain adaptive action recognition, we propose a Cross-modal Interactive Alignment (CIA) model that first supplements each modality with cross-modal transferable knowledge by mutual semantic refinement and then emphasizes transferable regions by exploiting the consensus of multiple modalities.

Figure 2 depicts the overview of the proposed CIA model. In both source domain S and target domain T, for each modality of RGB, Flow and Audio, we first use backbone (omitted in the figure) networks to encode the input into frame-level features $\boldsymbol{F}^S_{RGB}, \boldsymbol{F}^S_{Flow}, \boldsymbol{f}^S_{Audio}$, $\boldsymbol{F}^T_{RGB}, \boldsymbol{F}^T_{Flow}$ and $\boldsymbol{f}^T_{Audio}$. We omit the notation of the domain identifier in the following part of this section when the operation is identical on both domains. We then use two modules, named Mutual Complementarity module (MC) and Spatial Consensus module (SC), to allow feature interaction for exploiting the cross-modal complementarity and the cross-modal consensus, respectively. The MC module exploits cross-modal complementarity by enabling one modality to receive transferable semantic knowledge from other modalities, utilizing two gating functions (Sec. 3.1). Then the SC module emphasizes transferable spatial regions which share consensus among all modalities by a multi-scale correlation operation (Sec. 3.2). Finally, we adopt adversarial feature alignment on the SC outputs to minimize the discrepancy between source and target domains.

3.1. The Mutual Complementarity (MC) module

Different modalities excel in their unique perspectives for perceiving the input, and the MC module aims to leverage this cross-modal complementarity to enhance the transferability of each modality by selecting and absorbing domain-transferable knowledge from other modalities. Transferable semantic knowledge lies in the feature channels [62], however gaps between modalities prevent direct channel-wise fusion methods like max-pooling or summation. In our proposed MC, we instead use a “summarize and re-evaluate” operation to leverage cross-modal transferable information.

Figure 3 depicts the proposed MC by showcasing the workflow of modality M. The output of MC is a transferability-refined feature of modality M $\boldsymbol{F}_{rM} \in \mathbb{R}^{2c\times h\times w}$, which is the concatenation of two parts: a cross-refined feature $\boldsymbol{F}_{cM}$ and a self-refined feature $\boldsymbol{F}_{sM}$.

$\boldsymbol{F}_{cM}$ represents the feature of modality M refined by transferable knowledge from other modalities. For getting $\boldsymbol{F}_{cM}$, we first apply global average pooling on features of other modalities and concatenate them to obtain a cross-modal knowledge representation $\boldsymbol{f}^{in}_M$. With $\boldsymbol{f}^{in}_M$, we summarize the domain-transferable knowledge and re-evaluate the semantic transferability of modality M by a cross-gating function [16]:

$$\boldsymbol{t}_{cM} = \sigma\boldsymbol{W}^{in}_2(\delta(\boldsymbol{W}^{in}_1\boldsymbol{f}^{in}_M)), \tag{1}$$

$$\boldsymbol{F}_{cM} = \boldsymbol{F}_M \cdot \boldsymbol{t}_{cM}, \tag{2}$$

where $\boldsymbol{W}^{in}_1, \boldsymbol{W}^{in}_2$ are weight matrices, $\cdot$ is channel-wise multiplication, σ and δ denotes the sigmoid and ReLU activations, respectively. Here $\boldsymbol{t}_{cM}$ is the re-evaluation of semantic transferability of modality M by using cross-modal knowledge. $\boldsymbol{t}_{cM}$ serves as the “advice” from other modalities to emphasize the channels of $\boldsymbol{F}_M$ by channel-wise multiplication.

Although this gating mechanism is simple, it can learn nonlinear interaction between channels, while allowing multiple channels to be emphasized during the re-evaluation. This helps the gating operations to first summarize domain-transferable knowledge (using $\boldsymbol{W}^{in}_1$) and then re-weight the channels of $\boldsymbol{F}_M$ utilizing the summarized knowledge (using $\boldsymbol{W}^{in}_2$ and channel-wise multiplication).

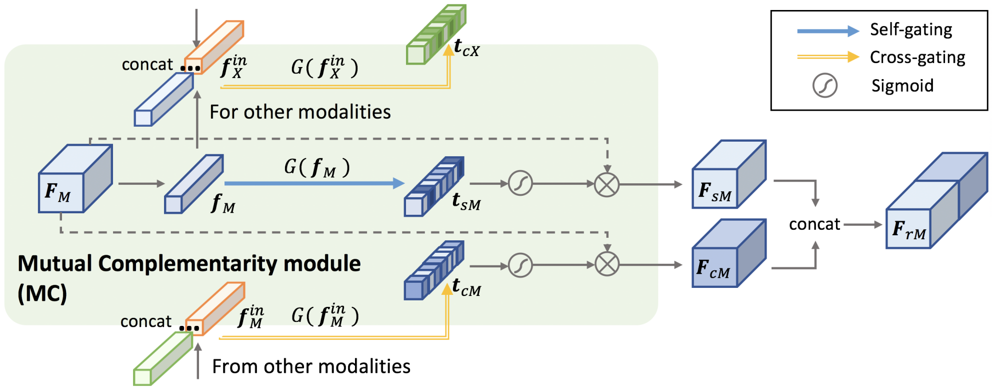

Figure 3. The Mutual Complementarity module (MC) showcased using modality M. M could be any modalities of RGB, Flow and Audio, also can be extended to other modalities if available, *e.g.*, depth.

When receiving the complementary knowledge from other modalities, it is also important for modality M to preserve unique information and modality characteristics of itself. Thus, in addition to cross-gating, we use a self-gating operation to perform a self re-evaluation of modality M:

$$\boldsymbol{t}_{sM} = \sigma \boldsymbol{W}_2^M(\delta(\boldsymbol{W}_1^M \boldsymbol{f}_M)), \quad \boldsymbol{F}_{sM} = \boldsymbol{F}_M \cdot \boldsymbol{t}_{sM}, \tag{3}$$

To summarize domain-transferable knowledge while preventing the domain adaptation model from overfitting on the source domain, the MC only introduces a small number of model parameters by leveraging bottleneck during gating. In other words, we reduce the dimension by a ratio r via making $\boldsymbol{W}_1 \in \mathbb{R}^{\frac{c}{r} \times c}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{c \times \frac{c}{r}}$. Finally, we get the transferability-refined feature of modality M by fusing the two refined features $\boldsymbol{F}_{sM}$ and $\boldsymbol{F}_{cM}$ via concatenation:

$$\boldsymbol{F}_{rM} = Concat(\boldsymbol{F}_{sM}, \boldsymbol{F}_{cM}). \tag{4}$$

We show the analysis of model parameters and computational complexity in the supplementary.

3.2. The Spatial Consensus (SC) module

To further enhance feature transferability by focusing on the most transferable spatial regions (*e.g.* foreground objects), previous works mainly use the spatial attention mechanism [18,27,55]. However, this introduces additional model parameters which will also be affected by domain shift. Different from spatial attention, we propose a Spatial Consensus (SC) module to highlight the transferable regions that have shared consensus among modalities.

Our idea to find transferable locations is letting multiple modalities work with "collective wisdom". Since features $\boldsymbol{F}_{rR}$ and $\boldsymbol{F}_{rF}$ encode different information, we first map these features into the same latent space to get transferability estimations from their own perspective. Then we compute the feature similarity using a correlation operation to measure whether two modalities share the same opinion about spatial transferability. For each location, the feature similarity is high only if two modalities both find this location to be transferable.

Since transferable regions vary in size in different samples, we compute the correlation of feature maps at different scales [30]: the features $\boldsymbol{H}_{rR}$ and $\boldsymbol{H}_{rF}$ are first downsampled by a factor of 2 k times, resulting in two groups of feature maps $\{\boldsymbol{H}_{rR}^0, \boldsymbol{H}_{rR}^1, ...\boldsymbol{H}_{rR}^k,\}$ and $\{\boldsymbol{H}_{rF}^0, \boldsymbol{H}_{rF}^1, ..., \boldsymbol{H}_{rF}^k,\}$. For each scale k, we compute the Pearson correlation coefficient on each spatial position (i,j) as:

$$\boldsymbol{C}^{k,(i,j)} = \frac{\boldsymbol{H}_{rR}^{k,(i,j)} * \boldsymbol{H}_{rF}^{k,(i,j)}}{\|\boldsymbol{H}_{rR}^{k,(i,j)}\|^2 \times \|\boldsymbol{H}_{rF}^{k,(i,j)}\|^2}, C^k \in \mathbb{R}^{\frac{w}{2^k} \times \frac{h}{2^k}} \tag{5}$$

where $*$ indicate dot product. It is important that SC contains only a small number of parameters so that most of the representation is learned in the MC while also preventing overfitting. To this end, we choose to use correlation instead of spatial attention [56].

Finally, all the correlation maps $\{\boldsymbol{C}^0, \boldsymbol{C}^1, ..., \boldsymbol{C}^k\}$ are upsampled to match the size as $\boldsymbol{F}_{rR}$ and then summed together to form a consensus map $\boldsymbol{C}$. The consensus map $\boldsymbol{C}$ is then used as a spatial weight map for the weighted average pooling of $\boldsymbol{F}_{rR}$ and $\boldsymbol{F}_{rF}$. We also add residual connections following [15,53], forming feature vectors $\boldsymbol{f}_{rR}$ and $\boldsymbol{f}_{rF}$. Since MC already involves audio information and $\boldsymbol{f}_{rA}$ does not contain spatial dimensions, $\boldsymbol{f}_{rA}$ is not used in this module. During training, the SC module will encourage the network to extract features such that the spatial correlation becomes higher for locations more helpful for domain alignment.

3.3. Adversarial Domain Alignment

We apply adversarial domain alignment on three transferability enhanced features $\boldsymbol{f}_{rR}$, $\boldsymbol{f}_{rF}$ and $\boldsymbol{f}_{rA}$, individually. Denote the two-layer MLP based discriminator as D, the discriminator loss can be written as:

$$\mathcal{L}_{fd} = \sum_{M\in\{rR,rF,rA\}} \sum_{\boldsymbol{f}_M\in S,T} -d\log(D_M(\boldsymbol{f}_M)) - (1-d)\log(1-D(\boldsymbol{f}_M)) \tag{6}$$

where d is the binary domain label, S, T denotes the source and target domains respectively, and $\boldsymbol{f}_M$ represents one of the features in $\{\boldsymbol{f}_{rR}, \boldsymbol{f}_{rF}, \boldsymbol{f}_{rA}\}$.

We average the frame-wise features to form video-level features $\boldsymbol{v}_{rR}$, $\boldsymbol{v}_{rF}$ and $\boldsymbol{v}_{rA}$ and fuse them as $\boldsymbol{v}_{mm}$. The domain alignment is also done on the video-level features $\boldsymbol{v}_{rR}$, $\boldsymbol{v}_{rF}$ and $\boldsymbol{v}_{rA}$ and its loss is denoted as $\mathcal{L}_{vd}$.

On the source domain, we apply the standard classification loss on the fused video-level feature $\boldsymbol{v}_{mm}$:

$$\mathcal{L}_y = -\sum_{\boldsymbol{v}_{mm}\in S} \boldsymbol{y}\log P(G_M(\boldsymbol{v}_{mm})), \tag{7}$$

where G_M represents the linear action classifier for the corresponding feature.

As a result, our full loss function is a combination of $\mathcal{L}_y$, $\mathcal{L}_{fd}$ and $\mathcal{L}_{vd}$:

$$\mathcal{L} = \lambda_y\mathcal{L}_y + \lambda_{fd}\mathcal{L}_{fd} + \lambda_{vd}\mathcal{L}_{vd} \tag{8}$$

4. Experiments

4.1. Dataset and Implementation Details

We validate our proposed CIA model on three representative domain adaptive action recognition datasets: UCF-HMDB [26, 48] (**U-H**) is one widely used dataset that contains 12 action classes. We use the full version [6] in our experiments. **H** $\rightarrow$ **U** indicates the source dataset is HMDB while the target dataset is UCF, and vice versa. We also use the EPIC-Kitchens-55 (**E55**) as another benchmark dataset. To make a fair comparison with [24, 36, 47], we follow the same setting as [36]. Class-wise action recognition accuracy is used as the evaluation metric on these two datasets.

Additionally, EPIC-Kitchens-100 [10] (**E100**) is a newly released dataset with fine-grained actions taken from the first-person perspective. This dataset is extremely challenging because (1) source and target actions are performed by different individuals in different kitchens. (2) The first-person viewpoint often makes the action happen in a non-salient region, and (3) the annotation is fine-grained. There are 16115/26115 training videos for source/target domains and 7906 clips as the target-validation split. 97 verb classes, 300 noun classes form a total of 3369 fine-grained action classes. We further add experiments on this dataset since its large-scale and fine-grained property makes it more suitable for analyzing model performance. Following the protocol in [10], we use the accuracy of verb, noun and action as the evaluation metric.

Implementation Details For a fair comparison, we use two backbones for feature extraction: I3D [4] pretrained on Kinetics and TBN [23] pretrained on Kinetics then fine-tuned on the source training set of the according dataset.

Modality	Backbone	Method	U→H	H→U
RGB	R-TRN	TA^{3}N [6]	78.33	81.79
	R-TRN	TCoN [38]	87.24	89.06
	I3D	SAVA [9]	82.20	91.20
	I3D-TRN	TA^{3}N [6]	82.78	91.77
Flow	I3D-TRN	TA^{3}N [6]	82.50	90.89
R+F	I3D	Avg$^\diamond$	83.61	91.07
	I3D	G-blend [54]	84.72	91.24
	I3D	MMTM [21]	85.83	92.47
	I3D	MM-SADA [36]	84.20	91.10
	I3D	STCDA [47]	83.10	92.10
	I3D	Kim *et al.* [24]	84.70	92.80
	I3D	CIA source only$^\diamond$	86.11	92.47
	I3D	CIA (Ours)$^\diamond$	**88.33**	**94.05**
	I3D	Concat*	86.11	92.99
	I3D	CIA source only*	85.83	93.52
	I3D	CIA (Ours)*	**90.56**	**94.22**
	I3D-TRN	TA^{3}N [6]*	89.17	92.81
	I3D-TRN	CIA (Ours)*	89.72	93.17
	I3D-TRN	CIA +TA^{3}N*	**91.94**	**94.57**
	I3D	CIA target only*	96.83	99.12

Table 1. Performance comparison on the UCF-HMDB (**U-H**)) dataset. $^\diamond$ refers to averaging the outputs from each modality classifier, while * means concatenate features of different modalities.

The MC processes the feature with dimension $c = 1024$, and the ratio for gating bottleneck is $r = 16$. We use either average or concatenate as the late fusion methods based on datasets. For all experiments, we train the model on 4 NVIDIA-V100 GPUs. Other dataset-specific details can be found in the supplementary.

4.2. Comparison with state-of-the-art

We compare our CIA model with the following methods:

- Multi-modal UDA action recognition methods. We compare with three recent methods **MM-SADA** [36], **STCDA** [47] and **Kim *et al.*** [24]. These methods show state-of-the-art performance in the UDA action recognition task.
- Single-modal UDA action recognition methods [6, 9, 20, 29, 33, 38, 42]. For better comparison, we follow [10] to enable **TA^{3}N** [6] with multi-modality input and use TRN [61] on the backbone for temporal feature fusion.
- Multi-modal fusion methods for other tasks. To better evaluate our CIA's ability on using multi-modal information in the scope of domain adaptation, other than direct fusion via average (**Avg**) or concatenation (**Concat**), we add comparison with previous multi-modal fusion methods **G-blend** [54] and **MMTM** [21]. Since [21, 54] are not originally designed for domain adaptation, we use their method on the same adversarial alignment framework with our method for a fair comparison.

Method	D1→D2	D1→D3	D2→D1	D2→D3	D3→D1	D3→D2	mean
Ours Source only	43.2	42.5	43.0	48.0	43.0	55.5	45.9
MMD [33]	46.6	39.2	43.1	48.5	48.3	55.2	46.8
AdaBN [29]	47.0	40.3	44.6	48.8	47.8	54.7	47.2
MCD [42]	46.5	43.5	42.1	51.0	47.9	52.7	47.3
DAAA [20]	50.0	43.5	46.5	51.5	51.0	53.7	49.4
MM-SADA [36]	49.5	44.1	48.2	52.7	50.9	56.1	50.3
Kim *et al.* [24]	50.3	46.3	49.5	52.0	51.5	56.3	51.0
STCDA [47]	52.0	45.5	49.0	52.5	**52.6**	55.6	51.2
CIA (Ours)	**52.5**	**47.8**	**49.8**	**53.2**	52.2	**57.6**	**52.2**
Ours target only	71.6	73.6	63.3	73.6	63.3	71.6	69.5

Table 2. Performance comparison on the EPIC-Kitchens-55 (**E55**) dataset.

Results on U-H dataset are shown in Table 1. From the table, because of the inherent difficulty of video data, multi-modal methods generally surpass single modality methods [6, 9, 38]. Meanwhile, previous multi-modal fusion works **G-blend** [54] and **MMTM** [21] do not perform well in the domain adaptation setting, suggesting that our proposed CIA model better suits the task of domain adaptation. Our method significantly outperforms previous state-of-the-art multi-modal works **MM-SADA**, **STCDA** and **Kim *et al.***. Compared with **Kim *et al.***, we can increase the accuracy from 84.70 to 88.33 on **U**→ **H** and 92.80 to 94.05 on **H**→ **U**. This indicates the superiority of our CIA model in leveraging multi-modal interaction compared with self-supervised learning.

We also validate different late fusion methods by comparing average$^{\diamond}$ and concatenation*. We found that using concatenation for late modality fusion can be more helpful. Using TRN [61] as a more sophisticated temporal aggregation method, our method outperforms TA^{3}N on both datasets. Since our method can be flexibly fitted into any domain adaptation framework, we can further enhance TA^{3}N by adding our model, achieving 91.94 and 94.57 on the two datasets.

Results on E55 dataset are illustrated in Table 2. We average the outputs of individual modality classifiers as the late fusion method for a fair comparison with prior works. Using cross-modal self-supervision, MM-SADA, STCDA and Kim *et al*. cannot perform as good as our proposed method. This proves our assumption that simultaneously optimizing cross-modal alignment and cross-domain alignment can distract the modal from minimizing the domain gap. However, by interacting before alignment, our method can better leverage the cross-modal complementarity and cross-modal consensus, thus boosting the mean accuracy by up to 1% compared with the previous state-of-the-art.

Results on E100 dataset Table 3 demonstrates the performance comparison with state-of-the-art methods on the challenging **E100** validation set. We average the scores of each modality for late fusion when implementing methods

Modality	Backbone	Method	Verb	Noun	Action
R+F	I3D	Source only	39.28	22.28	11.62
	I3D	MM-SADA [36]	40.41	23.92	12.80
	I3D	Source only	40.17	22.89	12.27
	I3D	CIA (Ours)	**42.35**	**24.49**	**14.25**
	TBN	Source only	42.41	27.26	16.03
	TBN	DAAA [20]	42.99	27.38	16.32
	TBN	Source only	42.98	27.49	16.44
	TBN	CIA (Ours)	**43.93**	**27.54**	**17.01**
	TBN-TRN	Source only	43.78	26.65	16.70
	TBN-TRN	TA^{3}N [6]	44.88	27.41	17.39
	TBN-TRN	Source only	44.12	27.12	16.86
	TBN-TRN	CIA (Ours)	**45.23**	**27.75**	**18.02**
R+F+A	TBN-TRN	Source only	46.67	27.57	19.00
	TBN-TRN	TA^{3}N [6]	47.43	28.40	19.42
	TBN-TRN	Source only	47.69	28.48	19.61
	TBN-TRN	CIA (Ours)	**48.34**	**29.50**	**20.30**
	TBN	Source only	47.10	28.30	18.66
	TBN	DAAA [20]	47.96	29.08	19.19
	TBN	Source only	48.22	29.86	19.73
	TBN	CIA (Ours)	**49.08**	**30.36**	**20.49**

Table 3. Performance comparison on the EPIC-Kitchens-100 (**E100**) validation set. R, F and A refers to RGB, Flow and Audio modalities, respectively. We show each method together with its source only performance in the row above.

on the I3D backbone, while we use concatenation for methods on other backbones. Using RGB and Flow modalities and the same backbone, our proposed method performs favorably against the state-of-the-art method MM-SADA [36] by 1.45% in terms of the accuracy of action. When using RGB, Flow and audio modalities, our method can show more significant improvements over previous works on all of the verb, noun and action metrics.

4.3. Visualization

To better understand the proposed CIA model, in Figure 4 we show the Grad-CAM [43] visualizations of activation maps before and after cross-modality feature refinement by the MC module. From these cases we can clearly see the benefit of feature interaction with other modalities: in (a-1) and (a-2), other modalities help the RGB modality

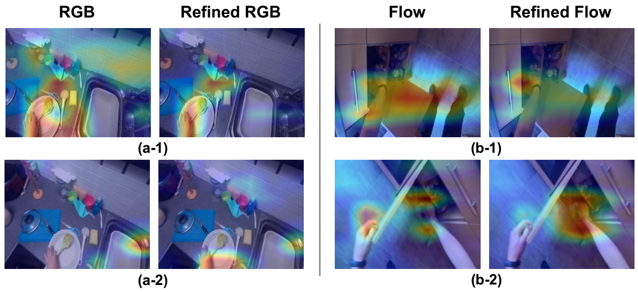

Figure 4. Grad-CAM [43] visualizations of features before and after cross-modality feature refinement by MC. The ground-truth actions are: (a-1) take spoon, (a-2) move spoon, (b-1) take garlic, (b-2) take oil. (a-1) and (a-2) show RGB activation maps (left) and the activation map of RGB modality refined by other modalities (right). Similarly, (b-1) and (b-2) depict the activation maps of the Flow modality alone and Flow refined by other modalities.

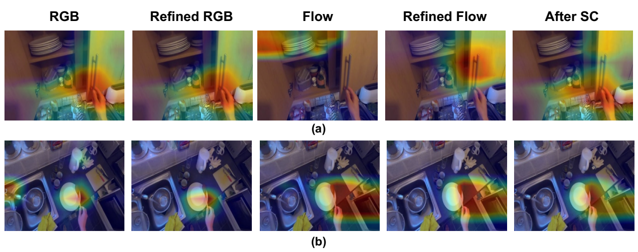

Figure 5. Grad-CAM [43] visualizations of RGB, refined RGB, Flow, refined Flow and fused modality after SC. The ground-truth action labels are: (a) open cupboard, (b) put down spoon.

to put more focus on the hand by suppressing the attention on other objects. In (b-1), the refined Flow modality transfers its focus from foot to hand, and in (b-2) from left hand to right hand. These examples strongly prove that cross-modal transferable knowledge helps each modality to perform better on the target domain.

We also visualize the activation maps after the SC module to qualitatively evaluate its effectiveness. In the action "put down spoon" shown in Figure 5(b), the RGB modality is guided by other modalities to ignore the tap, and the refined Flow feature becomes more focused in the center. And finally our SC module can find the best focus by taking advantage of consensus from all modalities.

4.4. Ablation Study

Contribution of each module In this section, we conduct an ablation study on the **E100** validation set to examine the contribution brought by each module. We test our method w/o the MC or SC module, and also test whether to use self- or cross-refined features within the MC module.

Results can be seen in Table 4. Compared with the base setting (1st row), both self-refinement (2nd row) and cross-refinement (3rd row) benefit from the "summarize and re-evaluation" operation, while combining the self- and cross-

MC	SC	Verb	Noun	Action
×	×	47.96	29.08	19.19
Self-refinement	×	48.01	29.31	19.56
Cross-refinement	×	48.48	29.48	19.67
✓	×	48.62	29.96	19.98
×	✓	48.66	29.79	19.83
✓	✓	49.08	30.36	20.49

Table 4. Ablation study on Mutual Complementarity module (MC) and Spatial Consensus module (SC) of our CIA model.

Setting	Module	Verb	Noun	Action
Source only	Avg	47.10	28.30	18.66
	Att [56]	47.32	28.85	19.21
	SC	**47.85**	**29.18**	**19.55**
Domain Adaptation	Avg	47.96	29.08	19.19
	Max	48.11	29.59	19.48
	Att [56]	48.08	29.46	19.39
	TADA [55]	47.79	29.69	19.59
	SC †	48.39	29.70	19.62
	SC	**48.66**	**29.79**	**19.83**
Action Recognition	Avg	72.43	51.36	40.90
	Att [56]	72.89	**53.00**	42.20
	SC	**73.09**	52.50	**42.28**

Table 5. Performance comparison of our SC module with other approaches on the **E100** validation set.

refinement our MC (4th row) gets a more obvious increase in accuracy. This strongly proves that self- and cross-refinement provide mutual promotion to leverage multi-modal transferable information for better domain adaptation. With only MC or SC, the performance is not favorable against their combined version, indicating that our MC and SC can cooperate well to leverage both cross-modal complementarity and consensus for minimizing the domain gap.

Different design options of SC We also test different design options of our proposed SC. The SC module aims to spatially re-weight the features based on the transferability of each location. We compare with the most widely adopted feature fusion methods: spatial max pooling (**Max**) and average pooling (**Avg**). Other than these direct fusion methods, we consider two methods based on spatial attention mechanisms, one for general purpose (**Att** [56]) and one for domain adaptation (**TADA** [55]), to generate a spatial attention map for each modality. Weighted average is used to fuse the features based on the attention maps. **SC †** is a simplified version of our SC which computes the correlation of feature maps only at a single scale.

Table 5 shows the comparison on the **E100** validation set. In the domain adaptation setting, simply replacing SC with max or average pooling on each spatial location negatively affects the performance. This indicates that max and aver-

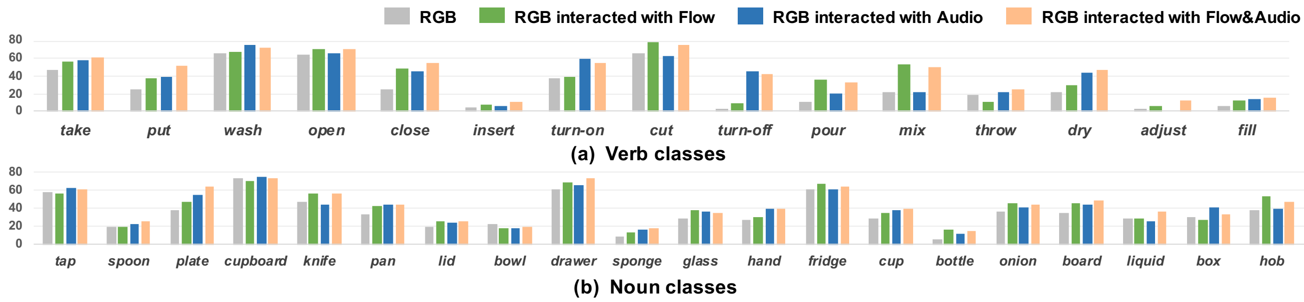

Figure 6. Per-class accuracy of several most frequent verbs (a) and nouns (b) of the **E100** validation dataset.

age pooling do not do well in putting the focus on the transferable regions. The usefulness of multi-scale correlation compared with single-scale correlation is proved, as SC can outperform SC †. Without fully exploiting the multi-modal knowledge, Att and TADA with adversarial alignment cannot find transferable regions as good as our SC. Our SC gets the best performance among these options in the source-only and domain adaptation settings, showing that the spatial consensus among modalities is more domain-invariant.

Due to the lack of labels on the target domain, we cannot show target-only results. Instead, we show an "action recognition" setting by both training and testing models on the source domain. From Table 5, Att outperforms our SC in Noun accuracy since it can learn modality-specific spatial weights when no domain gap exists. From the comparison under different settings, we can see that when domain gap hinders the learning of spatial weight, generating modality-specific spatial weight becomes even more challenging. In this case, our consensus-based SC shows superiority in highlighting transferable regions. However, when no domain gap exists, our SC becomes sub-optimal as we cannot emphasize different regions for different modalities, showing the limitation of our method.

4.5. Contribution of different modalities

To validate the contribution of each modality, in Table 6 we show the results of one modality before and after interacting with other modalities. From the table, we can clearly see the benefit brought by information interaction among multiple modalities. We can also see different modalities have different influences on verb and noun. For example, in the bottom block of Table 6, RGB brings more improvements for Audio in the noun accuracy, and Flow guides the Audio modality to better classify the verbs.

To further validate the enhancement brought by modality interaction, per-class accuracy for RGB modality interacted with different modalities can be seen in Figure 6 (referring to rows 1,2,4,5 of Table 6). In Figure 6(a), for the verbs like "wash", "turn-on" and "turn-off", RGB modality interacted with Audio modality can have a significant performance boost. We think this is because the unique sounds of water and switch are very similar in both source and target domains. Information from the Flow modality helps RGB in discriminating verbs like "open", "cut" and "mix". This is expected since Flow contains more transferable information of the motion and thus complements the RGB modality in predicting verbs. A similar conclusion can be derived from the performance of noun classes, *e.g.* "tap" and "sponge".

Modality	Module	Verb	Noun	Action
RGB	-	30.88	22.98	10.23
(interact with Flow)	MC	39.17	24.94	13.88
(interact with Flow)	MC + SC	40.69	25.22	14.63
(interact with Audio)	MC	40.48	25.64	15.51
(interact with Flow, Audio)	MC	45.38	27.25	17.43
(interact with Flow, Audio)	MC + SC	45.21	27.85	17.80
Flow	-	42.02	21.15	12.90
(interact with RGB)	MC	42.52	24.54	15.32
(interact with RGB)	MC + SC	42.90	25.34	15.81
(interact with Audio)	MC	46.57	23.37	15.95
(interact with RGB, Audio)	MC	46.02	26.14	17.68
(interact with RGB, Audio)	MC + SC	46.28	26.30	17.75
Audio	-	33.34	14.82	8.64
(interact with RGB)	MC	40.10	22.26	13.80
(interact with Flow)	MC	43.80	21.20	14.26
(interact with RGB, Flow)	MC	45.11	24.66	16.27

Table 6. Results of single modality before and after interacting with different modalities on the **E100** validation set are shown to validate the contribution of each modality.

5. Conclusion

In this work, we propose a novel CIA model for multi-modal domain adaptive action recognition. Our CIA model uses two modules to enable the cross-modality feature interaction, which leverages both cross-modal complementarity and cross-modal consensus to accurately learn the most transferable features across the source and target domains. Our method shows considerable improvements on multiple datasets over a variety of previous methods. Our proposed method also has great potential in other domain adaptation tasks, which we will explore in the future.

Acknowledgement This work is supported by JST AIP Acceleration Research Grant Number JPMJCR20U1, JSPS KAKENHI Grant Number JP20H04205 and the Spring GX program of the University of Tokyo.

References

[1] Nakul Agarwal, Yi-Ting Chen, Behzad Dariush, and Ming-Hsuan Yang. Unsupervised domain adaptation for spatio-temporal action localization. *BMVC*, 2020. 2

[2] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3722–3731, 2017. 2

[3] Minjie Cai, Feng Lu, and Yoichi Sato. Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14392–14401, 2020. 1, 2

[4] 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. 2, 5

[5] Chaoqi Chen, Jiongcheng Li, Zebiao Zheng, Yue Huang, Xinghao Ding, and Yizhou Yu. Dual bipartite graph learning: A general approach for domain adaptive object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2703–2712, 2021. 1

[6] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6321–6330, 2019. 1, 5, 6

[7] 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

[8] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3339–3348, 2018. 1

[9] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In *European Conference on Computer Vision*, pages 678–695. Springer, 2020. 5, 6

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

[11] Nazli Faraji Davar, Teofilo de Campos, David Windridge, Josef Kittler, and William Christmas. Domain adaptation in the context of sport video action recognition. In *Domain Adaptation Workshop, in conjunction with NIPS*, 2011. 2

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 2

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2

[14] Dayan Guan, Jiaxing Huang, Aoran Xiao, and Shijian Lu. Domain adaptive video segmentation via temporal consistency regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8053–8064, 2021. 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 3

[17] Haoshuo Huang, Qixing Huang, and Philipp Krahenbuhl. Domain transfer through deep activation matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 590–605, 2018. 2

[18] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769, 2018. 4

[19] 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

[20] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep domain adaptation in action space. In *BMVC*, volume 2, page 4, 2018. 1, 5, 6

[21] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13289–13299, 2020. 5, 6

[22] Guoliang Kang, Yunchao Wei, Yi Yang, Yueting Zhuang, and Alexander G Hauptmann. Pixel-level cycle association: A new perspective for domain adaptive semantic segmentation. *Advances in Neural Information Processing Systems*, 2020. 2

[23] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019. 2, 5

[24] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13618–13627, 2021. 1, 2, 5, 6

[25] Yu Kong, Zhengming Ding, Jun Li, and Yun Fu. Deeply learned view-invariant features for cross-view action recognition. *IEEE Transactions on Image Processing*, pages 3028–3037, 2017. 2

[26] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video

database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 5

[27] Congcong Li, Dawei Du, Libo Zhang, Longyin Wen, Tiejian Luo, Yanjun Wu, and Pengfei Zhu. Spatial attention pyramid network for unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 481–497. Springer, 2020. 2, 4

[28] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. Tea: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 909–918, 2020. 2

[29] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 5, 6

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[31] Mengyuan Liu, Hong Liu, and Chen Chen. Enhanced skeleton visualization for view invariant human action recognition. *Pattern Recognition*, 68:346–362, 2017. 2

[32] Yang Liu, Zhaoyang Lu, Jing Li, Tao Yang, and Chao Yao. Deep image-to-video adaptation and fusion networks for action recognition. *IEEE Transactions on Image Processing*, 29:3168–3182, 2019. 2

[33] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105, 2015. 1, 5, 6

[34] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 2

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

[36] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 122–132, 2020. 1, 2, 5, 6

[37] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018. 2

[38] Boxiao Pan, Zhangjie Cao, Ehsan Adeli, and Juan Carlos Niebles. Adversarial cross-domain action recognition with co-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11815–11822, 2020. 1, 2, 5, 6

[39] Barbara Plank and Alessandro Moschitti. Embedding semantic similarity in tree kernels for domain adaptation of relation extraction. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1498–1507, 2013. 2

[40] Fan Qi, Xiaoshan Yang, and Changsheng Xu. A unified framework for multimodal domain adaptation. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 429–437, 2018. 2

[41] Kuniaki Saito, Yoshitaka Ushiku, and Tatsuya Harada. Asymmetric tri-training for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 2988–2997. PMLR, 2017. 2

[42] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 5, 6

[43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6, 7

[44] Darsh Shah, Tao Lei, Alessandro Moschitti, Salvatore Romeo, and Preslav Nakov. Adversarial domain adaptation for duplicate question detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1056–1063, 2018. 2

[45] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Actor and observer: Joint modeling of first and third-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7396–7404, 2018. 2

[46] Sijie Song, Jiaying Liu, Yanghao Li, and Zongming Guo. Modality compensation network: Cross-modal adaptation for action recognition. *IEEE Transactions on Image Processing*, 29:3957–3969, 2020. 2

[47] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9787–9795, 2021. 1, 2, 5, 6

[48] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[49] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *European conference on computer vision*, pages 443–450. Springer, 2016. 1, 2

[50] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through self-supervision. *arXiv preprint arXiv:1909.11825*, 2019. 2

[51] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558, 2019. 2

[52] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 1, 2

[53] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2017. 4

[54] 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. 5, 6

[55] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5345–5352, 2019. 2, 4, 7

[56] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 4, 7

[57] Lijin Yang, Yifei Huang, Yusuke Sugano, and Yoichi Sato. Stacked temporal attention: Improving first-person action recognition by emphasizing discriminative clips. *arXiv preprint arXiv:2112.01038*, 2021. 2

[58] Chaohui Yu, Jindong Wang, Yiqiang Chen, and Meiyu Huang. Transfer learning with dynamic adversarial adaptation network. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 778–786. IEEE, 2019. 1

[59] Weichen Zhang, Dong Xu, Jing Zhang, and Wanli Ouyang. Progressive modality cooperation for multi-modality domain adaptation. *IEEE Transactions on Image Processing*, 30:3293–3306, 2021. 2

[60] Yang Zhang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2020–2030, 2017. 2

[61] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 5, 6

[62] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3

[63] Fan Zhu and Ling Shao. Enhancing action recognition by cross-domain dictionary learning. In *BMVC*. Citeseer, 2013. 2

[64] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 1, 2