Multi-Modal Temporal Attentive Adversarial Adaptation Network for Egocentric Action Recognition

Andrea Ferretti
s289677@studenti.polito.it

Davide Gariglio
s292964@studenti.polito.it

Pietro D'orto
s297340@studenti.polito.it

***Abstract*—In the past the Domain Adaptation task has been mainly explored related to images, while just in the recent years the research focused also on video-based problems. This is mainly because, at first, there were not large-scale video datasets available; these are more challenging given the higher number of domains, resulting in a bigger domain discrepancy and number of classes.**

The proposed work addresses a video-based Domain Adaptation method for Action Recognition applied to a large-scale egocentric video dataset by extending a pre-existing architecture to different modalities, applying an efficient sampling strategy and different temporal aggregation methods in order to extract relations between frames. The code is publicly available on Github.

I. Introduction

The general purpose of Domain adaptation (DA) is to address the domain shift problem, which means that the model trained on source labeled dataset do not generalize well to target dataset. There are several unsupervised DA (UDA) techniques that address the task in different ways; in this work we focus on a method that allows the model to generalize to target samples without access to any target labels.

In the past this problem has been explored mainly on image-based dataset; there are many approaches that are able to diminish the distribution gap between source and target domains while learning discriminative deep features [5], [6], [8], [9], [11], [12]. In the recent year instead, thanks to constant innovations in video-recording technologies, wearable camera devices became more and more popular; as a result of this, more egocentric video recordings became available and captured the attention of the computer vision researchers towards them.

The main goal is to modify the methods used for images and apply them to large-scale video datasets. Videos can suffer from domain discrepancy along both spatial and temporal direction; therefore there is the need of align embedded feature spaces along both directions.

In order to do so we extended a pre-existing architecture for unsupervised video DA [2] to different modalities, similar to [13]. An illustration of our proposed extension can be seen in Figure 1.

The purpose of the architecture is to extract features that are domain independent thanks to the **adversarial DA** approach, then it aligns them along both spatial and temporal dimensions according to their contribution in the action classification. This is possible with a weighting system which is able to recognize their importance (Section III-C).

In the proposed work we used a subset of three kitchens from the EPIC-Kitchens dataset [4], a large-scale egocentric video benchmark recorded by 32 partecipants in their native kitchen environments which correspond to different domains.

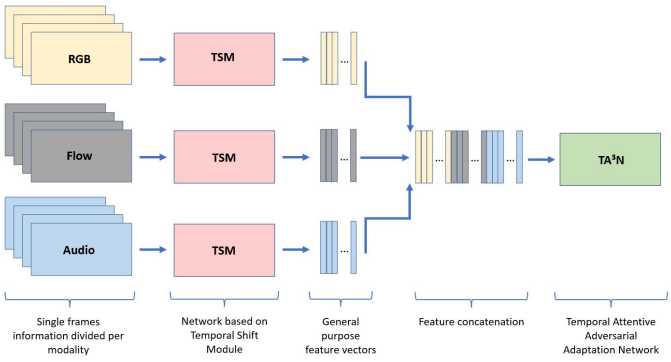

Fig. 1: Multi-modal extension for *Temporal Attentive Adversarial Adaptation Network*. Input frames are sent into *TSM* which has the goal of extracting general purpose features; then, by concatenating them, these are used as input for TA^3N.

II. Related Works

A. Action-Recognition

In recent years, several architectures based on 2D [15] [16] [7] [10] or 3D [1] CNN have been proposed for egocentric action recognition. One key point in video recognition is the temporal modeling. The 2D CNNs perform temporal modeling independent of 2D spatial convolutions. On the other hand, 3D CNNs learn space and time information jointly using 3D convolutions. Therefore, different methods for temporal aggregation have been proposed:

Temporal Segment Networks (TSN) [15] is based on the idea of long-range temporal structure modeling. It performs sparse temporal sampling followed by temporal aggregation (averaging) of soft-max scores across samples. Each modality is trained independently, with late fusion of modalities by averaging their predictions.

In *Temporal Binding Network (TBN)* [7] instead, modalities are fused before temporal aggregation with shared modality and fusion weights over time. *Temporal Shift Module (TSM)* [10], shifts parts of the channels along the temporal dimension, both forward and backward, facilitating information exchanged among neighboring frames, while *Temporal Relation Module (TRM)* [16] is designed to learn and reason about temporal dependencies between video frames at multiple time scales.

The *Two-Stream Inflated 3D ConvNets (I3D)* model was presented in [1]; the proposed approach starts with a 2D image classification architecture and is extended by inflating all its filters and pooling kernels into 3D, making it possible to learn seamless spatio-temporal feature extractors from video.

B. Unsupervised Video-Based Domain Adaptation

Most UDA approaches aim to find a common feature space between the source and target domains. The models are therefore optimized with a combination of classification and domain losses [3].

Unlike image-based UDA, video-based UDA is still an under-explored area. Compared with images, video source and target domains also differ along the temporal dimension.

Recently, few works have been proposed for video-based UDA. The key idea of these methods is to achieve temporal alignment by aligning both frame and video-level features through different learning mechanism such as adversarial learning [2] and contrastive learning [14].

In [2], the Temporal Attentive Adversarial Adaptation Network (*TA^{3}N*) is proposed; it is composed by two main modules (spatial and temporal) and integrates discriminators after them, in order to align the model across domains while learning spatial and temporal dynamics. In the same work is presented also a domain attention mechanism based on entropy criterion to generate the domain attention value for each relation feature, resulting in domain discriminative features.

Instead of using adversarial learning, [14] contains an end-to-end temporal contrastive learning framework named CoMix with background mixing and target pseudo-labels.

There are also a few works integrating multiple modality data for video-based DA, the first and most important one to mention is *MM-SADA* [13]; the correspondence of multiple modalities is exploited as a self-supervised alignment in addition to adversarial alignment. In detail, by focusing on two modalities (RGB and Optical Flow), they incorporate a self-supervision alignment classifier that determines whether modalities are sampled from the same or different actions to learn modality correspondence. This module receive as input the concatenated features from both modalities without any labels, encouraging features that generalise to both domains. The alignment of the domain statistics is achieved by adversarial training, with a domain discriminator per modality.

III. Method

In the following sections we present the main steps and issues we addressed in our analysis, starting from the choice of the network up to the extension of [2] to finally obtain a ***Multi-Modal Temporal Attentive Adversarial Adaptation Network*** (*MM-TA^{3}N*) (Fig. 2).

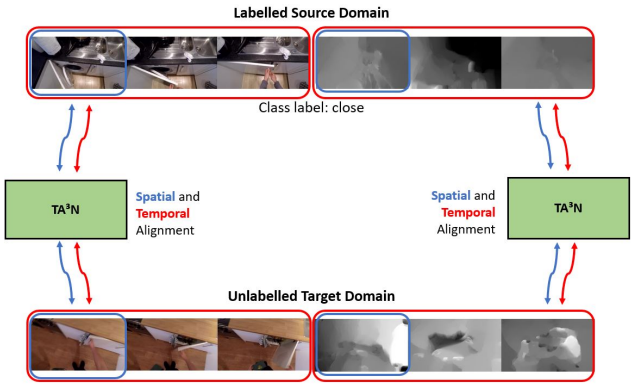

Fig. 2: Multi-Modal Action Recognition. In order to enhance the UDA performances, frames coming from different modalities are taken as input for the *Temporal Attentive Adversarial Adaptation Network* (TA^{3}N) (Section III-C).

A. Networks and sampling strategies

In the first step the focus is, given two networks (*I3D* [1] and *TSM* [10]) with pre-trained model weights, to implement and test the sampling strategies required for those to work in different modalities. In this phase, the key part of this work is to study more deeply the networks already existing and the concept of egocentric action recognition in practice.

Given the two networks (*I3D* and *TSM*), the way in which the frames are loaded and passed to the models is respectively:

- **Dense sampling**: in this strategy the frames taken from the video are divided in clips; for each one of them, a certain number of consecutive frames is sampled resulting in multiple local portions taken into account.
- **Uniform sampling**: given an entire video, this strategy uniformly samples a certain number of frames. In this way, given relatively long duration videos, the model can receive a wider view of the input avoiding analyzing multiple local portions that can potentially contain non-important features for the final classification.

After the implementation of the sampling strategies, the aim is to extend the networks to be **multi-modal** by building a branch for every modality considered and combining the prediction from each one of them. In table II can be seen that exploiting more modalities for the same videos (such as *RGB* and *Optical Flow*) leads to better performances since every modality can provide different and substantial information about the frames.

For this part, the features used were for both *I3D* [1] and *TSM* [10], while for the next parts only the *TSM* [10] features will be used.

B. Temporal aggregation

Once the different frames have been sampled, it is necessary to extract the key information that are present between them.

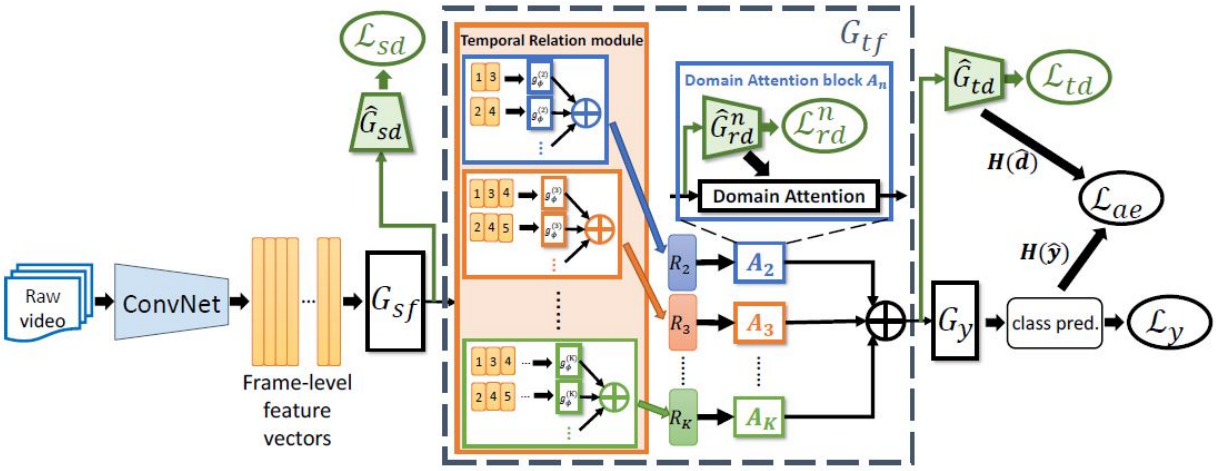

Fig. 3: Temporal Attentive Adversarial Adaptation Network (TA^{3}N). In the temporal relation module, time-ordered frames are used to generate *K-1* relation feature representations $\mathbf{R} = \{R_2; ...; R_K\}$, where R_n corresponds to the n-frame relation (the numbers in this figure are examples of time indices). After attending with the domain predictions from relation discriminators G^n_{rd}, the relation features are summed up to the final video representation.

In order achieve this goal, the architecture must aggregate the input samples according to some specific methods; in our work we mainly focused on the following two strategies:

- **Average pooling:** This strategy is the simplest and naive one, which is a key component of many models. It simply makes the mathematical average between different features taken from every frame. It is a basic approach because there are a lot of relational informations lost during the average calculations, therefore the results obtained are lower. For this reason we will consider this approach as a baseline.
- **Temporal relation module (TRM):** This aggregation strategy instead is able to extract the relations between the different frames at multiple scales [16]. The intuition is that the human brain can recognize actions by observing them across time, therefore it is necessary to capture temporal relations present between the frames at multiple time scales. For this reason, the TRM takes the sampled frames and analyze them in temporal order, generating *K-1* relation feature representations $\mathbf{R} = \{R_2, ..., R_k\}$. Every R_n corresponds to the *n*-frame relational features extracted from the n ordered frames; these are finally combined together according to the attention weighting mechanism, presented in [2], since not every feature contributes in the same way to the action classification.

As shown in [2] [16], the presence of the *TRM* allows to achieve higher performances with respect to the *Average pooling* aggregation strategy since it is designed to learn temporal dynamics from the video frames. We performed a test to assess the positive contribution of the *TRM* applied to the dataset under analysis; the results are reported in Table III.

C. Temporal Attentive Adversarial Adaptation Network

Regarding the structure of the model used for the inference, in this work we adopted the ***Temporal Attentive Adversarial Adaptation Network*** *(TA^{3}N)* (Fig. 3) [2]. This architecture aims to extract and align domain independent features by means of adversarial discriminators while learning temporal dynamics. It is mainly composed by two modules:

- **Spatial module:** It consists of multilayer perceptrons (MLP); the goal is to convert the general purpose features extracted by the DCNN (in this case TSM [16]) into task-driven feature vectors.
- **Temporal module:** It consists of a temporal aggregation component (Section III-B) followed by an attention mechanism. The purpose is to aggregate the frame-level feature vectors extracted by the *spatial module* in order to retrieve informations along the temporal dimension. Once the temporal features have been learned by the *temporal module*, these are combined together according to an attention weighting mechanism. The reason behind this is due to an unequal importance, and so a different contribution, of the features for the action classification.

In order to transfer the knowledge learnt from the *source-domain* into the *target-domain*, the architecture, inspired by *DANN* [6], uses *adversarial DA* by means of ***domain discriminators***: $\hat{G}_{sd}$, $\hat{G}_{rd}$, $\hat{G}_{td}$. These components are placed respectively after the spatial, relational and temporal modules.

They consists of a *Gradient Reversal Layer* (*GRL*) and a domain classifier. During the forward phase the *GRL* behaves like an identity function, while during the back-propagation it inverts the gradients; in this way the feature generators are optimized to gradually align the feature distribution between the two domains.

Finally, a fully-connected layer G_y converts the video-level features from G_{tf} the into the final predictions, which generate the class prediction loss $\mathcal{L}_y$. There is also and attentive entropy loss $\mathcal{L}_{ae}$ which is the combination of domain entropy $H(\hat{d})$ and the class entropy $H(\hat{y})$; the former is generated by the domain classifier inside the temporal domain discriminator $\hat{G}_{td}$, while the latter is produced by the final action classifier G_y. The use of $\mathcal{L}_{ae}$ enhances the centainty of videos that are more similar across domains, leading the architecture to better performances.

The network adopted is explored in all of its components by performing an ablation study, in which every one of the different *discriminators* is used and tested alone in order to measure its contribution. Eventually, a final test is performed by using all the three adversarial components with the addition of the *domain attention mechanism*. All the obtained results are reported in the Section IV-B.

D. Multi-Modal TA3N

Our proposed extension aims at improving the overall accuracy by merging different modalities for the videos considered. This is due to the fact that every modality brings important information about the frames and isolating them leads to inefficient estimations. Therefore, adapting the model's pipeline in such a way that all these different aspects are merged together effectively leads to an improvement in the accuracy for the *UDA Egocentric Action Recognition* task.

To do so, the first important step is to retrieve the frame-level features for the different modalities and link them all together. In this work the features are extracted individually for each modality with TSM [10] and then the fusion is performed by concatenating them in order to obtain a single vector containing every frame informations regarding *RGB*, *Optical Flow* and *Audio*, in order to . In Table V the accuracy obtained with *TA*^{3}*N* and *MM-TA*^{3}*N* on the subset from the *EPIC-Kitchens* Dataset [4] are compared, showing the positive contribution of the *multi-modal* extension.

IV. Experiments

A. Dataset

EPIC-Kitchens [4], as mentioned, is a large-scale egocentric dataset containing recording collected by 32 participants, belonging to 10 nationalities, in their native kitchens.

The recordings, which include both video and audio, not only show the typical interactions the participants have with their kitchen utensils but more importantly, through first-person views, show the natural multi-tasking they do and also show the many different ways they perform a variety of daily tasks.

The annotation is unique in that, since to each participant was asked to watch their videos, after completing all recordings, and narrate the actions carried out. Thus reflecting true intention of the participants, and the authors have developed ground-truths based on these.

As done by [13], in the proposed work the three largest kitchens in number of training action instances were taken under analysis. These corresponds to *Participant 08*, *Participant 01* and *Participant 22*, whose domains are respectively *D1*, *D2* and *D3* (Fig. 4).

In Fig. 5 it is reported the action classes ('put', 'take', 'open', 'close', 'wash', 'cut', 'mix', and 'pour') distribution across the three domains. A large class imbalance and differing distribution of verbs are present, adding new challenges for the DA task. We also report in table I the number of action segments of train and test sets for each domain, highlighting the different cardinalities.

Fig. 4: Portion of EPIC Kitchens Dataset used to evaluate the model, containing 3 domains.

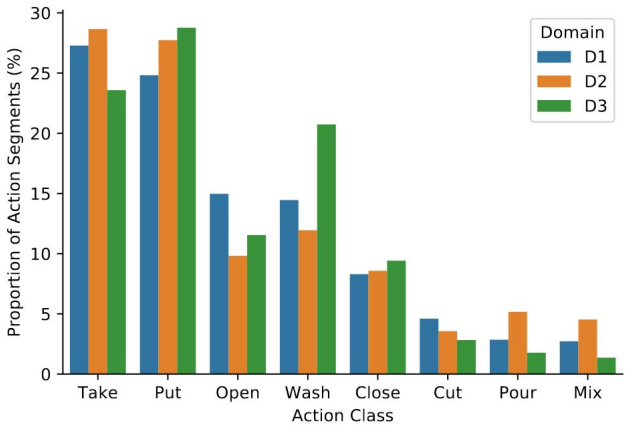

Fig. 5: Action distribution per domain for the 8 classes.

Domain	D1	D2	D2
Ref. EPIC Kitchen	P08	P01	P22
Training Action Segments	1543	2495	3897
Test Action Segments	435	750	974

TABLE I: Number of train and test action segments present in each domain.

B. Results

In the following section we present the results achieved for the different steps presented in Section III, in particular:

the choice of the network and the relative sampling frame aggregation strategies (Section III-B).

Finally, we compare the results of the *Temporal Attentive Adversarial Adaptation Network (TA^{3}N)* against the proposed multi-modal extension *MM-TA^{3}N*.

1) Networks and sampling strategy: In Section III-A we presented two different networks with their relative sampling strategies, describing the functionalities and their possible advantages. First we compared *I3D* and *TSM* performances with the respective sampling strategy and without any DA technique, so with the same *source* and *target* domain; the purpose was to understand which of the two proposed networks could have been the most suitable in order to extract frame-level feature vectors for our work. The results shown in table II highlight that *TSM*, thanks to the temporal-shifting mechanism and the global frame view given by the *uniform sampling* method, is able to extract more important features with respect to *I3D* and its *dense sampling* strategy, leading the former classifier to higher accuracy for every modality and also for an initial multi-modal features combination (*RGB + Flow*); in fact, the difference goes from a minimum of 7.375 up to 8.414.

Network	Sampling	RGB (%)	Flow (%)	RGB + Flow (%)
I3D	Dense	53.78	57	60.381
TSM	Uniform	62.194 (+8.414)	64.403 (+7.403)	67.756 (+7.375)

TABLE II: Single modal and multi-modal mean accuracy compared for *I3D* and *TSM* with the relative gain obtained by using the latter network.

2) Temporal aggregation strategies: As presented in Section III-B, we analyzed also the contribution of the temporal aggregation methods associated to a network and applied to different modalities. In table III the mean accuracy for the different temporal aggregation strategies are shown and, as mentioned before, *TRM* highly outperforms *average pooling* in RGB modality while in the *Flow* modality the performances are slighlty lower.

Network	Aggregation Strategy	RGB (%)	Flow (%)
I3D	AvgPool	58.000	63.333
TSM	TRM	68.000 (+10.000)	62.667(-0.666)

TABLE III: Mean accuracy across the three domains for different networks with aggregation strategies implemented; inside the parenthesis it is reported the relative gain.

3) TA^{3}N ablation study: In this step we adopted TA^{3}N [2], described in section III-C, as the architecture for our analysis. Here the focus is on analyzing the impact of the different discriminators by performing an ablation study. The results in table IV are obtained using the best hyper-parameters found in Section IV-C and show that using all of the discriminators with the domain attention mechanism achieves the highest possible gain.

4) TA^{3}N and MM-TA^{3}N comparison: The last tests performed are with our proposed extension (Section III-D). Here the models are trained using all the discriminators and the domain attention mechanism. In table V the results obtained show that integrating multiple modalities significantly improves the accuracy obtained. In this case the gain is of +9.986 with *Optical Flow* and *Audio* more than the single modality used by *TA^{3}N* (in this case *RGB*).

Model	Accuracy (%)
TA^{3}N (RGB)	36.428
MM-TA^{3}N (RGB + Flow)	40.863 (+4.435)
MM-TA^{3}N (RGB + Flow + Audio)	46.414 (+9.986)

TABLE V: *TA^{3}N* and *MM-TA^{3}N* accuracy compared by adding different modalities. The results are obtained by averaging on all the domains considered.

C. Hyper-parameters search

When performing the ablation study on *TA^{3}N* (Section III-C), some important parameters have been tested and modified in order to find the best combination to improve the overall accuracy. In Table IV the results of this study are obtained by training and testing the model with the best hyper-parameters found. The parameters tested are the weighting for the discrepancy loss α, the trade-off weighting for each domain loss β_0, β_1, β_2, the weighting for the attentive entropy loss γ and the *batch size (bs)*. To find the combination that leads to the best possible results, an extensive grid-search is performed for every possible domain and different values of those parameters. The values tested can be seen in table VI. The best values obtained are: $\alpha = -1, \beta_0 = 0.5, \beta_1 = 0.75, \beta_2 = 0.5, \gamma = 0.003$ (for *Temporal Pooling*), $\gamma = 0.03$ (for *TRM*), $bs = 64$.

Parameter	Values
α	[-1, 0, 0.5, 1]
β_0	[0.5, 0.75, 1]
β_1	[0.5, 0.75, 1]
β_2	[0.5, 0.75, 1]
γ	[0.003, 0.03, 0.3]
bs	[64, 128, 256]

TABLE VI: Hyper-parameters search for the ablation study.

V. Conclusion

We analyzed and proposed a multi-modal extension for an existing unsupervised video domain adaptation architecture, studying different aspects related to the task under analysis taking into account many key factors. Videos can suffer from domain discrepancy along both spatial and temporal directions: by analyzing only the single frames, the time relations between them are not considered, leading to sub-optimal performances. In our work we highlighted the importance of the **temporal module** in the unsupervised video DA task, comparing a naive approach (*Temporal Pooling*) with an efficient temporal aggregation strategy (*TRM*) already present in literature, showing how the architecture can benefit from it.

Component	Aggregation	Accuracy RGB (%)
Source Only	Temporal Pooling	34.650 (-)
	Temporal Relation	**35.655** (-)
$\hat{G}_{sd}$	Temporal Pooling	35.500 (+0.850)
	Temporal Relation	**36** (+0.345)
$\hat{G}_{td}$	Temporal Pooling	35.758 (+1.110)
	Temporal Relation	**36.260** (+0.605)
$\hat{G}_{rd}$	Temporal Pooling	35.966 (+1.316)
	Temporal Relation	**36.300** (+0.645)
All $\hat{G}_{d}$	Temporal Pooling	36.032 (+1.382)
	Temporal Relation	**36.364** (+0.709)
All $\hat{G}_{sd}$ + *Domain Attention*	Temporal Pooling	36.216 (+1.566)
	Temporal Relation	**36.428** (+0.773)

TABLE IV: *TA*^{3}*N* ablation study's accuracy. In the parenthesis is the gain which indicates the improvement from the source only setting for the same aggregation method. The highest values for each component are highlighted.

Moreover, we combined the best network, sample and temporal aggregation strategies in the chosen architecture which exploits *adversarial learning* using different domain discriminators and a *weighting attention mechanism* to evaluate the contribution of the most significant features, disregarding low level importance features.

Finally we studied and implemented a **multi-modal approach**, showing the diversity of informations that different modalities can bring to the model and highlighting how the performances can be enhanced by considering all of them.

REFERENCES

[1] 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] 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.

[3] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. In *Domain Adaptation in Computer Vision Applications*, pages 1–35, 2017.

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

[5] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[6] 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.

[7] 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.

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

[9] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *arXiv preprint arXiv:1603.04779*, 2016.

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

[11] 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. PMLR, 2015.

[12] 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.

[13] 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.

[14] Aadarsh Sahoo, Rutav Shah, Rameswar Panda, Kate Saenko, and Abir Das. Contrast and mix: Temporal contrastive video domain adaptation with background mixing. *Advances in Neural Information Processing Systems*, 34:23386–23400, 2021.

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

[16] 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.