POLITECNICO DI TORINO
Repository ISTITUZIONALE

Domain generalization using action sequences for egocentric action recognition

Original
Domain generalization using action sequences for egocentric action recognition / Nasirimajd, A., Plizzari, C., Peirone, S.A., Ciccone, M., Averta, G., Caputo, B.. - In: PATTERN RECOGNITION LETTERS. - ISSN 0167-8655. - ELETTRONICO. - 196:(2025), pp. 213-220. [10.1016/j.patrec.2025.06.010]

Availability:
This version is available at: 11583/3001251 since: 2025-06-25T11:21:03Z

Publisher:
Elsevier

Published
DOI:10.1016/j.patrec.2025.06.010

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

27 November 2026

Pattern Recognition Letters 196 (2025) 213–220

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Domain generalization using action sequences for egocentric action recognition☆

Amirshayan Nasirimajd *, Chiara Plizzari, Simone Alberto Peirone, Marco Ciccone, Giuseppe Averta, Barbara Caputo

Politecnico di Torino, Department of Control and Computer Engineering, Corso Castelfidardo, 34/d, Turin, 10138, TO, Italy

ARTICLE INFO

Editor: Maria De Marsico

Keywords:
Egocentric action recognition
Domain adaptation
Domain generalization

ABSTRACT

Recognizing human activities from visual inputs, particularly through a first-person viewpoint, is essential for enabling robots to replicate human behavior. Egocentric vision, characterized by cameras worn by observers, captures diverse changes in illumination, viewpoint, and environment. This variability leads to a notable drop in the performance of Egocentric Action Recognition models when tested in environments not seen during training. In this paper, we tackle these challenges by proposing a domain generalization approach for Egocentric Action Recognition. Our insight is that action sequences often reflect consistent user intent across visual domains. By leveraging *action sequences*, we aim to enhance the model's generalization ability across unseen environments. Our proposed method, named SeqDG, introduces a visual-text sequence reconstruction objective (SeqRec) that uses contextual cues from both text and visual inputs to reconstruct the central action of the sequence. Additionally, we enhance the model's robustness by training it on mixed sequences of actions from different domains (SeqMix). We validate SeqDG on the EGTEA and EPIC-KITCHENS-100 datasets. Results on EPIC-KITCHENS-100, show that SeqDG leads to +2.4% relative average improvement in cross-domain action recognition in unseen environments, and on EGTEA the model achieved +0.6% Top-1 accuracy over SOTA in intra-domain action recognition. Code and Data: Github.com/Ashayan97/SeqDG

1. Introduction

Egocentric vision offers a unique (first-person) perspective to capture human activities, which serves as a foundational tool in many fields, including assistive technologies [1] and robotics applications [2]. Among many egocentric vision tasks [3], Egocentric Action Recognition (EAR) represents a crucial component to build comprehensive models of human actions.

A major challenge in deploying action recognition models in the real-world lies in their ability to generalize to environments and scenarios that significantly differ from those seen during training (*domain shift*) [4,5]. Indeed, EAR models often become overly dependent on specific environmental cues present in the training data [6], which could limit their ability to recognize actions when performed in new, unseen environments [7]. Moreover, unique behavioral traits may influence the way similar actions are performed and add an extra layer of complexity in action recognition. A key insight we can leverage to improve generalization is that human actions do not occur in isolation but sequentially and are naturally goal-driven [8], making action sequences crucial to understanding individual actions [9]. By exploiting the additional context provided by surrounding actions, we can in turn reduce the impact of confounding factors, such as changes in the environment or user behavior. As an example, consider two people preparing an omelet, each in their own kitchen, generally perform the same steps, one after the other: they take eggs from the refrigerator, break them, add spices or other ingredients, and finally fry the omelet. Kitchens and users may be different, but the steps needed to prepare the omelet are the same. A similar example is shown in Fig. 1: despite different contexts, the sequence of actions – opening the fridge, picking up milk, opening the lid, and pouring milk – are the same. We validate these observations by looking at the number of repeated action sequences in one of the most significant egocentric vision datasets, EPIC-KITCHENS-100 [10]. Samples are annotated with a *Verb* and a

☆ This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them. This study also acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support.

* Corresponding author.
E-mail address: amirshayan.nasirimajd@studenti.polito.it (A. Nasirimajd).

https://doi.org/10.1016/j.patrec.2025.06.010
Received 20 September 2024; Received in revised form 3 June 2025; Accepted 12 June 2025
Available online 24 June 2025
0167-8655/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

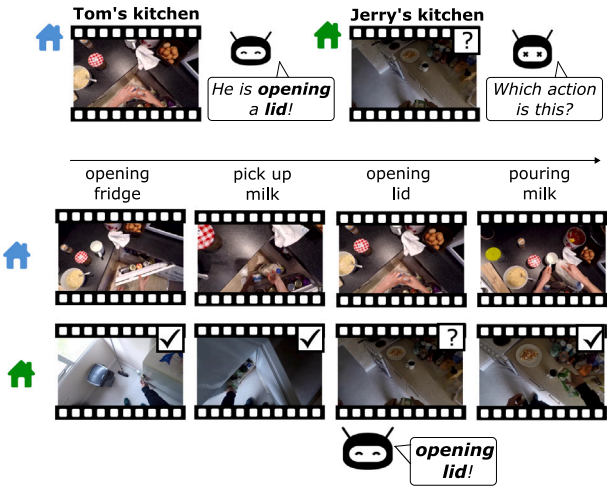


Fig. 1. Egocentric Action Recognition (EAR) models struggle in classifying the same action when it is performed in different environments (top). Despite different visual contexts, the sequence of actions – opening the fridge, picking up milk, opening lid and pouring milk – remains consistent. Thus, considering previous and subsequent actions within a sequence proves beneficial for improving cross-domain robustness of the EAR models.

Noun label, as well as their combination, indicated as *Action*. Actions are recorded in 45 different kitchens, by 37 different participants. Despite the large diversity in environments and participants, we observe a significant number of repeated action sequences: out of 32k samples, 5882, 4649, and 1852 sequences of length up to 5 are repeated for the Verb, Noun, and Action categories, respectively (Fig. 2).

In this work, we claim that *action sequences* can help generalization in two ways: (i) although the specific execution can differ – such as how eggs are cracked or ingredients are mixed – the overall sequence of actions forms consistent patterns across different subjects; (ii) because these sequences are independent of the environment's layout or appearance, they can be used to adapt to and recognize individual actions in visually diverse scenarios, mitigating the negative effect of domain shift. We hence propose **SeqDG**, a Domain Generalization method for EAR that leverages the similarity of action sequences across different visual domains to improve the generalization of action recognition models. SeqDG integrates **SeqRec**, a visual-text sequence reconstruction objective that utilizes contextual information from both text and visual modalities to reconstruct a sequence's masked action. This serves as a *proxy task* that explicitly pushes the model to capture temporal dependencies and relational cues across the sequence. In addition, we introduce **SeqMix**, a data augmentation technique that mixes actions that share the same label but come from different domains to make the model more robust to visual changes.

We note the distinction of Domain Generalization (DG) from the Unsupervised Domain Adaptation (UDA) setting, where unlabeled target samples are available during training [11–13]. We do not only demonstrate the advantages of our approach for cross-domain generalization but also outperform standard UDA methods and advanced sequence-based solutions, despite not accessing target data during training.

In summary, we present the following contributions:

- we highlight the role of action sequences in cross-domain EAR, guided by the principle that human actions follow patterns that are repeated across different environments;
- we propose SeqDG, a DG approach for EAR combining a visual-text reconstruction objective with mixed sequences across different domains, to encourage the model to better exploit action sequences;
- we validate SeqDG on EGTEA and EPIC-KITCHENS-100, observing SOTA performance and demonstrating that effective usage of sequence helps EAR in unseen domains.

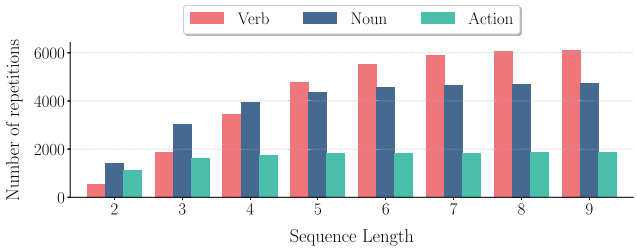


Fig. 2. Number of repeated Verb, Noun and Action sequences in the EPIC-KITCHENS-100 dataset. For each sequence length, we report the number of sub-sequences of any smaller (or equal) length that repeat across domains.

2. Related works

2.1. Egocentric action recognition

CNN-based approaches for EAR can be divided into two main categories, using either 3D [14,15] or 2D [16,17] convolutions. The goal is to adapt CNNs to the temporal nature of the videos, allowing for better modeling of the temporal context, which is crucial for many tasks, e.g., action segmentation [16,18] and action recognition [19,20], specifically in EAR setting [9,21]. Additionally, various approaches in EAR have explored integrating multiple modalities, a common strategy in video understanding to enhance model robustness [22,23]. For instance, some works focus on the fusion of audio and visual information [24–26] or other non-audiovisual modalities, like IMU [27]. More recently, transformer-based approaches have been proposed to model temporal relations through self-attention [28]. They also improved the processing of different modalities – especially text – enabling the use of both vision and text to enhance long-term video understanding [29]. In most EAR approaches, actions are classified independently, i.e., the temporal context from surrounding actions is not considered. In contrast, sequence-based approaches [9] integrate temporal context to improve action recognition. We design SeqDG on the intuition that temporal context may help EAR models, leveraging text to represent action sequences in a domain-agnostic way.

2.2. Domain adaptation and generalization

Unsupervised Domain Adaptation (UDA) combines training on labeled source data with unlabeled target data to align their representations. In adversarial approaches, alignment is achieved using adversarial classifiers. Among these, TA3N [30] uses multiple adversarial branches to align the distributions at different temporal scales, while MM-SADA [11] learns to align source and target distributions and features of different modalities. Similarly, CIA [31] combines adversarial training with a gating mechanism to leverage cross-modal synergies. Other works explored contrastive alignment across different modalities and domains [13] or category-aware alignment [12].

Domain Generalization (DG) learns robust domain-agnostic representations without target data. VideoDG [32] aligns local temporal features across domains as they generalize better compared to global features. RNA [33] introduces a cross-domain and cross-modality norm alignment loss to learn equally from all domains and modalities. EgoZAR [5] proposes leveraging domain-agnostic representations of the specific locations associated with particular actions (activity-centric zones) to enhance the generalization across different environments. CIR [4] implements cross-instance reconstruction of each video representation from other domains videos.

Previous works in DG/UDA investigated EAR on individual actions. Here, we study a new paradigm for DG in EAR that leverages action sequences, exploiting the observation that they repeat in different domains in accordance with the user' intentions, and may help EAR models in unseen environments.

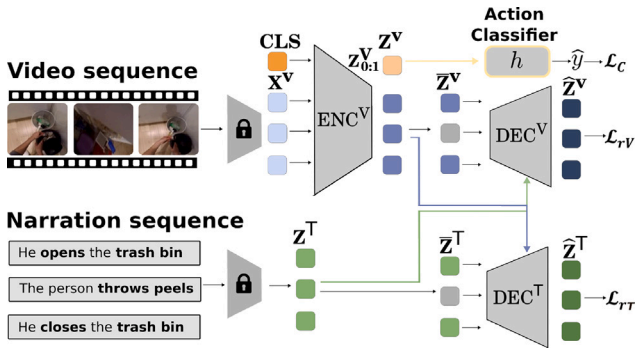


Fig. 3. SeqDG architecture. We are given visual and textual inputs $\boldsymbol{X}^V$ and $\boldsymbol{X}^T$. A classification token CLS is appended to the visual input for classification. Visual inputs are fed to an encoder $\texttt{ENC}^{\texttt{V}}$, resulting in intermediate visual embeddings $\boldsymbol{Z}_i^V$, while textual features are passed through an identity function to get $\boldsymbol{Z}_i^T$. The latter are masked ($\bar{\boldsymbol{Z}}_i^V$ and $\bar{\boldsymbol{Z}}_i^T$) and fed to two separate decoders $\texttt{DEC}^{\texttt{V}}$ and $\texttt{DEC}^{\texttt{T}}$ for visual and text reconstruction ($\mathcal{L}_{rV}$ and $\mathcal{L}_{rT}$). The transformed CLS token $\boldsymbol{Z}_{0:1}^v$ is fed to classifier h for action classification ($\mathcal{L}_C$).

2.3. Masked Autoencoders for video applications

Masked Autoencoders (MAEs) [34] learn to reconstruct missing parts of input data, which demonstrates to be an effective pre-training objective for various downstream tasks, with applications also to videos [35,36] and multi-modal vision-language inputs [37]. Most related to our work is Voltron [38], which reconstructs video patches at the pixel level conditioned by textual narrations and generates captions for input clips. Unlike previous approaches, which reconstruct masked inputs at the pixel level [35–38], SeqDG is designed to reconstruct masked actions in a sequence to achieve high-level reasoning about human activities.

3. Proposed method

Our goal is to increase the robustness of action recognition models in the presence of domain shift, by leveraging the similarity of action sequences across different domains. Relying solely on visual information might be problematic as visual data is susceptible to variations that may not be related to the actions themselves, e.g., different environments, backgrounds, lighting conditions, etc. In contrast, textual narrations are less tied to the domain and inherently more robust to domain shift [4]. We propose a visual-text sequence reconstruction objective (**SeqRec**, Section 3.2), that reconstructs a masked action of the sequence using the information of surrounding actions from both the text and visual modalities. Additionally, we train the network with sequences that combine actions from various domains (**SeqMix**, Section 3.3). This approach is designed to explicitly model different distributions of actions within sequences during training, thereby increasing performance on out-of-distribution test data. We call our overall approach **SeqDG**. Formally, we define SeqDG as a Domain Generalization approach for EAR, in which the model is trained on action samples from a set of N_s different *source* domains $\{D_1^s, \dots, D_{N_s}^s\}$ and evaluated on a set of N_t unseen *target* domains $\{D_1^t, \dots, D_{N_t}^t\}$, under the assumption that there is no overlap between the different domains. An illustration of the overall method is shown in Fig. 3.

3.1. Leveraging sequences of actions

SeqDG leverages sequences of actions rather than individual actions to mitigate the negative effect of domain shift: instead of relying solely on the visual content of a single clip, which may be affected by confusing cues such as unfamiliar backgrounds, occlusions, or user-specific behavior, the model can reason about the relations between the different actions in the sequence to infer the correct action. We define an *action sequence* S_i as the ordered set of W actions centered around action a_i, represented by:

$$S_i = \{a_{i-W/2}, \dots, a_i, \dots, a_{i+W/2}\}, \tag{1}$$

where each action is a tuple $a_i = (v_i, t_i, y_i, d_i)$ consisting of a *video clip* v_i of $\mathcal{F}$ frames, a *free-form text narration* t_i as a textual description of the action, an action label y_i, and d_i identifies one of the N_s/N_t source/target domains, respectively.

We use pre-trained and frozen visual and text features extractors on the video clip v_i and text narration t_i to obtain the visual $\boldsymbol{X}_i^V \in \mathbb{R}^{W \times D_V}$ and textual $\boldsymbol{X}_i^T \in \mathbb{R}^{W \times D_T}$ features for each action a_i in the sequence. Here, W represents the length of the sequence, and D_V and D_T correspond to the feature size for the visual and textual input respectively.

To encourage the model to share information between the actions in the sequence, we exploit the self-attention mechanism of transformers [39], which naturally models relations between elements in a sequence. These are built as a stack of multi-head attention, feed-forward (FF) layers, residual connection, and normalization layers (LN).

Let $\boldsymbol{H}^l$ be the features encoded at the lth layer. Features produced at each layer of the transformer encoder are:

$$\boldsymbol{H}_{attn}^l = \texttt{LN}(f_{SA}(\boldsymbol{H}^l) + \boldsymbol{H}^l), \tag{2}$$

$$\boldsymbol{H}^{l+1} = \texttt{LN}(\texttt{FF}(\boldsymbol{H}_{attn}^l) + \boldsymbol{H}_{attn}^l), \tag{3}$$

where $f_{SA}(\cdot)$ is the self-attention operator defined as:

$$f_{SA} = \sigma\left(\frac{q(\boldsymbol{H}^l)k(\boldsymbol{H}^l)}{\sqrt{D}}\right)v(\boldsymbol{H}^l), \tag{4}$$

where q, k and v are learnable projections, D is the size of the input features and $\sigma(\cdot)$ is the softmax function.

We feed $\boldsymbol{X}_i^V$ to a vanilla transformer encoder $\texttt{ENC}^{\texttt{V}}$ to aggregate information from all the actions in the sequence. As transformers are permutation invariant, we employ positional encoding to preserve information about the order of actions in the sequence. These are learnable vectors $p \in \mathbb{R}^{W \times D_p}$, which are added to the action features to indicate their absolute position inside the sequence. Finally, we append a learnable classification token to the end of the sequence, denoted as $\texttt{CLS} \in \mathbb{R}^{D_V}$, as in [9]. The inputs $\boldsymbol{X}_i^V$ and $\boldsymbol{X}_i^T$ are transformed as follows:

$$\begin{aligned} \boldsymbol{Z}_i^V &= \texttt{ENC}^{\texttt{V}}\left([\boldsymbol{X}_i^V + p_i,\ \texttt{CLS}]\right), \\ \boldsymbol{Z}_i^T &= \mathcal{I}\left(\boldsymbol{X}_i^T\right), \end{aligned} \tag{5}$$

where $\boldsymbol{Z}_i^V \in \mathbb{R}^{(W+2)\times D}$ and $\boldsymbol{Z}_i^T \in \mathbb{R}^{W \times D}$ are the final output embeddings encoding the relationships between different actions in the sequence and $\mathcal{I}$ indicates the identity function. This process is illustrated in Fig. 3.

3.2. Sequence reconstruction

We mask the visual or textual features of the central action in the sequence and learn how to reconstruct it using the surrounding actions. As shown in Section 4.5, this increases the benefits of using sequences across domains. Formally, given the visual $\boldsymbol{Z}_i^V$ and textual $\boldsymbol{Z}_i^T$ features, we zero out the central actions:

$$\begin{aligned} \bar{\boldsymbol{Z}}_i^V &= \{z_{i-W/2}^V, \dots, \bar{z}_i^V, \dots, z_{i+W/2}^V\}, \\ \bar{\boldsymbol{Z}}_i^T &= \{z_{i-W/2}^T, \dots, \bar{z}_i^T, \dots, z_{i+W/2}^T\}, \end{aligned} \tag{6}$$

where $\bar{z}_i^V$ and $\bar{z}_i^T$ represent the masked features. Two separate transformer decoders, $\texttt{DEC}^{\texttt{V}}$ and $\texttt{DEC}^{\texttt{T}}$, are trained on $\bar{\boldsymbol{Z}}_i^V$ and $\bar{\boldsymbol{Z}}_i^T$ to reconstruct the original visual and textual input $\boldsymbol{Z}_i^V$ and $\boldsymbol{Z}_i^T$ respectively. We learn to reconstruct the masked visual features with the guidance

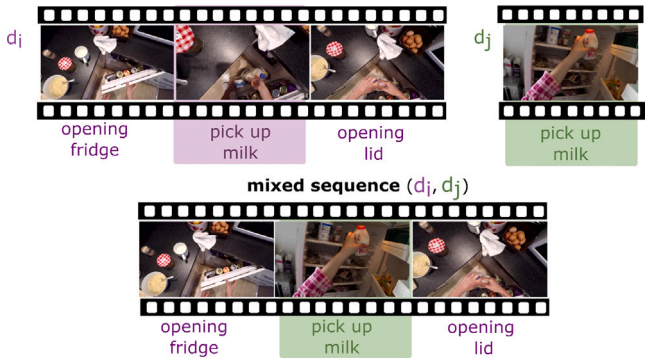


Fig. 4. SeqMix. Given a sequence from domain d_i, we replace action a_i in the sequence with a different action a_j from another sequence belonging to a different domain d_j.

of the unmasked textual features and vice versa. To do this, during the decoding process, the masked visual features are combined with the unmasked textual features using a cross-attention mechanism [39]. The resulting reconstructed vectors $\hat{\boldsymbol{Z}}_i^V$ and $\hat{\boldsymbol{Z}}_i^T$ are:

$$\hat{\boldsymbol{Z}}_i^V = \texttt{DEC}^{\texttt{V}}(\bar{\boldsymbol{Z}}_i^V, \boldsymbol{Z}_i^T), \qquad \hat{\boldsymbol{Z}}_i^T = \texttt{DEC}^{\texttt{T}}(\bar{\boldsymbol{Z}}_i^T, \boldsymbol{Z}_i^V). \tag{7}$$

In the decoding stage, each decoder layer operates similarly to Eqs. (2) and (4), where f_{SA} is replaced by the cross-attention operator f_{CA}. At each visual and text decoder layer $\boldsymbol{K}_V^l$ and $\boldsymbol{K}_T^l$, cross-attention f_{CA} is defined as:

$$f_{CA}(\boldsymbol{K}_V^l, \boldsymbol{K}_T^l) = \sigma\left(\frac{q(\boldsymbol{K}_V^l)k(\boldsymbol{K}_T^l)}{\sqrt{D}}\right)v(\boldsymbol{K}_V^l) \tag{8}$$

where q, k and v are learnable projections of the input features, $\boldsymbol{K}_V^l$ and $\boldsymbol{K}_T^l$ are features decoded at the lth layer by the visual and textual decoder respectively, D is the size of the input features and $\sigma(\cdot)$ is the softmax function. The decoders $\texttt{DEC}^{\texttt{V}}$ and $\texttt{DEC}^{\texttt{T}}$ are trained by minimizing the L2 distance between the original and the reconstructed visual features ($\mathcal{L}_{rV}$) and the cross-entropy loss between reconstructed textual features and the original text narrations ($\mathcal{L}_{rT}$) respectively.

Alternative strategies could be employed to integrate visual and textual modalities in the reconstruction process, for example, by concatenating their embeddings or by using co-attention. We opt for cross-attention as it can be integrated nicely in SeqDG architecture by adding cross-attention layers after each decoder layer. This way, each decoder layer implements temporal reasoning over the sequence independently within each modality, while cross-attention allows modalities to peek at each other. This approach effectively integrates cross-modal cues in the reconstruction process, without explicitly mixing the two modalities, which might hurt the unimodal reconstruction objective of SeqDG.

3.3. SeqMix: Sequence mix augmentation

Sequences $\mathcal{S}_i$ are composed of actions from the same video, thus sharing a similar visual appearance of the environment. We generate new "mixed sequences" of actions from different source domains to train the network. The rationale behind this approach is to increase the diversity of the training sequences, thereby enhancing the model's ability to generalize to sequences from different domains [40].

During training, given an action sequence $\mathcal{S}_i$, we randomly replace (with probability $p = 0.5$) an action a_i in the sequence with a different action a_j from another sequence $\mathcal{S}_j$. The two actions are from different visual domains but share the same action label:

$$(a_i, a_j): \quad d_i \neq d_j \ \wedge \ y_i = y_j \tag{9}$$

where d_i, d_j and y_i, y_j are the domain and action labels of the two samples respectively. Both d_i and d_j are domains from the source split of the dataset, i.e. $d_i, d_j \in \{\mathcal{D}_1^s, \dots, \mathcal{D}_{N_s}^s\}$. An example of SeqMix is shown in Fig. 4.

3.4. Training losses

For action recognition, we add a classification head, denoted as h, which receives the transformed classification token $\boldsymbol{Z}_{0:1}^v$ as input, producing the action classification logits $\hat{y}_i$ as output:

$$\hat{y}_i = h(\boldsymbol{Z}_{0:1}^v). \tag{10}$$

The classification head is trained using cross-entropy loss on the central action of the sequence ($\mathcal{L}_C$) using the corresponding ground truth label $\hat{y}_i$, while the visual and textual reconstruction losses ($\mathcal{L}_{rV}$, $\mathcal{L}_{rT}$) are defined as the mean squared error loss between the reconstructed features $\hat{\boldsymbol{Z}}_i^V$ and $\hat{\boldsymbol{Z}}_i^T$ and the original unmasked features $\boldsymbol{Z}_i^V$ and $\boldsymbol{Z}_i^T$, respectively. The network is trained jointly using the sum of the classification and reconstruction losses. Overall, the training objective of SeqDG is defined as:

$$\mathcal{L} = \mathcal{L}_C + \lambda_{rV}\mathcal{L}_{rV} + \lambda_{rT}\mathcal{L}_{rT} \tag{11}$$

where λ_{rV} and λ_{rT} weight the two reconstruction losses for the visual and textual features.

3.5. Inference

The encoder–decoder architecture, trained with text narrations through the reconstruction task, encourages the visual encoder $\texttt{ENC}^V$ to more effectively integrate information from surrounding actions. At test time, sequences made of only video clips v_i are processed by the visual encoder $\texttt{ENC}^V$ and classifier h. The prediction is performed on the central action and a sliding window mechanism is used for sequential testing. Textual annotations are not used at inference time.

4. Experiments

4.1. Datasets

We evaluate SeqDG on the UDA benchmark of the EPIC-KITCHENS-100 dataset [10], a large collection of human activities recorded in a kitchen environment. EPIC-KITCHENS-100 provides over 100 h of videos recorded in 45 different kitchens. The dataset consists of two splits, *source* and *target*, containing labeled and unlabeled samples respectively, with videos captured in distinct environments by different participants, thus making the shift between the two splits particularly significant. Each visual domain $\mathcal{D}_i$ in the dataset corresponds to a different kitchen in which actions were collected. Actions are annotated with *(verb, noun)* pairs from a set of 97 verbs and 300 nouns. Models are evaluated in terms of Top-1 and Top-5 accuracy for Verb, Noun and Action predictions. The latter is a combination of the Verb and Noun labels and is used to evaluate the ability of the network to predict both. We also evaluate SeqDG on the EGTEA [42] dataset, which consists of 28 h of videos with about 10k annotated actions from a set of 106 labels. Videos are captured by 32 subjects over 86 unique sessions, which makes it a valuable testbed to evaluate the better generalization offered by sequence-based reasoning.

4.2. Setting

For EPIC-KITCHENS-100, we consider two different experimental settings. In the *Intra-Domain* setting, the train and validation/test samples are from the same visual domains, i.e. clips have been recorded in the same kitchens in both splits, whereas, in the *Cross-Domain* setting, they are from different kitchens (*location shift*) or from the same kitchen but recorded after a long temporal interval of several years (*temporal shift*). In this setting, we further distinguish between UDA methods that use unlabeled target data during training and DG methods that only learn from source data.

Table 1
Experimental results on the EPIC-KITCHENS-100 [41] UDA benchmark, comparing both UDA and DG methods in the *Cross-Domain* setting (target validation set). Models are evaluated in terms of Top-1 and Top-5 Verb, Noun and Action accuracy (%). For each method, we indicate with ▲ the improvement over the Source Only.

Method	Backbone	Sequence	Modalities			Top-1 accuracy (%)			Top-5 accuracy (%)		
			RGB	Flow	Audio	Verb	Noun	Action	Verb	Noun	Action
UDA											
Source Only	TBN-TRN [24]	–	✓	✓	✓	46.7	27.8	19.2	75.4	48.3	42.2
TA3N [30]	TBN-TRN [24]	–	✓	✓	✓	48.5	28.9	19.6 (▲ +0.4)	76.0	50.1	43.4 (▲ +1.2)
Source Only	TBN-TRN [24]	–	✓	✓	✓	47.1	27.4	19.0	75.3	49.4	41.8
MM-SADA [11]	TBN-TRN [24]	–	✓	✓	✓	48.4	28.3	19.3 (▲ +0.3)	77.6	50.6	43.4 (▲ +1.6)
Source Only	TBN-TRN [24]	–	✓	✓	✓	47.6	28.4	19.6	–	–	–
CIA [31]	TBN-TRN [24]	–	✓	✓	✓	48.3	29.5	20.3 (▲ +0.7)	–	–	–
Source Only	TBN-TRN [24]	–	✓	✓	✓	46.7	26.7	18.2	75.3	48.4	41.3
RNA [33]	TBN-TRN [24]	–	✓	✓	✓	**50.8**	29.1	20.0 (▲ +1.8)	80.8	52.1	46.0 (▲ +4.7)
DG											
Source Only	TBN-TRN [24]	–	✓	✓	✓	47.2	27.4	19.0	75.3	49.4	41.8
MM-SADA (SS) [11]	TBN-TRN [24]	–	✓	✓	✓	47.8	27.9	19.2 (▲ +0.2)	77.1	49.8	42.9 (▲ +1.1)
Source Only	TBN-TRN [24]	–	✓	✓	✓	46.7	26.7	18.2	75.3	48.4	41.3
RNA [33]	TBN-TRN [24]	–	✓	✓	✓	50.7	27.9	19.8 (▲ +1.6)	80.6	51.3	45.3 (▲ +4.0)
Source Only	TBN-TRN [24]	–	✓	✓	✓	49.0	29.1	19.4	80.7	52.4	45.8
CIR [4][a](No text)	TBN-TRN [24]	–	✓	✓	✓	49.4	29.1	19.4 (▲ +0.0)	80.6	53.4	46.8 (▲ +1.0)
CIR [4][a]	TBN-TRN [24]	–	✓	✓	✓	48.8	29.0	19.4 (▲ +0.0)	81.0	53.2	46.9 (▲ +1.1)
Source Only	TBN-TRN [24]	–	✓	✓	✓	49.0	29.1	19.4	80.7	52.4	45.8
EgoZAR [5]	TBN-TRN [24]	–	✓	✓	✓	50.0	29.5	20.3 (▲ +0.9)	**81.1**	**53.6**	**46.9** (▲ +1.1)
Source Only	TBN-TRN [24]	✓	✓	✗	✓	38.3	22.4	14.3	72.3	42.6	36.4
MTCN [9]	TBN-TRN [24]	✓	✓	✗	✓	39.8	25.5	14.8 (▲ +0.5)	75.0	46.6	40.0 (▲ +3.6)
Source Only	TBN-TRN [24]	✓	✓	✗	✓	38.7	23.8	14.8	75.3	45.3	38.8
SeqDG	TBN-TRN [24]	✓	✓	✗	✓	41.3	26.8	16.9 (▲ +2.1)	75.8	48.7	41.5 (▲ +2.7)
Source Only	TBN-TRN [24]	✓	✓	✓	✓	46.4	26.6	18.2	76.8	51.9	42.2
SeqDG (No text)	TBN-TRN [24]	✓	✓	✓	✓	47.5	28.9	19.5 (▲ +1.3)	78.9	50.6	44.1 (▲ +1.9)
SeqDG	TBN-TRN [24]	✓	✓	✓	✓	49.1	**29.8**	**20.6** (▲ +2.4)	79.7	52.6	45.8 (▲ +3.6)

[a] Reproduced results from [5].

4.3. Implementation details

For EPIC-KITCHENS-100, we use pre-extracted features using the TBN [24] architecture pre-trained on the source split of the dataset, which are the ones officially provided for that setting [41]. Features are extracted from 25 uniformly sampled clips for each input action and have dimension 1024. For EGTEA, we use the SlowFast [15] pre-trained features released by the authors of MTCN [9] for a fair comparison with previous methods. These features are extracted from 10 clips for each action and have dimension 2304. For EPIC-KITCHENS-100 experiments, we sample 5 clips out of the 25 available, while all 10 clips are used for EGTEA. A Temporal Relation Network (TRN) [43] layer is used to summarize information across different clips for each action. For text, we use the provided narrations from the EPIC-KITCHENS-100 dataset, which consists of simple verb and noun combinations.

We encode the narrations using the BERT pre-trained model [44]. Textual features have size 768. To align the sizes of the textual and visual features, we project the latter using a fully connected layer with output size 768. For EPIC-KITCHENS-100, we use separate classification tokens for the Verb and Noun labels and apply the cross entropy classification loss to each. SeqDG is trained with the SGD optimizer, batch size 32 and learning rate 0.005. Training lasts 100 and 50 epochs for EPIC-KITCHENS-100 and EGTEA respectively. Learning rate is decreased by a factor 10 at epochs 50 and 75 for EPIC-KITCHENS-100, and at epochs 25 and 38 for EGTEA.

4.4. Comparison with state-of-the-art

In line with previous works [31,33], we report each method with the corresponding baseline reported in the original paper, indicated as *Source Only*, which allows to compare relative improvements of each method.

Cross-Domain Analysis. We compare in Table 1 SeqDG with previous state-of-the-art methods in UDA and DG on the EPIC-KITCHENS-100 dataset. For a fair comparison, we include results across all modalities, i.e., RGB, Optical Flow, and Audio, and results on RGB and Audio only, to fairly compare with MTCN [9]. TA3N [30], MM-SADA [11] and CIA [31] use data from the target domain to perform adversarial domain alignment. RNA [33] learns to align features from different modalities and domains. Additionally, CIR [4] and EgoZAR [5] can only use source train data, where CIR represents an action as a weighted combination of actions from other scenarios and locations and EgoZAR extracts domain-agnostic representations for activity-centric zones. SeqDG demonstrates better or comparable performance than these methods, in terms of both absolute accuracy and relative improvements compared to the baseline of each method, i.e., the corresponding Source Only. In particular, when considering the combination of verb and noun (Action accuracy), SeqDG achieves 20.6% accuracy, while the best performing method, EgoZAR, achieves 20.3%.

When looking at relative improvements with respect to the Source Only, SeqDG improves by up to 2.4%, while the method with the highest improvement in this case, RNA, improves by up to 1.8%. Overall, we observe that SeqDG performs best in the Noun and Action metrics, with state-of-the-art Top-1 accuracy. Slightly lower improvements are observed in the Verb metrics, where RNA achieves slightly better results (50.8% vs. 49.1%). Intuitively, we observe that manipulated objects tend to remain the same between consecutive actions. For example, an apple is first washed, then peeled and finally cut. Differently, verbs may be more diverse and their temporal order may vary, making the reconstruction objective more challenging. Note that, unlike these methods, SeqDG does not use the target domain and can also work with only one modality, differently from multi-modal approaches like RNA.

Table 2
Results on the Intra-Domain setting of EPIC-KITCHENS-100 using RGB and Audio modalities.

Method	Top-1 Acc. (%)			
	Backbone	Verb	Noun	Action
Source Only	TBN-TRN [24]	60.0	42.8	30.8
MTCN [9]	TBN-TRN [24]	60.6	48.2	33.0
MTCN[a] [9]	TBN-TRN [24]	61.5	49.3	34.3
Source Only	TBN-TRN [24]	60.1	42.8	31.1
SeqDG	TBN-TRN [24]	**63.9**	49.1	36.1
SeqDG[a]	TBN-TRN [24]	63.6	**49.7**	**36.3**

[a] Test time predictions filtering using a LM [9].

Table 3
Results on the EGTEA dataset in terms of Top-1 accuracy (%) and Mean Class accuracy (%).

Method	Backbone	Top-1 Acc. (%)	Mean class Acc. (%)
Kapidis et al. [45]	MFNet [46]	68.9	60.5
Lu et al. [47]	I3D [14]	68.6	60.5
SlowFast [15]	SlowFast [15]	70.4	61.9
Min et al. [48]	I3D [14]	68.5	62.8
Shiota et al. [49]	SlowFast [15]	69.3	60.9
MTCN [9]	SlowFast [15]	72.5	64.8
SeqDG	SlowFast [15]	74.0	66.5
MTCN[a] [9]	Slowfast [15]	73.5	65.8
SeqDG[a]	Slowfast [15]	**74.1**	**66.9**

[a] Test time predictions filtering using a LM [9].

Table 4
Ablation study on the different components of SeqDG on EPIC-KITCHENS-100both on RGB-only and all modalities (RGB, Flow, Audio).

Setting	Seq.	SeqMix	$\mathcal{L}_{rV}$	$\mathcal{L}_{rT}$	Top-1 Acc. (%)		
					Verb	Noun	Action
RGB-only	–	–	–	–	33.5	21.6	11.6
RGB, Flow, Audio	–	–	–	–	46.4	26.6	18.2
RGB-only	✓	–	–	–	32.9	22.7	12.0
RGB, Flow, Audio	✓	–	–	–	46.6	28.1	19.3
RGB-only	✓	✓	–	–	33.9	23.2	12.1
RGB, Flow, Audio	✓	✓	–	–	47.5	29.1	19.7
RGB-only	✓	✓	✓	–	33.5	23.7	12.4
RGB, Flow, Audio	✓	✓	✓	–	47.9	28.6	19.8
RGB-only	✓	✓	–	✓	34.2	23.6	12.2
RGB, Flow, Audio	✓	✓	–	✓	47.7	29.2	19.8
RGB-only	✓	✓	✓	✓	34.3	24.2	12.8
RGB, Flow, Audio	✓	✓	✓	✓	**49.1**	**29.8**	**20.6**

The closest approach to SeqDG is MTCN [9]. Despite not being designed for Domain Generalization, its sequence-based architecture is similar to ours. We compare MTCN and SeqDG using the same set of modalities, i.e. RGB and Audio. The small improvements of MTCN in the *Cross-Domain* setting compared to its baseline (+0.5%) show that naively using information from neighboring actions is not sufficient across different domains. Instead, SeqDG is built to explicitly exploit action sequences to improve generalization, which results in consistent improvements over MTCN (14.8% vs. 16.9%).

Finally, to demonstrate that the improvements of SeqDG are not merely due to the integration of text in the reconstruction process, we also include a variant of our approach that does not use text entirely (SeqDG (No text)). We observe that even this variant of SeqDG significantly improves over the baseline (+1.3%), underlining the importance of exploiting action sequences in a cross-domain scenario.

Intra-Domain Analysis. In Table 2 we also evaluate the performance of SeqDG in the intra-domain setting of EPIC-KITCHENS-100, i.e. the visual domains are shared between the source and target domains. This setting better highlights SeqDG's improvements in exploiting action sequences for action recognition, even when domain shift is not present. Due to the similarity that exists between MTCN and SeqDG in using sequence, we compare SeqDG with MTCN using the same modalities, i.e. RGB and Audio. Additionally, MTCN uses Language Model (LM) at test time to filter predictions by looking for the most likely sequence using the top-k predictions for each action. To ensure a fair comparison, we adopt the same approach at test-time for both MTCN and SeqDG, which we refer to as MTCN† and SeqDG†. Overall, SeqDG outperforms MTCN by 3.1% when not using text information at test-time, and by 2% when using it. We observe that the effect of filtering predictions at test time using text is more limited with SeqDG† compared to MTCN†, indicating that the network has already learned to effectively use text in the reconstruction process.

We report in Table 3 results of SeqDG on the EGTEA dataset. SeqDG outperforms all existing methods, achieving SOTA results on both Top-1 accuracy (%) and Mean Class accuracy (%). Also in this setting, we observe limited benefits by filtering predictions at test time using text, which further confirms SeqDG learns to effectively exploit text during its training process. Most notably, SeqDG without filtering outperforms MTCN with predictions filtering.

Table 5
Multi-modal performance of SeqDG on EPIC-KITCHENS-100.

Setting	Modalities			Top-1 Acc. (%)		
	RGB	Flow	Audio	Verb	Noun	Action
Source Only	✓	–	–	33.5	21.6	11.6
SeqDG	✓	–	–	34.3	24.2	12.8
Source Only	✓	✓	–	42.1	24.3	15.3
SeqDG	✓	✓	–	45.0	28.4	18.0
Source Only	✓	–	✓	38.7	23.8	14.8
SeqDG	✓	–	✓	41.3	26.8	16.9
Source Only	✓	✓	✓	46.4	26.6	18.2
SeqDG	✓	✓	✓	**49.1**	**29.8**	**20.6**

Table 6
Comparison of different LMs for SeqDG(RGB-only).

Language model	Top-1 Acc. (%)		
	Verb	Noun	Action
CLIP [50]	34.4	23.2	12.3
MiniLM [51]	**34.8**	23.6	12.4
BERT [44]	34.3	**24.2**	**12.8**

Table 7
Comparison of different visual backbones for SeqDG on EPIC-KITCHENS-100 [41] using RGB and Flow modalities.

Method	Backbone	Top-1 accuracy (%)		
		Verb	Noun	Action
Source Only	TSN-TRN [16]	44.0	25.7	15.8
SeqDG	TSN-TRN [16]	46.0	27.4	17.6 (▲ +1.8)
Source Only	I3D-TRN [14]	46.9	25.8	16.6
SeqDG	I3D-TRN [14]	**48.7**	**28.7**	19.3 (▲ +2.7)
Source Only	TSM-TRN [17]	45.7	26.5	17.0
SeqDG	TSM-TRN [17]	48.5	28.3	**19.4** (▲ +2.4)
Source Only	TBN-TRN [24]	42.1	24.3	15.3
SeqDG	TBN-TRN [24]	45.0	28.4	18.0 (▲ +2.7)

4.5. Ablations

SeqDG components. In Table 4, we illustrate the impact of the different components of SeqDG. Specifically, we compare the use of sequences and individual actions for action recognition, the effect of SeqMix, and the two cross-modal reconstruction losses, $\mathcal{L}_{rV}$ and $\mathcal{L}_{rT}$. We find that using sequences improves results by 1.1% using RGB-only, and 1.5% with all modalities (RGB, Flow, and Audio) on Nouns and

Table 8
Ablation study of different sequence lengths on Top-1 accuracy (%) for action classification of EPIC-KITCHENS-100 (RGB-only) using SeqDG.

Method	Seq. Length	Top-1 accuracy (%)		
		Verb	Noun	Action
Source Only	1	33.5	21.6	11.6
SeqDG	3	34.0 (▲ +0.5)	23.2 (▲ +1.6)	12.1 (▲ +0.5)
SeqDG	5	34.3 (▲ +0.8)	**24.2** (▲ **+2.6**)	**12.8** (▲ **+1.2**)
SeqDG	7	**34.8** (▲ **+1.3**)	23.3 (▲ +1.7)	12.6 (▲ +1.0)

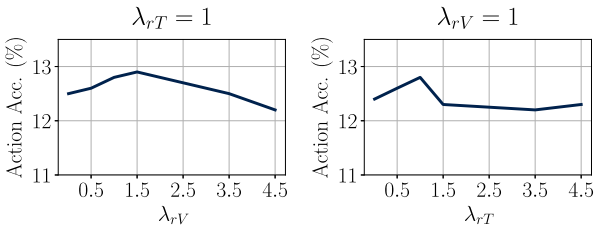


Fig. 5. Parameter analysis of the weights associated with the visual and textual reconstruction losses of SeqDG on EPIC-KITCHENS-100 (RGB-only).

0.4% using RGB-only, and 1.1% with all modalities (RGB, Flow, and Audio) on Action. SeqMix further improves results for both Verbs and Nouns by up to 1%. Mixing actions from different sequences is a key component of SeqDG to improve generalization, as it encourages the encoder to focus more on the semantics of the actions rather than their visual appearance, which is more dependent on the domain. Using the reconstruction losses on textual features ($\mathcal{L}_{rT}$) and visual features ($\mathcal{L}_{rV}$) further improves performance.

Ablation on different modalities. Other modalities, such as Audio and Optical flow, may also experience domain shift in forms other than visual appearance. For example, objects' shape and material may affect how interactions are captured by optical flow and audio, respectively. We show the multi-modal performance of SeqDG in Table 5, observing even larger performance improvements compared to the single modality setting. This not only indicates that our approach can be easily extended to other modalities, but also shows that the complementary information brought by them helps in the reconstruction process.

Language Model. We evaluate different Language Models for SeqDG(RGB only) to extract the textual features used in the reconstruction process in Table 6. Despite the LMs having very different model sizes, we observe comparable results for Verb and Noun accuracy. We remark that text is only used to guide the reconstruction, but it is not part of the inference process.

Backbone. In Table 7, we present the results of SeqDG on TSN [16], I3D [14], TSM [17], and TBN [24], using the same subset of modalities (RGB, Flow) available for all models to ensure a fair comparison. For each backbone, we compare the performance of the Source Only baseline with SeqDG. Critically, across *all* backbone architectures, SeqDG consistently outperforms the Source Only baseline, highlighting that the method is effective and agnostic to the specific visual encoder used.

Sequence length. In Table 8, we analyze the performance of SeqDG with different sequence lengths (W). We observe the best performance with $W = 5$. With shorter sequences, the model receives limited temporal information from both past and future actions. Conversely, increasing the sequence length beyond a certain point may not consistently improve performance, as it can introduce misleading actions that are weakly related to the target action and the overall sequence context, making them less useful for sequence reconstruction (SeqRec).

Variation on hyper-parameters Fig. 5 shows the effect of different loss weights for the visual and textual reconstruction losses, when one of two is fixed and the other is varying. Overall, we observe the best performance when the losses weight is equal to one ($\lambda_{rV} = \lambda_{rT} = 1.0$). These results can be interpreted as higher weights for the SeqRec components, leading to less focus on classification, while the lower values for λ_{rV} and λ_{rT} reduce the knowledge distillation capability of SeqDG.

5. Conclusion and future works

In this work, we propose a method that leverages sequence consistency across diverse environments to improve generalization in egocentric action recognition tasks. We validate the effectiveness of our approach through extensive experiments on widely-used egocentric action recognition datasets, achieving state-of-the-art performance. Future work could focus on identifying action patterns of varying lengths that generalize well across different scenarios while distinguishing user-specific behaviors. This would enhance our model's adaptability to more dynamic environments.

CRediT authorship contribution statement

Amirshayan Nasirimajd: Writing – review & editing, Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **Chiara Plizzari:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Simone Alberto Peirone:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Conceptualization. **Marco Ciccone:** Writing – review & editing, Supervision. **Giuseppe Averta:** Writing – review & editing, Supervision. **Barbara Caputo:** Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Marco Ciccone reports financial support was provided by FAIR - Future Artificial Intelligence Research. Giuseppe Averta reports financial support was provided by FAIR - Future Artificial Intelligence Research. Barbara Caputo reports financial support was provided by FAIR - Future Artificial Intelligence Research. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.patrec.2025.06.010.

Data availability

The dataset we used is publicly available.

References

[1] E. OhnBar, K. Kitani, C. Asakawa, Personalized dynamics models for adaptive assistive navigation systems, in: CORL, 2018.
[2] H.S. Park, J.-J. Hwang, Y. Niu, J. Shi, Egocentric future localization, in: CVPR, 2016.
[3] C. Plizzari, G. Goletto, A. Furnari, S. Bansal, F. Ragusa, G.M. Farinella, D. Damen, T. Tommasi, An outlook into the future of egocentric vision, in: IJCV, 2024.
[4] C. Plizzari, T. Perrett, B. Caputo, D. Damen, What can a cook in italy teach a mechanic in india? action recognition generalisation over scenarios and locations, in: ICCV, 2023.
[5] S.A. Peirone, G. Goletto, M. Planamente, A. Bottino, B. Caputo, G. Averta, Egocentric zone-aware action recognition across environments, Pattern Recognit. Lett. (2025).

[6] M. Kowal, M. Siam, M.A. Islam, N.D. Bruce, R.P. Wildes, K.G. Derpanis, A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information, in: CVPR, 2022.
[7] A. Torralba, A.A. Efros, Unbiased look at dataset bias, in: CVPR, 2011.
[8] R.P. Cooper, T. Shallice, Hierarchical schemas and goals in the control of sequential behavior., 2006.
[9] E. Kazakos, J. Huh, A. Nagrani, A. Zisserman, D. Damen, With a little help from my temporal context: Multimodal egocentric action recognition, in: BMVC, 2021.
[10] D. Damen, H. Doughty, G.M. Farinella, A. Furnari, J. Ma, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, M. Wray, Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100, in: IJCV, 2022.
[11] J. Munro, D. Damen, Multi-modal domain adaptation for fine-grained action recognition, in: CVPR, 2020.
[12] X. Song, S. Zhao, J. Yang, H. Yue, P. Xu, R. Hu, H. Chai, Spatio-temporal contrastive domain adaptation for action recognition, in: CVPR, 2021.
[13] D. Kim, Y.-H. Tsai, B. Zhuang, X. Yu, S. Sclaroff, K. Saenko, M. Chandraker, Learning cross-modal contrastive features for video domain adaptation, in: ICCV, 2021.
[14] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: CVPR, 2017.
[15] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: ICCV, 2019.
[16] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, L. Van Gool, Temporal segment networks: Towards good practices for deep action recognition, in: ECCV, 2016.
[17] J. Lin, C. Gan, S. Han, Tsm: Temporal shift module for efficient video understanding, in: ICCV, 2019.
[18] C. Quattrocchi, A. Furnari, D. Di Mauro, M.V. Giuffrida, G.M. Farinella, Synchronization is all you need: Exocentric-to-egocentric transfer for temporal action segmentation with unlabeled synchronized video pairs, in: ECCV, 2024.
[19] Z. Tu, J. Zhang, H. Li, Y. Chen, J. Yuan, Joint-bone fusion graph convolutional network for semi-supervised skeleton action recognition, in: IEEE TMM, 2022.
[20] Y. Liu, J. Yuan, Z. Tu, Motion-driven visual tempo learning for video-based action recognition, in: IEEE TIP, 2022.
[21] S.A. Peirone, F. Pistilli, A. Alliegro, G. Averta, A backpack full of skills: Egocentric video understanding with diverse task perspectives, in: CVPR, 2024.
[22] G. Radevski, D. Grujicic, M. Blaschko, M.-F. Moens, T. Tuytelaars, Multimodal distillation for egocentric action recognition, in: ICCV, 2023.
[23] M. Ramazanova, A. Pardo, H. Alwassel, B. Ghanem, Exploring missing modality in multimodal egocentric datasets, 2024, arXiv preprint arXiv:2401.11470.
[24] E. Kazakos, A. Nagrani, A. Zisserman, D. Damen, Epic-fusion: Audio-visual temporal binding for egocentric action recognition, in: ICCV, 2019.
[25] F. Xiao, Y.J. Lee, K. Grauman, J. Malik, C. Feichtenhofer, Audiovisual slowfast networks for video recognition, 2020, arXiv preprint arXiv:2001.08740.
[26] J. Chalk, J. Huh, E. Kazakos, A. Zisserman, D. Damen, Tim: A time interval machine for audio-visual action recognition, in: CVPR, 2024.
[27] M. Zhang, Y. Huang, R. Liu, Y. Sato, Masked video and body-worn IMU autoencoder for egocentric action recognition, in: ECCV, 2024.
[28] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, Vivit: A video vision transformer, in: ICCV, 2021.
[29] Y. Sun, H. Xue, R. Song, B. Liu, H. Yang, J. Fu, Long-form video-language pre-training with multimodal temporal contrastive learning, in: NeurIPS, 2022.
[30] M.-H. Chen, Z. Kira, G. AlRegib, J. Yoo, R. Chen, J. Zheng, Temporal attentive alignment for large-scale video domain adaptation, in: ICCV, 2019.
[31] L. Yang, Y. Huang, Y. Sugano, Y. Sato, Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition, in: CVPR, 2022.
[32] Z. Yao, Y. Wang, J. Wang, S.Y. Philip, M. Long, Videodg: Generalizing temporal relations in videos to novel domains, in: IEEE TPAMI, 2021.
[33] M. Planamente, C. Plizzari, S.A. Peirone, B. Caputo, A. Bottino, Relative norm alignment for tackling domain shift in deep multi-modal classification, in: IJCV, 2024.
[34] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: CVPR, 2022.
[35] Z. Tong, Y. Song, J. Wang, L. Wang, Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, in: NeurIPS, 2022.
[36] L. Wang, B. Huang, Z. Zhao, Z. Tong, Y. He, Y. Wang, Y. Wang, Y. Qiao, Videomae v2: Scaling video masked autoencoders with dual masking, in: CVPR, 2023.
[37] X. Geng, H. Liu, L. Lee, D. Schuurmans, S. Levine, P. Abbeel, Multimodal masked autoencoders learn transferable representations, 2022, arXiv preprint arXiv:2205.14204.
[38] S. Karamcheti, S. Nair, A.S. Chen, T. Kollar, C. Finn, D. Sadigh, P. Liang, Language-driven representation learning for robotics, 2023, arXiv preprint arXiv:2302.12766.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: NeurIPS, 2017.
[40] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, Mixup: Beyond empirical risk minimization, 2017, arXiv preprint arXiv:1710.09412.
[41] D. Damen, H. Doughty, G.M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, M. Wray, Scaling egocentric vision: The EPIC-KITCHENS dataset, in: ECCV, 2018.
[42] Y. Li, M. Liu, J.M. Rehg, In the eye of beholder: Joint learning of gaze and actions in first person video, in: ECCV, 2018.
[43] B. Zhou, A. Andonian, A. Oliva, A. Torralba, Temporal relational reasoning in videos, in: ECCV, 2018.
[44] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint arXiv:1810.04805.
[45] G. Kapidis, R. Poppe, E. van Dam, L. Noldus, R. Veltkamp, Multitask learning to improve egocentric action recognition, in: ICCVW, 2019.
[46] Y. Chen, Y. Kalantidis, J. Li, S. Yan, J. Feng, Multi-fiber networks for video recognition, in: ECCV, 2018.
[47] M. Lu, D. Liao, Z.-N. Li, Learning spatiotemporal attention for egocentric action recognition, in: ICCVW, 2019.
[48] K. Min, J.J. Corso, Integrating human gaze into attention for egocentric activity recognition, in: WACV, 2021.
[49] T. Shiota, M. Takagi, K. Kumagai, H. Seshimo, Y. Aono, Egocentric action recognition by capturing hand-object contact and object state, in: WACV, 2024.
[50] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: ICML.
[51] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, M. Zhou, Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, in: NeurIPS, 2020.