POLITECNICO DI TORINO
Repository ISTITUZIONALE

Egocentric zone-aware action recognition across environments

Original
Egocentric zone-aware action recognition across environments / Peirone, S.A., Goletto, G., Planamente, M., Bottino, A., Caputo, B., Averta, G.. - In: PATTERN RECOGNITION LETTERS. - ISSN 0167-8655. - 188:(2025), pp. 140-147. [10.1016/j.patrec.2024.12.008]

Availability:
This version is available at: 11583/2995679 since: 2025-01-08T09:50:14Z

Publisher:
Elsevier

Published
DOI:10.1016/j.patrec.2024.12.008

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)

29 November 2026

Pattern Recognition Letters 188 (2025) 140–147

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Egocentric zone-aware action recognition across environments

Simone Alberto Peirone *,[1], Gabriele Goletto[1], Mirco Planamente, Andrea Bottino, Barbara Caputo, Giuseppe Averta

Department of Control and Computer Engineering, Polytechnic of Turin Corso Castelfidardo, 34/d, Turin, 10138, Italy

ARTICLE INFO

Editor: Mauricio Pamplona Segundo

MSC:
68T45
68T05

Keywords:
First person (egocentric) vision
Domain generalization
Multimodal learning
Video analysis and understanding
Deep learning

ABSTRACT

Human activities exhibit a strong correlation between actions and the places where these are performed, such as washing something at a sink. More specifically, in daily living environments we may identify particular locations, hereinafter named *activity-centric zones*, which may afford a set of homogeneous actions. Their knowledge can serve as a prior to favor vision models to recognize human activities. However, the appearance of these zones is scene-specific, limiting the transferability of this prior information to unfamiliar areas and domains. This problem is particularly relevant in egocentric vision, where the environment takes up most of the image, making it even more difficult to separate the action from the context. In this paper, we discuss the importance of decoupling the domain-specific appearance of activity-centric zones from their universal, domain-agnostic representations, and show how the latter can improve the cross-domain transferability of Egocentric Action Recognition (EAR) models. We validate our solution on the EPIC-Kitchens-100 and Argo1M datasets.

1. Introduction

The privileged perspective offered by egocentric vision has proven highly effective in tracking human activities in daily life, thanks to the camera constantly following the wearer [1,2]. While providing an advantageous viewpoint on ongoing activities, the first-person perspective also brings the background remarkably close to the camera, inherently increasing its prominence in the field of view compared to third-person videos.

In this context, the concept of environmental affordance plays a pivotal role in connecting the wearer's activity with the underlying physical space. Specifically, the notion of *affordance* has been extensively studied in neuroscience and cognitive psychology since the seminal work of Gibson [3]. Affordances describe the potential actions or uses suggested by the physical characteristics of objects or the surrounding environment. This concept has recently gained attention in egocentric vision [4,5]. In particular, the work of Nagarajan et al. [6] refers to environmental affordances as *activity-centric zones*, defined as spatial locations, affording a coherent set of interactions, e.g. a sink or a stove in a kitchen.

This prompts us to explore whether and how activity-centric zones are currently exploited for egocentric video understanding models, connecting human actions with the persistent underlying environment. In particular, we demonstrate that the co-occurrence of specific actions in certain locations, predominantly present in egocentric vision, leads action recognition models to naturally learn a relationship between the actions and the locations in which they occur, exploiting it for making context-aware predictions.

This phenomenon, observable in the training data, is commonly known in computer vision as *co-occurrence bias* [7] and, in our case, it aids the model in identifying a limited set of potential actions based on what is visible in the camera's field of view. For instance, when a user looks at a sink, it is more likely that the action being performed is washing rather than cooking. This process mirrors how people, in their daily lives, use their understanding of objects and tools to navigate unfamiliar environments and identify the activities the environment can afford.

However, despite their ability to exploit activity-centric locations observed during training, we demonstrate that current Egocentric Action Recognition (EAR) networks lack a mechanism to explicitly model the contribution of the environment in their inference process on unobserved zones. In other words, the co-occurrence bias, which aids the model in autonomously learning environmental affordances, leads to confusion in predictions as soon as the appearance of the zone changes.

Indeed, extensive datasets of egocentric videos, such as EK100 [8] and Ego4D [9], have a number of actions (i.e. verb-noun combinations)

* Corresponding author.
E-mail addresses: simone.peirone@polito.it (S.A. Peirone), gabriele.goletto@polito.it (G. Goletto), mirco.planamente@polito.it (M. Planamente), andrea.bottino@polito.it (A. Bottino), barbara.caputo@polito.it (B. Caputo), giuseppe.averta@polito.it (G. Averta).
[1] These authors contributed equally.

https://doi.org/10.1016/j.patrec.2024.12.008
Received 21 February 2024; Received in revised form 13 September 2024; Accepted 14 December 2024
Available online 24 December 2024
0167-8655/© 2024 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

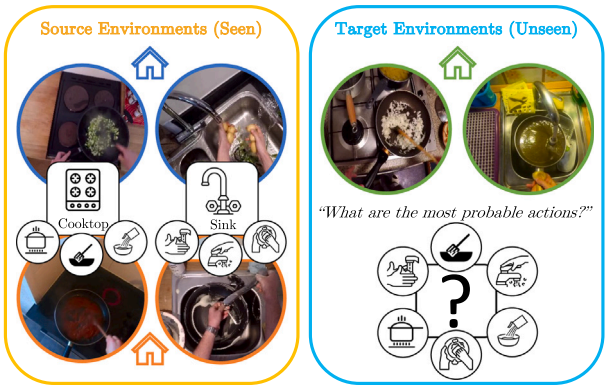


Fig. 1. The actions a person performs in a scene are closely related to the specific places where they are performed (*environmental affordances* [6]). Current egocentric action recognition models learn these correlations during training, but struggle when faced with an unfamiliar environment, losing context.

greatly exceeding the number of environments in which they were recorded. As a consequence, models trained on these datasets overly depend on appearance-based features, such as visual representations of objects and tools, for recognizing actions [7]. This reliance leads the models to effectively recognize activity-centric zones only when tested with data from the same training environments, struggling to exploit the environmental affordances in new domains, as shown in Fig. 1.

To address this issue, this work focuses on universal representations of the activity-centric zones, which - we show - have the potential to assist RGB models in removing domain-specific biases from the encoding of activity-centric zones. By leveraging a domain-agnostic representations of these locations, we aim to isolate the domain-specific representation of the activity-centric zones in EAR models and replace them with a more general, domain-independent equivalent, resulting in more general EAR models. The main goal of our work is to address two key questions: *how can we detect and identify these locations in real world conditions?* And, *can we use a domain-agnostic representation of these locations to improve the generalization capability of first person action recognition models?*

We evaluate our approach on the EPIC-Kitchens-100 (EK100) [8] and Argo1M [10] datasets in a Domain Generalization (DG) setting, where multiple source domains are available at training time but no target data can be accessed. In summary, this paper presents the following contributions:

- we shed light on the side-effect of the co-occurrence bias in egocentric video processing, which steers models in indirectly learning domain-dependent information about the environments (i.e. domain-specific activity-centric zones);
- we propose EgoZAR, an architecture which adopts more general representations of activity-centric zones to improve action recognition performance on unseen domains, enabling models to leverage the *environmental affordances* even in unknown zones;
- we demonstrate with extensive experiments on the EK100 and Argo1M datasets how replacing domain-specific environmental representations with their universal counterparts can help action recognition on unseen environments, achieving state-of-the-art Domain Generalization results on EK100 and competitive performance on Argo1M.

2. Related works

Objects affordances and activity-centric zones. James J. Gibson defined the term *affordances* in 1979 in the field of cognitive psychology [3] referring to the physical properties of an object (or environment) that support certain human actions and interactions. The concept of affordance is now being widely explored in computer vision [11,12], robotic manipulation [13,14] and navigation [15], and human–computer interaction [16]. Most of the previous works on the topic focus on human-object interactions [5,17,18], object grasping [19] and affordance detection [20]. The concept of affordances has also been recently generalized to scenes. Most notably, EGO-TOPO [6] extracts environmental affordances from egocentric videos and builds a topological map of the locations of the environment. These *so-called* activity-centric zones represent the main spatial regions in which actions may occur, driving interest towards their use in action recognition. More recently, Mur-Labadia et al. [4] built EPIC-Aff, a dataset based on EK100 providing multi-label pixel-wise affordance annotations with the camera pose.

Egocentric action recognition. Action recognition is one of the most studied tasks in egocentric vision [1]. The first architectures used in this context usually come from the third-person literature and fall into the categories of 2D CNN-based methods [21,22] and 3D CNN-based methods [23,24]. LSTM and its variants [25,26] followed this first wave to better encode temporal information. The most popular technique is the multi-modal approach [27,28], especially in EK100 competitions [8], to combine the complementary information provided by different modalities, e.g. RGB and optical flow. However, although optical flow has proven to be a strong modality for the action recognition task, it is computationally expensive. As shown in Crasto et al. [29], the use of optical flow limits the application of several methods in online scenarios, pushing the community either towards single-stream architectures [26,30], or to investigate alternative modalities [31,32].

Video Domain Adaptation. The goal of Unsupervised Domain Adaptation (UDA) is to close the gap between a labeled source domain and an unlabeled target domain. This task has been studied in detail in the context of image classification [33,34]. UDA for video analysis has been primarily focused on extending existing techniques to include the temporal dimension [35] and/or the multi-modal nature of videos [27].

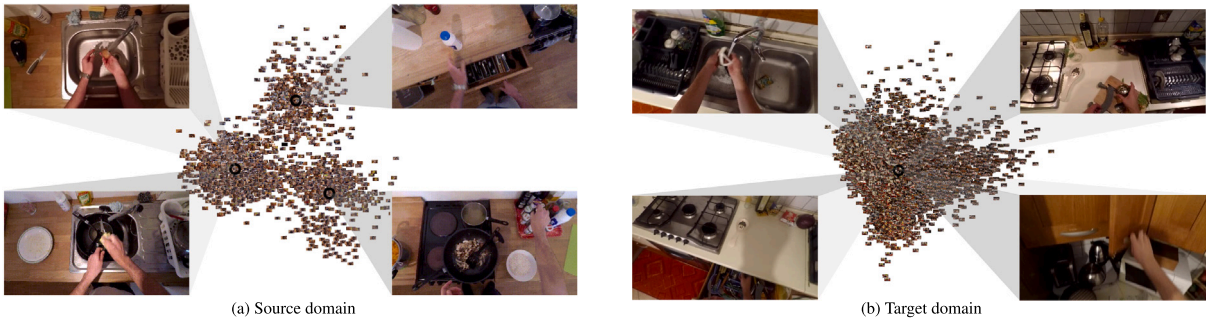
Unlike UDA, the goal of DG [36] is to improve generalization to out-of-distribution data without requiring access to the target data, using only data from one or more source training domains. DG has been studied in different contexts from object recognition [37], to semantic segmentation [38] and face recognition [39]. Applications to video are more scarce. Among these, RNA-Net [40] improves modalities cooperation on unseen scenarios by aligning feature norms. VideoDG [41] learns to align the local temporal features across different domains. CIR [10] reconstructs samples from different domains to learn more domain-agnostic representations. Unlike previous approaches, ours is the first to emphasize the importance of activity-centric zones in improving domain generalization.

3. Proposed method

Activity-centric zones provide useful insights into which actions are most likely to occur at a given location in the environment. However, exploiting these insights across different domains is not straightforward and requires models to reason about the location while ignoring its appearance. We provide more intuitions behind this behavior in Section 3.1. We describe how to extract domain-agnostic representations for activity-centric zones in Section 3.2 and show how these features can be integrated in an action recognition pipeline in Section 3.3.

3.1. Intuition

In egocentric vision, cameras are often positioned very close to the actions and the surrounding environments, causing RGB models to focus strongly on the environment. We observe empirical evidence of this phenomenon looking at the feature space of an EAR model (Fig. 2). Actions occurring in the same activity-centric zone are mapped to the



(a) Source domain (b) Target domain

Fig. 2. Feature space of an EAR model. On the left (Fig. 2(a)), the features obtained from a model trained and tested in the same environment are well separated based on the location where the actions are taking place. On the right (Fig. 2(b)), when the same model is used in a different environment, this clustering effect is not present anymore and different locations are mapped to the same region of the feature space.

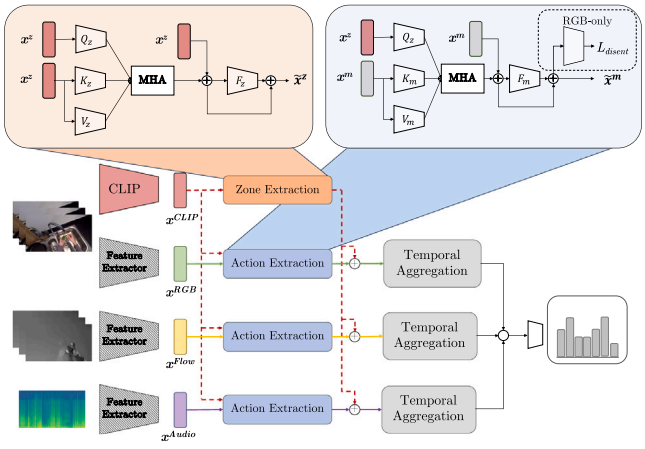


Fig. 3. Architecture of EgoZAR. Each modality m is processed using a separate features extractor to obtain the features x^m. CLIP features are then adopted both in the Zone and Action Extraction modules. This approach helps the network to focus on the activity-centric zones while minimizing the impact of environmental bias from the training domains. After temporal aggregation, the contributions from the different modalities y_M are combined to produce the final prediction of the action label.

same region of the feature space, regardless of their action label, which suggests that the model has learned a positive correlation between the environment and the set of actions that can be performed at a given location. This clustering in feature space highlights the importance of the environment for action recognition. However, this phenomenon does not transfer easily to new domains, as activity-centric zones are strongly coupled to their appearance, and models struggle to recognize the former (the semantic of the location) while ignoring the latter (its appearance). Indeed, comparing the feature spaces on the train and test data, which belong to different visual domains, reveals that the clustering by activity-centric zones is no longer present when evaluating test data, as shown on the right side of Fig. 2. To overcome this limitation, it is essential to allow the model to learn these activity-centric priors without the domain appearance bias, which is inherently present in video datasets and results from limited variability in the number of environments and locations represented. This would allow models to embed this prior knowledge about the distribution of actions in a given location while avoiding the negative influence of domain-specific biases that hinder generalization. Based on these observations, we identify two main challenges. First, the inclusion of domain-agnostic representations of the activity-centric zones into EAR models. Second, the development of a strategy for training an action recognition model that uses these domain-agnostic representations to leverage the contextual information provided by the environment.

3.2. Extracting activity-centric zone features

We propose a method that leverages visual-language models trained on large-scale image datasets as a *zone recognition model* to detect the activity-centric zones from the video stream. Indeed, being trained on millions of (image, caption) pairs sourced from the internet, these models are intrinsically able to recognize and generate similar features for the same location, e.g. a sink or a stove in a kitchen, across different environments. We use the features obtained from the zone recognition model to (i) include a domain-agnostic information from the environment, and (ii) remove the environment information from the input features of the action recognition model. We adopt an unsupervised clustering algorithm on the features of the zone recognition model to discover clusters in the features space that correspond to different locations in which the actions occur.

Given a dataset of egocentric human actions, we define each sample x_i as a triplet $x_i = (\mathbf{x}_i^m, \mathbf{x}_i^z, y_i)$, where $\mathbf{x}_i^m \in \mathbb{R}^{N \times D_m}$ and $\mathbf{x}_i^z \in \mathbb{R}^{N \times D_z}$ represent respectively the features extracted from an action recognition model $\mathcal{M}$ and a zone recognition model $\mathcal{Z}$ from N uniformly sampled clips across the input video segment. Additional implementation details on the features extraction processes are presented in Section 4.1. Zone features from all the samples in the training dataset $\mathbf{x}_i^s$ are averaged over the clip dimension and clustered using K-Means in the euclidean features space. This results in a set of K prototypes that represent the centers of the clusters $\mathbf{c}_k \in \mathbb{R}^{D_z}$, each corresponding to a different location. During training, each sample is assigned to the closest cluster using euclidean distance to obtain the corresponding activity-centric zone pseudo-label $y_i^z = \min_k \|\mathbf{x}_i^z - \mathbf{c}_k\|_2$.

3.3. Integration of the activity-centric zones

To integrate the prior information provided by the zone recognition model we propose EgoZAR (see Fig. 3). Our proposed architecture introduces two attention-based modules, namely the Zone Extraction (ZE) and Action Extraction (AE) modules, to explicitly separate the input features into two components, encoding zone and motion clues respectively. The ZE module extracts the relevant zone-related information from the zone features $\mathbf{x}_i^z$, while the AE module encourages the action recognition features $\mathbf{x}_i^m$ to ignore the zone and domain appearance biases they incorporate. These modules are implemented using Multi-Head Attention followed by a linear projection and a residual connection. Queries are computed from the zone features while keys and values are obtained from the zone or action features for the ZE and AE modules respectively. Formally, the updated features $\tilde{\mathbf{x}}_i^z$ and $\tilde{\mathbf{x}}_i^m$ are computed as follows:

$$\mathbf{o}_i^z = \mathbf{x}_i^z + \sigma\left(\frac{Q_z(\mathbf{x}_i^z)K_z(\mathbf{x}_i^z)^T}{\sqrt{D_z}}\right)\cdot V_z(\mathbf{x}_i^z), \qquad \tilde{\mathbf{x}}_i^z = \mathbf{o}_i^z + F_z(\mathbf{o}_i^z), \tag{1}$$

$$\mathbf{o}_i^m = \mathbf{x}_i^m + \sigma\left(\frac{Q_m(\mathbf{x}_i^z)K_m(\mathbf{x}_i^m)^T}{\sqrt{D_z}}\right)\cdot V_m(\mathbf{x}_i^m), \qquad \tilde{\mathbf{x}}_i^m = \mathbf{o}_i^m + F_m(\mathbf{o}_i^m), \tag{2}$$

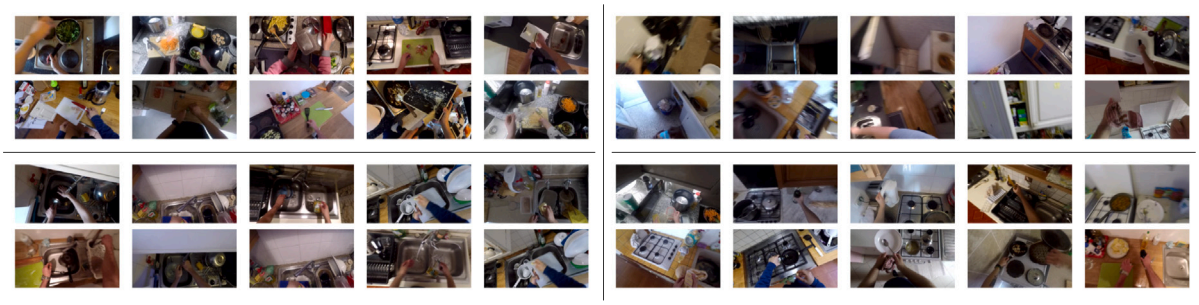


Fig. 4. Clusters obtained with K-Means (K = 4) on the CLIP features of EK100, showing how the same locations, for example sinks and stoves, but different kitchens are clustered together.

where Q, K and V represent the queries, keys and values projections of the features and F is a linear projection. Then, the updated features are concatenated on the clips dimension and fed to a TRN [42] layer:

$$\mathbf{x}_i = TRN\left([\tilde{\mathbf{x}}_i^z, \tilde{\mathbf{x}}_i^m]\right), \tag{3}$$

where $\mathbf{x}_i \in \mathbb{R}^{D_m}$ and TRN is implemented as a linear projection, followed by a Batch Normalization layer, a ReLU activation and a dropout layer. Finally, features $\mathbf{x}_i$ are fed to a linear classifier that outputs the action logits $\tilde{\mathbf{y}}_i$.

Disentanglement of the action features. The objective of the AE module is to leverage the clues brought by the zone features to remove the appearance component of the motion features. To encourage this behavior, we introduce an adversarial classifier on top of the output of the AE module $\tilde{\mathbf{x}}_i^m$. The objective of the classifier is to recognize the activity-centric zone from the motion features. As a result, these features are pushed to discard any residual zone information. The classifier is implemented as a two-layers MLP with hidden size 256, Batch Normalization and ReLU activations. The classifier outputs the activity-centric zone logits $\tilde{\mathbf{y}}_i^d$.

3.4. Training and inference

EgoZAR architecture is trained jointly using Cross Entropy loss on the action logits $\tilde{\mathbf{y}}_i$ and on the output of the adversarial activity-centric zone classifier $\tilde{\mathbf{y}}_i^d$ with the supervision of the zone pseudo-labels. Other modalities, such as optical flow and audio, are less impacted by environmental bias, even though they can still benefit from the contextual features extracted from the zone recognition model. As an example, an audio model aware of its proximity to a sink can more easily understand if sounds are linked to activities like washing, leveraging contextual clues for inference. When training with multiple input modalities, the network is replicated for each modality and the *modality-specific* action logits are averaged before computing the loss. In this context, the network is trained with a double Cross Entropy loss on both the fused (averaged) logits as well as on the *modality-specific* logits. The Disentanglement Cross Entropy loss is computed just on the RGB modality, as it is the modality most affected by domain appearance biases.

4. Experiments and results

4.1. Experimental setup

Dataset. We evaluate EgoZAR on the Unsupervised Domain Adaptation (UDA) benchmark subset from EK100 [8], a large dataset of fine-grained activities in a kitchen environment. The dataset includes two forms of domain shift: (i) each participant records its actions in a different kitchen (*location shift*) and (ii) the source and target splits partially share environments, although they are separated by a time interval (*time shift*). Our approach focuses on Domain Generalization (DG), thus the target split is not used during the training process. Each action is annotated using a *(verb, noun)* pair from a combination of 97 verb classes and 300 noun classes. Performances are reported using Top-1 and Top-5 accuracies for verbs, nouns and actions on the target validation set of EK100.

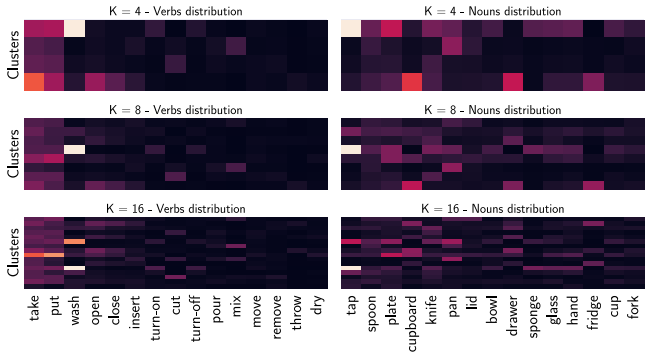


Fig. 5. Cluster-wise verbs and nouns distributions showing quite distinct functional dependencies between clusters and the corresponding actions. Lighter colors indicate higher density.

We also evaluate EgoZAR on Argo1M [10], a large scale egocentric vision dataset for Domain Generalization across different scenarios and locations. Argo1M consists of 10 splits, in each of which a specific scenario and location are not seen during training and only used for evaluation. Performance are reported using Top-1 Accuracy.

Implementation and training details. For EK100, RGB, optical flow and audio features are extracted using the TBN architecture [43] finetuned on the source train split on EK100, following the protocol described in Damen et al. [8]. The projection layers and the attention modules are trained for 30 epochs, using the SGD optimizer with weight decay $1e-5$ and momentum 0.9. The learning rate is initially set to $1e-3$ and reduced by a factor 0.1 after epochs 10 and 20. For the zone recognition model, we adopt various variants of CLIP [44] and SWAG [45]. For Argo1M, we reuse the same hyperparameters as CIR [10] and learning rate $1e-6$. RGB and zone features are extracted using SlowFast [24] and CLIP ViT-L/14 respectively. The features extractors and the zone recognition models are not updated during training.

4.2. Analysis of unsupervised environment clustering

Our approach identifies the locations in which actions are being performed through an unsupervised clustering of the features extracted with the zone recognition model. Fig. 4 shows samples from clusters derived from the EK100 dataset. The data was clustered using K-Means with a value of K set to 4, employing the L2 distance metric on features extracted with CLIP ViT-L/14. We observe that functionally similar locations are naturally clustered together in the CLIP's features space.

Table 1
Results on the validation set of EPIC-Kitchens-100 dataset in Unsupervised Domain Adaptation (UDA) and Domain Generalization (DG) settings using RGB, Optical Flow and Audio. To ensure a fair comparison we report the Source Only performance for each method. Our results are averaged over three runs. Best in **bold**. Second best is underlined.

Method	Modality	EAR Network	Setting	Top-1 Accuracy (%)			Top-5 Accuracy (%)			Mean Accuracy (%)
				Action	Verb	Noun	Action	Verb	Noun	
Source Only	RGB-Flow-Audio	TBN-TRN	–	19.20	46.70	27.78	42.12	75.42	48.27	38.95
TA3N [35]	RGB-Flow-Audio	TBN-TRN	UDA	19.61	48.44	28.87	43.36	75.95	50.12	41.25 (▲ +2.30%)
Source Only	RGB-Flow-Audio	TBN-TRN	–	18.99	47.14	27.35	41.82	75.27	49.36	43.32
MM-SADA [27]	RGB-Flow-Audio	TBN-TRN	DG	19.15	47.76	27.93	42.90	77.07	49.77	44.10 (▲ +0.78%)
MM-SADA [27]	RGB-Flow-Audio	TBN-TRN	UDA	19.25	48.44	28.26	43.41	77.56	50.59	44.59 (▲ +1.27%)
Source Only	RGB-Flow-Audio	TBN-TRN	–	18.29	46.79	26.79	41.36	75.39	48.44	42.84
RNA [40]	RGB-Flow-Audio	TBN-TRN	DG	19.81	50.75	27.92	46.76	80.64	51.37	46.21 (▲ +3.37%)
RNA [40]	RGB-Flow-Audio	TBN-TRN	UDA	20.05	**50.82**	29.19	46.04	80.89	52.18	46.53 (▲ +3.69%)
Source Only	RGB-Flow-Audio	TBN-TRN	–	19.61	47.69	28.48	–	–	–	–
CIA [46]	RGB-Flow-Audio	TBN-TRN	UDA	20.30	48.34	29.50	–	–	–	–
Source Only	RGB-Flow-Audio	TBN-TRN	DG	19.96	50.27	29.04	46.74	81.74	52.14	46.65
Gradient Blending [47]	RGB-Flow-Audio	TBN-TRN	DG	20.26	50.18	29.60	46.86	**81.82**	52.57	46.88 (▲ +0.23%)
Source Only	RGB-Flow-Audio	TBN-TRN	–	19.41	49.09	29.17	45.89	80.72	52.42	46.16
CIR [10] (w/o text)[a]	RGB-Flow-Audio	TBN-TRN	DG	19.41	49.45	29.13	46.82	80.64	53.49	46.49 (▲ +0.33%)
CIR [10][a]	RGB-Flow-Audio	TBN-TRN	DG	19.43	48.82	29.08	46.94	81.07	53.25	46.43 (▲ +0.27%)
Source Only	RGB-Flow-Audio	TBN-TRN	–	19.41	49.09	29.17	45.89	80.72	52.42	46.16
EgoZAR (RN50)	RGB-Flow-Audio	TBN-TRN	DG	20.32	50.05	29.53	46.95	81.18	53.65	46.95 (▲ +0.79%)
EgoZAR (ViT-L/14)	RGB-Flow-Audio	TBN-TRN	DG	**21.83**	50.41	**31.99**	**50.06**	81.27	**58.13**	**48.95** (▲ +2.79%)

[a] Reproduced.

Table 2
Contribution of the attention and disentanglement components of EgoZAR across different input modalities.

	Top-1 Accuracy (%)			Top-5 Accuracy (%)			Mean
	Action	Verb	Noun	Action	Verb	Noun	Acc. (%)
RGB	10.91	33.76	21.80	36.97	75.40	43.72	37.09
+ Attn.	13.17	36.13	24.32	41.76	76.71	49.22	40.22
+ Disent.	**13.63**	**37.33**	**25.06**	**42.46**	**77.18**	**50.34**	**41.00**
Flow	13.05	44.69	20.57	35.51	77.44	40.50	38.63
+ Attn.	**16.80**	**46.02**	**25.92**	**43.97**	**79.00**	**51.37**	**43.85**
Audio	8.18	32.36	13.78	27.07	70.47	31.89	30.63
+ Attn.	**14.97**	**39.74**	**23.65**	**40.34**	**75.90**	**47.99**	**40.43**

Additionally, we integrate these qualitative observations with the *per-cluster* verbs and nouns distributions, computed for different number of clusters, as shown in Fig. 5. The plot confirms the presence of functional dependencies between the clusters and the labels distributions, with the exception of some verbs, e.g., *take* and *put* that are not tied to specific locations.

4.3. EK100 results

We present a comparison of state-of-the-art methods on EK100 in Table 1, comparing EgoZAR with MM-SADA [27], Gradient Blending [47], RNA [40] and CIR [10] in the DG setting and with TA3N [35] and CIA [46] in UDA. To account for variations in the network architectures used by these models and ensure a fair comparison, we report each model with its *Source Only* performance, corresponding to standard training using cross-entropy only. This dataset poses significant challenges, with limited improvements and difficulties in comparing the results of different methods, e.g. *Source Only* of Gradient Blending outperforms TA3N or MM-SADA. Comparison of UDA and DG approaches leads to similar observations, as the gap between the best approaches in these two settings is quite small. We also evaluate CIR on EK100, using the same *Source Only* as EgoZAR. CIR benefits from the detailed narrations in Argo1M, the dataset for which it was originally proposed, while EK100 narrations are less descriptive, mostly a simple concatenation of the verb and noun labels, limiting its performance.

Our solution shows significant improvements over the previous SOTA without access to target data. We observe noticeable gains especially in action and noun accuracy, indicating that EgoZAR enables better reasoning about the manipulated objects and their interactions in key locations across environments. Indeed, the domain-agnostic clues introduced by the zone features reduce the negative effect of appearance biases, focusing less on the environment and helping the model in recognizing the same objects in different domains. Overall, EgoZAR achieves a considerable improvement over the previous SOTA, without access to the target data.

Single modality training. The integration of the zone information may be also beneficial for modalities that lack visual clues of the environment. These modalities suffer less from environmental biases but can benefit from the integration of zones information. We show the effect of the integration of the AE and SE modules in unimodal AR models in Table 2, observing a significant improvement compared to the baselines, especially on the noun metric. For RGB, the modality most affected by visual domain bias, we report results using both the attention and disentanglement modules of EgoZAR. Compared to the baseline, we observe an overall improvement of +3.13% using the attention modules and +3.91% when also the disentanglement loss is introduced. Notably, the better Top-1 verb accuracy indicates that the network is leveraging the contextual and domain-agnostic location clues provided by CLIP features to identify the action being performed.

Comparison of different zone recognition models. We evaluate in Table 4 the impact of different backbones for features extraction, using multiple CLIP and SWAG [45] variants. The latter is a ViT architecture trained for image classification using weak supervision of hashtags and is the current SOTA for scene classification on Places-365 [53], suggesting it could be useful in our context to recognize the activity-centric zones. Even the least capable model (CLIP RN50) considerably outperforms the baseline, proving the effectiveness of the attention modules of EgoZAR, and larger models consistently provide higher average accuracy. Additionally, EgoZAR shows robust performance improvements using different zone recognition models. We attribute the better performance of CLIP compared to SWAG to the former being trained on more descriptive captions using an image-language contrastive objective, compared to the weak supervision of the hashtags used in the training process of SWAG.

Ablation on the number of clusters. We analyze in Table 5, the impact of different number of clusters. All configurations exceed the performance of attention modules alone and we observe similar performances across

Table 3
Top-1 accuracy on ARGO1M [10]. Best results in **bold**, second best underlined.

	DG strategies							Ga US-PNA	Cl US-MN	Kn IND	Sh IND	Bu US-PNA	Me SAU	Sp COL	Co JPN	Ar ITA	Pl US-IN	Mean
	D	A	M	P	R	T	Z											
Random								8.00	10.64	9.13	14.36	9.55	13.04	8.35	10.13	9.86	15.68	10.84
ERM								20.75	22.35	18.69	22.14	20.73	23.51	18.97	24.81	22.75	23.29	21.80
CORAL[a] [48]	✓							22.14	22.55	19.07	24.01	22.18	24.31	19.16	25.36	23.89	25.96	22.86
DANN[a] [34]	✓	✓						22.42	23.85	19.27	22.89	22.23	23.70	18.64	25.86	23.86	23.28	22.60
MMD[a] [49]	✓							22.42	23.60	19.66	24.46	22.08	24.64	19.59	25.87	23.84	24.78	23.09
Mixup [50]			✓					21.97	22.21	19.90	23.81	21.45	24.35	19.01	25.90	23.85	24.41	22.69
BoDA[a] [51]	✓							22.17	22.78	19.62	22.94	21.46	23.97	19.18	25.68	23.92	24.90	22.66
DoPrompt[a] [52]				✓				21.92	22.77	20.40	23.67	22.75	24.67	18.24	25.04	24.74	25.24	22.94
CIR w/o text [10]					✓			23.39	24.52	21.02	26.62	24.64	**27.00**	19.66	25.42	25.71	30.17	24.81
CIR [10]					✓	✓		24.10	25.51	20.46	**27.78**	**24.93**	26.83	**19.75**	**26.34**	25.67	**30.94**	**25.23**
EgoZAR[a]							✓	**24.53**	**26.12**	**21.70**	25.82	24.05	24.88	18.91	26.02	**26.05**	29.94	24.80

[a] Domain labels required during training. *D*: distribution matching, *A*: adversarial learning, *M*: label-wise mix-up, *P*: domain-prompts, *R*: reconstruction, *T*: video-text association, *Z*: activity-centric zone learning.

Table 4
Comparison of models for zone features extraction using the cross attention component. Experiments conducted with RGB only.

Arch.	Top-1 Acc. (%)			Top-5 Acc. (%)			Mean
	Action	Verb	Noun	Action	Verb	Noun	Acc. (%)
Baseline	10.91	33.76	21.79	36.97	75.40	43.72	37.09
SWAG [45]							
ViT/B-16	11.60	34.24	22.58	39.35	75.44	46.80	38.34
ViT/L-16	12.06	34.24	23.64	40.19	75.59	48.10	38.97
CLIP [44]							
RN-50	11.69	35.23	22.43	38.54	75.69	45.56	38.19
ViT-B/32	11.56	34.38	22.44	38.26	75.00	45.51	37.86
ViT-B/16	11.88	34.82	22.87	39.30	75.23	46.73	38.47
ViT-L/14	**13.17**	**36.13**	**24.32**	**41.76**	**76.71**	**49.22**	**40.22**

Table 5
Ablation on the number of clusters for disentanglement of unimodal RGB model.

K	Top-1 Accuracy (%)			Top-5 Accuracy (%)			Mean
	Action	Verb	Noun	Action	Verb	Noun	Acc. (%)
–	13.17	36.13	24.32	41.76	76.71	49.22	40.22
2	**13.71**	37.10	24.97	41.96	77.04	49.62	40.73
4	13.63	**37.33**	**25.06**	**42.46**	77.18	**50.34**	**41.00**
8	13.44	37.12	24.37	42.30	76.76	49.92	40.65
16	13.08	36.68	24.12	42.00	76.86	49.57	40.38
32	13.29	**37.33**	24.59	42.33	**77.36**	49.89	40.80

a large set of values. We attribute this behavior to two factors. First, the number of locations in EK100 is limited and mostly dominated by sinks and stoves, which occur frequently. Second, having more clusters means that the larger clusters are broken into smaller chunks, although the disentanglement objective, which encourages the network to become more confused about the locations, remains the same. Unless otherwise specified, we set $K = 4$ for all disentanglement experiments.

4.4. ARGO1M results

Argo1M [10] features actions from a collection of different scenarios (e.g., *Cooking* and *Sport*) and locations (e.g., *United States* and *India*). While certain scenarios, such as *cooking* or *cleaning*, benefit from EgoZAR's activity-centric zones, others, such as *shopping*, are less suitable due to the even distribution of the same actions across different locations in the environment. The heterogeneity in Argo1M's data distribution required some minor adjustments to the clustering process adopted in EgoZAR. Indeed, activity-centric zones are typically associated with the scenario, e.g. a sink and an oven in the kitchen, but the location can introduce confounding factors. For example, kitchens in the USA can differ significantly from those in Saudi Arabia. To account for this, we clustered zone features separately by location and then merged clusters from different locations based on similarity of action distributions. This approach clusters zones that are visually distinct but support similar actions and may represent the same underlying activity-centric zone.

Despite these limitations, Argo1M can be considered the largest and most diverse setting for DG in egocentric vision and a valuable addition to our analysis (Table 3). EgoZAR outperforms all previous methods (CORAL [48], DANN [34], MMD [49], MixUp [50], BoDA [51] and DoPrompt [52]) and is *on-par* with CIR [10], which was specifically designed to deal with the many different scenarios and locations of Argo1M. CIR recombines samples from different scenarios and locations to learn a more agnostic representation that does not depend on the context in which the action occurs. On the contrary, we argue that the role of the environment is crucial in action recognition across different domains, and therefore build EgoZAR to leverage the location information while reducing the impact of the appearance bias. The advantage of EgoZAR compared to CIR is more evident in settings like EK100 where environmental affordances are more dominant, as discussed in Section 4.3.

5. Limitations and future works

EgoZAR heavily relies on the assumption that human activities are highly correlated with the locations in which they occur. This behavior is very evident in datasets like EK100, and more noisy in other datasets such as Argo1M. This may impact the applicability of EgoZAR to new datasets and partially reduce its effectiveness. Also, our approach requires to tune the number of clusters which is a very *dataset-dependent* parameter. Future works could focus on making the clustering process more flexible to discover the activity-centric zones in a completely unsupervised way, without using any prior knowledge on the number of clusters.

6. Conclusions

In this paper, we showcase the impact of environmental affordances in action recognition. We argue that leveraging this environmental information is significantly influenced by their appearance, strongly limiting the generalization ability to other areas and domains. We propose EgoZAR, a method to exploit zone-recognition models as source of a domain-agnostic information on the activity-centric zones where the actions are taking place, and to replace domain-specific appearance of activity-centric zones. Extensive experiments on EK100 show the effectiveness of EgoZAR, achieving SOTA performance and highlighting how the integration of zone information may help in action recognition.

CRediT authorship contribution statement

Simone Alberto Peirone: Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Gabriele Goletto:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Mirco Planamente:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Andrea Bottino:** Writing – review & editing, Writing – original draft, Supervision. **Barbara Caputo:** Writing – review & editing, Writing – original draft, Supervision. **Giuseppe Averta:** Writing – review & editing, Writing – original draft, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Gabriele Goleto reports financial support was provided by Ministry of Education and Merit. Barbara Caputo, Giuseppe Averta reports financial support was provided by FAIR - Future Artificial Intelligence Research. Mirco Planamente reports was provided by National Interuniversity Consortium for Informatics. 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.

Acknowledgments

This study was supported in part by the CINI Consortium through the VIDESEC project and 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. G. Goletto is supported by PON "Ricerca e Innovazione" 2014–2020 – DM 1061/2021 funds.

Data availability

The authors have used publicly available datasets.

References

[1] 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, IJCV (2024) 1–57.
[2] G. Goletto, M. Planamente, B. Caputo, G. Averta, Bringing online egocentric action recognition into the wild, Robot. Automat. Lett. 8 (4) (2023) 2333–2340.
[3] J.J. Gibson, The ecological approach to visual perception: classic edition, Psychology Press, 2014.
[4] L. Mur-Labadia, J.J. Guerrero, R. Martinez-Cantin, Multi-label affordance mapping from egocentric vision, in: ICCV, 2023.
[5] H. Luo, W. Zhai, J. Zhang, Y. Cao, D. Tao, Learning affordance grounding from exocentric images, in: CVPR, 2022.
[6] T. Nagarajan, Y. Li, C. Feichtenhofer, K. Grauman, Ego-topo: Environment affordances from egocentric video, in: CVPR, 2020.
[7] K.K. Singh, D. Mahajan, K. Grauman, Y.J. Lee, M. Feiszli, D. Ghadiyaram, Don't judge an object by its context: learning to overcome contextual bias, in: CVPR, 2020.
[8] D. Damen, H. Doughty, G.M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Moltisanti, J. Munro, T. Perrett, W. Price, et al., Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100, IJCV (2022) 1–23.
[9] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al., Ego4d: Around the world in 3,000 hours of egocentric video, in: CVPR, 2022.
[10] 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: CVPR, 2023.
[11] M. Hassanin, S. Khan, M. Tahtali, Visual affordance and function understanding: A survey, ACM Comput. Surv. 54 (3) (2021) 1–35.
[12] T. Nagarajan, C. Feichtenhofer, K. Grauman, Grounded human-object interaction hotspots from video, in: ICCV, 2019.
[13] P. Ardón, È. Pairet, R.P. Petrick, S. Ramamoorthy, K.S. Lohan, Learning grasp affordance reasoning through semantic relations, Robot. Automat. Lett. 4 (4) (2019) 4571–4578.
[14] A. Hämäläinen, K. Arndt, A. Ghadirzadeh, V. Kyrki, Affordance learning for end-to-end visuomotor robot control, in: IROS, 2019.
[15] M. Wang, R. Luo, A.Ö. Önol, T. Padir, Affordance-based mobile robot navigation among movable obstacles, in: IROS, 2020.
[16] V. Kaptelinin, B. Nardi, Affordances in HCI: toward a mediated action perspective, in: Conference on Human Factors in Computing Systems, 2012.
[17] Z. Yu, Y. Huang, R. Furuta, T. Yagi, Y. Goutsu, Y. Sato, Fine-grained affordance annotation for egocentric hand-object interaction videos, in: WACV, 2023.
[18] H. Kjellström, J. Romero, D. Kragić, Visual object-action recognition: Inferring object affordances from human demonstration, CVIU 115 (1) (2011) 81–90.
[19] P. Mandikal, K. Grauman, Learning dexterous grasping with object-centric visual affordances, in: ICRA, 2021.
[20] T.-T. Do, A. Nguyen, I. Reid, Affordancenet: An end-to-end deep learning approach for object affordance detection, in: ICRA, 2018.
[21] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: NeurIPS, 2014.
[22] J. Lin, C. Gan, S. Han, Tsm: Temporal shift module for efficient video understanding, in: ICCV, 2019.
[23] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: CVPR, 2017.
[24] C. Feichtenhofer, H. Fan, J. Malik, K. He, Slowfast networks for video recognition, in: ICCV, 2019.
[25] S. Sudhakaran, O. Lanz, Convolutional long short-term memory networks for recognizing first person interactions, in: ICCVW, 2017.
[26] M. Planamente, A. Bottino, B. Caputo, Self-supervised joint encoding of motion and appearance for first person action recognition, in: ICPR, 2021.
[27] J. Munro, D. Damen, Multi-modal domain adaptation for fine-grained action recognition, in: CVPR, 2020.
[28] A. Furnari, G. Farinella, Rolling-unrolling LSTMs for action anticipation from first-person video, TPAMI 43 (11) (2020) 4021–4036.
[29] N. Crasto, P. Weinzaepfel, K. Alahari, C. Schmid, MARS: Motion-augmented RGB stream for action recognition, in: CVPR, 2019.
[30] J. Zhao, C.G. Snoek, Dance with flow: Two-in-one stream action detection, in: CVPR, 2019.
[31] E. Kazakos, A. Nagrani, A. Zisserman, D. Damen, Slow-fast auditory streams for audio recognition, in: ICASSP, 2021.
[32] C. Plizzari, M. Planamente, G. Goletto, M. Cannici, E. Gusso, M. Matteucci, B. Caputo, E2(GO)motion: Motion augmented event stream for egocentric action recognition, in: CVPR, 2022.
[33] M. Long, Y. Cao, J. Wang, M. Jordan, Learning transferable features with deep adaptation networks, in: ICML, 2015.
[34] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks, JMLR 17 (1) (2016) 2096–2030.
[35] 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.
[36] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, P. Yu, Generalizing to unseen domains: A survey on domain generalization, Trans. Knowl. Data Eng. 35 (8) (2022) 8052–8072.
[37] P. Koniusz, Y. Tas, F. Porikli, Domain adaptation by mixture of alignments of second-or higher-order scatter tensors, in: CVPR, 2017.
[38] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: CVPR, 2016.
[39] Y. Shi, X. Yu, K. Sohn, M. Chandraker, A.K. Jain, Towards universal representation learning for deep face recognition, in: CVPR, 2020.
[40] M. Planamente, C. Plizzari, S.A. Peirone, B. Caputo, A. Bottino, Relative norm alignment for tackling domain shift in deep multi-modal classification, IJCV (2024) 1–21.
[41] Z. Yao, Y. Wang, J. Wang, S.Y. Philip, M. Long, Videodg: Generalizing temporal relations in videos to novel domains, TPAMI 44 (11) (2021) 7989–8004.
[42] B. Zhou, A. Andonian, A. Oliva, A. Torralba, Temporal relational reasoning in videos, in: ECCV, 2018.
[43] E. Kazakos, A. Nagrani, A. Zisserman, D. Damen, EPIC-fusion: Audio-visual temporal binding for egocentric action recognition, in: ICCV, 2019.
[44] 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, 2021.
[45] M. Singh, L. Gustafson, A. Adcock, V. de Freitas Reis, B. Gedik, R.P. Kosaraju, D. Mahajan, R. Girshick, P. Dollár, L. Van Der Maaten, Revisiting weakly supervised pre-training of visual perception models, in: CVPR, 2022.
[46] L. Yang, Y. Huang, Y. Sugano, Y. Sato, Interact before align: Leveraging cross-modal knowledge for domain adaptive action recognition, in: CVPR, 2022.
[47] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard? in: CVPR, 2020.

[48] B. Sun, K. Saenko, Deep coral: Correlation alignment for deep domain adaptation, in: ECCV, 2016.
[49] H. Li, S. Jialin Pan, S. Wang, A.C. Kot, Domain generalization with adversarial feature learning, in: CVPR, 2018.
[50] Y. Wang, H. Li, A.C. Kot, Heterogeneous domain generalization via domain mixup, in: ICASSP, 2020.
[51] Y. Yang, H. Wang, D. Katabi, On multi-domain long-tailed recognition, generalization and beyond, in: ECCV, 2022.
[52] Z. Zheng, X. Yue, K. Wang, Y. You, Prompt vision transformer for domain generalization, in: arXiv Preprint , 2022, arXiv:2208.08914.
[53] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: NeurIPS, 2014.