What can a cook in Italy teach a mechanic in India?
Action Recognition Generalisation Over Scenarios and Locations

Chiara Plizzari♣*   Toby Perrett♦   Barbara Caputo♣   Dima Damen♦

♣ Politecnico di Torino, Italy   ♦ University of Bristol, United Kingdom

Abstract

We propose and address a new generalisation problem: can a model trained for action recognition successfully classify actions when they are performed within a previously unseen scenario and in a previously unseen location? To answer this question, we introduce the Action Recognition Generalisation Over scenarios and locations dataset (ARGO1M), which contains 1.1M video clips from the large-scale Ego4D dataset, across 10 scenarios and 13 locations. We demonstrate recognition models struggle to generalise over 10 proposed test splits, each of an unseen scenario in an unseen location. We thus propose CIR, a method to represent each video as a Cross-Instance Reconstruction of videos from other domains. Reconstructions are paired with text narrations to guide the learning of a domain generalisable representation. We provide extensive analysis and ablations on ARGO1M that show CIR outperforms prior domain generalisation works on all test splits. Code and data: `https://chiaraplizz.github.io/what-can-a-cook/`.

1. Introduction

A notable distinction between human and machine intelligence is the ability of humans to generalise. We can see an example of the action "cut" performed by a cook in Italy, and recognise the same action performed in a different geographic *location*, *e.g.* India, despite having never visited. We can also recognise actions within new *scenarios*, such as a mechanic cutting metal, even if we are unfamiliar with the tools they use.

This problem is known as domain generalisation [62], where a model trained on a set of labelled data fails to generalise to a different distribution in inference. The gap between distributions is known as *domain shift*. To date, works have focused on generalising over visual domain shifts [25, 46, 31, 10, 39]. In this paper, we introduce the *scenario shift*, where the same action is performed as part

*Work carried during Chiara's research visit to the University of Bristol

Figure 1: Problem statement and samples from the ARGO1M dataset. The same action, e.g. "cut", is performed differently based on the *scenario* and the *location* in which it is carried out. We aim to generalise so as to recognise the same action within a new scenario, *unseen* during training, and in an *unseen location*, *e.g.*, Mechanic (🔧) in India (🇮🇳).

of a different activity, impacting the tools used, objects interacted with, goals and behaviour. We combine this with the location shift, generalising over both simultaneously.

In Fig. 1, the action "cut' is performed using a knife whilst cooking (🍴), pliers whilst building (🏗) and scissors for arts and crafts (✂). Tools are not specific for a scenario and can vary over locations – e.g. in Fig. 1, seaweed sheets are cut with scissors while cooking in Japan. Generalising would be best achieved by learning the notion of "cutting" as separating an object into two or more pieces, regardless of the tool or background location. Successful generalisation can thus enable recognising metal being "cut" by a mechanic in India using an angle grinder (Fig. 1 Test).

Our investigation is enabled by the recent introduction of the Ego4D [17] dataset of egocentric footage from around the world. We curate a setup specifically for action generalisation, called ARGO1M. It contains 1.1M action clips of 60 classes from 73 unique scenario/location combinations.

To tackle the challenge of ARGO1M, we propose a new method for domain generalisation. We represent each video

as a weighted combination of other videos in the batch, potentially from other domains. We refer to this as Cross-Instance Reconstruction (CIR). Through reconstruction, the method learns domain generalisable video features. CIR is supervised by a classification loss and a video-text association loss. To summarise, our key contributions are:

- We curate the Action Recognition Generalisation dataset (ARGO1M) from videos and narrations from Ego4D. ARGO1M is the first to test action generalisation across both scenario and location shifts, and is the largest domain generalisation dataset across images and video.

- We introduce CIR, a domain generalisation method which exploits Cross-Instance Reconstruction and video-text pairing to learn generalisable representations.

- We test CIR on the proposed ARGO1M, showing that it consistently outperforms baselines and recent domain generalisation approaches on 10 test sets.

2. Related Work

In this section, we review datasets and methods for Domain Generalisation. **Domain Generalisation (DG)** aims to generalise to any unseen target domain, where data from the target domain are not available during training [62]. We note the distinction from the **Domain Adaptation** setting, where unlabelled target samples are available during training [31, 44, 22]. Adaptation is out of scope for this paper.

2.1. Domain Generalisation (DG) datasets

Table 1 presents a comparison of vision datasets used for domain generalisation. Existing image datasets present a stylistic shift. For example, common objects in photos, paintings, clipart, cartoons and sketches [25, 49, 36], or common categories across datasets [46]. Location shift was explored in [2] which contains animals photographed in different locations. Image DG works typically test on a number of these benchmarks [19]. For video, shifts include cross-dataset [8], synthetic-to-real [8], viewpoint [9], location [31] and the passage of time [11].

Compared to prior works, ARGO1M is 21× the largest video DG dataset and 1.8× image DG dataset. Importantly, ARGO1M introduces the scenario shift, which it tests alongside the location shift, with many more domains (up to 64 training domains and 10 test domains).

2.2. Domain Generalisation (DG) Methods

Previous approaches for DG are mostly designed around image data [4, 51, 27, 13, 28, 3]. *Feature-based alignment* between training domains can be used to learn domain-invariant representations [27, 45, 16, 55]. This can be achieved using a domain-adversarial network [16] or by minimising distances such as Maximum Mean Descrepancy (MMD) [27, 18]. This has recently been extended

	Dataset	Samples		Domains		
		# Samples	# Cls	# Train	# Test	Domain Shift
Images	PACS [25]	9,991	7	3	4	Style
	VLCS [46]	10,729	5	3	4	N/A
	OfficeHome [49]	15,588	65	3	4	Style
	TerraIncognita [2]	24,788	10	3	4	Loc
	DomainNet [36]	586,575	345	5	6	Style
Videos	UCF-HMDB [8]	3809	12	1	2	N/A
	Kinetics-Gameplay [8]	49,998	30	1	2	Realism
	MM-SADA [31]	10,094	8	2	3	Loc
	EPIC-Kitchens [11]	48,139	86	11	1	Time Gap
	ARGO1M	1,050,371	60	54-64	10	(Scenario, Loc)

Table 1: **Datasets for DG.** ARGO1M tests combined scenario and location shifts, and is the largest in # of samples & # of domains.

in [55], which handles class and domain imbalance with a weighted loss. *Data-based* methods augment training data to prevent overfitting [51, 50, 62, 59, 6, 32, 7, 53, 54]. For example, data augmentation such as Mixup [57] has been shown to improve accuracy on unseen data. *Meta-Learning* methods simulate the distribution shift between seen and unseen environments [24, 1, 13, 26, 29] using meta-train and meta-test domains. *Self-Supervision* [4, 3] has been shown to learn generalisable representations, with unsupervised pretext tasks better capturing the shared knowledge among multiple sources. A recent trend is to learn *domain prompts* from visual [60, 42] or text information [33, 58], or utilise *cross-modal supervision* [30]. For example, Do-Prompt [60] learns training domain-specific prompts, and predicts prompts for test samples as linear combinations of training prompts. There are limited works on video domain generalisation. [39] relies on multi-modal alignment, and [56] uses adversarial data augmentation.

For our comparative analysis, we extend a representative selection of prior works [53, 27, 45, 16, 55, 60] to the large number of training domains in ARGO1M, and showcase their limitation experimentally.

2.3. Cross-Attention for Reconstruction

The task of predicting masked tokens within one video is now common in many representation learning approaches, e.g. [15]. We differ from these works in reconstructing from other videos in the batch. Such cross-instance attention has been used to reconstruct query instances from examples of each class for few-shot learning [12, 37]. In [38], few-shot instances are reconstructed from samples of head classes. In cross-modal retrieval [35], reconstruction through cross-attention learns better video-text representations through a caption generation task. Differently from prior works, we reconstruct each video as a learned weighted combination of videos from *various domains*.

3. ARGO1M Benchmark

In this section, we detail how we curated the ARGO1M dataset from videos of the Ego4D [17] dataset.
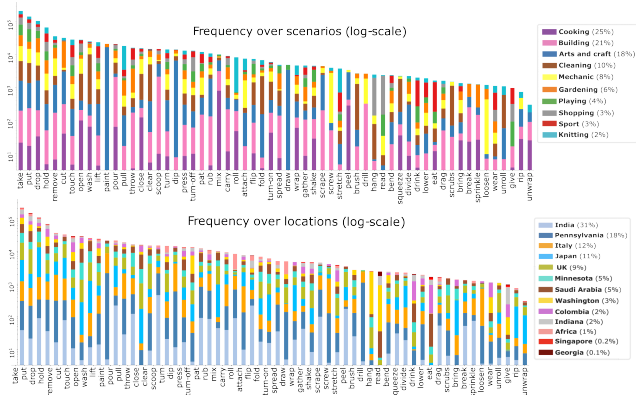**Ego4D Background.** Ego4D [17] contains untrimmed ego-

Figure 2: Frequency (log-scale) of the 60 classes in ARGO1M across scenarios (top) and locations (bottom) - % in legend. Scenarios and locations are linearly scaled within each bar.

centric videos totaling 3,670 hours collected from 8 non-US countries and 5 US states. These represent a variety of daily life scenarios (*e.g.* playing cards, cooking, fixing the car). Each video is associated with metadata reflecting the geographic location and the scenario it captures. Within each video, timestamp-level narrations of actions are provided.

ARGO1M Metadata. The high-level scenario descriptions in Ego4D are free-form and at times missing. We exclude repetitive scenarios such as "talking" or "on a screen", as well as videos with missing or multiple scenarios. We then manually cluster the free-form descriptions into 10 scenarios. These are: *Cooking* (🍳), *Building* (🏗), *Arts and crafts* (✂), *Cleaning* (🧹), *Mechanic* (🔧), *Gardening* (🪴), *Playing* (🃏), *Shopping* (🛒), *Sport* (🤸), *Knitting* (🧶). As an example, the free-form descriptions "Car mechanic", "Getting the car fixed" and "Bike mechanic" are clustered into *Mechanic*.

Similarly, while text narrations offer the ground-truth for the action in each video clip, they are also free-form sentences. We extract action labels by parsing the text narrations using spaCy [20]. We take verbs as actions and convert these to closed-vocabulary classes, using modified clustering from [10] for the additional vocabulary. We have 60 action classes shown in Fig. 2. The distribution is long-tailed, and each action class appears in multiple scenarios and in multiple locations. On average each class appears in 8 scenarios and 11 locations.

In summary, ARGO1M contains 1,050,371 video clips. Each video *clip* is captured in a given *scenario* (out of 10) and geographic *location* (out of 13), with associated *text narration* and *action class* (out of 60). For example, the caption, "#Camera wearer (C) cuts the lemon strand." is associated to a clip recorded in "Italy" and capturing "Gardening" scenario, with associated action label "cut".

ARGO1M Splits. We curate 10 distinct train/test splits to evaluate generalisation over scenarios and locations. We
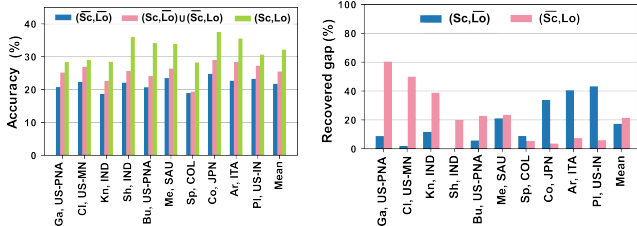
(a) Accuracy without samples from the test scenario or location ($\overline{\text{Sc}}$, $\overline{\text{Lo}}$) as well as (Sc, $\overline{\text{Lo}}$)∪($\overline{\text{Sc}}$, Lo) and (Sc Lo).

(b) % of drop recovered when adding examples from either scenario (Sc, $\overline{\text{Lo}}$) or location ($\overline{\text{Sc}}$, Lo).

Figure 3: Analysis of scenario and location shifts on ARGO1M.

select these 10 test splits so *all scenarios* are covered. For each scenario, we select the location with the largest number of samples to form the test split for robust evaluation. Given paired scenario and location (**Sc, Lo**), the corresponding training split excludes all samples from the scenario (**Sc**) as well as all samples from the location (**Lo**). We show later in this section that these 10 splits present a variety of combined scenario/location shift properties.

The selected test splits and their [number of samples] are: *Gardening* in *Pennsylvania* (**Ga, US-PNA**[1]) [16,410], *Cleaning* in *Minnesota* (**Cl, US-MN**) [22,008], *Knitting* in *India* (**Kn, IND**) [13,250], *Shopping* in *India* (**Sh, IND**) [11,239], *Building* in *Pennsylvania* (**Bu, US-PNA**) [99,865], *Mechanic* in *Saudi Arabia* (**Me, SAU**) [11,700], *Sport* in *Colombia* (**Sp, COL**) [16,453], *Cooking* in *Japan* (**Co, JPN**) [82,128], *Arts and crafts* in *Italy* (**Ar, ITA**) [36,812], *Playing* in *Indiana* (**Pl, US-IN**) [17,379].

ARGO1M Domain Shift Analysis. We analyse the impact of scenario and location shifts on the 10 test splits in ARGO1M by varying whether samples from the test scenario and/or location appear during training.

For all experiments we use Empirical Risk Minimization (ERM) (*i.e.* standard cross entropy training) - see Section 5 for full experimental details. We present this early analysis so as to understand the domain shift in ARGO1M . We take the default setting (1) where no examples from the test scenario or the location appear during training. We denote this as ($\overline{\text{Sc}}$, $\overline{\text{Lo}}$), where overline indicates samples are excluded from the training split. We compare this against cases where (2) the training split also includes samples showcasing either the test scenario or the test location but not both, i.e. (Sc, $\overline{\text{Lo}}$)∪($\overline{\text{Sc}}$, Lo), and (3) samples from the test scenario in the test location are included, i.e. (Sc, Lo). In Figure 3a, performance improves from (1) → (2) with a bigger improvement (2) → (3). This demonstrates that generalisation is particularly challenging when the combined test scenario and location do not appear during training.

Next, we analyse how much the scenario and location shifts individually contribute to this drop in performance.

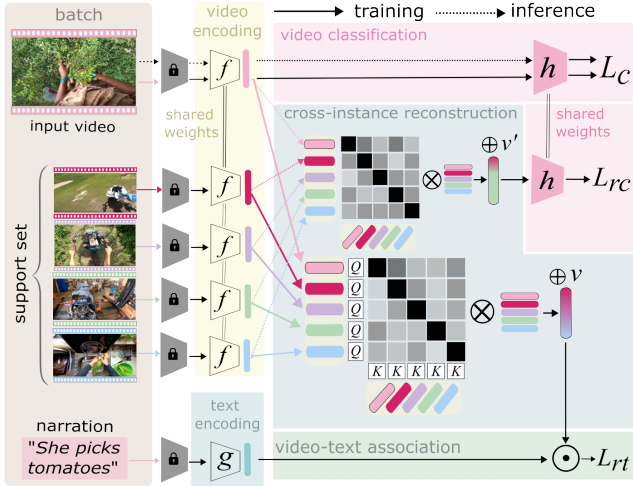[1]We use ISO country codes and US state codes.

Figure 4: **CIR**. One clip and corresponding narration are shown along with the support set of other clips in the batch. Video $f(v)$ and text $g(t)$ embeddings are extracted using trained encoders on top of a frozen model. Cross entropy $\mathcal{L}_c$, and two CIR objectives $\mathcal{L}_{rt}$ and $\mathcal{L}_{rc}$ are minimized. For $\mathcal{L}_{rt}$, query Q and key K projections are learnt for clips in the batch, followed by self-masking. Weights are multiplied by $f(v)$, and the reconstructed $\oplus v$ is paired with the corresponding narration. For $\mathcal{L}_{rc}$, $\oplus v'$ is classified using the classifier h. At inference, only the video classifier h is used.

We show the fraction of the drop recovered against (3) when introducing training samples from either the test scenario (**Sc**, $\overline{\textbf{Lo}}$) or the test location ($\overline{\textbf{Sc}}$, **Lo**). Fig. 3b shows the impact of scenario and location varies widely for each test split. For example, on (**Sh, IND**), training with the test scenario *shopping* does not help, whereas the location *India* does. Conversely, on (**Ar, ITA**), training with *arts and crafts* recovers 40% of the drop, whereas the location does not help. This showcases that both shifts are interesting and that our 10 test splits offer the diversity to study both.

4. Method

We propose Cross-Instance Reconstruction (CIR) to represent an action as a weighted combination of actions from other scenarios and locations. We first formulate the input to our method in Section 4.1, then focus on our proposed CIR in Section 4.2. We detail training in Section 4.3 and inference in Section 4.4.

4.1. Proposed Setting

Each training sample is a video clip v with a free-form text narration t and an action class label y: (v, t, y). During testing, we only require an input video clip, to predict the action label. We use $\hat{y}$ to refer to the predicted label.

We consider a composite function to classify actions:

$$\hat{y} = h \circ f(v) \qquad (1)$$

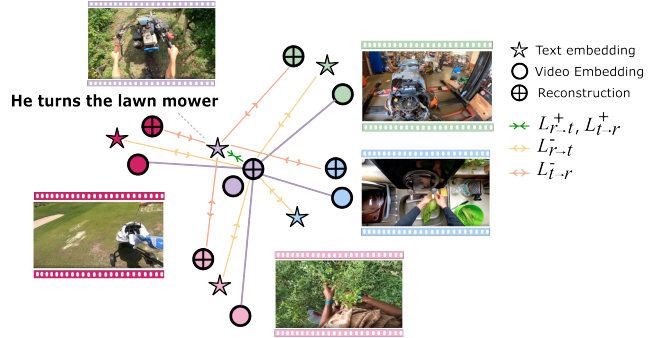where f is an encoder which learns a video representation

Figure 5: **Video-text association.** The reconstructed clip $\oplus v'_i$ (violet) is paired with its text representation. The reconstruction-text loss $\mathcal{L}_{r \to t}$ has the reconstruction $\oplus v'_i$ as positive and other text narrations as negatives, and the text-to-reconstruction loss $\mathcal{L}_{t \to r}$ has other reconstructions $\oplus v'_j$ as negative.

suitable for domain generalisation, while h specialises in learning an action classifier from that representation.

In addition to the cross-entropy loss $\mathcal{L}_c$ on h, we train the domain generalisable representation f using two losses; one cross-modal and another classification loss.

4.2. Cross-Instance Reconstruction (CIR)

Our main premise in cross-instance reconstruction (CIR) is to encourage cross-domain representations of actions, where domains are scenarios and locations. In doing so, these representations can be domain generalisable, as it reconstructs the same action from samples of other domains.

We learn-to-reconstruct any video clip from *other* video clips in the randomly sampled batch, which we call the support set S. We *jointly* reconstruct all video clips in the batch, at the feature level. Each video clip appears in the support set of every other video clip in the batch. Before outlining the training objectives, we first describe the reconstruction process.

We learn two projection heads, which we term the query and key heads, Q and K, in line with standard works [48], along with a layer norm L. We calculate the correlation between each pair of video clips, v_i and v_j, in the training batch as:

$$c_{ij} = L(Q(f(v_i))) \cdot L(K(f(v_j))) \qquad (2)$$

The resulting weights c_{ij} are softmaxed and self-masked to avoid trivial reconstructions from the sample itself. The reconstructed representation $\oplus v_i$ is a weighted combination of all embeddings in its support set, using the weights c_{ij}:

$$\forall i: \quad \oplus v_i = \sum_{j \in S} \frac{\exp(c_{ij}) f(v_j)}{\sum_{k \in S} \exp(c_{ik})} \qquad (3)$$

We directly weight $f(v)$ – this is analogous to using the identity matrix for the value head in standard attention.

4.3. Training CIR

Fig. 4 gives an overview of CIR which we detail next. We intend for reconstructions to learn to generalise, and backpropogate this ability to the video encoder f (Eq. 1). We propose two reconstructions, each guided by a different objective. The video-text association reconstruction ($\oplus v$ in Fig. 4) uses text narrations so these cross-instance reconstructions are associated with the video clip's semantic description. The classification reconstruction ($\oplus v'$ in Fig. 4) is trained to recognise the clip's action class.

For the **video-text association reconstruction** $\oplus v_i$, we use contrastive learning to push $\oplus v_i$ towards the embedding of the text narration associated with the video, *e.g.* "He turns the lawn mower". Given a batch of video-text pairs with corresponding reconstructions $\mathcal{B} = \{(v_i, \oplus v_i, t_i)\}_{i=1}^{B}$, the resulting objective is formulated using Noise Contrastive Estimation [34] over both reconstruction-text and text-reconstruction pairs. Specifically, the reconstruction-text loss considers the reconstruction $\oplus v_i$ as the anchor and the negatives as other text narrations in the batch, such that:

$$\mathcal{L}_{r \to t}(\oplus v_i, g(t_i)) = -\frac{1}{B} \sum_{i}^{B} \log \frac{\exp\left(s(\oplus v_i, g(t_i))/\tau\right)}{\sum_{j}^{B} \exp\left(s(\oplus v_i, g(t_j))/\tau\right)} \quad (4)$$

where $s(\cdot, \cdot)$ is the cosine similarity, g is the text encoder, $g(t_i)$ is the encoded text narration, and τ is a learnable temperature. The analogous loss $\mathcal{L}_{t \to r}$ considers $g(t_i)$ as anchor and other reconstructions as negatives. We showcase these in Fig. 5. Both are combined to form our reconstruction-text association loss $\mathcal{L}_{rt} = \mathcal{L}_{r \to t} + \mathcal{L}_{t \to r}$.

Note that we avoid pairing this reconstruction with the video embedding $f(v_i)$, instead of the text narration $g(t_i)$, as it may convey domain-knowledge (*i.e.* scenario and location), which might bias the reconstruction to videos from the same scenario or location. Instead, the associated narration offers an instance-level description of the action, which guides the reconstruction.

Our **classification reconstruction** $\oplus v_i'$ forms the input to the classifier h, so as to recognise the action class such that $\hat{y}' = h(\oplus v')$. We train with cross-entropy loss, which we term $\mathcal{L}_{rc}$ to imply classifying reconstructions. We share the weights between the classifier for videos and for reconstructions. Additionally, for this reconstruction, we compute weights with cross-product attention: $c_{ij}' = f(v_i) \cdot f(v_j)$, *i.e.* by replacing c with c' in Eq. 3. We thus do not learn additional query and key projections. We ablate these decisions in Section 5.2.

We combine our two losses with the cross-entropy video classification loss $\mathcal{L}_c$ (see Section 4.1). Our overall training objective is:

$$\mathcal{L} = \mathcal{L}_c + \lambda_1 \mathcal{L}_{rt} + \lambda_2 \mathcal{L}_{rc}. \quad (5)$$

where λ_1 and λ_2 weight the two reconstruction losses.

4.4. Inference

Once training concludes, f is capable of extracting domain generalisable representations that maintain action class knowledge without domain bias. Accordingly, at test time, only video clips v_i from the test split are processed by the encoder f and classifier h. We do not require any narration during inference, and there is no reconstruction – *i.e.* each clip is classified independently.

5. Experiments

We test the ability of CIR to generalise over scenarios and locations by comparing it against baseline and state-of-the-art domain generalisation methods adapted for our setting. We then show ablations on its different components, and visualise its impact with qualitative examples.

Dataset and metrics. We use the ARGO1M dataset introduced in Section 3 for all experiments. We report top-1 accuracy for each test split, as well as mean accuracy.

Baselines. We first compare our method with the Empirical Risk Minimisation (ERM) baseline [47], as is standard practice in DG works [4, 19]. This is cross-entropy ($\mathcal{L}_c$) without a generalisation objective. We then compare against 6 methods for DG, all trained jointly with $\mathcal{L}_c$.

Most DG methods do require domain labels during training. We thus provide these labels when required and mark these methods with (*). At test time, all methods only use video clip input, and are not aware of any domain knowledge. Our baselines, ordered by publication year, are:

- CORAL* [45]: two mean and covariance distances are minimised. These are the distances between means and covariances of video representations from different scenarios, and the distances between means and covariances from different locations.
- DANN* [16]: 2-fully connected layers form an adversarial network to predict the location. A separate adversarial network predicts the scenario.
- MMD* [27]: same as CORAL w/ MMD distances [18].
- Mixup [53]: training data is augmented by performing linear interpolations of samples and labels. Note that Mixup is distinct from CIR as it focuses only on pairs of videos selected randomly, rather than reconstructing from all videos in the batch based on visual similarity. Additionally, Mixup changes the output label, while in CIR the video class label is maintained.
- BoDA* [55]: minimises distances between domains, similar to MMD, weighted by both domain size and class size, in an effort to handle imbalance.
- DoPrompt* [60]: learns one domain prompt for each scenario and location to be appended to visual features before classification.

We also provide random chance averaged over 10 trials.

	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											
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* [45]	✓						22.14	22.55	19.07	24.01	22.18	24.31	19.16	25.36	23.89	_25.96_	22.86
DANN* [16]	✓	✓					_22.42_	_23.85_	19.27	22.89	22.23	23.70	18.64	25.86	23.86	23.28	22.60
MMD* [27]	✓						_22.42_	23.60	19.66	_24.46_	22.08	24.64	_19.59_	25.87	23.84	24.78	_23.09_
Mixup [53]			✓				21.97	22.21	19.90	23.81	21.45	24.35	19.01	_25.90_	23.85	24.41	22.69
BoDA* [55]	✓						22.17	22.78	19.62	22.94	21.46	23.97	19.18	25.68	23.92	24.90	22.66
DoPrompt* [60]				✓			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)					✓		23.39	24.52	21.02	26.62	24.64	27.00	19.66	25.42	25.71	30.17	24.81
CIR					✓	✓	**24.10**	**25.51**	**20.46**	**27.78**	**24.93**	**26.83**	**19.75**	**26.34**	**25.67**	**30.94**	**25.23**

Table 2: Top-1 accuracy on ARGO1M. Best results in **bold**, second best underlined (omitting CIR w/o video-text association loss, which is greyed out but given for direct comparison showcasing strong performance w/o narrations). ∗: 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.

	Cl US-MN	Bu US-PNA	Co JPN	Ar ITA	Pl US-IN	Mean
CIR (ours)	25.51	**24.93**	26.34	**25.67**	**30.94**	**26.68**
$-\mathcal{L}_{rt}$	24.83	24.80	25.06	25.38	29.50	25.91
$-\mathcal{L}_{rc}$	23.13	23.53	25.87	24.95	26.59	24.81
$-\mathcal{L}_{rt} - \mathcal{L}_{rc}$	22.35	20.73	24.81	22.75	23.29	22.78
$\oplus v$ cross-product	**25.66**	24.84	25.42	25.41	30.67	26.40
$\oplus v'$ learnt att.	22.58	22.55	25.85	24.53	25.35	24.17
$\oplus v = \oplus v'$	23.47	23.33	25.53	24.06	28.74	25.03
$h \neq h'$	24.47	23.12	**26.74**	24.74	27.37	25.29

Table 3: Ablation on CIR, showing the contribution of the two reconstructions and alternative design choices.

SL	SS	OL	OS	Cl US-MN	Bu US-PNA	Co JPN	Ar ITA	Pl US-IN	Mean
✓	✓	✗	✓	25.01	24.86	25.73	**25.99**	30.69	26.46
✓	✓	✓	✗	25.00	25.05	26.07	25.62	30.98	26.55
✓	✓	✗	✗	24.87	24.68	25.77	25.38	30.07	26.15
✗	✓	✓	✓	24.89	**25.13**	26.05	25.80	30.47	26.47
✓	✗	✓	✓	25.22	24.99	26.34	25.84	30.25	26.53
✗	✗	✓	✓	25.17	24.97	**26.36**	25.61	30.31	26.48
✓	✓	✓	✓	25.51	24.93	26.34	25.67	**30.94**	**26.68**

Table 4: Effect of masking samples in the support set used for reconstruction. Columns indicate whether the query can (✓) or cannot (✗) attend to samples from the **S**ame **S**cenario/**L**ocation (**SS**, **SL**) or **O**ther **S**cenario/**L**ocation (**OS**, **OL**) based on the domains they belong to. Note that CIR (bottom) does not use any masking.

Implementation details. We use SlowFast features [14], pre-trained on Kinetics [5], provided with the videos of Ego4D [17]. We represent the action by concatenating three features, forming a 6912-D vector, as in [61], taken from the action's onset as associated with the narration, halfway to the next action, and before the start of the next action. For text features (512-D) we use the frozen text encoder of the pre-trained CLIP-ViT-B-32 model [41]

f is implemented as 2 fully connected layers of hidden dimension 4096 and output dimension 512, with a ReLU activation function and a Batch Normalisation layer [21]. g is implemented as 2 fully connected layers with 512 hidden dimension and a ReLU activation function. The dimension of query and key embeddings for reconstruction is 128.

We use a batch size of 128 for all experiments and methods, and train for 50 epochs using the Adam optimiser [23]. The learning rate is set to $2e^{-4}$ for CIR, decaying by a factor of 10 at epochs 30 and 40. We set $\lambda_1 = 1$ and $\lambda_2 = 0.5$ (Eq. 5). Ablation on hyperparameters is in the Supplementary. Training takes 8 hours on one Nvidia P100 GPU.

5.1. Results

Table 2 shows CIR outperforms all previous approaches, on every test split, by up to 4.9%, and is on average 2.1% better than the second best method. Compared to the ERM baseline, CIR outperforms by 3.4% on average and up to 7.7%. The improvement varies across splits, with the small-est improvements occurring on harder splits – those with lower ERM baselines, *e.g.* (**Kn**, **IND**) and (**Sp**, **COL**).

CIR does not use any domain labels during training, which is a common strategy for other methods (marked by ∗ in Table 2), but instead assumes access to textual narrations. We also report results of CIR without text (*i.e.* without $\mathcal{L}_{rt}$) or domain labels showcasing strong average performance for CIR with less supervision than other methods.

The second best performing method varies per split, showcasing the complexity of the problem as well as the need for multiple test splits to properly assess domain generalisation approaches. Methods that learn domain invariant visual features by matching distributions or via domain prompts seem to struggle with the scenario shift proposed in ARGO1M. Results of CIR show that reconstruction and usage of text narrations are an effective alternative.

5.2. Ablations

We use the 5 largest test splits for all ablation results.
CIR Ablation. CIR has two reconstruction objectives, and three architectural choices for reconstruction, which are ablated in Table 3. For the two objectives, the one with the largest impact differs per split, with the classification reconstructions ($\mathcal{L}_{rc}$) performing better on average (shown by
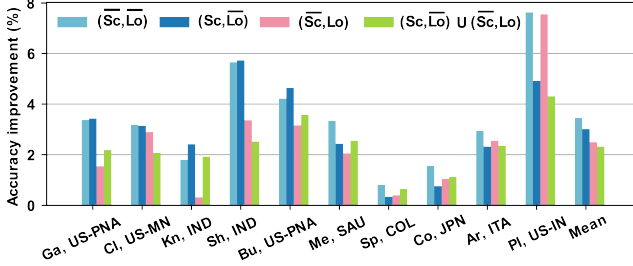
Figure 6: Accuracy improvement of CIR over ERM using the same training: (1) neither the test scenario nor location appears in training ($\overline{\text{Sc}}$,$\overline{\text{Lo}}$), (2) w/ scenario samples ((Sc,$\overline{\text{Lo}}$), (3), w/ location samples($\overline{\text{Sc}}$,Lo)), and (4) w/ both ((Sc, $\overline{\text{Lo}}$)∪($\overline{\text{Sc}}$, Lo)).

worse results when it is excluded). Both outperform the baseline (- $\mathcal{L}_{rc}$ - $\mathcal{L}_{rt}$) without reconstruction by a large margin. We also ablate other decisions in the reconstruction. Recall that $\oplus v$ is computed using learnt attention, while $\oplus v'$ is computed using cross-product attention. We show the impact of reversing each of these decisions. Finally, we show that sharing the same reconstruction ($\oplus v' = \oplus v$) and not sharing the classifier ($h \neq h'$) produces worse results.

Attention Masking. CIR reconstructs each clip from others in the batch. On average, a batch contains 11% videos from the same scenario, 9% from same location and 3% from both. We do not restrict which samples to attend to, only avoiding reconstruction from the sample itself. In Table 4, we ablate possible masks of **S**ame **S**cenario/**L**ocation (**SS**, **SL**) or **O**ther **S**cenario/**L**ocation (**OS**, **OL**). Results obtained without masking are best on average, followed by results where the same/other scenario is masked. On certain splits, masking improves performance. For example, masking out samples from different locations helps for (**Ar, ITA**). We do not use masking (which avoids the need for domain labels) but showcase its potential value when additional knowledge of the domain shift can be utilised.

Effect of scenarios and locations on CIR. Figure 6 shows the top-1 accuracy improvement of CIR over ERM when both methods have access to samples from test scenarios and locations. Four cases are evaluated: ($\overline{\text{Sc}}$,$\overline{\text{Lo}}$), (Sc,$\overline{\text{Lo}}$), ($\overline{\text{Sc}}$,Lo), and (Sc, $\overline{\text{Lo}}$)∪($\overline{\text{Sc}}$, Lo). CIR improves over ERM in every case and every split. The improvement is largest on the hardest case ($\overline{\text{Sc}}$,$\overline{\text{Lo}}$).

Support-Set Size. In Table 5 we show how CIR is affected by the size of the batch, which determines the size of the support set used for reconstruction. CIR is relatively stable over a range of sizes, with slightly worse performance for very small or very large batch sizes.

Text models. We compare the CLIP-ViT-B-32 text encoder to other pre-trained language models in Table 6. Results are comparable for different language models.

CIR exploits text narrations to help overcome domain shifts. Table 7 shows the benefit of this approach, and that merely adding video-text association to existing methods is

	Cl US-MN	Bu US-PNA	Co JPN	Ar ITA	Pl US-IN	Mean
16	23.90	22.99	**26.04**	23.87	28.46	25.05
64	23.89	24.36	**26.54**	24.98	28.97	25.75
128	**25.51**	24.93	26.34	25.67	**30.94**	**26.68**
256	25.00	**24.97**	26.52	**25.96**	30.61	26.61
2048	24.66	24.73	25.48	25.53	30.27	26.14

Table 5: Effect of varying the batch size on CIR.

LM	Cl US-MN	Bu US-PNA	Co JPN	Ar ITA	Pl US-IN	Mean
CLIP-ViT-B-32 [40]	**25.51**	24.93	26.34	25.67	**30.94**	**26.68**
all-mpnet-base-v2 [43]	25.15	25.01	26.30	**25.73**	30.71	26.58
all-miniLM-L6-v2 [52]	25.08	**25.36**	**26.36**	25.45	30.50	26.55

Table 6: Comparison of pre-trained text models.

	T	Cl US-MN	Bu US-PNA	Co JPN	Ar ITA	Pl US-IN	Mean
ERM		22.35	20.73	24.81	22.75	23.29	22.78
MMD*		23.60	22.08	25.87	23.84	24.78	24.03
Mixup		22.21	21.45	**25.90**	23.85	24.41	23.56
CIR		**24.52**	**24.64**	25.42	**25.71**	**30.17**	**26.09**
ERM	✓	23.32	23.30	25.84	24.31	27.32	24.82
MMD*	✓	23.69	23.43	25.90	24.27	27.66	24.99
Mixup	✓	23.94	22.94	25.45	24.71	28.52	25.11
CIR	✓	**25.51**	**24.93**	**26.34**	**25.67**	**30.94**	**26.68**

Table 7: Impact of adding text to existing DG methods. T indicates text supervision. * requires additional domain label supervision.
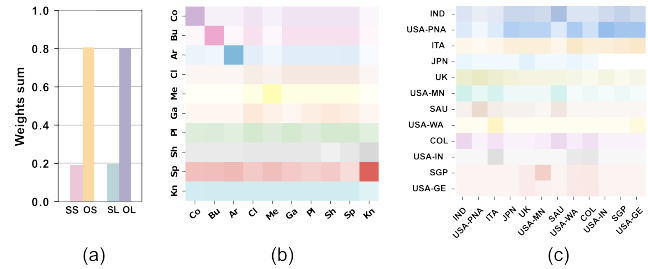

Figure 7: Analysis of attention during reconstruction. (a) Normalised sum of attention weights over SS, OS, SL, OL. (b) Cross-scenario attention (c) Cross-location attention.

insufficient. We add the text association loss L_{rt}, acting directly on video representations (*i.e.* no reconstruction) to existing DG methods. We compare MMD, which performs second best after CIR, and which requires domain labels. We also provide results for ERM and Mixup which do not require domain labels, and thus have the same level of supervision as CIR. Importantly, CIR *without text* is better than other methods *with text*.

5.3. CIR Analysis

Figure 7 analyses how videos attend to other videos during reconstruction-text association. (a) shows that videos primarily attend to other scenarios and locations, which helps to learn representations that generalise across domain shifts. (b) shows attention between scenarios, with some strong self-attention (*e.g.* cooking) as well as cross-attention (*e.g.* sport attending to knitting). Certain scenarios attend
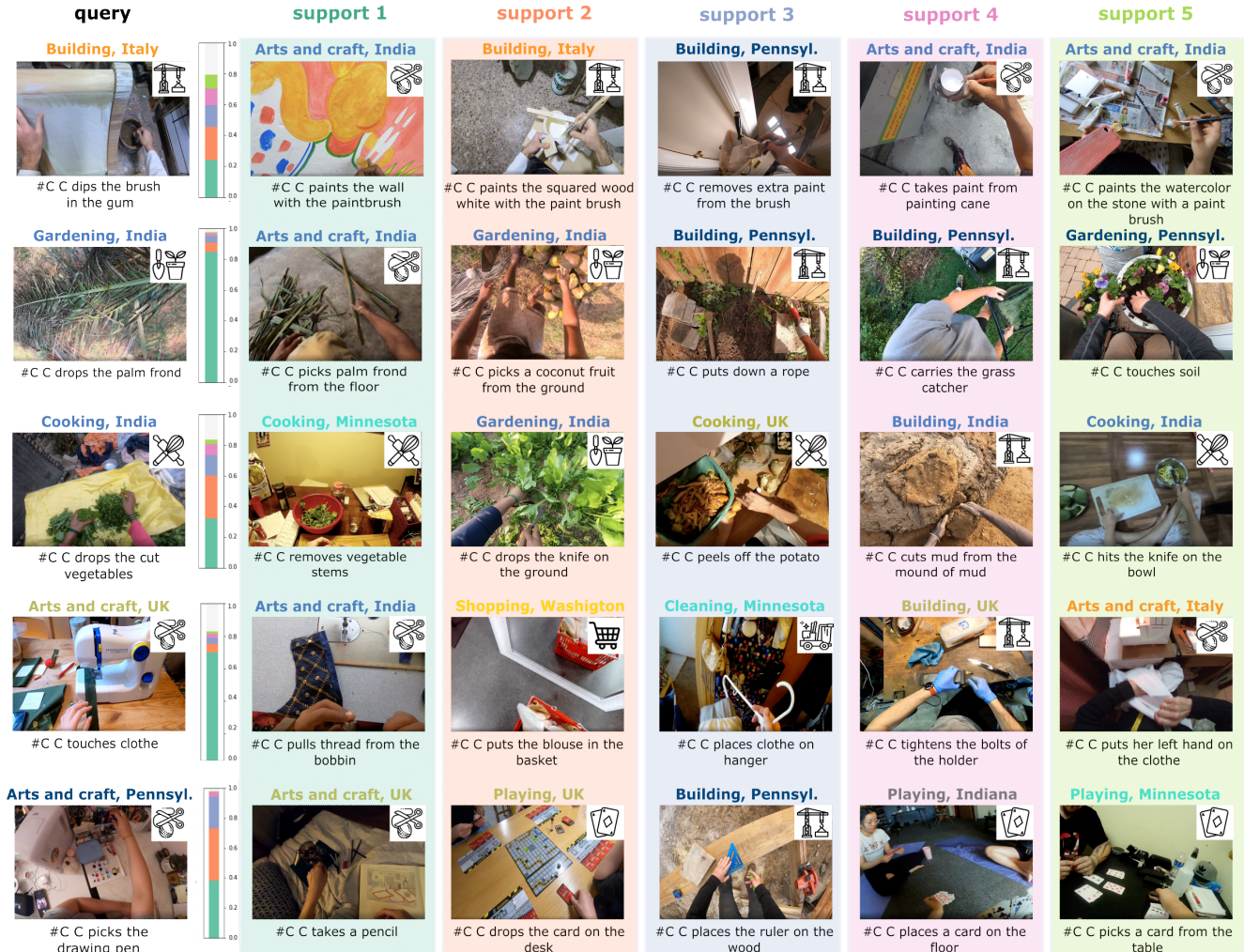
Figure 8: CIR weights for reconstruction. Five examples of cross-instance reconstruction from the training set. The query video is shown on the left. For each video, we show its corresponding scenario/location/narration. For each query, the bar shows the score of the j-th support video (colour-matched) with white indicating the sum of the remaining scores from other samples.

evenly to all scenarios (*e.g.* playing). (c) shows attention between locations, which has fewer strong entries, suggesting that knowledge from all locations is helpful.

We show selected samples of our reconstructions during training in Fig. 8. The Top-5 support set videos with the highest weights in the reconstruction (right) to the query video (left) obtained via CIR (c_{ij}, Section 4.2) are shown. CIR is able to attend to samples belonging to other scenarios, other locations, and both. For example, in the top row, a video of painting from a 'Building' scenario in Italy is reconstructed using examples of 'Arts and Crafts' in India, as well as 'Building' from Italy.

6. Conclusion

In this paper, we introduced ARGO1M, a dataset for Action Recognition Generalisation Over scenarios and locations. We hypothesise that it is plausible to learn actions in a way that generalises to new scenarios (*e.g.* an action 'cut'

in cooking can be used to recognise 'cut' by a mechanic) in new locations (*e.g.* the action 'cut' in Italy can be used to recognise 'cut' in India), as motivated by our paper's title. We propose a method to reconstruct a video using samples from other scenarios and locations. In doing so, the learnt representation is generalisable to test splits with different scenarios and/or locations. CIR consistently improves over baselines, and we offer extensive analysis and ablations.

The problem posed by ARGO1M is both practical and challenging. We hope this paper will foster further research on domain generalisation, which is under-explored in videos.

References

[1] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. Metareg: Towards domain generalization using meta-regularization. In *NeurIPS*, 2018. 2

[2] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, 2018. 2

[3] Silvia Bucci, Antonio D'Innocente, Yujun Liao, Fabio M Carlucci, Barbara Caputo, and Tatiana Tommasi. Self-supervised learning across domains. *IEEE TPAMI*, 44(9):5516–5528, 2021. 2

[4] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *CVPR*, 2019. 2, 5

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 6

[6] Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. Compound domain generalization via meta-knowledge encoding. In *CVPR*, 2022. 2

[7] Chaoqi Chen, Luyao Tang, Feng Liu, Gangming Zhao, Yue Huang, and Yizhou Yu. Mix and reason: Reasoning over semantic topology with data mixing for domain generalization. In *NeurIPS*, 2022. 2

[8] Min-Hung Chen, Zsolt Kira, and Ghassan AlRegib. Temporal attentive alignment for video domain adaptation. In *ICCV*, 2019. 2

[9] Jinwoo Choi, Gaurav Sharma, Manmohan Chandraker, and Jia-Bin Huang. Unsupervised and semi-supervised domain adaptation for action recognition from drones. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1717–1726, 2020. 2

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, 2018. 1, 3

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 130(1):33–55, 2022. 2

[12] Carl Doersch, Ankush Gupta, and Andrew Zisserman. Crosstransformers: spatially-aware few-shot transfer. In *NeurIPS*, 2020. 2

[13] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, 2019. 2

[14] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 6

[15] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. In *NeurIPS*, 2022. 2

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016. 2, 5, 6

[17] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *CVPR*, 2022. 1, 2, 6

[18] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 13(1):723–773, 2012. 2, 5

[19] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021. 2, 5

[20] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. 2017. 3

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 6

[22] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *ICCV*, 2021. 2

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[24] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018. 2

[25] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *ICCV*, 2017. 1, 2

[26] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *ICCV*, 2019. 2

[27] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *CVPR*, 2018. 2, 5, 6

[28] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *ECCV*, 2018. 2

[29] Yiying Li, Yongxin Yang, Wei Zhou, and Timothy Hospedales. Feature-critic networks for heterogeneous domain generalization. In *ICML*, 2019. 2

[30] Seonwoo Min, Nokyung Park, Siwon Kim, Seunghyun Park, and Jinkyu Kim. Grounding visual representations with texts for domain generalization. In *ECCV*, 2022. 2

[31] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *CVPR*, 2020. 1, 2

[32] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *CVPR*, 2021. 2

[33] Hongjing Niu, Hanting Li, Feng Zhao, and Bin Li. Domain-unified prompt representations for source-free domain generalization. *arXiv preprint arXiv:2209.14926*, 2022. 2

[34] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[35] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, Joao Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021. 2

[36] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, 2019. 2

[37] Toby Perrett, Alessandro Masullo, Tilo Burghardt, Majid Mirmehdi, and Dima Damen. Temporal-relational crosstransformers for few-shot action recognition. In *CVPR*, 2021. 2

[38] Toby Perrett, Saptarshi Sinha, Tilo Burghardt, Majid Mirhemdi, and Dima Damen. Use your head: Improving long-tail video recognition. In *CVPR*, 2023. 2

[39] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. In *WACV*, 2022. 1, 2

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 2021. 7

[41] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. Association for Computational Linguistics, 2019. 6

[42] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022. 2

[43] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020. 7

[44] Xiaolin Song, Sicheng Zhao, Jingyu Yang, Huanjing Yue, Pengfei Xu, Runbo Hu, and Hua Chai. Spatio-temporal contrastive domain adaptation for action recognition. In *CVPR*, 2021. 2

[45] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016. 2, 5, 6

[46] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR*, 2011. 1, 2

[47] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE trans. neural netw.*, 10(5):988–999, 1999. 5

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[49] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *CVPR*, 2017. 2

[50] Riccardo Volpi and Vittorio Murino. Addressing model vulnerability to distributional shifts over image transformation sets. In *ICCV*, 2019. 2

[51] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, 2018. 2

[52] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020. 7

[53] Yufei Wang, Haoliang Li, and Alex C Kot. Heterogeneous domain generalization via domain mixup. In *ICASSP*, 2020. 2, 5, 6

[54] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *AAAI*, 2020. 2

[55] Yuzhe Yang, Hao Wang, and Dina Katabi. On multi-domain long-tailed recognition, generalization and beyond. In *ECCV*, 2022. 2, 5, 6

[56] Zhiyu Yao, Yunbo Wang, Jianmin Wang, S Yu Philip, and Mingsheng Long. Videodg: generalizing temporal relations in videos to novel domains. *IEEE TPAMI*, 44(11):7989–8004, 2021. 2

[57] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2

[58] Xin Zhang, Yusuke Iwasawa, Yutaka Matsuo, and Shixiang Shane Gu. Amortized prompt: Lightweight fine-tuning for clip in domain generalization. *arXiv preprint arXiv:2111.12853*, 2021. 2

[59] Yabin Zhang, Minghan Li, Ruihuang Li, Kui Jia, and Lei Zhang. Exact feature distribution matching for arbitrary style transfer and domain generalization. In *CVPR*, 2022. 2

[60] Zangwei Zheng, Xiangyu Yue, Kai Wang, and Yang You. Prompt vision transformer for domain generalization. *arXiv preprint arXiv:2208.08914*, 2022. 2, 5, 6

[61] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, 2018. 6

[62] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization in vision: A survey. *IEEE TPAMI*, 2021. 1, 2