frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 26 September 2022
DOI 10.3389/fcomp.2022.876846

Check for updates

OPEN ACCESS

EDITED BY
Elisa Ricci,
University of Trento, Italy

REVIEWED BY
Yang Wang,
University of Manitoba, Canada
Massimiliano Mancini,
University of Tübingen, Germany

*CORRESPONDENCE
Diala Lteif
dlteif@bu.edu

SPECIALTY SECTION
This article was submitted to Computer Vision, a section of the journal Frontiers in Computer Science

RECEIVED 15 February 2022
ACCEPTED 31 August 2022
PUBLISHED 26 September 2022

CITATION
Majumdar SS, Jain S, Tourni IC, Mustafin A, Lteif D, Sclaroff S, Saenko K and Bargal SA (2022) Ani-GIFs: A benchmark dataset for domain generalization of action recognition from GIFs.
Front. Comput. Sci. 4:876846.
doi: 10.3389/fcomp.2022.876846

COPYRIGHT
© 2022 Majumdar, Jain, Tourni, Mustafin, Lteif, Sclaroff, Saenko and Bargal. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Ani-GIFs: A benchmark dataset for domain generalization of action recognition from GIFs

Shoumik Sovan Majumdar, Shubhangi Jain, Isidora Chara Tourni, Arsenii Mustafin, Diala Lteif*, Stan Sclaroff, Kate Saenko and Sarah Adel Bargal

Department of Computer Science, Boston University, Boston, MA, United States

Deep learning models perform remarkably well for the same task under the assumption that data is always coming from the same distribution. However, this is generally violated in practice, mainly due to the differences in data acquisition techniques and the lack of information about the underlying source of new data. Domain generalization targets the ability to generalize to test data of an unseen domain; while this problem is well-studied for images, such studies are significantly lacking in spatiotemporal visual content—videos and GIFs. This is due to (1) the challenging nature of misalignment of temporal features and the varying appearance/motion of actors and actions in different domains, and (2) spatiotemporal datasets being laborious to collect and annotate for multiple domains. We collect and present the first synthetic video dataset of Animated GIFs for domain generalization, *Ani-GIFs*, that is used to study the domain gap of videos vs. GIFs, and animated vs. real GIFs, for the task of action recognition. We provide a training and testing setting for *Ani-GIFs*, and extend two domain generalization baseline approaches, based on data augmentation and explainability, to the spatiotemporal domain to catalyze research in this direction.

KEYWORDS

domain generalization, domain adaptation, video action recognition, GIFs, transfer learning, explainability

1. Introduction

Deep neural networks allow us to learn representations for a variety of computer vision tasks when large amounts of labeled data are available, but are susceptible to a *domain shift*, when applied to unseen data of new domains at test time. Solutions such as further fine-tuning the network on new data, are not always efficient or trivial, and data collection and annotation are expensive and time-consuming processes, setting obstacles to the application and generalization of the existing models to other domains.

Domain adaptation attempts to address these shortcomings, by training a network on labeled data from a single (Pan et al., 2010; Baktashmotlagh et al., 2016; Long et al., 2016) or multiple (Duan et al., 2012; Jhuo et al., 2012; Yang and Hospedales, 2014; Liu et al., 2016; Xu et al., 2018) source domains, and on a related but different target domain, to learn more transferable representations. Since labeled data are often limited and hard to obtain, unsupervised domain adaptation (Long et al., 2015, 2018; Ganin et al., 2016; Sun and Saenko, 2016; Wilson and Cook, 2020) is of most interest, aiming to leverage the few or no labeled samples. A more complex problem is deep domain generalization (Muandet et al., 2013; Ghifary et al., 2015; Li et al., 2017, 2018b), in which the model is completely unaware of the target domain, and does not see any samples from the target distribution during training. These methods have been widely explored for images, but the scarcity of work and applications in videos serves as a motivation for our current approach.

Our paper comes to address the crucial need to build high-quality benchmark video datasets, in multiple domains, to objectively measure the performance of these techniques. This is because well-defined, rich in features, labeled datasets allow for a universal evaluation of the different methods (Ponce et al., 2006; Torralba and Efros, 2011; Russakovsky et al., 2015; Beery et al., 2018; Recht et al., 2019). Given the arduity of real-world data collection and labeling, synthetic data have grown in popularity, as they can be generated in abundance, introducing a substantial domain gap when compared to other domains (Ros et al., 2016; Rössler et al., 2018; Cruz et al., 2020; Kong et al., 2020; Scheck et al., 2020).

Our focus is on videos, and, more specifically, on Animated GIFs (Eppink, 2014), in which this gap is identified in both space and time (unlike in images, which suffer only from spatial domain shift). Temporal features can be misaligned between domains, which makes the problem more challenging, and significantly under-explored. GIFs are videos that are short in duration, designed to repeat (or re-play), and do not include audio. They typically illustrate a certain action, and have the ability to express a broad spectrum of emotions, aiming at performance of affect and conveyance of cultural knowledge (Miltner and Highfield, 2017). GIFs are created by sampling frames from a video and are extensively used nowadays on the internet, especially in social networks and online communication (Tolins and Samermit, 2016; Jiang et al., 2018). Animated GIFs are synthetically generated and tend to exaggerate or emphasize action motion. In this work, we aim to answer the following questions: *How large is the domain gap between (1) videos and GIFs, and (2) animated and real GIFs?*

We propose the first synthetic domain generalization Animated GIFs dataset, *Ani-GIFs*, designed for the task of action recognition in videos. To our knowledge, no other synthetic GIFs dataset exists designed explicitly for spatiotemporal domain generalization, as depicted in Table 1. Figure 1 presents sample examples from *Ani-GIFs*, and contrasts it with GIFs of the real domain from the Kinetics GIFs dataset. We evaluate domain generalization baselines on *Ani-GIFs* using an I3D action recognition model (Carreira and Zisserman, 2018).

In order to verify the model robustness on our benchmark and the suitability of the dataset for testing domain adaptation and domain generalization methods, we employ the data augmentation approach proposed by Volpi and Murino (2019) for images and extend it to GIF (video) frames. We define a series of content-preserving frame transformations (e.g., contrast enhancement, sharpness/color adjustment), which do not alter the content of the frames, but only the way it is presented. Starting with the identity transformation, we apply a set of concatenated data transformations, given as tuples of a specific size, to the training data, in an alternating process of augmenting the samples with a uniformly selected tuple from the set, and training the model to choose the one among those applied which maximizes the model loss, using a random-search algorithm for selection, so as to strengthen our model.

TABLE 1 Comparing our proposed benchmark to existing ones for spatiotemporal action recognition.

Dataset	Domain	Dataset size	Number of classes
Kinetics-600	Real videos	70,901	600
HMDB-51	Real videos	6,849	51
Ani-GIFs	Synthetic GIFs *(animated, cartoon, graphics)*	17,095	536

Dataset size represents the number of samples, videos or GIFs, in each dataset. Ani-GIFs is the first synthetic GIFs dataset to our knowledge designed explicitly to study the domain gap between videos vs. GIFs, and animated vs. real GIFs for the task of domain generalization.

We also extend an explainability-based domain generalization technique initially proposed for images (Zunino et al., 2020) to the spatiotemporal domain. Explainability, i.e., using the correct evidence for prediction, is utilized to bridge the gap between the real and the synthetic domains. The black-box nature of deep neural network models creates highly non-linear feature representations that make it difficult to understand what causes models to make certain classification decisions. We use the extended saliency-based explainability approach to identify regions in the image that contribute the most to the model's predictions. We leverage these spatiotemporal saliency tubes to guide the model in focusing on image regions, where a particular action is being performed, as opposed to focusing on domain-specific details, that do not necessarily generalize across domains.

To summarize, our contributions are: providing (1) a spatiotemporal dataset, (2) a training and testing setting, (3) a spatiotemporal baseline, (4) an augmentation-based spatiotemporal training strategy, and (5) an explainability-based spatiotemporal training strategy, to enable research addressing the challenging domain generalization problem.

Our paper is organized as follows: First, we discuss the related work on GIF and video datasets, state-of-the-art methods for domain generalization, domain adaptation, data augmentation (Section 2), and explainability. Second, we describe our dataset and the processes of collection and annotation (Section 3). We then analyze the selected baseline

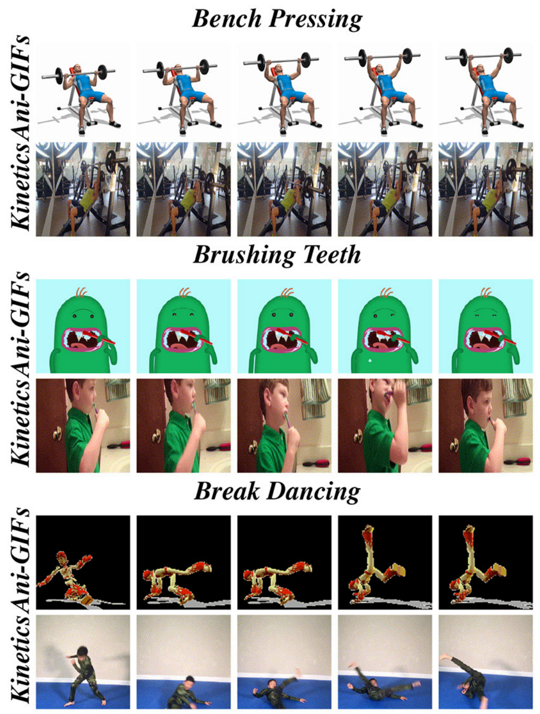

FIGURE 1
This figure highlights the spatiotemporal domain gap between *Ani-GIFs*, our proposed benchmark dataset, and GIFs of the real domain—from the Kinetics dataset (Kay et al., 2017)—for three classes: *Bench Pressing*, *Brushing Teeth*, and *Break Dancing*. This illustrates the domain gap between real vs. animated frames.

methods for the task of action recognition (Section 4) and evaluate the performance presenting the experimental results of our approach (Section 5), before concluding our work (Section 6). Our dataset and baseline implementations will be made publicly available upon acceptance.

2. Related work

2.1. Domain adaptation

Domain adaptation tackles the problem of domain shift between one or more source domains to a different but related target domain. The case where unlabeled data from the target are available for training is addressed by Unsupervised Domain Adaptation (UDA). UDA methods can be categorized into: divergence-based (Long et al., 2017; Saito et al., 2018), adversarial-based (Ganin et al., 2016; Tzeng et al., 2017; Liu et al., 2021), and reconsrtuction-based (image-level translation) methods (Hoffman et al., 2018; Murez et al., 2018). Divergence-based methods focus on minimizing a divergence criterion between the source and target distributions, like the Maximum Mean Discrepancy (MMD) (Long et al., 2017). Adversarial-based approaches focus on making features from different domains indistinguishable. Semi-Supervised Domain Adaptation (SSDA) addresses the other case where a few target labels are provided. In addition, other image domain adaptation methods can be applied to cross-domain tasks, like domain generalization, UDA, and SSDA (Nam et al., 2021).

2.2. Video domain adaptation

The problem of domain adaptation in video action recognition is still under-explored, despite the extensive work in this area for image classification and object recognition. Two approaches are introduced by Jamal et al. (2018), Action Modeling on Latent Subspace (AMLS), which models the videos as points or sequences of points in a latent space, and uses adaptive kernels to learn from source domain points to target domain point sequences, and Deep Adversarial Action Adaptation (DAAA), an adversarial learning framework built to minimize the domain shift. In a most recent work, Chen et al. (2019), a variety of alignment and learning techniques are proposed or extended to minimize domain discrepancy in videos along the spatial and temporal directions. In Chen et al. (2020), the authors propose a generative adversarial network, VideoGAN, which uses an X-shape generator to preserve the intra-video consistency during translation of video data across different domains, and a color-based loss, to tune the color distribution of each translated frame and bridge the domain gap.

2.3. Domain generalization

In domain generalization methods, a relaxed approach is adopted in learning distributions of source domains to generalize to unseen domains, without prior knowledge of the target distribution. Common DG methods can be categorized into: domain agnostic/invariant model learning (Muandet et al., 2013; Ghifary et al., 2015; Dou et al., 2019), self-supervision based (Kim et al., 2021), data-augmentation based (Volpi et al., 2018; Yao et al., 2019), and feature-augmentation based DG methods (Li et al., 2021).

2.4. Video domain generalization

Several techniques have been introduced to solve this problem with deep models (Muandet et al., 2013; Li et al., 2017, 2018a; Motiian et al., 2017), and with important results for a variety of datasets and data types, but the area is significantly under-explored with respect to video datasets, due to the complexity of entangling spatial and temporal domain shifts. In Yao et al. (2019, 2021), the only recent prominent work in this area, the authors present the Adversarial Pyramid Network (APN), a network capturing the videos' local-, global-, and multi-layer cross-relation features. They also extend an adversarial data augmentation method in Volpi et al. (2018), ADA, to videos. Their improved approach, namely Robust Adaptive Data Augmentation (RADA), uses robust regularization to improve the robustness of APN to various adversarial perturbations derived from the relational features at multiple levels. Given the reliability of RADA on those relational features, it is intimately coupled with the proposed APN architecture and does not perform as well on other non-hierarchical models, contrary to other model-agnostic domain generalization approaches.

2.5. Video domain adaptation/generalization datasets

Several existing datasets built for Video analysis tasks are or could be extended to solve the problem of domain shift in action videos, but few new video datasets have been introduced exclusively for the task of domain adaptation or generalization for video action recognition, and are all depicting real actions. The Gameplay dataset (Chen et al., 2019) is a collection of videos of length 1–10 s in 91 categories from two video games. Selecting 30 overlapping categories between Gameplay and Kinetics (Kay et al., 2017; Carreira et al., 2018), the authors create the Kinetics-Gameplay dataset, observing a significant domain shift in the distributions of virtual and real data. In the same work, all relevant and overlapping categories between existing video datasets UCF101 (Soomro et al., 2012) and HMDB51 (Kuehne et al., 2011) are combined in UCF-HMDB$_{\text{full}}$, a large-scale collection of videos of length 1–33 s in 12 classes, used in evaluating several state-of-the-art video domain adaptation methods (Ganin and Lempitsky, 2014; Long et al., 2017; Li et al., 2018c; Saito et al., 2018). For domain generalization, Yao et al. (2019, 2021) propose four video domain generalization benchmarks, UCF-HMDB, Something-Something, PKU-MMD, and NTU, built from existing action recognition videos, in which they divide the source and target domains according to different datasets, consequences of actions, and camera views, to test their method's performance. In parallel, datasets with a focus on more specific tasks such as autonomous driving (Yu et al., 2018) and medical diagnosis (Cheplygina et al., 2017) have been introduced, allowing for domain adaptation evaluation in a variety of sub-domains.

2.6. GIF datasets and analysis techniques

There is an abundance of GIF datasets collected and available in the literature. TGIF (Li et al., 2016) is a dataset of 100K animated GIFs from Tumblr and 120K natural language descriptions obtained *via* crowdsourcing, serving as a benchmark for the task of visual content captioning, namely in the generation of natural language descriptions for animated GIFs or video clips. In Vid2GIF (Gygli et al., 2016), a robust framework, RankNet, is proposed, to learn the content in videos most frequently selected for creating popular animated GIFs, and produce a ranked list of segments according to their suitability, generalizing this ability to other tasks such as video highlight detection. To this purpose, a dataset of 120K user-generated animated GIFs with their corresponding video sources is collected, that is one to two orders of magnitude larger than existing datasets in video highlight detection. GIF Super-Resolution (Wang et al.) is an approach proposed to tackle the problem of slow download speed of GIFs, by using the first and last high-resolution frames of a GIF and a low-resolution representation of it, to reconstruct a GIF easier to process. To this purpose, the authors create GIFSR, a dataset of 1,000 GIFs in 5 categories: Emotion, Action, Scene, Animation, and Animal. In GIFGIF+ (Chen et al., 2017), an emotions GIF dataset is introduced, consisting of 23,544 GIFs over 17 emotion categories, as the authors propose a novel method for animated GIFs collection, to explore the problem of automatic analysis of emotions in GIFs. Similarly, in Jou et al. (2014), 4,000 GIFs are collected, with scores for 17 discrete emotions, and are used in a computational analysis and evaluation of emotion prediction on animated GIFs. However, all these datasets were designed to be used for tasks other than domain adaptation or generalization.

2.7. Data augmentation

Data augmentation is widely used as a model domain generalization improvement technique in computer vision, to obtain more information from the training dataset, and reduce the gap between this and the unseen validation set, preventing the model from performing poorly in evaluation (Shorten and Khoshgoftaar, 2019). When applied on image datasets, data augmentation techniques exploit the spatial properties of the data, and can range from image manipulations, such as geometric or color transformations, rotation, or blurring (Ciregan et al., 2012; Wan et al., 2013; Sato et al., 2015), to feature space augmentation (DeVries and Taylor, 2017), adversarial training techniques (Moosavi-Dezfooli et al., 2016; Volpi et al., 2018; Zajac et al., 2019), and GAN-based approaches (Bowles et al., 2018). Expanding the objective to videos, the proposed methods augment the dataset in both spatial and temporal dimensions, in domain generalization approaches for tasks such as semantic segmentation (Budvytis et al., 2017) and video action recognition (Yao et al., 2019, 2021).

2.8. Explainability

Explainability techniques were initially developed as a diagnostic tool to visualize and explain a model's behavior. GradCAM (Selvaraju et al., 2017) is a gradient-based approach that uses gradients flowing into a target layer to compute coarse localization maps at that layer. In recent work on explainability, Zunino et al. (2020) use an explainability-based training strategy on images to boost model performance. We extend this to the spatiotemporal domain by computing saliency tubes using GradCAM (Selvaraju et al., 2017) in space and time.

3. Our dataset: *Ani-GIFs*

In this section, we introduce our benchmark dataset together with conducted collection and filtration procedures. Our dataset focuses on actions occurring in Animated GIFs, in mirror classes of the Kinetics-600 dataset.

We propose *Ani-GIFs* as a domain generalization benchmark, acting as the target domain in a domain generalization approach from a **real** source domain of actions performed by human characters, to a **synthetic** target domain of actions performed by animated/cartoon/graphical characters. As the real domain dataset, we are using the GIFs from the existing Kinetics dataset (Kay et al., 2017), and we collect the GIFs in the synthetic domain, forming the proposed dataset, *Ani-GIFs*.

3.1. Data collection

We created the *Ani-GIFs* dataset by collecting animated GIFs using the Bing search engine. For each action class in the Kinetics-600 dataset, we set up an automated script to search and download datapoints. Three search keywords were used, the first being "animated" or "cartoon" or "graphics", the second being the action class, and the third being "GIF". For example, for the action class "Applauding", we performed three separate searches: "animated Applauding gif", "cartoon Applauding gif", and "graphics Applauding gif". The keywords "animated", "graphics", and "cartoon" were used synonymously as means of maximizing the number of retrievals from the search engine. We then collected GIFs from each separately. Each of the three collection processes, for all 600 classes, took approximately 100 h to complete.

3.2. Filtration and annotation

After collecting the animated GIFs, we performed extensive filtering. The first stage of filtering was combining search results of animated, cartoons, and graphics and removing duplicates.

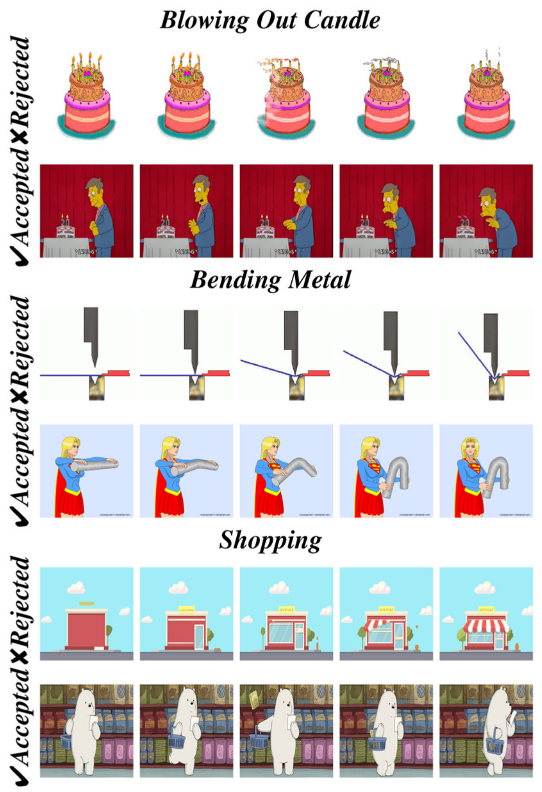

FIGURE 2
In this figure, we can see that first GIFs, in *Blowing out candles* and *Bending metal* classes, were rejected as the actions are not performed by any character. We also rejected GIFs in the *Shopping* class, as the action was not relevant to the class (i.e., no shopping action was observed).

The second stage was performed manually by four graduate students. This stage involved ensuring a downloaded video was indeed: (1) a GIF, (2) performed by an animated, cartoon or graphics character, and (3) depicting the exact class action in Kinetics-600. Annotation was performed based on the action classes only, and not on the type of synthetic domain. Figure 2 provides examples of collected animated GIFs which were rejected or accepted during the filtering process.

3.3. Correspondence with Kinetics-600

Ani-GIFs is designed to have one-to-one correspondence with the classes of Kinetics-600, to act as a domain generalization benchmark. 60 classes from Kinetics-600 did not have corresponding animated GIFs after filtration. Examples for such classes that do not typically have associated animated GIFs, are: *Arranging Flowers, Changing Oil, Curling Hair, Feeding Goats, Making Jewellery, Sharpening Knives, Putting On Sari*. Therefore, the resulting *Ani-GIFs* dataset has 536 classes, and

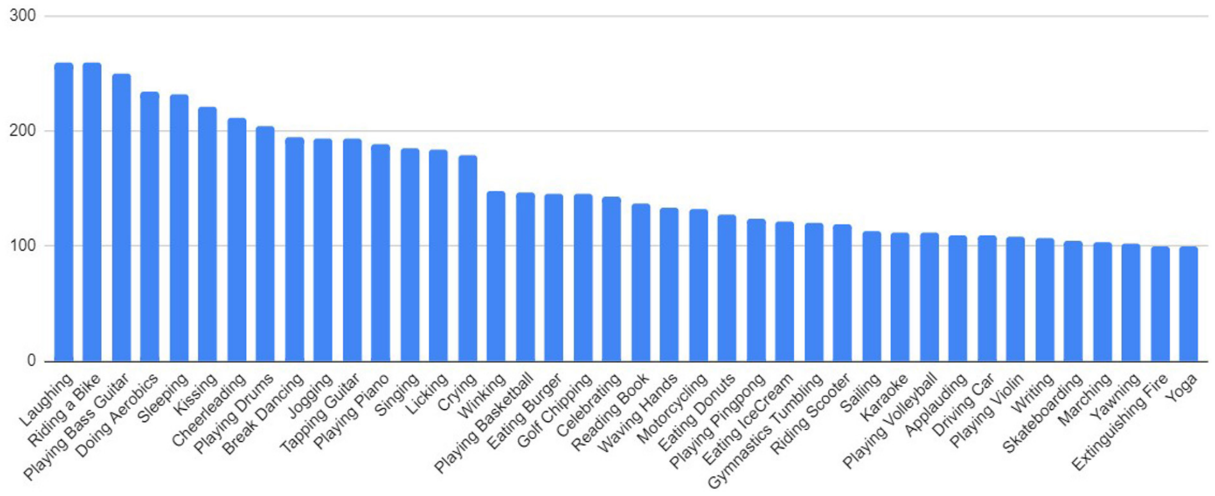

FIGURE 3
The 40 most frequent classes after filtration of the *Ani-GIFs* dataset. Classes that have the highest frequency belong to actions with a large number of associated GIFs, e.g., common actions and emotions. We choose these forty classes as a subset for GIFs domain adaptation.

17,095 animated GIFs in total, all intersecting with Kinetics-600. Figure 3 shows the number of GIF samples per class in the *Ani-GIFs* dataset for the forty top-frequency classes.

3.4. Subset for domain adaptation

While our dataset is designed for the task of GIF domain generalization, we identify a subset of *Ani-GIFs* for the task of GIF domain adaptation for action recognition. The subset consists of the forty classes having the highest frequency. This would allow for standard testing of domain adaptation, i.e., from Real to Animated GIFs and vice versa.

4. Spatiotemporal domain generalization

In this work we address the challenging problem of single-source domain generalization for spatiotemporal GIFs. At training time, we only have access to a single source domain, and at test time we have access to a different target domain that is unseen at training time. We focus on the real videos/GIFs source domain and the animated GIFs target domain. While the problem of attributing an action to an animated spatiotemporal progression is trivial for humans, it is a significantly challenging task for machine learning models that have only been trained on real video data. The gap between the two domains in this problem setting is large. The two domains exhibit significant variations in color templates, as animated GIFs tend to only have a few colors in all frames, while real videos or GIFs have a significantly richer color template. Moreover, animated GIFs tend to have a smaller level of detail, in contrast to real videos or GIFs. At the same time, animated GIFs exhibit a faster speed for actions than real videos or GIFs, i.e., while the difference in motion between subsequent frames in real videos is usually small even after sub-sampling, the difference between subsequent frames in GIFs is significantly larger. We demonstrate how large this domain gap is experimentally in Section 5.

To reduce this huge domain gap, we use a GIF version of the Kinetics dataset—Kinetics GIFs (Kay et al., 2017)—as the source domain in our data augmentation baseline experiments. Samples in Kinetics GIFs are GIFs produced from original Kinetics videos, which have a fixed length of 40 frames and a significantly smaller resolution, typically of 400 by 400 pixels. After training the model on Kinetics GIFs we evaluate it on *Ani-GIFs* to obtain a baseline performance, that is then compared to applying domain generalization techniques.

We also use the AVA-Kinetics Localized Human Actions Video Dataset (Li et al., 2020) to extend the explainable training strategy of Zunino et al. (2020) on images to the spatiotemporal domain to achieve better evidence for domain generalization. The dataset is an extension of the Kinetics dataset with AVA-style bounding boxes and atomic actions, which makes it suitable as a train set in our explainability-based approach. AVA-Kinetics has more than 230k clips labeled with one of 80 AVA action classes, which are manually mapped to their corresponding top related Kinetics classes.

4.1. Data augmentation approach

We extend the work of Volpi and Murino (2019) on images and develop a spatiotemporal data augmentation approach for animated GIFs. Data augmentation is a very powerful

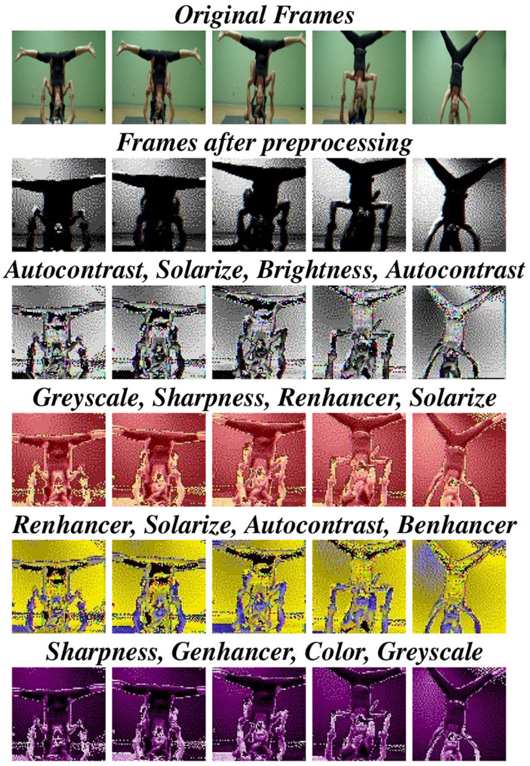
Original Frames
Frames after preprocessing
Autocontrast, Solarize, Brightness, Autocontrast
Greyscale, Sharpness, Renhancer, Solarize
Renhancer, Solarize, Autocontrast, Benhancer
Sharpness, Genhancer, Color, Greyscale

FIGURE 4
This figure presents frames sampled equally from a video in class "Yoga". The first set of frames belongs to the Kinetics GIFs dataset, and is followed by the frames after Batch Normalization is applied. The subsequent set of images depicts the frames after tuples of transformations, chosen by the random search approach, are applied to them.

technique to create additional representations and increase the generalization ability of a model to domains that are unseen at training time. We artificially inflate the dataset by applying transformations in space and time. Following Volpi and Murino (2019), we apply a set of image transformations $\mathcal{T}$ from the Python library Pillow, to compute the augmented versions of each GIF. We consider transformation tuples of length four, i.e., four transformations are applied concurrently to a GIF of the training set for every augmentation. The pool of transformations is (intensity in parenthesis): auto-contrast (20), sharpness (20), brightness (20), color (20), contrast (20), gray scale conversion (1), R-channel enhancer (30), G-channel enhancer (30), B-channel enhancer (30), solarize (20).

Starting with a model pre-trained on the Kinetics-600 dataset, and the transformations set $\mathcal{T}$ containing only identity transformations, we perform a fine-tuning process to identify the tuple of transformations that the model is most vulnerable to. Vulnerability of the model is defined to be the tuple of transformations that leads to the highest value of cross-entropy loss when applied to the input batches. At every iteration of the training process, we randomly sample a tuple from our set of vulnerable transformations $\mathcal{T}$, and apply those to our training batches with their associated intensity values. We train our model using Stochastic Gradient Descent to minimize the cross-entropy loss. The transformations set is updated every 200 training iterations, using random search.

The identification process targets adding one tuple of transformations to the set of known vulnerable transformations $\mathcal{T}$ using a random search approach. At every iteration of the random search, four transformations are randomly sampled with repetition, from the pool of transformations at random intensity values to create a tuple. While extending the augmentation approach to account for temporal shifts, all four transformations of the tuple are applied to all frames of the input batches. This ensures that the same transformation is performed on all frames of the video to obtain a single augmented instance. The vulnerability of the model to this tuple of transformation is then determined by evaluating the cross-entropy loss. At the end of 50 iterations of the searching process, the tuple of transformations that led to the highest cross-entropy loss is identified and added to our set of vulnerable transformations $\mathcal{T}$, along with its intensity value. In subsequent iterations of the standard training process, this identified tuple of transformations is available to be randomly sampled from our set $\mathcal{T}$ and applied to the training batches for training with the Adam optimizer. In Figure 4, we show images after different tuples of transformations applied to frames that were equally sampled from a video in class "Yoga" taken from the Kinetics GIFs dataset. Transformations are applied after Batch Normalization.

4.2. Explainability approach

We extend and apply a saliency-based spatiotemporal explainability approach (Zunino et al., 2020) on our dataset. At training time, saliency maps for the ground-truth class are periodically computed as saliency tubes in space and time. As training progresses, we have access to these regions as bounding box co-ordinates for the input batch. Saliency maps are computed using the GradCAM (Selvaraju et al., 2017) algorithm after the last block of the feature extactor layer l of the model. We estimate saliency on the last spatial layer as it models higher level spatial patterns, that are most correlated with the target label. If the peak saliency does not fall within the ground-truth region, we enforce that by utilizing a multiplicative binary 3D-mask (saliency tube) that is applied to the forward activations of layer l. This mask contains a value of 1 for pixels that lie within the spatiotemporal region of interest and 0 otherwise. We run the saliency estimation periodically every 200 batches, and train using the Adam optimizer.

TABLE 2 Study 1 results. Top-1 and top-5 test accuracies of our baseline algorithm are given, from various training on different testing domains.

Source (train) domain	Target (test) domain	Spatiotemporal augmentation	Test accuracy (%)	
			Top-1	Top-5
Kinetics	Kinetics	✘	71.70	90.40
Kinetics	Kinetics GIFs	✘	21.12	40.86
Kinetics GIFs	Kinetics GIFs	✘	23.10	46.28
Kinetics GIFs	*Ani-GIFs*	✘	1.95	6.09
Kinetics GIFs	*Ani-GIFs*	✔	2.91	8.44

The difference in the reported accuracies between rows 1 and 2 demonstrates the existing domain gap from Kinetics to Kinetics GIFs, and in rows 3 and 4 the domain shift between Kinetics GIFs and *Ani-GIFs*, with the latter dataset used in its entirety for measuring accuracy while testing. The increase from row 4 to 5 shows the gain in accuracy yielded by extending and applying the spatiotemporal data augmentation algorithm for domain generalization on the training dataset, Kinetics GIFs.

TABLE 3 Study 2 results. Top-1 and top-5 test accuracies of our baseline algorithms are given, from fine tuning on 12 classes of the AVA-Kinetics dataset, on the Ani-GIFs dataset, the testing domain.

Source (train) domain	Target (test) domain	Approach	Test accuracy (%)	
			Top-1	Top-5
AVA-Kinetics	*Ani-GIFs*	Random Augmentation	8.56	34.47
AVA-Kinetics	*Ani-GIFs*	Spatiotemporal Augmentation	11.88	42.74
AVA-Kinetics	*Ani-GIFs*	RADA	12.55	34.63
AVA-Kinetics	*Ani-GIFs*	Explainability	17.31	58.78

The difference in the reported accuracies between rows 1 and 2 shows that our spatiotemporal augmentation approach outperforms random augmentation on the training dataset, AVA-Kinetics. Row 3 shows that spatiotemporal augmentation outperforms Robust Adaptive Data Augmentation (RADA) in the top-5 test accuracy. Furthermore, we show that our explainability approach in row 4 outperforms all reported augmentation baselines, verifying the effectiveness of explainability for domain generalization.

5. Experiments

In this section, we start by experimentally demonstrating the huge domain gap between real videos vs. GIFs of the same videos, and real videos vs. animated GIFs. We then demonstrate how spatiotemporal domain generalization can reduce the gap in the latter scenario. Results of this first study are shown in Table 2. Furthermore, we conduct a second study to emphasize the effectiveness of our spatiotemporal and explainability-based approaches over baseline domain generalization and report results in Table 3.

5.1. Datasets

In the domain gap study, which we call Study 1 (Table 2), we use the Kinetics-600 dataset as the source domain for both training and fine-tuning, and the (*real*) Kinetics GIFs vs. *Ani-GIFs* datasets as target domains. As for the second study (Table 3), we use AVA-Kinetics as the source domain to fine-tune a baseline model pre-trained on Kinetics-600 using our spatiotemporal domain generalization algorithms. We then use *Ani-GIFs* as the target domain to compare the performance of our algorithms with baseline domain generalization approaches.

5.2. Experimental setup

We choose the I3D model architecture (Carreira and Zisserman, 2018) as the first baseline for the spatiotemporal training and testing of our videos and animated GIFs, because of its increased transferability and its ability to capture a fine-grained temporal structure of actions. This model builds upon state-of-the-art image classification architectures, expanding their filters and pooling kernels (and optionally their parameters) into 3D, hence learning seamless spatiotemporal features from videos while leveraging successful ImageNet architecture designs and their parameters. More specifically, starting from a 2D architecture, all the filters and pooling kernels are inflated with an additional temporal dimension. The model is trained on 64-frame video snippets at 25 frames per second, processing all video frames at test time, and learning high temporal resolution features.

While training, we perform certain preprocessing on the input frames that aims to improve quality by suppressing unwanted noise in the frames, and enhancing important features. Animated GIFs are preprocessed frame-wise—each frame was rescaled such that its shorter side has length of 224 pixels. Realignment was followed by center cropping, resulting in a frame of size 224 by 224. Hence, during training, each training sample has a fixed size of (40, 224, 224, 3). The number of frames in *Ani-GIFs* samples may though vary, so we upsampled frames for animated GIFs that had less than 9 frames, and subsampled frames of animated GIFs that had more than 60 frames, such that the chosen frames have equal spacing in time. All values were rescaled to the [-1, 1] interval.

The models were trained on four *Nvidia TITAN V* GPUs for 60 epochs with a batch size of 32 samples. We start with an I3D model that is pre-trained on Kinetics videos (Piergiovanni, 2018).

In Study 1, we start with the model trained on Kinetics GIFs using the I3D model architecture, and fine-tune it further with the random search approach and Adam optimizer (Diederik P. Kingma, 2014). The fine-tuning process was performed on three *Nvidia TITAN V* GPUs, in batches of eight animated GIFs. The model was tuned for 600 random search iterations. We used the same upsampling and subsampling criteria as in the training process, which resulted in every animated GIF having a fixed shape of (40, 224, 224, 3). Every frame was similarly preprocessed with realignment, center cropping and rescaling. In order to augment the animated GIFs, we made sure the same transformations are applied to the entire batch of input GIFs, resulting in a batch with a shape of (8*40, 224, 224, 3).

The second study uses AVA-Kinetics for fine-tuning, to compare our spatiotemporal augmentation and explainability algorithms against a baseline with random data augmentations and Yao et al. (2021)'s Robust Adaptive Data Augmentation (RADA). While incorporating the saliency-based approach for our training, we start with a model that is pre-trained on the Kinetics dataset. This pre-trained model uses the same I3D architecture as in Study 1 and is further tuned on the AVA Kinetics dataset using the GradCAM saliency algorithm. We filter the AVA Kinetics dataset and use only 12 classes that are a one-to-one mapping to classes in Ani-GIFs. To ensure a balanced training setting, we sample data from these 12 classes such that every class contains 2000 training datapoints. The fine-tuning process was performed on three *Nvidia TITAN V* GPUs, in batches of 8 videos. We run the saliency estimation every 200 batches. We also apply RADA (Yao et al., 2021) at the same iteration frequency to ensure a fair comparison, and follow the authors' implementation which is available online. We maintain the same data preprocessing as in our augmentation experiment and make sure that the entire batch undergoes the same preprocessing which results in batches with the shape (8*40,224,224,3). All models were trained and fine-tuned using the Adam optimizer with the following hyperparameters: learning rate = 10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.

5.3. Experimental results

The results of our two studies are presented in Tables 2, 3. In Table 2, we begin with two experiments demonstrating the domain gap within videos, and also between videos and GIFs, both from the same (real) domain. The first row of the table reports the results of training and testing processes on Kinetics-600 real videos (Kay et al., 2017), with a 71.7% top-1 accuracy, and the second row reports the outcome of testing the same model on the GIFs version of the Kinetics-600 dataset (Gituma, 2019), similarly in the real domain.

We mark the significant accuracy drop, to a 21.12% top-1 accuracy, which we can attribute to the frame sampling process in GIFs, or the difference in GIF frames' speed, in comparison to videos, between the source and target domains, in the second variation of the model application. We then train a model on the Kinetics GIFs dataset (Gituma, 2019) and test on GIFs from the same dataset and, hence, domain. This, as we can observe, increases the model performance to a higher top-1 accuracy of 23.1%, compared to the previous experiment, as expected when training and testing within the same domain. This result is given in row 3 of Table 2, while rows 4 and 5 show how our domain generalization baseline performs, when trained on the Kinetics GIFs dataset and tested on the *Ani-GIFs* dataset, with and without data augmentation. We can see how our proposed data augmentation approach gives an absolute improvement of 0.96% in the top-1 accuracy and 2.35% in the top-5 accuracy and can serve as an initial baseline for *Ani-GIFs*.

We further compare our spatiotemporal augmentation approach against random and adversarial data augmentation and demonstrate the results in Table 3. We use the same I3D model architecture as in Study 1, pre-trained on Kinetics-600 and fine-tuned on AVA-Kinetics using the same fine tuning process as earlier. We show in rows 1 and 2 of Table 3 that our spatiotemporal approach outperforms random augmentation by 3.32% and 8.27% in the top-1 and top-5 accuracies respectively. In rows 2 and 3, while RADA (Yao et al., 2021) gets a slight top-1 accuracy improvement of less than 1% over our approach, we show that the latter none-the-less outperforms it by 8.11% in the top-5 accuracy. Furthermore, our proposed explainability-based approach outperforms all three augmentation baselines in both the top-1 and top-5 accuracies, which emphasizes the effectiveness of explainability in domain generalization.

5.4. Explainability for spatiotemporal domain generalization

We utilize explainability as a visualization tool for evaluating the generalization capability of models for domain generalization on spatiotemporal data. We show that a model is able to generalize an action across various domains, in our case real vs. animated GIFs for the task of action recognition.

Typically, classification accuracy is reported to summarize the recognition capability of models on classification datasets. However, it alone is not indicative as to whether the models have learnt to generalize an action across the source and target domains. For example, it may be that the model is correctly classifying a sample based on the wrong cues. Figure 5 illustrates examples of poor generalization ability of the baseline model

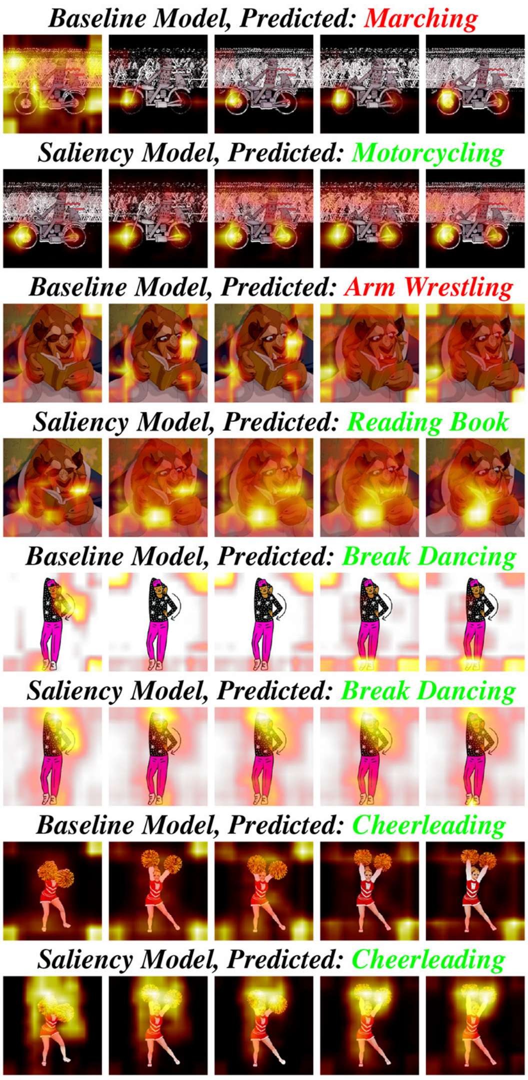
Baseline Model, Predicted: Marching
Saliency Model, Predicted: Motorcycling
Baseline Model, Predicted: Arm Wrestling
Saliency Model, Predicted: Reading Book
Baseline Model, Predicted: Break Dancing
Saliency Model, Predicted: Break Dancing
Baseline Model, Predicted: Cheerleading
Saliency Model, Predicted: Cheerleading

FIGURE 5
This figure presents the visualizations of predictions of four different GIFs from the AniGIFs dataset (GIF frames equally sampled) demonstrating evidence of the baseline model without domain generalization and the model trained with the explainability-based domain generalization approach (Section 4.2). The top two examples show that for some GIFs the explainability approach boosts both the model accuracy and generalization ability. And the bottom two examples show that even when making a correct prediction, the baseline model does not use discriminative cues to make that prediction. In contrast, the model trained with the explainability-based domain generalization approach accurately highlights the correct action-specific cues.

from the source, AVA-Kinetics, to the target domain, *Ani-GIFs*, compared against the saliency model trained with domain adaptation using the explainability approach. We use GradCAM to visualize saliency on different GIFs from the *Ani-GIFs* dataset. In addition, we report evaluation results in Table 3 of our explainability approach on the Ani-GIFs dataset, for an I3D model pre-trained on Kinetics and fine-tuned on 12 classes of AVA-Kinetics using saliency. The results show how using explainability for domain generalization outperforms all three augmentation baselines by a maximum of 8.75% in the top-1 accuracy and 24.31% in the top-5 accuracy.

6. Conclusion

We introduce the first domain generalization GIFs Dataset, *Ani-GIFs*, designed for the task of video action recognition in a synthetic domain, which consists of 536 classes, mirroring the classes in the real domain of the Kinetics GIFs dataset. We discuss the collection and filtration process, provide the results of evaluating a domain generalization baseline, trained on Kinetics GIFs, and an explainability-based domain generalization model, trained on the AVA-Kinetics Localized Human Actions Video Dataset, and also evaluate the baselines after extending and applying an existing image data augmentation technique. Our results show that it is evident that the domain gap in the temporal space is a great challenge. Current domain generalization techniques for images, when extended to Videos/GIFs, showcase a performance improvement, although small enough to highlight the need for better methods tailored toward the temporal dimension. Our dataset serves as a benchmark to catalyze the development and testing of state-of-the-art domain generalization techniques tailored for videos and animated GIFs, and as a motivation for further exploration and enrichment of the existing GIF datasets, to span different domains for the tasks of domain adaptation and domain generalization.

Data availability statement

The urls of the full set of unfiltered images of our dataset supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

SS, KS, and SB: study conception and design. SM, SJ, IT, and AM: data collection and filtration. SM, SJ, IT, AM, and DL: implementation and model training. AM and DL: analysis and interpretation of results. SM, SJ, IT, AM, DL, SS, KS, and SB: draft manuscript preparation. All authors reviewed the results and approved the final version of the manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Baktashmotlagh, M., Harandi, M., and Salzmann, M. (2016). Distribution-matching embedding for visual domain adaptation. *J. Mach. Learn. Res.* 17, 3760–3789. doi: 10.5555/2946645.3007061

Beery, S., Van Horn, G., and Perona, P. (2018). "Recognition in terra incognita," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Cham: Springer International Publishing), 456–473. doi: 10.1007/978-3-030-01270-0_28

Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., et al. (2018). Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*. doi: 10.48550/arXiv.1810.10863

Budvytis, I., Sauer, P., Roddick, T., Breen, K., and Cipolla, R. (2017). "Large scale labelled video data augmentation for semantic segmentation in driving scenarios," in *Proceedings of the IEEE International Conference on Computer Vision Workshops* (Venice), 230–237. doi: 10.1109/ICCVW.2017.36

Carreira, J., and Zisserman, A. (2018). Quo vadis, action recognition? A new model and the kinetics dataset. *arXiv preprint arXiv:1705.07750*. doi: 10.1109/CVPR.2017.502

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., and Zisserman, A. (2018). A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*. doi: 10.48550/arXiv.1808.01340

Chen, J., Li, Y., Ma, K., and Zheng, Y. (2020). Generative adversarial networks for video-to-video domain adaptation. *arXiv preprint arXiv:2004.08058*. doi: 10.1609/aaai.v34i04.5750

Chen, M.-H., Kira, Z., and AlRegib, G. (2019). Temporal attentive alignment for video domain adaptation. *arXiv preprint arXiv:1905.10861*. doi: 10.1109/ICCV.2019.00642

Chen, W., Rudovic, O. O., and Picard, R. W. (2017). "Gifgif+: Collecting emotional animated gifs with clustered multi-task learning," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)* (New York, NY: IEEE), 510–517. doi: 10.1109/ACII.2017.8273647

Cheplygina, V., Pena, I. P., Pedersen, J. H., Lynch, D. A., Sørensen, L., and de Bruijne, M. (2017). Transfer learning for multicenter classification of chronic obstructive pulmonary disease. *IEEE J. Biomed. Health Inform.* 22, 1486–1496. doi: 10.1109/JBHI.2017.2769800

Ciregan, D., Meier, U., and Schmidhuber, J. (2012). "Multi-column deep neural networks for image classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (Providence, RI: IEEE), 3642–3649. doi: 10.1109/CVPR.2012.6248110

Cruz, S. D. D., Wasenmuller, O., Beise, H.-P., Stifter, T., and Stricker, D. (2020). "SVIRO: synthetic vehicle interior rear seat occupancy dataset and benchmark," in *The IEEE Winter Conference on Applications of Computer Vision* (Snowmass, CO), 973–982. doi: 10.1109/WACV45572.2020.9093315

DeVries, T., and Taylor, G. W. (2017). Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*. doi: 10.48550/arXiv.1702.05538

Diederik, P., and Kingma, J. B. (2014). ADAM: a method for stochastic optimization. *arXiv preprint arXiv*:1412.6980. doi: 10.48550/arXiv.1412.6980

Dou, Q., Coelho de Castro, D., Kamnitsas, K., and Glocker, B. (2019). "Domain generalization via model-agnostic learning of semantic features," in *Annual Conference on Neural Information Processing Systems, Vol. 32*, eds H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Vancouver, BC: Curran Associates), 6447–6458.

Duan, L., Xu, D., and Tsang, I. W.-H. (2012). Domain adaptation from multiple sources: a domain-dependent regularization approach. *IEEE Trans. Neural Netw. Learn. Syst.* 23, 504–518. doi: 10.1109/TNNLS.2011.2178556

Eppink, J. (2014). A brief history of the gif (so far). *J. Visual Cult.* 13, 298–306. doi: 10.1177/1470412914553365

Ganin, Y., and Lempitsky, V. (2014). Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*. doi: 10.48550/arXiv.1409.7495

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., et al. (2016). Domain-adversarial training of neural networks. *J. Mach. Learn. Res.* 17, 2096–2030. doi: 10.1007/978-3-319-58347-1_10

Ghifary, M., Bastiaan Kleijn, W., Zhang, M., and Balduzzi, D. (2015). "Domain generalization for object recognition with multi-task autoencoders," in *Proceedings of the IEEE International Conference on Computer Vision* (Santiago), 2551–2559. doi: 10.1109/ICCV.2015.293

Gituma, M. (2019). *The Kinetics Dataset Explorer Using GIFs*. Available online at: https://towardsdatascience.com/the-kinetics-dataset-explorer-using-gifs-8ceeebcbdaba (accessed February 24, 2019).

Gygli, M., Song, Y., and Cao, L. (2016). "Video2gif: automatic generation of animated gifs from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 1001–1009. doi: 10.1109/CVPR.2016.114

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., et al. (2018). "CyCADA: cycle-consistent adversarial domain adaptation," in *International Conference on Machine Learning* (PMLR) (Stockholm), 1989–1998.

Jamal, A., Namboodiri, V. P., Deodhare, D., and Venkatesh, K. (2018). "Deep domain adaptation in action space," in *BMVC* (Newcastle), 264.

Jhuo, I.-H., Liu, D., Lee, D., and Chang, S.-F. (2012). "Robust visual domain adaptation with low-rank reconstruction," in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (Providence, RI: IEEE), 2168–2175. doi: 10.1109/CVPR.2012.6247924

Jiang, J. A., Fiesler, C., and Brubaker, J. R. (2018). 'The perfect one? understanding communication practices and challenges with animated gifs. *Proc. ACM Hum. Comput. Interact.* 2, 1–20. doi: 10.1145/3274349

Jou, B., Bhattacharya, S., and Chang, S.-F. (2014). "Predicting viewer perceived emotions in animated GIFs," in *Proceedings of the 22nd ACM International Conference on Multimedia* (New York, NY), 213–216. doi: 10.1145/2647868.2656408

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*. doi: 10.48550/arXiv.1705.06950

Kim, D., Yoo, Y., Park, S., Kim, J., and Lee, J. (2021). "Selfreg: self-supervised contrastive regularization for domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC), 9619–9628. doi: 10.1109/ICCV48922.2021.00948

Kong, F., Huang, B., Bradbury, K., and Malof, J. (2020). "The synthinel-1 dataset: a collection of high resolution synthetic overhead imagery for building segmentation," in *The IEEE Winter Conference on Applications of Computer Vision* (Snowmass, CO), 1814–1823. doi: 10.1109/WACV45572.2020.9093339

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., and Serre, T. (2011). "HMDB: a large video database for human motion recognition," in *2011 International Conference on Computer Vision* (Barcelona: IEEE), 2556–2563. doi: 10.1109/ICCV.2011.6126543

Li, P., Li, D., Li, W., Gong, S., Fu, Y., and Hospedales, T. M. (2021). "A simple feature augmentation for domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC), 8886–8895. doi: 10.1109/ICCV48922.2021.00876

Li, A., Thotakuri, M., Ross, D. A., Carreira, J., Vostrikov, A., and Zisserman, A. (2020). The AVA-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*. doi: 10.48550/arXiv.2005.00214

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. (2017). "Deeper, broader and artier domain generalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 5542–5550. doi: 10.1109/ICCV.2017.591

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. (2018a). "Learning to generalize: meta-learning for domain generalization," in *Thirty-Second AAAI Conference on Artificial Intelligence* (New Orleans, LA). doi: 10.1609/aaai.v32i1.11596

Li, Y., Song, Y., Cao, L., Tetreault, J., Goldberg, L., Jaimes, A., et al. (2016). "TGIF: a new dataset and benchmark on animated gif description," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 4641–4650. doi: 10.1109/CVPR.2016.502

Li, Y., Tian, X., Gong, M., Liu, Y., Liu, T., Zhang, K., et al. (2018b). "Deep domain generalization via conditional invariant adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 624–639. doi: 10.1007/978-3-030-01267-0_38

Li, Y., Wang, N., Shi, J., Hou, X., and Liu, J. (2018c). Adaptive batch normalization for practical domain adaptation. *Pattern Recogn*. 80, 109–117. doi: 10.1016/j.patcog.2018.03.005

Liu, H., Shao, M., and Fu, Y. (2016). "Structure-preserved multi-source domain adaptation," in *2016 IEEE 16th International Conference on Data Mining (ICDM)* (Montreal, QC: IEEE), 1059–1064. doi: 10.1109/ICDM.2016.0136

Liu, X., Guo, Z., Li, S., Xing, F., You, J., Kuo, C.-C. J., et al. (2021). "Adversarial unsupervised domain adaptation with conditional and label shift: infer, align and iterate," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10367–10376. doi: 10.1109/ICCV48922.2021.01020

Long, M., Cao, Z., Wang, J., and Jordan, M. I. (2018). "Conditional adversarial domain adaptation," in *Advances in Neural Information Processing Systems, Vol. 31*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montreal, QC: Curran Associates Inc), 1640–1650.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2016). "Unsupervised domain adaptation with residual transfer networks," in *Advances in Neural Information Processing System, Vol. 29*, eds D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Red Hook, NY, USA: Curran Associates Inc), 136–144.

Long, M., Zhu, H., Wang, J., and Jordan, M. I. (2017). "Deep transfer learning with joint adaptation networks," in *Proceedings of the 34th International Conference on Machine Learning*, eds D. Precup and Y. W. Teh (Sydney, NSW: PMLR), 2208–2217.

Long, M., Cao, Y., Wang, J., and Jordan, M. (2015). "Learning transferable features with deep adaptation networks," in *International Conference on Machine Learning* (Lille), 97–105.

Miltner, K. M., and Highfield, T. (2017). Never gonna gif you up: analyzing the cultural significance of the animated gif. *Soc. Media Soc*. 3, 2056305117725223. doi: 10.1177/2056305117725223

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. (2016). "Deepfool: a simple and accurate method to fool deep neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV). doi: 10.1109/CVPR.2016.282

Motiian, S., Piccirilli, M., Adjeroh, D. A., and Doretto, G. (2017). "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 5715–5725. doi: 10.1109/ICCV.2017.609

Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation," in *International Conference on Machine Learning* (Atlanta, GA), 10–18.

Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., and Kim, K. (2018). "Image to image translation for domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 4500–4509. doi: 10.1109/CVPR.2018.00473

Nam, H., Lee, H., Park, J., Yoon, W., and Yoo, D. (2021). "Reducing domain gap by reducing style bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Nashville, TN), 8690–8699. doi: 10.1109/CVPR46437.2021.00858

Pan, S. J., Tsang, I. W., Kwok, J. T., and Yang, Q. (2010). Domain adaptation via transfer component analysis. *IEEE Trans. Neural Netw*. 22, 199–210. doi: 10.1109/TNN.2010.2091281

Piergiovanni, A. (2018). *3D models trained on Kinetics*. Available online at: https://github.com/piergiaj/pytorch-i3d (accessed June 28, 2018).

Ponce, J. et al. (2006). "Dataset issues in object recognition,"' in *Toward Category-Level Object Recognition. Lecture Notes in Computer Science, Vol 4170*, eds J. Ponce, M. Hebert, C. Schmid, and A. Zisserman (Berlin; Heidelberg: Springer), 29–48. doi: 10.1007/11957959_2

Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. (2019). Do imagenet classifiers generalize to imagenet? *arXiv preprint arXiv:1902.10811*. doi: 10.48550/arXiv.1902.10811

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). "The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 3234–3243. doi: 10.1109/CVPR.2016.352

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2018). Faceforensics: a large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*. doi: 10.48550/arXiv.1803.09179

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. (2018). "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 3723–3732. doi: 10.1109/CVPR.2018.00392

Sato, I., Nishimura, H., and Yokoi, K. (2015). APAC: augmented pattern classification with neural networks. *arXiv preprint arXiv:1505.03229*. doi: 10.48550/arXiv.1505.03229

Scheck, T., Seidel, R., and Hirtz, G. (2020). "Learning from theodore: a synthetic omnidirectional top-view indoor dataset for deep transfer learning," in *The IEEE Winter Conference on Applications of Computer Vision* (Snowmass, CO), 943–952. doi: 10.1109/WACV45572.2020.9093563

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice), 618–626. doi: 10.1109/ICCV.2017.74

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 60. doi: 10.1186/s40537-019-0197-0

Soomro, K., Zamir, A. R., and Shah, M. (2012). *A Dataset of 101 Human Action Classes from Videos in the Wild*. Center for Research in Computer Vision.

Sun, B., and Saenko, K. (2016). "Deep coral: correlation alignment for deep domain adaptation," in *European Conference on Computer Vision* (Cham: Springer), 443–450. doi: 10.1007/978-3-319-49409-8_35

Tolins, J., and Samermit, P. (2016). GIFs as embodied enactments in text-mediated conversation. *Res. Lang. Soc. Interact*. 49, 75–91. doi: 10.1080/08351813.2016.1164391

Torralba, A., and Efros, A. A. (2011). "Unbiased look at dataset bias," in *CVPR 2011* (Colorado Springs, CO: IEEE), 1521–1528. doi: 10.1109/CVPR.2011.5995347

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 7167–7176. doi: 10.1109/CVPR.2017.316

Volpi, R., and Murino, V. (2019). "Addressing model vulnerability to distributional shifts over image transformation sets," in *Proceedings of the IEEE International Conference on Computer Vision* (Seoul), 7980–7989. doi: 10.1109/ICCV.2019.00807

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. (2018). "Generalizing to unseen domains via adversarial data augmentation," in *Advances in Neural Information Processing Systems, Vol. 31*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montreal, QC: Curran Associates Inc), 5334–5344.

Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., and Fergus, R. (2013). "Regularization of neural networks using dropconnect," in *International Conference on Machine Learning* (Atlanta, GA), 1058–1066.

Wang, Y., Cao, L., and Hellovera, A. Gif super-resolution.

Wilson, G., and Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol*. 11, 1–46. doi: 10.1145/3400066

Xu, R., Chen, Z., Zuo, W., Yan, J., and Lin, L. (2018). "Deep cocktail network: multi-source unsupervised domain adaptation with category shift," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 3964–3973. doi: 10.1109/CVPR.2018.00417

Yang, Y., and Hospedales, T. M. (2014). A unified perspective on multi-domain and multi-task learning. *arXiv preprint arXiv:1412.7489*. doi: 10.48550/arXiv.1412.7489

Yao, Z., Wang, Y., Wang, J., Yu, P., and Long, M. (2021). VideoDG: generalizing temporal relations in videos to novel domains. *IEEE Trans. Pattern Anal. Mach. Intell*. doi: 10.1109/TPAMI.2021.3116945

Yao, Z., Wang, Y., Du, X., Long, M., and Wang, J. (2019). Adversarial pyramid network for video domain generalization. *arXiv preprint arXiv:1912.03716*. doi: 10.48550/arXiv.1912.03716

Yu, F., Xian, W., Chen, Y., Liu, F., Liao, M., Madhavan, V., et al. (2018). BDD100K: a diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687*. doi: 10.48550/arXiv.1805.04687

Zajac, M., Zołna, k., Rostamzadeh, N., and Pinheiro, P. O. (2019). "Adversarial framing for image and video classification," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Honolulu, HI), 10077–10078. doi: 10.1609/aaai.v33i01.330110077

Zunino, A., Bargal, S. A., Volpi, R., Sameki, M., Zhang, J., Sclaroff, S., et al. (2020). Explainable deep classification models for domain generalization. *arXiv preprint arXiv:2003.06498*. doi: 10.1109/CVPRW53098.2021.00361