Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

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 | Sarah Adel Bargal

Department of Computer Science, Boston University, Massachussetts, USA

Correspondence
Sarah Adel Bargal, 111 Cummington Mall, MA USA 02215.
Email: sbargal@bu.edu

Summary

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 the 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 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:
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 [1,2,3] or multiple [4,5,6,7,8] 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 [9,10,11,12,13] is of most interest, aiming to leverage the few or no labeled samples. A more complex problem is Deep Domain Generalization [14,15,16,17], 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, as well-defined, rich in features, labeled datasets, allow for a universal evaluation of the different methods[18,19,20,21,22]. Given the arduous 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' data[23,24,25,26,27].

Our focus is on videos, and, more specifically, on Animated GIFs[28], 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[29]. GIFs are created by sampling frames from a video and are extensively used nowadays on the internet, especially in social networks and online communication[30,31]. Animated GIFs are synthetically generated and tend to posses exaggeration or anticipation of action motion. In this work, we aim to answer the following questions: *How large is the domain gap between (1) videos and GIFs, (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. 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[32].

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 in[33] 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.

We also extend an explainability-based domain generalization technique initially proposed for images[34] 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 decision. We use the extended saliency-based explainability approach to identify regions in the image that contribute most to the models 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 a spatiotemporal dataset, a training and testing setting, a spatiotemporal baseline, an augmentation-based spatiotemporal training strategy, and 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 dataseta, state-of-the-art methods for domain generalization, domain adaptation, data augmentation (Section 2), and explainability. We then describe our dataset and the processes of collection and annotation (Section 3). We analyze the selected baseline 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

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 in[36], 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,[37], a variety of alignment and learning techniques are being proposed or extended to minimize domain discrepancy in videos along the spatial and temporal directions: TemPooling, TemPooling with Adversarial Discriminator, TemRelation, TA2N and TA3N. In[38], the authors propose a generative

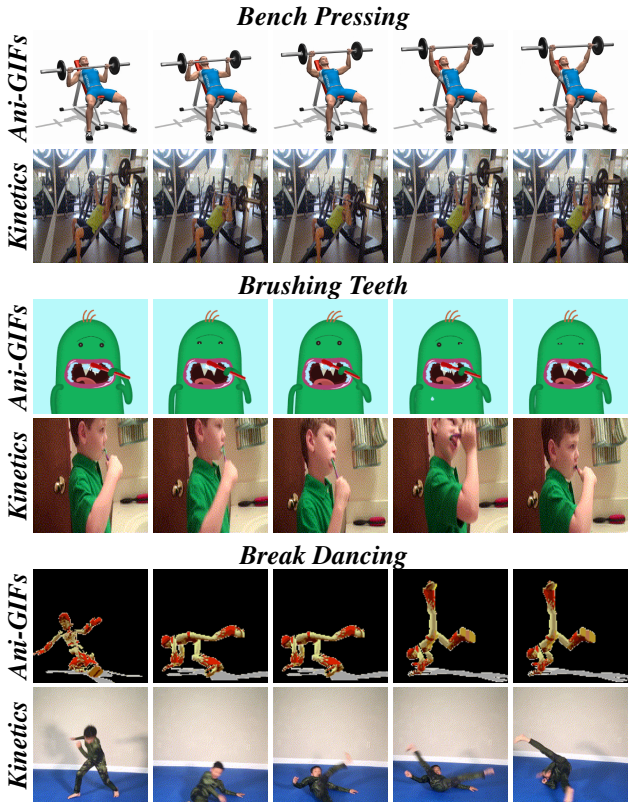

FIGURE 1: This figure highlights the spatiotemporal domain gap between *Ani-GIFs*, our proposed benchmark dataset, and GIFs of the real domain - from the Kinetics dataset[35] - for three classes: *Bench Pressing*, *Brushing Teeth* and *Break Dancing*. This illustrates the domain gap between real *vs.* animated frames.

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.

Video 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. Several techniques have been introduced to solve this problem with deep models[14,16,39,40], and with important results for a variety of datasets and data types, but the area is significantly underexplored with respect to video datasets, due to the complexity of entangling spatial and temporal domain shifts. In[41], 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.

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[37] is a collection of videos of length 1-10 seconds in 91 categories from two video games. Selecting 30 overlapping categories between Gameplay and Kinetics[35,42], 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[43] and HMDB51[44] are combined in UCF-HMDB$_{\text{full}}$, a large-scale collection of videos of length 1-33 seconds in 12 classes, used in evaluating several state-of-the-art video Domain Adaptation methods[45,46,47,48]. For Domain Generalization, in[41] the authors 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[49] and medical diagnosis[50] have been introduced, allowing for domain adaptation evaluation in a variety of sub-domains.

GIF Datasets and Analysis Techniques. There is an abundance of GIF datasets collected and available in the literature. TGIF[51] 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 generation of natural language descriptions for animated GIFs or video clips. In Vid2GIF[52], 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 the video highlight detection. GIF Super-Resolution[53] 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 1000 GIFs in 5 categories: Emotion, Action, Scene, Animation and Animal. In GIFGIF+[54], 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[55], 4,000 GIFs are collected, with scores for

17 discrete emotions, and are used in a computational analysis and evaluation of emotions prediction on animated GIFs. However all these datasets were designed to be used for tasks other than Domain Adaptation or Generalization.

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[56]. 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[57,58,59], to feature space augmentation[60], adversarial training techniques[61,62,63], and GAN-based approaches[64]. 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[65] and video action recognition[41].

Explainability. Explainability techniques were initially developed as a diagnostic tool to visualize and explain a model's behavior. GradCAM[66] 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.* [34] 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[66] 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 source domain of actions performed by human characters, **Real**, to a target domain of actions performed by animated/cartoon/graphical characters, **Synthetic**. As the Real domain dataset, we are using the GIFs from the existing Kinetics dataset[35], and we collect the GIFs in the Synthetic Domain, forming the proposed dataset, *Ani-GIFs*.

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. Three search terms 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'. We then collected GIFs from each separately. Each of the three collection processes, for all 600 classes, took approximately 100 hours to complete.

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. 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. Figure 2 provides examples of animated GIFs collected which were rejected or accepted during the filtering process.

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 17,095 animated GIFs in total, all intersecting with Kinetics-600. Figure 4 shows the number of GIF samples per class in the *Ani-GIFs* dataset for the top-frequency 100 classes.

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 from an animated GIF to Real.

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

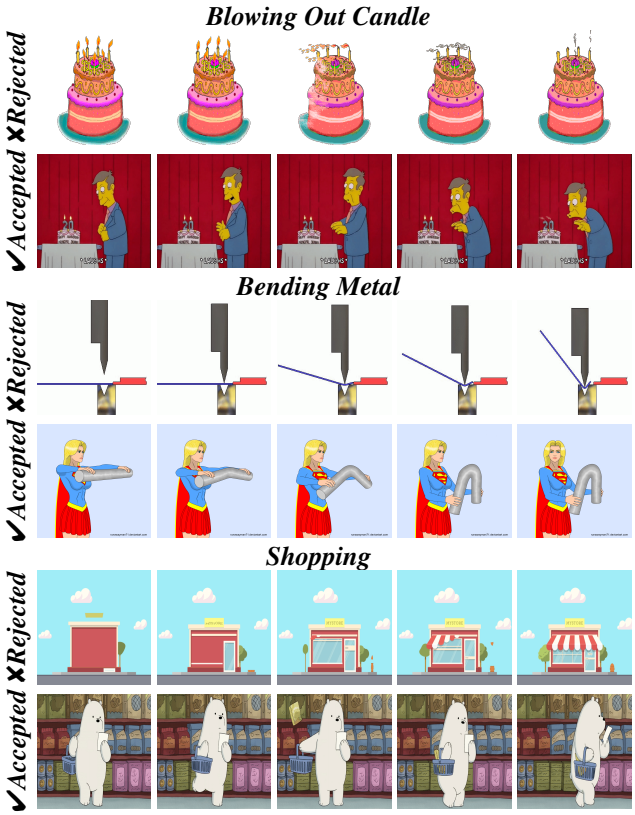

FIGURE 2: In this figure we can see that first GIFs, in *Blowing out candles* and *Bending metal* classes, were rejected as the actions are not performed by any character. We also rejected GIFs in the *Shopping* class, as the action was not relevant to the class (*i.e.* no shopping action is observed).

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[35] - as the source domain in our 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 are then compared to applying Domain Generalization techniques.

We also use the AVA-Kinetics Localized Human Actions Video Dataset[67] to extend the explainable training strategy of Zunino *et al.* [34] 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.

Data Augmentation Approach. We extend the work of Volpi *et al.* on images[33] and develop a spatiotemporal data augmentation approach for animated GIFs. Data Augmentation is a very powerful 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 *et al.* [33], 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), Bchannel enhancer (30), solarize (20).

Starting with a model pre-trained on the Kinetics-400 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 3, 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.

Explainability Approach. We extend and apply a saliency-based spatiotemporal explainability approach[34] 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[66] 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 co-related 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 withing the spatiotemporal region of interest and 0 otherwise. We run the saliency estimation periodically every 200 batches, and train using the Adam optimizer.

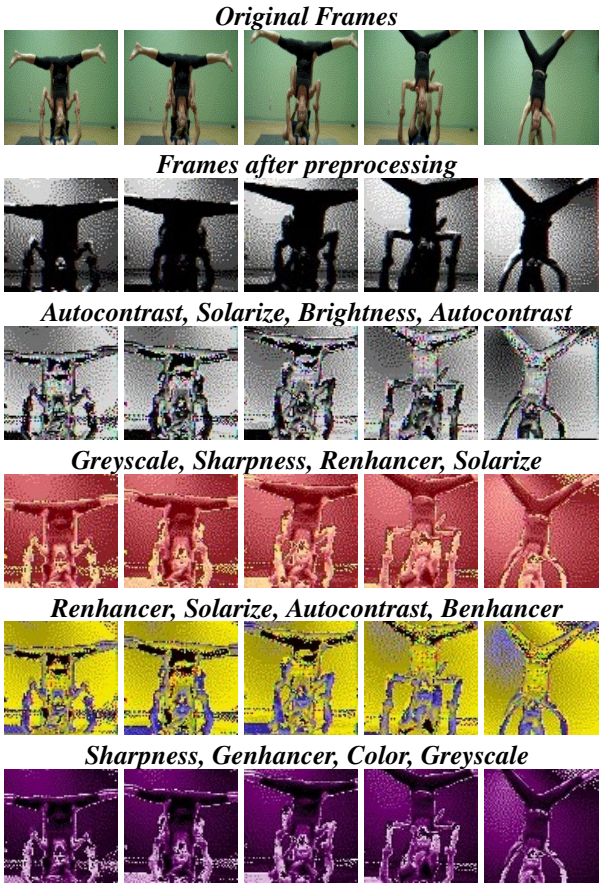

FIGURE 3: This figure presents frames sampled equally from a video in class 'Yoga'. The first set of frames belong to the Kinetics GIFs dataset, and are followed by the frames after Batch Normalization is applied. The subsequent set of images depicts the frames after tuples of transformations, chosen by the Random Search approach, are applied to them.

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.

Experimental Setup. We use the I3D model architecture[32] as the first baseline for the spatiotemporal training and testing of our videos and animated GIFs. 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 pre-processed 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, using Adam optimizer[68] with the following hyper parameters: learning rate = 10^{-4}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. We start with an I3D model that is pre-trained on Kinetics videos[69].

While extending the data augmentation approach to animated GIFs, we started with the model trained on the Kinetics GIFs using the I3D model architecture. The model was further fine-tuned with the Random Search approach and Adam optimizer[68]. 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, using the following hyper parameters: learning rate = 10^{-4}, $\beta_1 = 0.9$,

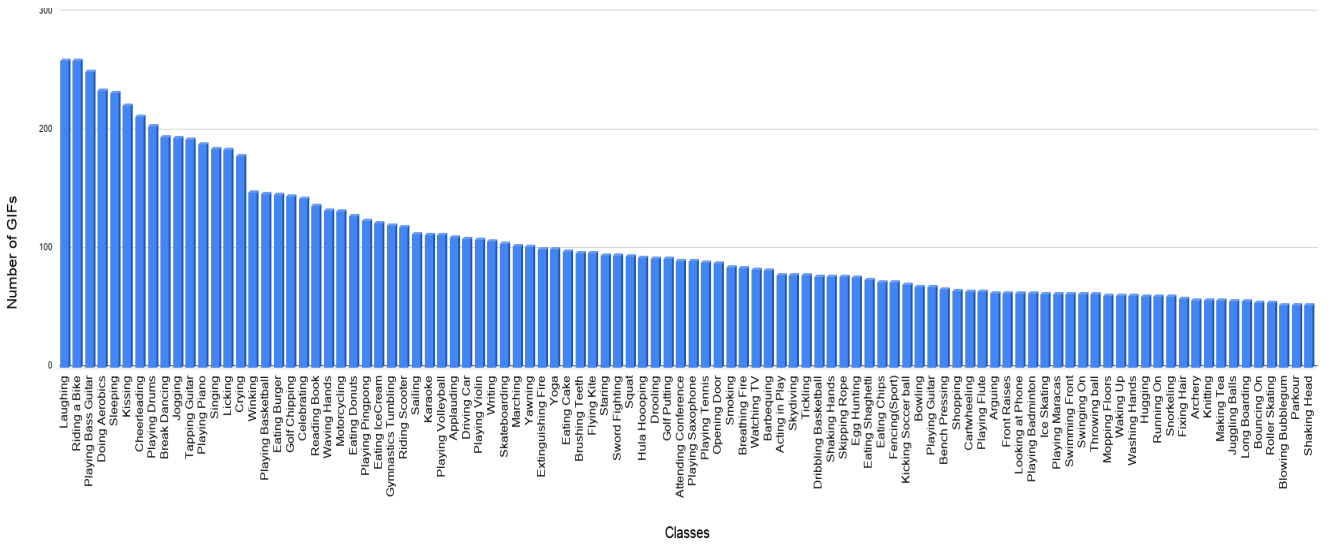

FIGURE 4: The 100 most-frequent classes after filtration of the *Ani-GIFs* dataset. Classes that have the highest frequency are those that belong to actions with a large number of associated GIFs, *e.g.* common actions and emotions. The forty classes of highest frequency are identified as a subset for GIFs domain adaptation.

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

TABLE 1: Our Experimental Results. Top-1 and top-5 test accuracies of our baseline algorithm are given, from various training on different testing domains. The difference in the reported accuracies between rows one and two demonstrates the existing domain gap from Kinetics to Kinetics GIFs, and in rows three and four 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 four to five shows the gain in accuracy yielded by extending and applying the spatiotemporal data augmentation algorithm for Domain Generalization on the training dataset, Kinetics GIFs.

and $\beta_2 = 0.999$. 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).

Experimental Results. The results of our experiments are presented in Table 1. 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[35], 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[70], similarly in the Real domain.

We mark the significant accuracy drop, to a 21.12% top-1 accuracy, which we can attribute to the frames' sampling process in GIFs, or the difference in GIFs 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[70] 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 1, 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*.

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, classification accuracy 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 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.

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, though small enough to highlight the need for better methods tailored towards 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.

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). 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 adaptation accurately highlights the correct action-specific cues.

References

1. Pan SJ, Tsang IW, Kwok JT, Yang Q. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 2010; 22(2): 199–210.

2. Baktashmotlagh M, Harandi M, Salzmann M. Distribution-matching embedding for visual domain adaptation. *The Journal of Machine Learning Research* 2016; 17(1): 3760–3789.

3. Long M, Zhu H, Wang J, Jordan MI. Unsupervised domain adaptation with residual transfer networks. In: ; 2016: 136–144.

4. Xu R, Chen Z, Zuo W, Yan J, Lin L. Deep cocktail network: Multi-source unsupervised domain adaptation with category shift. In: ; 2018: 3964–3973.

5. Yang Y, Hospedales TM. A unified perspective on multi-domain and multi-task learning. *arXiv preprint arXiv:1412.7489* 2014.

6. Duan L, Xu D, Tsang IWH. Domain adaptation from multiple sources: A domain-dependent regularization approach. *IEEE Transactions on neural networks and learning systems* 2012; 23(3): 504–518.

7. Jhuo IH, Liu D, Lee D, Chang SF. Robust visual domain adaptation with low-rank reconstruction. In: IEEE. ; 2012: 2168–2175.

8. Liu H, Shao M, Fu Y. Structure-preserved multi-source domain adaptation. In: IEEE. ; 2016: 1059–1064.

9. Wilson G, Cook DJ. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)* 2020; 11(5): 1–46.

10. Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research* 2016; 17(1): 2096–2030.

11. Long M, Cao Y, Wang J, Jordan M. Learning transferable features with deep adaptation networks. In: ; 2015: 97–105.

12. Long M, Cao Z, Wang J, Jordan MI. Conditional adversarial domain adaptation. In: ; 2018: 1640–1650.

13. Sun B, Saenko K. Deep coral: Correlation alignment for deep domain adaptation. In: Springer. ; 2016: 443–450.

14. Muandet K, Balduzzi D, Schölkopf B. Domain generalization via invariant feature representation. In: ; 2013: 10–18.

15. Ghifary M, Bastiaan Kleijn W, Zhang M, Balduzzi D. Domain generalization for object recognition with multi-task autoencoders. In: ; 2015: 2551–2559.

16. Li D, Yang Y, Song YZ, Hospedales TM. Deeper, broader and artier domain generalization. In: ; 2017: 5542–5550.

17. Li Y, Tian X, Gong M, et al. Deep domain generalization via conditional invariant adversarial networks. In: ; 2018: 624–639.

18. Torralba A, Efros AA. Unbiased look at dataset bias. In: IEEE. ; 2011: 1521–1528.

19. Beery S, Van Horn G, Perona P. Recognition in terra incognita. In: ; 2018: 456–473.

20. Recht B, Roelofs R, Schmidt L, Shankar V. Do imagenet classifiers generalize to imagenet?. *arXiv preprint arXiv:1902.10811* 2019.

21. Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision* 2015; 115(3): 211–252.

22. Ponce J, Berg TL, Everingham M, et al. Dataset issues in object recognition. In: Springer. 2006 (pp. 29–48).

23. Scheck T, Seidel R, Hirtz G. Learning from THEODORE: A Synthetic Omnidirectional Top-View Indoor Dataset for Deep Transfer Learning. In: ; 2020: 943–952.

24. Cruz SDD, Wasenmuller O, Beise HP, Stifter T, Stricker D. Sviro: Synthetic vehicle interior rear seat occupancy dataset and benchmark. In: ; 2020: 973–982.

25. Kong F, Huang B, Bradbury K, Malof J. The Synthinel-1 dataset: a collection of high resolution synthetic overhead imagery for building segmentation. In: ; 2020: 1814–1823.

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

27. Ros G, Sellart L, Materzynska J, Vazquez D, Lopez AM. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: ; 2016: 3234–3243.

28. Eppink J. A brief history of the GIF (so far). *Journal of Visual Culture* 2014; 13(3): 298–306.

29. Miltner KM, Highfield T. Never gonna GIF you up: Analyzing the cultural significance of the animated GIF. *Social Media+ Society* 2017; 3(3): 2056305117725223.

30. Jiang JA, Fiesler C, Brubaker JR. 'The Perfect One' Understanding Communication Practices and Challenges with Animated GIFs. *Proceedings of the ACM on human-computer interaction* 2018; 2(CSCW): 1–20.

31. Tolins J, Samermit P. GIFs as embodied enactments in text-mediated conversation. *Research on Language and Social Interaction* 2016; 49(2): 75–91.

32. Joao Carreira AZ. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *arXiv preprint arXiv:1705.07750* 2018.

33. Volpi R, Murino V. Addressing model vulnerability to distributional shifts over image transformation sets. In: ; 2019: 7980–7989.

34. Zunino A, Bargal SA, Volpi R, et al. Explainable deep classification models for domain generalization. *arXiv preprint arXiv:2003.06498* 2020.

35. Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* 2017.

36. Jamal A, Namboodiri VP, Deodhare D, Venkatesh K. Deep Domain Adaptation in Action Space.. In: ; 2018: 264.

37. Chen MH, Kira Z, AlRegib G. Temporal Attentive Alignment for Video Domain Adaptation. *arXiv preprint arXiv:1905.10861* 2019.

38. Chen J, Li Y, Ma K, Zheng Y. Generative Adversarial Networks for Video-to-Video Domain Adaptation. *arXiv preprint arXiv:2004.08058* 2020.

39. Motiian S, Piccirilli M, Adjeroh DA, Doretto G. Unified deep supervised domain adaptation and generalization. In: ; 2017: 5715–5725.

40. Li D, Yang Y, Song YZ, Hospedales TM. Learning to generalize: Meta-learning for domain generalization. In: ; 2018.

41. Yao Z, Wang Y, Du X, Long M, Wang J. Adversarial Pyramid Network for Video Domain Generalization. *arXiv preprint arXiv:1912.03716* 2019.

42. Joao Carreira ABHCH. A Short Note about Kinetics-600. *arXiv preprint arXiv:1808.01340* 2018.

43. Soomro K, Zamir AR, Shah M. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision* 2012; 2.

44. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB: a large video database for human motion recognition. In: IEEE. ; 2011: 2556–2563.

45. Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495* 2014.

46. Long M, Zhu H, Wang J, Jordan MI. Deep transfer learning with joint adaptation networks. In: JMLR. org. ; 2017: 2208–2217.

47. Li Y, Wang N, Shi J, Hou X, Liu J. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition* 2018; 80: 109–117.

48. Saito K, Watanabe K, Ushiku Y, Harada T. Maximum classifier discrepancy for unsupervised domain adaptation. In: ; 2018: 3723–3732.

49. Yu F, Xian W, Chen Y, et al. Bdd100k: A diverse driving video database with scalable annotation tooling. *arXiv preprint arXiv:1805.04687* 2018.

50. Cheplygina V, Pena IP, Pedersen JH, Lynch DA, Sørensen L, Bruijne dM. Transfer learning for multicenter classification of chronic obstructive pulmonary disease. *IEEE journal of biomedical and health informatics* 2017; 22(5): 1486–1496.

51. Li Y, Song Y, Cao L, et al. TGIF: A new dataset and benchmark on animated GIF description. In: ; 2016: 4641–4650.

52. Gygli M, Song Y, Cao L. Video2gif: Automatic generation of animated gifs from video. In: ; 2016: 1001–1009.

53. Wang Y, Cao L, Hellovera A. GIF Super-Resolution.

54. Chen W, Rudovic OO, Picard RW. Gifgif+: Collecting emotional animated gifs with clustered multi-task learning. In: IEEE. ; 2017: 510–517.

55. Jou B, Bhattacharya S, Chang SF. Predicting viewer perceived emotions in animated GIFs. In: ; 2014: 213–216.

56. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *Journal of Big Data* 2019; 6(1): 60.

57. Sato I, Nishimura H, Yokoi K. Apac: Augmented pattern classification with neural networks. *arXiv preprint arXiv:1505.03229* 2015.

58. Wan L, Zeiler M, Zhang S, Le Cun Y, Fergus R. Regularization of neural networks using dropconnect. In: ; 2013: 1058–1066.

59. Ciregan D, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. In: IEEE. ; 2012: 3642–3649.

60. DeVries T, Taylor GW. Dataset Augmentation in Feature Space. 2017.

61. Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2016. doi: 10.1109/cvpr.2016.282

62. Zajac M, Zołna K, Rostamzadeh N, Pinheiro PO. Adversarial Framing for Image and Video Classification. *Proceedings of the AAAI Conference on Artificial Intelligence* 2019; 33: 10077–10078. doi: 10.1609/aaai.v33i01.330110077

63. Volpi R, Namkoong H, Sener O, Duchi JC, Murino V, Savarese S. Generalizing to unseen domains via adversarial data augmentation. In: ; 2018: 5334–5344.

64. Bowles C, Chen L, Guerrero R, et al. GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks. 2018.

65. Budvytis I, Sauer P, Roddick T, Breen K, Cipolla R. Large scale labelled video data augmentation for semantic segmentation in driving scenarios. In: ; 2017: 230–237.

66. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ; 2017: 618–626.

67. Li A, Thotakuri M, Ross DA, Carreira J, Vostrikov A, Zisserman A. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214* 2020.

68. Diederik P. Kingma JB. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* 2014.

69. Piergiovanni A. 3D models trained on Kinetics. https://github.com/piergiaj/pytorch-i3d; 2018. [Online; accessed 28-Jun-2018].

70. Gituma M. The Kinetics Dataset Explorer Using GIFs. https://towardsdatascience.com/the-kinetics-dataset-explorer-using-gifs-8ceeebcbdaba; 2019. [Online; accessed 24-Feb-2019].