Using Temporal Information for Recognizing Actions from Still Images

Samitha Herath[1,4,5], Basura Fernando[3], Mehrtash Harandi[2,4]

Abstract

In this paper we raise two important question,"**1.** Is temporal information beneficial in recognizing actions from still images? **2.** Do we know how to take the maximum advantage from them?". To answer these question we propose a novel transfer learning problem, Temporal To Still Image Learning (*i.e.*, T2SIL) where we learn to derive temporal information from still images. Thereafter, we use a two-stream model where still image action predictions are fused with derived temporal predictions. In T2SIL, the knowledge transferring occurs from temporal representations of videos (*e.g.*, Optical-flow, Dynamic Image representations) to still action images. Along with the T2SIL we propose a new action still image action dataset and a video dataset sharing the same set of classes. We explore three well established transfer learning frameworks (*i.e.*, GANs, Embedding learning and Teacher Student Networks (TSNs)) in place of the temporal knowledge transfer method. The use of derived temporal information from our TSN and Embedding learning improves still image action recognition.

Keywords: still image action recognition, two-stream, optical-flow, dynamic-images

1. Introduction

Action classification from videos [1, 2] is a well-established, yet extremely challenging problem in computer vision. A related, less explored and arguably more difficult problem is to predict human actions from still images with applications ranging from sports analysis to situation recognition [3], autonomous driving [4], and event recognition [5]. The list of related subproblems includes predicting human object interactions [6], pose recognition [7], still image motion estimation [8] and Gait recognition [9] to name a few.

Human actions are spatial-temporal events and biological evidences suggest that mammalian brain has a dedicated section for understanding time evolution of object positions [10]. In computer vision, a similar observation can be made. In particular, deep models that explicitly incorporate temporal information towards their decisions form state-of-the-art solutions in video action recognition [2].

As humans, we naturally have the ability to deduce motion information from still images. For example, we can explain the motion of the three basket-ball actions shown in Fig. 1. The motion cues seem to provide us with extra information, leading to more accurate labeling of actions. Nonetheless, this knowledge about "temporal" information do not come free with still images. One conjecture here is that humans transfer the knowledge about temporal information by relating still images to prior experience (*e.g.*, a basketball game watched on TV).

Based on the above discussions, we postulate that still image action recognition can also benefit from temporal information. In shaping the postulate, the following observations came handy;

1. **Complementary property of temporal information.** The idea behind two-stream video action recognition [2] is to improve video action recognition by fusing predictions on appearance and temporal representations of videos. To give the reader a figure, the performance of the individual video frame model and the optical flow model is reported as 73.0% and 83.7%, respectively on the UCF-101 [11]. By a mere averaging of the predictions of the two models, an accuracy of 86.9% is attained. We expect a model exploiting temporal information can benefit from such complementary information to improve still image action predictions.
2. **Reduce pose ambiguity.** Human actions may be represented as a temporal collection of human poses. An still image action appearing with any such pose may cause ambiguities at recognition time. Temporal representations compress information of the evolution of the pose sequences. For example, a pose history pattern [12] can be observed with dynamic images (see Fig. 5). By properly employing temporal information, one can expect a lesser level of ambiguity from pose variations.
3. **Potential to recognize objects of interest.** Motion helps to emphasize visual elements connected with the action. For instance, in Fig. 5, the motion boundaries of the objects in motion are emphasized by the dynamic images. Similarly, the Optical-flow field images can capture distinct patterns near moving objects. The idea is to recognize such elements in an still action image by means of knowledge transferring (*e.g.*, by learning feature correlations). In principle this avoids the need for additional

[1]The Australian National University, Canberra, Australia.
[2]Monash University, Melbourne, Australia.
[3]Human-Centric AI Programme, Artificial Intelligence Initiative, A*STAR, Singapore.
[4]Data61/CSIRO, Australia.
[5]Corresponding author e-mail : samitha.herath@data61.csiro.au

Figure 1: Three basketball actions with similar appearance information. We humans however can easily deduce the corresponding motion information.

human/body-part annotated data to reject noisy background information as done in the state of the art deep solutions [13].

To verify our postulate, *i.e.*, if temporal information can help in recognizing actions from still images, we propose a new transfer learning problem where temporal information (*e.g.*, deep features of motion representations, class predictions of motion representations) is transferred to still action images. We call this transfer learning problem Temporal To Still Image Learning or T2SIL. For T2SIL, we introduce a new still image action recognition dataset. This is because, the existing still image action datasets and their protocols [14, 3, 15] do not facilitate our purpose empirically. Our new dataset constitutes of two parts, the still image actions and the corresponding video actions, providing us with the required temporal information to formulate the problem as a knowledge transfer one. Furthermore, we propose three established end-to-end learning paradigms to verify our postulate empirically. This includes **1.** Adversarial learning **2.** transfer learning by deep embedding, and **3.** transfer learning by Teacher Student Network(TSN) distillation. Our study suggests;

1- Actions speak louder than Images.. Out of the proposed three solutions, we observe improvements can be obtained with the deep embedding and TSN solutions. As surprising as it may sound, the adversarial paradigm is not able to transfer temporal information properly for the task in hand, demonstrating the difficulty of the problem considered in this paper (*i.e.*, T2SIL). For these reasons we answer our first question in the abstract as **"Probably Yes"**.

2- Well begun is half done.. As we will show later, the effect of employing temporal information in T2SIL is far less if action recognition from videos (*e.g.*, two-stream paradigm [2]) is taken into account. Since the transfer learning solutions used to solve T2SIL show state of the art performances in related problems (*e.g.*, Zero-Shot Learning [16], Domain adaptation [17], Model distillation [18]), we believe that our paper introduces a new, extremely challenging problem in computer vision. We think that the proposed problem demands attention of a wider community. Hence, we answer with **"Possibly No"** to our second question in the abstract.

We emphasize that Solving T2SIL is beneficial to some other problems. One prime example is motion prediction from still images (*e.g.*, Optical-flow prediction of Walker *et al.* [19], Newtonian motions study of Mottaghi *et al.* [8], and frame prediction of Xue *et al.* [20]). To conclude the introduction, we list our contributions below

1. We propose to use temporal information to improve still image action recognition.
2. We formulate this problem as a novel transfer learning problem.
3. We propose a new still image action dataset with a corresponding video dataset to evaluate T2SIL.
4. We propose three transfer learning solutions and show while adversarial feature generation is not helpful for T2SIL, improvements can be attained with deep embedding learning and TSN frameworks.

2. Related work

In this section, we first discuss state of the art still image action recognition solutions. We also highlight the differences of T2SIL from early-action and motion prediction problems. Then we briefly review various forms of knowledge transfer from related problems.

Still image action recognition

State of the art solutions for still image action recognition use deep convolutional networks in conjunction with object/human detectors to filter out noisy information [21, 22, 23, 13]. The part based models of Sharma *et al.* [21], Rosenfeld *et al.* [22] and Zhao *et al.* [23] use object parts as visual attributes to describe the actions in a still image. Gkioxari *et al.* [13], use contextual information from surrounding regions of the detected human and show that the contextual information are complementary to the action predictions (and hence beneficial).All aforementioned methods exploit different forms of appearance cues from the images, whereas, we propose a different path that exploits temporal knowledge to improve still image action recognition. Therefore, in principle the above methods can benefit from our postulate as well.

A somehow relevant study is the work of Chen *et al.* [24] where video frames are used to fill missing poses in still image training data (*i.e.*, dataset augmentation). Although this solution uses information from videos as in our case, it disregards valuable temporal information and targets a different problem.

Early action recognition and prediction

Early action recognition and prediction [25, 26] from videos are too temporally constrained problems similar to ours. The task in early action recognition is to recognize an action by observing few frames as possible. Action prediction is where the first few frames are used to predict the action class and the occurrence of a future peak-pose [25]. However, the proposed T2SIL is distinct to these problem from **1.** still action images do not have corresponding temporal representations at train time, and **2.** at inference time we are only available with a single action image (*i.e.*, no temporal information) instead of few video frames.

Motion prediction

Motion prediction studies ways of producing motion representations given image inputs. Predicted Optical-flow of Walker *et al.* [19] and anticipated Dynamic-images of Rodriguez *et al.* [27] are such work predicting motion representations. However, our work is distinct to them as we study using temporal information to improve the recognition of still image actions. In other words, we study the transfer of discriminative temporal information useful for recognition.

Knowledge Transfer

As will be shown shortly, the problem of interest in our paper can be formulated as a knowledge transfer problem. As such, it is natural to make use of methods targeting knowledge transfer to address T2SIL. We identify three dominating school of thoughts when knowledge transfer solutions are considered.

1. Adversarial Learning.

Feature matching by adversarial learning is the core idea behind various state of the art Domain adaptation(DA) solutions [28]. Furthermore, feature Generative Adversarial Networks(GANs) has shown remarkable success in Zero-Shot Learning(ZSL) [16]. The motive behind using feature GANs for T2SIL is to train a model that could generate the corresponding temporal features for a given still image instance. The temporal features are thereafter used to obtain class predictions, mimicking the temporal stream in two-stream networks [2].

2. Deep Embedding.

Embedding learning is proven to be a method of choice when one needs to transfer properties of one space to another(*e.g.*, ZSL [29], DA [17], Cross modal similarity learning [30]). In such solutions, distances between the similar class embeddings are reduced while dissimilar class embeddings are repulsed. Current trend achieves this by minimizing a loss defined on triplets [31] and will be our choice in this paper as well. This is because triplets allow local regions of the embedding space to be adjusted with better flexibility.

3. Teacher-Student Paradigm.

Feature generation and embedding learning are two solutions where the correspondences between the data are deemed at the feature level. A distinct approach to this is to transfer the classifier's perception across data forms. Teacher Student Networks(TSNs) is a framework that can be used for this purpose. TSNs were originally proposed for distilling the information from complex models to simpler ones [32]. However, TSNs have been successfully used in a variety of applications (*e.g.*,data security [33]) and more importantly, a closely related problem, cross modal supervision transferring [18].

Cross Modal Supervision Transfer(CMST) [18] is the study of transferring supervised knowledge across different image modalities (*e.g.*, thermal images, optical-flow, depth images, LIDAR point clouds). The proposed T2SIL can be considered as a difficult end of CMST for two reasons. **1.** still action images do not contain paired temporal representations to be exploited (unlike for example RGB-depth, video frame-optical flow). The only association available is the class label, which is relatively weaker. **2.** for a given still image action, there can be many correct temporal realizations. This is not the case with other modality pairs. As such, we can expect a successful solution for T2SIL might have usage for CMST problems as well.

3. Formulating T2SIL

We aim to improve still image action recognition by transferring relevant temporal information of videos to still images. Per definition, in still image action recognition, labels for still image human actions are available. In addition to that, we propose to make use of abundantly available video level human action annotation. To this end, during the training process, we use labeled video data in conjunction with the still image human actions to develop more accurate and robust still-image action models. In doing so, we use two forms of temporal representations, namely optical-flow [2] and "dynamic images" [34]. Optical flow provides movement information of two consecutive frames while "dynamic images" summarise the motion information of several frames into a single image using "rank pooling" [35] principle. Note that human actions of still images are also stored as RGB images and our goal is to improve still image action recognition by transferring relevant temporal information of videos using optical-flow and dynamic images.

Our Assumptions.

The proposed T2SIL is formulated around the following two closely related assumptions.

Assumption 1 : Despite being different, the two joint distributions $P(X_s, Y)$ and $P(X_t, Y)$ carry a latent relationship. Here, X_s is a random variable representing still image action data[6] and X_t is random variable representing temporal representations of video action data. The class labels are given by the random variable Y.

[6]The still/video action data in concern could be of the form of deep features, images, softmax scores *etc.*.

Both the motion (*i.e.*, the temporal information) and the poses (*i.e.*, the appearance) of a given human action is constrained by the same generator bounds (*e.g.*, the restrictions in the limb movements, object interactions). As a result we could visually observe close similarities between state of the art temporal representations (*e.g.*, motion boundaries of optical-flow, dynamic-images as in Fig. 5) and still action images. We consider such similarities as indicators for the existence of a latent relationship between the temporal and the appearance data.

Assumption 2: Learning the conditional generation of temporal information from still data (*i.e.*, $P(X_t|X_s)$) of a given action class is feasible.

To support this second assumption we refer to the capability of human cognition to predict the motion when given a still action pose (*see* our discussion connected with Fig.1 in the introduction). Furthermore, given the existence of a relationship between appearance and temporal data in actions (*see* Assumption 1), this second assumption basically describes the existence of a transfer learning solution that can uncover such a complex relationship.

Our Notation.

We use bold capital letters to denote matrices (*e.g.*, $\boldsymbol{X}$) and bold lower-case letters to denote column vectors (*e.g.*, $\boldsymbol{x}$). The notation $\|\cdot\|_2$ is used to denote the L2 norm of a vector and $[\cdot]_+$ denotes the term $\max(0,\cdot)$. We use $f(\cdot;\boldsymbol{\theta}) : \mathbb{R}^m \rightarrow \mathbb{R}^p$ to denote a mapping function with parameters $\boldsymbol{\theta}$ from $\mathbb{R}^m$ to $\mathbb{R}^p$. Let the training samples from the still images and temporal representations be $\{\boldsymbol{x}_i^s, y_i^s\}_{i=1}^{n_s}$, $\boldsymbol{x}_i^s \in \mathcal{X}^s$ and $\{\boldsymbol{x}_j^t, y_j^t\}_{j=1}^{n_t}$, $\boldsymbol{x}_i^t \in \mathcal{X}^t$, respectively. We use the small-scripts "s" and "t" to denote still image and temporal domains. Here, n_s and n_t are the number of labeled training instances available from each domain. In our experimental setup we consider the case where $y_i^s,\ y_j^t \in C = \{1, 2, 3, \cdots, c\}$, *i.e.*, where both the still image and temporal data have matching classes.

3.1. T2SIL with Adversarial Learning

In this part, we discuss our GAN [36] solution for T2SIL (see Fig. 2 for a conceptual diagram). A GAN consists of a generator function, f_g and a discriminator, D that compete in a two player min-max optimization. In the context of our problem, f_g tries to generate temporal features while D attempts to discriminate the generated temporal features from the real temporal representations from videos. Similarly, our idea is to generate a temporal representation, $\boldsymbol{x}^t \in \mathcal{X}^t \subset \mathbb{R}^t$ for a given still image instance, $\boldsymbol{x}^s \in \mathcal{X}^s \subset \mathbb{R}^s$. In particular, we define the generator function, $f_g : \mathcal{X}^s \times \mathcal{Z} \rightarrow \mathcal{X}^t$ to take a still image input, $\boldsymbol{x}^s$ and a random Gaussian noise $z \in \mathcal{Z} \subset \mathbb{R}^z$ to formulate the generated temporal feature distribution. The discriminator model accepts the generated features and real temporal features as its inputs.

The parameters of the proposed feature generation GAN is learned by optimizing the objective,

$$\min_{\boldsymbol{\theta}_g} \max_{\boldsymbol{\theta}_d} L_{gan} = \mathbf{E}[\log D(\boldsymbol{x}^t)] \ + \ \mathbf{E}[\log(1 - D(\tilde{\boldsymbol{x}}^t))], \qquad (1)$$

with $\tilde{\boldsymbol{x}}^t = f_g(\boldsymbol{x}^s, z)$. Here, the parameters $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_d$ parameterizes the models f_g and D, respectively. As elaborated in Fig. 2 we realize the generator and the discriminator models with multi-layer perceptions. Nevertheless, optimization of the loss in equation (1) does not guarantee that the generated features are discriminative. Since our goal here is to use classifier predictions from the generated temporal features, we propose to make use of a multi-class classification loss along the model,

$$L_{aux} = -\mathbf{E}_{\tilde{\boldsymbol{x}}^t}[\log P(y|\tilde{\boldsymbol{x}}^t; \boldsymbol{\theta}_{aux})]. \qquad (2)$$

Here, y is the class label of the input still image to the generator. Hence, $P(y|\tilde{\boldsymbol{x}}^t; \boldsymbol{\theta}_{aux})$ denotes the probability of the generated temporal features being correctly classified by an auxiliary classifier, $f_{aux}(\cdot; \boldsymbol{\theta}_{aux}) : \mathbb{R}^t \rightarrow \mathbb{R}^{|C|}$. The discriminative loss over generated samples also helps reducing the mode collapsing of GANs [36]. The auxiliary classifier parameters, $\boldsymbol{\theta}_{aux}$ are in fact pre-trained with real temporal instances $\{\boldsymbol{x}_j^t, y_j^t\}_{j=1}^{n_t}$ and will be kept constant during training the GAN. The motive behind this is to avoid parameters $\boldsymbol{\theta}_{aux}$ being misdirected with poor quality generated features by the GAN (especially at the beginning of the training). With the described adversarial and the auxiliary classifier loss, our final training objective reads as,

$$\min_{\boldsymbol{\theta}_g} \max_{\boldsymbol{\theta}_d} L_{gan} + \lambda L_{aux}. \qquad (3)$$

Here, λ is a parameter to control the discriminative nature of the generated samples.

3.1.1. Inference protocol with T2SIL Adversarial Learning

At inference time our objective is to label a given still action image (*i.e.*, deep features, $\boldsymbol{x}^s \in \mathbb{R}^s$ from a given still action image). For this we incorporate the auxiliary classifier's (*i.e.*, $f_{aux}(\cdot, \boldsymbol{\theta}_{aux})$) softmax predictions for an input $f_g(\boldsymbol{x}^s)$ (*see* Fig. 2 for the block diagram). We shall call these predictions as the derived temporal predictions. Additionally, we use the softmax predictions from a deep model trained using only still action images. This still image deep model is equivalent to the spatial stream of two-stream networks [2]. Thereafter, we follow the two-streamed prediction fusion by averaging the derived temporal predictions and the still image model predictions.

3.2. T2SIL with Deep Embedding

The idea behind deep embedding learning is to pull similar class instances closer in the embedding space and push dissimilar class instances apart [31]. In knowledge transfer, deep embedding is used either **1.** to learn a shared latent space containing instances from all modalities [30] or **2.** to embed instances from one modality in the other [37]. The latter method is particularly used when one needs to preserve the properties (*e.g.*, inter class similarities) of a modality during the knowledge transferring.

In this paper we are interested in transferring the properties of temporal representations of actions to still action instances.

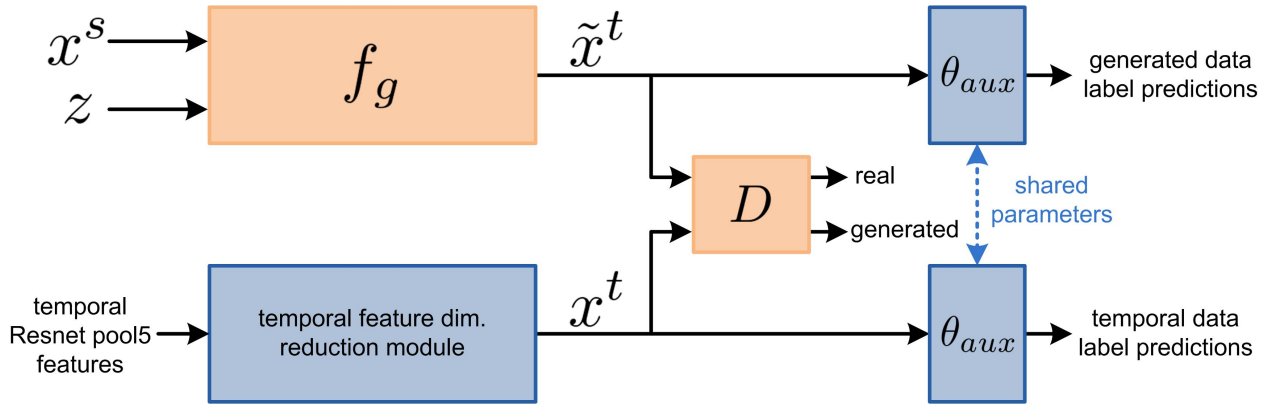

Figure 2: The schematic diagram of the proposed feature generation GAN. Here, we lean an temporal feature generator, f_g which takes still action image features as inputs. We train this generator in the adversarial framework.

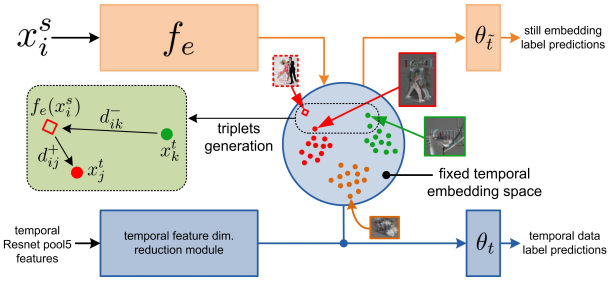

Figure 3: The proposed deep embedding solution with the triplet loss. Here, we learn an embedding function, f_e from the still image action inputs to the temporal embedding space. The temporal embedding space is kept fixed while training the embedding function, f_e.

Hence, we learn an embedding $f_e(\boldsymbol{x}_i^s)$ of a still image action instance, $\boldsymbol{x}_i^s \in \mathbb{R}^s$ to the fixed temporal representations space (see Fig. 3 for a conceptual diagram). We realize the embedding function $f_e : \mathbb{R}^s \rightarrow \mathbb{R}^t$ with a multi-layer perceptron. We define the distance between a still image embedding $f_e(\boldsymbol{x}_i^s)$ and a temporal embedding, $\boldsymbol{x}_j^t$ as $d_{ij} := \|f_e(\boldsymbol{x}_i^s) - \boldsymbol{x}_j^t\|_2$.

Our goal is to bring still image embeddings closer to the co-class temporal embeddings while pushing away from differently labeled temporal embeddings. Formally we achieve this by minimizing a triplets loss,

$$\sum_{i,j,k} L_{trip}(\boldsymbol{x}_i^s, \boldsymbol{x}_j^t, \boldsymbol{x}_k^t) = [d_{ij}^+ - d_{ik}^- + \alpha]_+. \tag{4}$$

Here, we use the superscript "+" and "-" to denote positive and negative pairs, respectively. That is, for a given $\boldsymbol{x}_i^s$, $d_{ij}^+ = \|f_e(\boldsymbol{x}_i^s) - \boldsymbol{x}_j^t\|_2$ where $y_i^s = y_j^t$. Similarly, $d_{ik}^- = \|f_e(\boldsymbol{x}_i^s) - \boldsymbol{x}_k^t\|_2$ where $y_i^s \neq y_k^t$. The parameter α is a tunable margin kept fixed during training. To keep the triplets loss computation feasible and to improve convergence we use the semi-hard negative mining heuristic [31] to form the triplets. To be specific, for a given $\boldsymbol{x}_i^s$ and $\boldsymbol{x}_j^t$, we sample the nearest dissimilar class instance to $f_e(\boldsymbol{x}_i^s)$ satisfying $d_{ik}^- > d_{ij}^+$ as $\boldsymbol{x}_k^t$.

3.2.1. Inference protocol with Deep Embeddings

At inference time we first obtain a temporally meaningful embedding for a given still action image (*i.e.*, deep features, $\boldsymbol{x}^s \in \mathbb{R}^s$ from a given still action image). For this we use the learned embedding projection model, $f_e(\cdot)$ to the fixed temporal space. The still image embedding in the temporal space is labeled using a classifier, $f_{\tilde{t}}(\cdot, \boldsymbol{\theta}_{\tilde{t}}) : \mathbb{R}^t \rightarrow \mathbb{R}^{\|C\|}$ (*see* Fig. 3 for the schematic). This temporal embedding classifier is trained discretely to our embedding model at training time. Thereafter, similarly to our adversarial solution we average the softmax scores of the temporal embedding classifier and the still image deep model. The still image deep model is trained only using still action images and acts similar to the spatial stream of two-stream networks.

3.3. T2SIL with Teacher-Student Networks

The Teacher-Student Network (TSN) is a transfer learning solution listed under deep model distillation. The foremost use of TSN distillation is to boil down a lager network (referred to as the Teacher model) into a network with less number of parameter (*i.e.*, the Student model) [32]. Our idea here is to use TSN distillation to train a student model, $f_{stu} : \mathbb{R}^s \rightarrow \mathbb{R}^{|C|}$, that takes a still image, $\boldsymbol{x}_i^s$ as the input but outputs the response of the teacher model, $f_{tch} : \mathbb{R}^t \rightarrow \mathbb{R}^{|C|}$. The teacher model f_{tch} is trained on temporal representations. By doing this, we transfer the perception of a temporal representation to a still image input.

For the sake of discussion, consider the outputs of the models f_{stu} and f_{tch} to be softmax scores. We use the parameters $\boldsymbol{\theta}_{stu}$ and $\boldsymbol{\theta}_{tch}$ to parameterize the student and teacher models, respectively (see Fig. 4 for a schematic.). As in a typical TSN, we pretrain the teacher model with labeled temporal representations, $\{\boldsymbol{x}_j^t, y_j^t\}_{j=1}^{n_t}$ and keep them fixed during the model distillation. The student model parameters, $\boldsymbol{\theta}_{stu}$ are learned by minimizing,

$$\min_{\boldsymbol{\theta}_{stu}} \sum_{\substack{i=0 \\ j:y_j^t=y_i^s}}^{n_s} H(f_{tch}(\boldsymbol{x}_j^t), f_{stu}(\boldsymbol{x}_i^s)) \tag{5}$$

Here, H is the cross-entropy loss given by,

$$H(\boldsymbol{y}_i, \boldsymbol{y}_j) = -\sum_{k=1}^{|C|} \boldsymbol{y}_i^{(k)} \log\ \boldsymbol{y}_j^{(k)}, \tag{6}$$

for $\boldsymbol{y}_i$, $\boldsymbol{y}_j \in \mathbb{R}^{|C|}$. Note that the $(\boldsymbol{x}_i^s, \boldsymbol{x}_j^t)$ pairs for minimization in equation (5) should be selected such that they represent the same action class.

3.3.1. Inference protocol for the TSN solution

At inference time we formulate a two-stream framework using the trained student network and a deep network trained using only still action images. For a given still action test image, $\boldsymbol{x}^s$ the student model's softmax prediction, $f_{stu}(\boldsymbol{x}^s)$ acts as the temporal prediction. Formally, we refer to this as the derived temporal prediction. We obtain the final softmax prediction by averaging this derived temporal prediction with the still image deep network's (*i.e.*, spatial stream) softmax prediction.

4. The Shared Still-Video Action Classes dataset

In this section, we discuss the details of the data we have gathered to verify our postulate. Our new still image action dataset consists of 12,347 still images representing 40 action classes. We coin this dataset *SSVAC-40-Still*. Along with this we also gather the corresponding video dataset called SSVAC-40-Videos with labeled video human actions. Both still and video collections have the same set of classes.

4.1. Composition of SSVAC-40 dataset

Each action class of SSVAC-40-Still dataset contains at least 190 still action images. We created SSVAC-40-Still dataset by collecting images from popular still image collections such as Stanford40 [14], MPII [15], imSitu [3], ImageNet [38] and from the web (see Fig. 6 for details). The dataset in particular contains human actions with face and hand (*e.g.*, brushing teeth, clap, wave), indoor ball games (*e.g.*, basket ball, basket ball dunk, dribble, volley ball spiking) , athletics (*e.g.*, high jump, long jump, pole vault), indoor gymnastics (*e.g.*, balance beam, parallel bars, pommel horse, uneven bars, trampoline jumping), boats and water (*e.g.*, kayaking, rowing, surfing).

SSVAC-40-Video data consists of videos from HMDB-51 [39] and UCF-101 [11]. Out of the total 40 classes, 9 classes are in common with the HMDB-51 dataset (*e.g.*, wave, walk, run, dribble) while the rest are from the UCF-101 dataset (*e.g.*, balance beam, brushing teeth, fencing, kayaking). We use optical-flow and dynamic images [34] as motion representations of the videos.

Dynamic image extraction: We use window size of 10 and a stride of 6 to create dynamic images. Following [34], we extract six dynamic images from the middle part of the video.

Optical-flow extraction: In all our experiments, we use the Optical-flow field images from [40]. To have a similar temporal extent as dynamic images, we use the mid 60 optical-flow frames of each video for training our models. Optical-flow inputs for networks compose of stacks of 10 consecutive optical flow frames.

4.2. Train test splits for SSVAC-40 dataset.

SSVAC-40-Still : We make use of the provided standard train and test splits from Stanford40 [14] and MPII [15] datasets. Instances from MPII dataset are only included in the training and validation sets as there are no publicly available test annotations. Images from imSitu and ImageNet are divided into train-val-test sets randomly. We manually check all images to minimize label inconsistencies and to remove multiple action occurrences in an image. Our dataset contains 6,821 training still action images, 728 validation images and 4,798 testing still action images (see Fig. 5 for samples).

SSVAC-40-Videos : We use the training set of UCF-101 (split1) and HMDB-51 (split1) and select only those videos containing action classes of SSVAC-40-Still dataset. The data consists 3520 train videos and 1412 test videos. Note that when training our temporal feature extraction models, we use test video samples from UCF-101 and HMDB-51 to validate our temporal CNN models (see Fig. 5 for samples).

5. Experiments

In this section, we discuss our results from the proposed experiments on SSVAC-40 data. We will first discuss two baseline experiments that we conduct on SSVAC-40-Still data. The first baseline uses data from the training set of the still action images and reports performance on validation and testing set of the still image collection. The second one uses video frames as additional training data (*i.e.*, data augmentation). For both baseline experiments, we use SGD optimizer with a momentum and a batch size of 128. We tune hyper-parameters of our experiments (*e.g.*, learning rate, learning rate decay interval, batch normalization decay) to obtain the best validation set performance.

As pre-processing, we use random crops of size 224×224 and horizontal flips of the still action images. We scale the images to have a length of 256 on its minimum dimension prior to cropping. At test time we use center-crops. As we use ImageNet pretrained models, we center the images using the mean RGB images of ImageNet data [38]. For dynamic images we use the same pre-processing approach. For Optical-flow, we center the fields using 128 as the mean. We maintain a consistent cropping/flipping for the Optical-flow stack.

Baseline 1: Resnet-50 [41] has shown excellent performances in image classification. For this reason we consider Resnet-50 as a baseline for SSVAC-40-Still action recognition. The model parameters are initialized with pre-trained weights on ImageNet data [38]. We will refer to this model as Resnet-Still.

Baseline 2: We also evaluate our models by augmenting still image data with video frames from SSVAC-40-Video. We first perform key-frame mining to find a set of informative frames from each video using x-means clustering [42]. A Resnet-50 model fine-tuned on UCF-101 video frames is used for feature extraction for clustering. These cluster centers are used as informative instances from each video (*e.g.*, key poses of an action). The frames nearest to the cluster centers (we allow a maximum

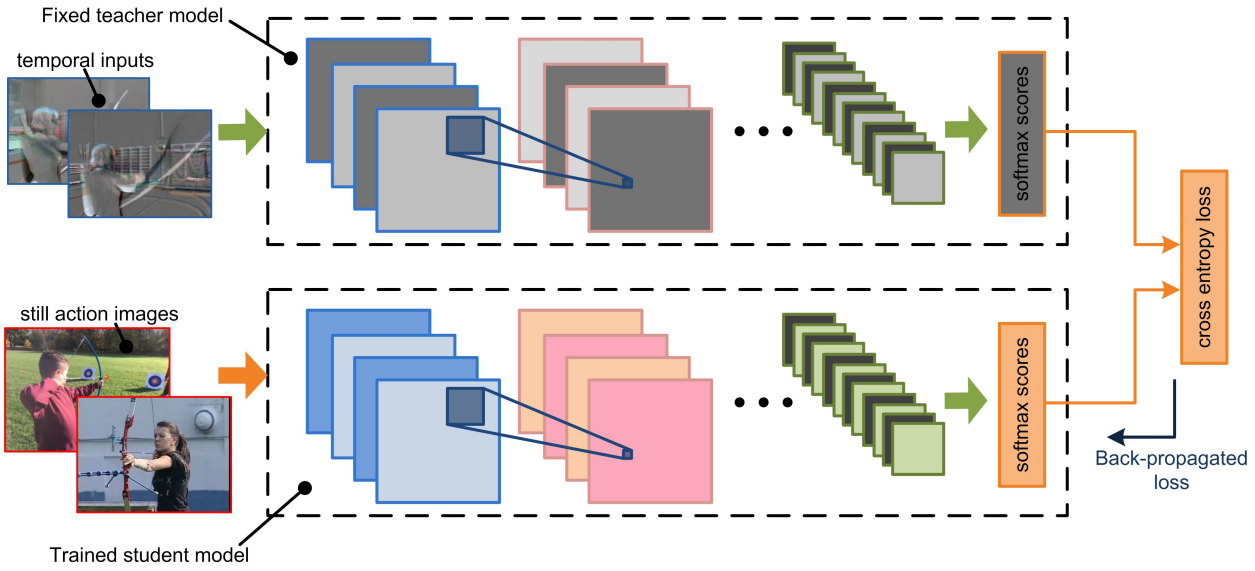

Figure 4: A schematic diagram for the proposed TSN solution for T2SIL. Here, the teacher model receives temporal representation inputs while the student model is fine-tuned to replicate the teacher's response with still action image inputs.

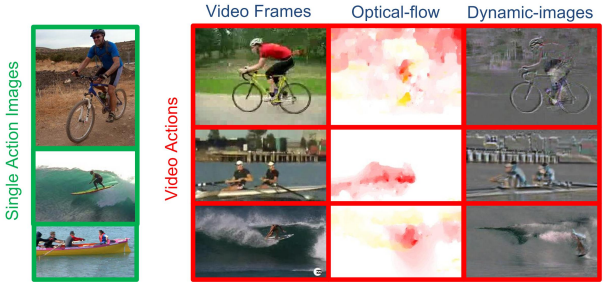

Figure 5: Still image action samples and video temporal representations (*i.e.*., optical-flow and dynamic images) from the proposed SSVAC-40 dataset.

of 5 clusters) are used as the key-frames. This results in around 15,000 new training samples from the videos. We compare the performances of the data augmentation experiment with the still image baseline in Table 1.

It is interesting to see that the data augmentation baseline gives lower performance compared to Resnet-Still model. One reason for this might be the domain shift presented in video data. For instance, still action images usually tend to be rich in context and center the subjects compared to video frames. Hence, naive data augmentation does not guarantee to improve the performance. This is analogous to the degrading of the *target* domain performance with *source* + *target* training in Domain adaptation [43].

5.1. T2SIL transfer learning experiments

Moving on to our solutions, we consider optical-flow and dynamic images [34] as video temporal representations. On video action recognition both these forms of temporal representations are exploited by the state of the art solutions. Another benefit is brought with their physical construction being similar to images. This allows us to use state of the art deep networks for feature extraction.

For temporal feature extraction, we train two separate Resnet-50 models with optical-flow and dynamic images from the SSVAC-40-Videos data. We train these two models for the best performance on the corresponding test data as our ultimate solution is not evaluated on video data. In the remaining sections we will refer to these two models as Resnet-Opt. and Resnet-Dyn. when needed.

5.1.1. T2SIL with Adversarial Learning.

We realize our feature generator and the discriminator models with multi-layer perceptrons. Instead of feeding Gaussian noise explicitly to the generator, we use dropouts and Gaussian noise layers within the generator network. We give detailed descriptions of the models in section 6. We use still image features from Resnet-Still as input to the generator network. GANs tend to work better with compact features [44]. Hence, as shown in Fig. 2 we use an additional dimensionality reduction model with tanh activations. This model takes pool5 features as input from the temporal Resnet model (*i.e.*, Resnet-Dyn. or Resnet-Opt. depending on the experiment). This dimensionality reduction module is pre-trained along with the auxiliary classifier and held fixed during the GAN training. We use RMSProp optimizer for training the GAN. We tune the dimensionality of the features for each data modality and the optimizer learning rate to obtain the best accuracy on the validation data.

We use the auxiliary classifier to obtain the class predictions on generated temporal features from still image inputs. These predictions are analogous to the temporal stream predictions in two-stream networks [2]. As such and in order to fuse predictions from the temporal and spatial streams, we average the softmax scores from the auxiliary classifier and the Resnet-Still predictions.

Table 1: Baseline experiments and temporal feature extraction model performances on SSVAC-40 data.

Model name	Train test data		Accuracy%	
	Trained on	Val/Test on	Val	Test
Still image	still images (6821)	still images (728/4798)	88.2	84.2
Aug. Still+Frame	still+frame images (6821+15077)	still images (728/4798)	87.5	82.7

5.1.2. T2SIL with Deep Embeddings.

We realize the embedding function with a multi-layer perceptron. Features from the dimensionality reduction module (see Fig. 3) with tanh activations are used for temporal embeddings. We give detailed descriptions of the models in section 6. We train a separate classifier to label embedded features of still images. The fused prediction for a given still image instance is obtained by averaging the softmax scores from this classifier and the Resnet-Still.

Remark 1 (Fixing the still image embedding space). *An alternative to having a fixed temporal embedding space is to instead map temporal instances to the still image embedding space. Thereafter, a classifier is trained on both sets of embeddings in the still feature space. However, in our preliminary experiments we find this approach to be not as good as fixing the temporal space.*

5.1.3. T2SIL with Teacher-Student Networks.

The proposed TSN solution uses two separate Resnet-50 networks (see Fig. 4) as the teacher and the student. We initialize the student network with the weights from Resnet-Still. The teacher model is initialized with the corresponding temporal Resnet-50 model (*i.e.*, Resnet-Dyn. or Resnet-Opt.). In our TSN experiments, we observe the best results when only the final three convolution layers of the student Resnet-50 are finetuned. For a given still image action, the fine-tuned student model is expected to produce softmax scores similar to the corresponding temporal model. Hence, the two-stream fusion for this solution is performed by averaging the student model predictions with Resnet-Still predictions. We use RMSProp optimizer and tune the learning rate and the softmax temperature parameter [32] to obtain the best validation performance on the fused predictions.

Remark 2 (Sampling pairs for TSN). *Typically, TSN distillation is performed with the same instance input to both teacher and student networks. However, in the proposed TSN solution for T2SIL, the student model receives a still image input while the teacher gets a randomly sampled co-class instance from the temporal data (*e.g.*, a dynamic image or optical-flow input).*

In Table 2 we report the accuracies of the proposed solutions. Here, we first report the results of the derived temporal streams from still images as "Derived temporal". In the set "Fusion with Still" we report the performance when the derived temporal predictions are averaged with the original still image softmax predictions. We observe an improvement by fusion for

Table 2: Accuracy in still image action classification with the proposed T2SIL with only the derived temporal stream and after the two-stream fusion.

Knowledge transfer method	Transfer data modality			
	Optical-flow		Dynamic images	
	Val	Test	Val	Test
Derived temporal				
GAN	89.3	83.7	88.6	83.4
Triplets Embed.	89.3	84.3	89.0	83.6
TSN	88.4	82.9	89.3	84.3
Fusion with Still				
GAN	89.0	84.0	88.6	83.7
Triplets Embed.	89.0	84.2	89.4	84.4
TSN	89.4	84.4	89.6	84.7

both optical-flow and dynamic images from the TSN solution. The triplets embedding learning has shown an improvement of 0.2% in test accuracy when dynamic images are used as the temporal information source. Note, that we report the improvement as the difference in percentage accuracies from the still image baseline in Table 1. We observe the highest improvement of 0.5% in test performance with the TSN dynamic images experiments.

5.2. Remarks on the performances

Feature generation with GANs have shown excellent performance for related transfer learning problems [28, 16]. Surprisingly, we do not observe improvement from our GAN solution. A reason behind this might be the insufficient training data for the GAN solution to produce a decent generator. In comparison, the most successful method is the TSN solution. Perhaps the way TSN is trained might be better suited for the issue of limited data. This is because TSN training uses a large number of paired combinations of data points as inputs and potentially exposed to larger variations than GAN solution during training.

In comparison to distance losses, the softmax-cross entropy loss has shown better generalization when used to train deep networks with labeled data (e.g, comparison of softmax features with distance loss trained features in [45]). Such properties of the softmax-cross entropy loss could have had favourable effects on the TSN solution. Furthermore, given labeled data from both temporal and image domains, the proposed T2SIL approach can be described as a Cross-Modality Supervision Transfer, CMST[7]. TSN solution has shown decent performance in CMST literature (e.g, [18]). The distinguishable property of

[7]We discuss this in the related work section in detail.

TSN from the other two solutions is that the student model is trained to reproduce the classifier scores of a model trained on temporal data. Hence, we conjecture that a solution attempting to learn the labeling function of temporal data is more suitable for T2SIL.

Although we see improvement from the proposed T2SIL, this is not on par the improvements video action recognition receives from two-stream framework [2]. Hence, we think the proposed T2SIL should be considered in the eye of a larger transfer learning community to reap more benefits. Furthermore, we observe that T2SIL affects different classes in different ways[8]. And by using different temporal data modalities (*e.g.*, LSTM features) one might be able to reach better improvements, implying a novel research problem with many theoretical and practical potentials.

5.3. Class-wise analysis on TSN solution

In Fig. 7 we compare the class-wise accuracy improvements of the still image data when the student model is used for classification. Out of the two data modalities we observe the dynamic images solution has improved in 21 classes out of 40. In Fig. 8, when fused with still image predictions this reduces to 16 (*i.e.*, the classes "long jump", "table tennis shot", "walking with dog", "walk", "wave" have shown negative impact from fusion). However, an overall improvement in the average of per-class accuracies can be seen in the fused solution.

Dynamic images and optical-flow are two video representations that capture temporal information. Hence, it is reasonable to expect some similar improvement patterns in them. We observe similar positive improvements in the two student model's for the classes "hula hoop", "ice dancing", "long jump", "walk", and "kick ball". One particular interesting thing about these classes are that they all involve full body motions. However, we observe some exceptions to this statement as well (*e.g.*, "basketball dunk", "pommel horse").

From this analysis, it is clear that T2SIL affects different classes in different ways. And by using different temporal data modalities (*e.g.*, LSTM features) one might be able to increase the throughput (*e.g.*, Optical-flow student model shows a bias to be better in many short and repetitive actions, "cricket shot", "hula hoop", "kayaking", "clap", "dribble", "walk"). This is another indication for the complexity of T2SIL.

5.4. Video Frame based Experiment

Still action images and video action frames contains notable visual differences. That is a video action frame image usually tends to be low in context, resolution and may not center the subject due to camera motions. In contrast, still image actions tends to be rich in context and resolution while in most cases the subject is in the center of the image. These differences bring an inherent domain disparity between video frames and still action images. As we observed in our dataset augmentation baseline (*see* Baseline 2) this domain disparity screens us in improving performance when additional video action frames are used as still image training data. Our proposed, TSN solution, with T2SIL framework attempts to work around this domain disparity by learning to transfer "still action images ⟶ temporal representations".

Table 3: Performance of proposed solutions when mid-frame from SSVAC-40-Videos are used as still images. The baseline performance on testing mid video frames with Resnet-50 was 76.8%.

Solution	Optical-flow		Dynamic images	
	Temp.	Fused	Temp.	Fused
GAN	75.5	76.0	75.7	76.2
Triplets Embed.	75.5	76.8	76.4	77.3
TSN	79.3	79.4	78.1	79.7

Here we describe an experiment where such domain disparity is non-existing. In other words, we replace our still action images with sampled mid-frame images from videos. For this experiment we use the mid-frame train/test set images of the proposed SSVAC-40-Videos datasets. In Table 3 we report the performance of our proposed three solutions with mid-frame images. Here we observe the baseline performance of a Resnet-50 trained on mid-video frames of SSVAC-40-Videos train data to be 76.8%. For all experiment setups we use the same training hyper-parameters as our still action image experiments.

We observe that our TSN solution yields a significant gain in accuracy (+2.9% to be specific) from the baseline model after two-stream fusion. This experiment demonstrates that tangible gains can be achieved by our TSN solution if domain shift, as considered in our work and in particular in SSVAC-40-Still dataset, does not exist. Although this is a less challenging setup than our original T2SIL we consider this experiment as an indication that temporal information can be used to gain improvements when used for still image action recognition.

5.5. Upper bound with an Oracle

It might be interesting to answer the question "**How much improvement can we get if we have a perfect solution for T2SIL?**"

To answer this question, we perform an oracle experiment. The assumption here is that we have learned a model that could perfectly derive temporal information for a given still image action. To simulate such a model, for a given still image action we pick a co-class temporal representation (*e.g.*, a dynamic image, optical-flow inputs). However, it is important to select the temporal representation from the examples that weren't used for training the corresponding temporal model. Thereafter, the prediction for still image is obtained by averaging the predictions of the still image and the picked temporal instance. In Table 4, we report the improvement results of 100 oracle test runs for each data modality from the still image baseline in Table 1. Although, our oracle experiment contains utopic assumptions (*e.g.*, a perfect temporal information derivation) it indicates us of an upper bound for what might be achieved from T2SIL. This oracle experiment demonstrates that with dynamic images, we can potentially improve the still image action recognition by 5.8%. In our effort to improve results using T2SIL, we manage

[8] Shortly we will provide a class-wise analysis of improvements brought by our TSN solution.

to get an improvement of 0.5% which is 10 times lower than what we can aim for. Therefore, we believe that this problem should be explored in a large community to reap the maximum benefits.

Table 4: Improvement in still image action classification accuracy from the baseline performance (*see* Table 1), Test:84.2% and Val:88.2%.

Knowledge transfer method	Transfer data modality			
	Optical-flow		Dynamic images	
	Val	Test	Val	Test
Oracle test	3.4 ± 0.7	4.4 ± 0.3	4.7 ± 0.7	5.8 ± 0.3

6. Further Details of Experiments

In this section we give details of the training hyper-parameters and network model details for the three transfer learning solutions. All network and training hyper-parameters are tuned for the best performance in the validation set of SSVAC-40-Still.

6.1. Network and Hyper-parameters : T2SIL with Adversarial Learning

In Table 5 we report the structure of the networks used for the adversarial solution. We report the training hyper-parameters for our GAN solution in Table. 6.

6.2. T2SIL with Deep Embedding

In Table 7 we report the structure of the networks used for the deep embedding learning solution. We report the training hyper-parameters for our embedding learning solution in Table. 8.

6.3. T2SIL with Teacher-Student Networks (TSN)

In Table. 9 we report the training hyper-parameters for the proposed TSN solution.

7. Conclusion

Remarks on Observations : In this paper, we raise two novel questions to the problem of action recognition from still images. To paraphrase, first we question whether still image action recognition could benefit from temporal information. To answer this question we proposed a novel transfer learning problem, T2SIL. Our experiments show that adversarial feature generation is not helpful for T2SIL. However, we see improvements with deep embedding learning and a proposed teacher student network solution. Out of the proposed three transfer learning solutions, the teacher student network gives the best performance. Hence, we conclude the answer to be "Yes" to this first question.

Secondly, we question if existing transfer learning frameworks are capable of reaping the maximum benefits from T2SIL. Although helpful, comparing to video action recognition (*e.g.*, two-stream fusion [2]), the improvements are not stellar. Hence, we answer this second question with a "No". We think the

Table 5: **Network details for the GAN solution.** All Leaky ReLU activations have a slope of 0.2 for negative inputs. The dropout rate for all Dropout layers are 0.5. The Gaussian noise layers have a mean of 0.0 and a standard deviation of 0.01. The output dimension of the fully connected layers are given by *dim*. In Table 6 we report the specific training hyper-parameters and layer parameters.

Layer	Layer name	Activation	B. Norm.
Generator network			
0	Input		
1	FC - dim = d0	Leaky ReLU	yes
2	Dropout	-	-
3	Gaussian noise	-	-
4	FC - dim = d0	Leaky ReLU	yes
5	Dropout	-	-
6	Gaussian noise	-	-
7	FC - dim = d1	Tanh	yes
Discriminator network			
0	Input		
1	FC - dim = d2	Leaky ReLU	yes
2	FC - dim = d2	Leaky ReLU	yes
3	FC - dim = d2	Leaky ReLU	yes
4	FC - dim = 1	Sigmoid	no
Auxiliary classifier			
1	FC - dim = 40	-	no
Dimensionality reduction network			
0	input		
1	FC - dim = d1	Tanh	yes
2	Dropout	-	-
3	FC - dim = 40	-	no

Table 6: Training hyper-parameters and network layer details for the GAN solution.

Data modality	d0	d1	d2	λ	Learning rate
Optical-flow	256	128	256	1.0	0.001
Dynamic images	256	128	256	1.0	0.001

Table 7: **Network details for the Triplets solution.** All Leaky ReLU activations have a slope of 0.2 for negative inputs. The dropout rate for all Dropout layers are 0.5. The Gaussian noise layers have a mean of 0.0 and a standard deviation of 0.01. The output dimension of the fully connected layers are given by *dim*. In Table 8 we report the specific training hyper-parameters and layer parameters.

Layer	Layer name	Activation	B. Norm.
Embedding network			
0	Input		
1	FC - dim = d0	Leaky ReLU	yes
2	Dropout	-	-
3	Gaussian noise	-	-
4	FC - dim = d0	Leaky ReLU	yes
5	Dropout	-	-
6	Gaussian noise	-	-
7	FC - dim = d1	Tanh	yes
Embedding classifier			
0	Input		
1	FC - dim = 40	-	no
Dimensionality reduction network			
0	Input		
1	FC - dim = d1	Tanh	yes
2	Dropout	-	-
3	FC - dim = 40	-	no

T2SIL problem will open new research directions within the transfer learning community and can be considered as a challenging problem in computer vision.

Limitations and Future Extensions : In its current formulation, T2SIL framework has certain limitations. First and foremost, we require labeled video data for transferring the knowledge to still image action data. Furthermore, we assume that both still and temporal data belongs to the same set of classes in our experiments. As such, this limits the use of abundantly available video data (*e.g.*, unlabeled) for current T2SIL. In future, we intend to explore unsupervised-T2SIL as well as new data setups where complete overlap between still and temporal data classes are not present.

The proposed SSVAC-40 data is limited in size and annotations. Lack of data samples in the proposed dataset might be the reason for somewhat lower performance for GAN solution. Furthermore, state of the art still action recognition solutions use complex visual cues (*e.g.*, body part annotations) to attain improvements. However, due to the unavailability of such expensive annotations in SSVAC-40-Still data, we are unable to benefit from the complex visual cues for T2SIL. As such, we intend to extend the proposed SSVAC-40 dataset by size and annotations in the future. With such modifications, we can explore a wide range of solutions to solve the proposed task.

References

[1] H. Yang, C. Yuan, B. Li, Y. Du, J. Xing, W. Hu, S. J. Maybank, Asymmetric 3d convolutional neural networks for action recognition, Pattern Recognition 85 (2019) 1 – 12. 1

[2] K. Simonyan, A. Zisserman, Two-stream convolutional networks for action recognition in videos, in: Proc. Advances in Neural Information Processing Systems (NIPS), 2014, pp. 568–576. 1, 2, 3, 4, 7, 9, 10

[3] M. Yatskar, L. Zettlemoyer, A. Farhadi, Situation recognition: Visual semantic role labeling for image understanding, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 1, 2, 6

[4] E. Ohn-Bar, M. M. Trivedi, Are all objects equal? deep spatio-temporal importance prediction in driving videos, Pattern Recognition 64 (2017) 425 – 436. 1

[5] L. Wang, Z. Wang, Y. Qiao, L. Van Gool, Transferring deep object and scene representations for event recognition in still images, in: Int. Journal of Computer Vision, 2017. 1

[6] Y.-W. Chao, Z. Wang, Y. He, J. Wang, J. Deng, Hico: A benchmark for recognizing human-object interactions in images, in: Proc. Int. Conference on Computer Vision (ICCV), 2015. 1

[7] J. Walker, K. Marino, A. Gupta, M. Hebert, The pose knows: Video forecasting by generating pose futures, in: Proc. Int. Conference on Computer Vision (ICCV), 2017, pp. 3332–3341. 1

[8] R. Mottaghi, M. Rastegari, A. Gupta, A. Farhadi, what happens if... learning to predict the effect of forces in images, in: Proc. European Conference on Computer Vision (ECCV), 2016, pp. 269–285. 1, 2

[9] S. Yu, R. Liao, W. An, H. Chen, E. B. Garca, Y. Huang, N. Poh, Gaitganv2: Invariant gait feature extraction using generative adversarial networks, Pattern Recognition 87 (2019) 179 – 189. 1

[10] L. Shmuelof, E. Zohary, Dissociation between ventral and dorsal fmri activation during object and action recognition, Neuron 47 (3) (2005) 457–470. 1

[11] K. Soomro, A. R. Zamir, M. Shah, Ucf101: A dataset of 101 human actions classes from videos in the wild, arXiv preprint arXiv:1212.0402. 1, 6

[12] A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates, IEEE Trans. Pattern Analysis and Machine Intelligence 23 (3) (2001) 257–267. 1

[13] G. Gkioxari, R. Girshick, J. Malik, Contextual action recognition with r*cnn, in: Proc. Int. Conference on Computer Vision (ICCV), 2015. 2

[14] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, L. Fei-Fei, Human action recognition by learning bases of action attributes and parts, in: Proc. Int. Conference on Computer Vision (ICCV), 2011. 2, 6

[15] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. 2, 6

[16] Y. Xian, T. Lorenz, B. Schiele, Z. Akata, Feature generating networks for zero-shot learning, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 2, 3, 8

[17] L. A. Pereira, R. da Silva Torres, Semi-supervised transfer subspace for domain adaptation, Pattern Recognition 75 (2018) 235 – 249. 2, 3

[18] S. Gupta, J. Hoffman, J. Malik, Cross modal distillation for supervision transfer, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 2827–2836. 2, 3, 8

[19] J. Walker, A. Gupta, M. Hebert, Dense optical flow prediction from a static image, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2443–2451. 2, 3

[20] T. Xue, J. Wu, K. Bouman, B. Freeman, Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks, in: Proc. Advances in Neural Information Processing Systems (NIPS), 2016, pp. 91–99. 2

[21] G. Sharma, F. Jurie, C. Schmid, Expanded parts model for semantic description of humans in still images, IEEE Trans. Pattern Analysis and Machine Intelligence 39 (1) (2017) 87–101. 2

[22] A. Rosenfeld, S. Ullman, Action classification via concepts and attributes, in: Proc. Int. Conference on Pattern Recognition (ICPR), 2018, pp. 1499–1505. 2

[23] Z. Zhao, H. Ma, S. You, Single image action recognition using semantic body part actions, in: Proc. Int. Conference on Computer Vision (ICCV), 2017, pp. 3411–3419. 2

[24] C.-Y. Chen, K. Grauman, Watching unlabeled video helps learn new human actions from very few labeled snapshots, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 572–579. 2

[25] V. Bloom, V. Argyriou, D. Makris, Linear latent low dimensional space

Table 8: Training hyper-parameters for the Triplets solution.

Data modality	d0	d1	α	Learning rate	Sampling
Optical-flow	256	256	1	0.0002	4 classes × 4 instances
Dynamic images	128	256	0.5	0.0002	4 classes × 4 instances

Table 9: Training hyper-parameters for the TSN solution.

Data modality	Softmax temperature	Learning rate
Optical-flow	2	0.0002
Dynamic images	10	0.0002

for online early action recognition and prediction, Pattern Recognition 72 (2017) 532 – 547. 3

[26] J. Hu, W. Zheng, L. Ma, G. Wang, J. Lai, J. Zhang, Early action prediction by soft regression, IEEE Trans. Pattern Analysis and Machine Intelligence (2018) 1–1. 3

[27] C. Rodriguez, B. Fernando, H. Li, Action anticipation by predicting future dynamic images, in: Proc. European Conference on Computer Vision (ECCV), 2018, pp. 0–0. 3

[28] Y. Ganin, V. Lempitsky, Unsupervised domain adaptation by backpropagation, in: Proc. Int. Conference on Machine Learning (ICML), 2015, pp. 1180–1189. 3, 8

[29] R. Socher, M. Ganjoo, C. D. Manning, A. Ng, Zero-shot learning through cross-modal transfer, in: Proc. Advances in Neural Information Processing Systems (NIPS), 2013, pp. 935–943. 3

[30] N. Gao, S.-J. Huang, Y. Yan, S. Chen, Cross modal similarity learning with active queries, Pattern Recognition 75 (2018) 214 – 222. 3, 4

[31] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 815–823. 3, 4, 5

[32] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531. 3, 5, 8

[33] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, K. Talwar, Semi-supervised knowledge transfer for deep learning from private training data, arXiv preprint arXiv:1610.05755. 3

[34] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, Action recognition with dynamic image networks, in: IEEE Trans. Pattern Analysis and Machine Intelligence, 2017. 3, 6, 7

[35] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, T. Tuytelaars, Rank pooling for action recognition, IEEE Trans. Pattern Analysis and Machine Intelligence 39 (4) (2016) 773–787. 3

[36] A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier gans, in: Proc. Int. Conference on Machine Learning (ICML), 2017. 4

[37] L. Zhang, T. Xiang, S. Gong, Learning a deep embedding model for zero-shot learning, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 2021–2030. 4

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge 115 (3) (2015) 211–252. 6

[39] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, T. Serre, Hmdb: a large video database for human motion recognition, in: Proc. Int. Conference on Computer Vision (ICCV), 2011, pp. 2556–2563. 6

[40] C. Feichtenhofer, A. Pinz, A. Zisserman, Convolutional two-stream network fusion for video action recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 6

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. 6

[42] D. Pelleg, A. Moore, X-means: Extending k-means with efficient estimation of the number of clusters, in: Proc. Int. Conference on Machine Learning (ICML), 2000. 6

[43] K. Saenko, B. Kulis, M. Fritz, T. Darrell, Adapting visual category models to new domains, in: Proc. European Conference on Computer Vision (ECCV), 2010, pp. 213–226. 7

[44] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434. 7

[45] S. Horiguchi, D. Ikami, K. Aizawa, Significance of softmax-based features in comparison to distance metric learning-based features, IEEE Trans. Pattern Analysis and Machine Intelligence (2019) 1–1. 8

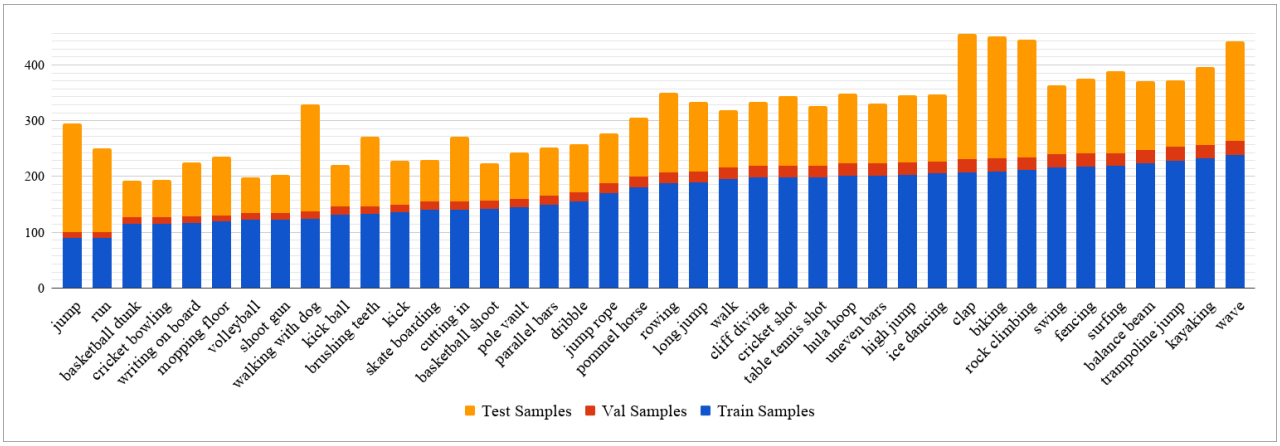

Figure 6: The class composition of the proposed still image action dataset SSVAC-40-Still.

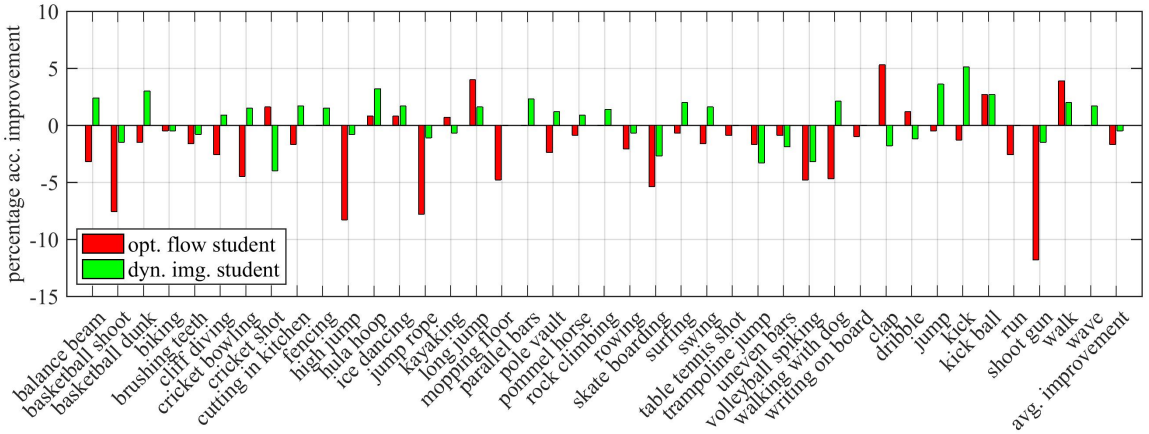

Figure 7: Improvement in per class accuracies for the still image predictions from the student model. The improvement is measured w.r.t. the baseline still image model's performance.

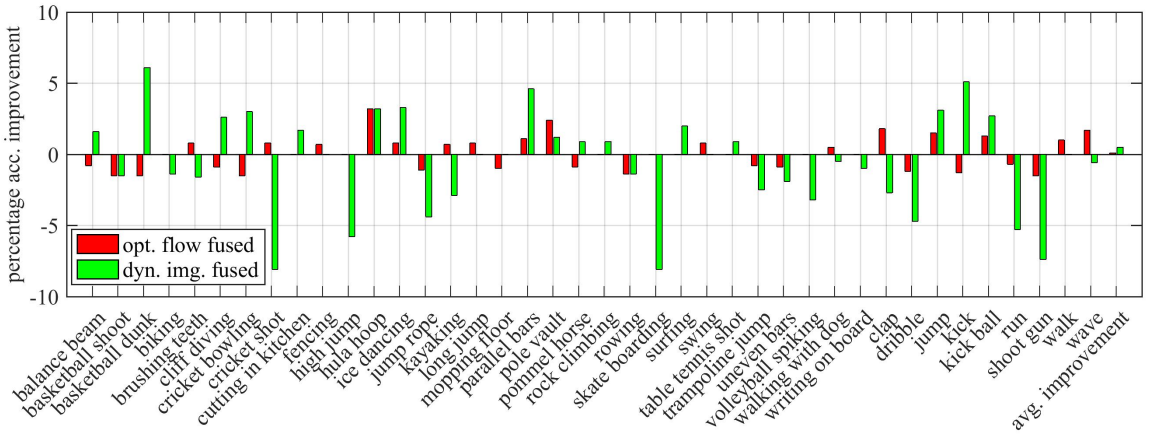

Figure 8: Improvement in per class accuracies when the predictions of the still image and the student model are averaged. The improvement is measured w.r.t. the baseline still image model's performance.