GliTr: Glimpse Transformers with Spatiotemporal Consistency for Online Action Prediction

Samrudhdhi B Rangrej[1] Kevin J Liang[2] Tal Hassner[2] James J Clark[1]
[1]McGill University [2]Meta AI
samrudhdhi.rangrej@mail.mcgill.ca

Abstract

*Many online action prediction models observe complete frames to locate and attend to informative subregions in the frames called glimpses and recognize an ongoing action based on global and local information. However, in applications with constrained resources, an agent may not be able to observe the complete frame, yet must still locate useful glimpses to predict an incomplete action based on local information only. In this paper, we develop Glimpse Transformers (GliTr), which observe only narrow glimpses at all times, thus predicting an ongoing action and the following most informative glimpse location based on the partial spatiotemporal information collected so far. In the absence of a ground truth for the optimal glimpse locations for action recognition, we train GliTr using a novel spatiotemporal consistency objective: We require GliTr to attend to the glimpses with features similar to the corresponding complete frames (*i.e. *spatial consistency) and the resultant class logits at time t equivalent to the ones predicted using whole frames up to t (*i.e. *temporal consistency). Inclusion of our proposed consistency objective yields $\sim 10\%$ higher accuracy on the Something-Something-v2 (SSv2) dataset than the baseline cross-entropy objective. Overall, despite observing only $\sim 33\%$ of the total area per frame, GliTr achieves 53.02% and 93.91% accuracy on the SSv2 and Jester datasets, respectively.*

1. Introduction

Recent models such as TSM [37], Swin-B [38], or VideoMAE [53] have achieved impressive performance on video action recognition benchmarks, but they often make several assumptions that limit their use for certain applications. For example, the aforementioned models operate in an offline manner, assuming the full clip (i.e. after the action has concluded) is available to make a decision. Offline models are often inefficient in online settings, where action recognition must be performed based on the incomplete clip seen up until the current time. For example, the performance of Swin-B drops by ∼30% on the Something-Something-v2 (SSv2) dataset when only the first 70% frames are observed [52].

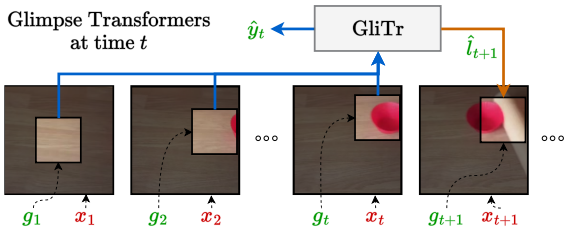


Figure 1: We propose **Glimpse Transformers (GliTr)**, an online action prediction model that only attends to the most informative glimpses (g_t) in the frames (x_t). While never observing frames completely, GliTr predicts label $\hat{y}_t$ (*i.e.* an estimate of ongoing action at time t) and the next glimpse location $\hat{l}_{t+1}$ based solely on the glimpses observed up to t.

Another common assumption is the requirement of complete spatial information over time. But, due to spatial redundancy, it is enough to observe only small but informative subregions of the full frames to make an accurate prediction. Several approaches [6, 25, 40, 62, 64] primarily process narrow regions called "glimpses". However, these approaches still require the entire frame to determine informative glimpses. While using a lightweight model for this "global" view reduces the overall computational cost, it still requires having a wide field of view initially, which does not come free. High-resolution, large FOV cameras are expensive, require more power, and consume more bandwidth to transmit data. It is essential to minimize such costs in certain high-risk time-sensitive applications, such as mobile drones for disaster recovery, monitoring at-risk animals in the wild, or real-time translation of sign language.

We thus develop an inexpensive model that predicts informative glimpse locations *without* observing whole frames, therefore obviating the need for high-resolution, large FOV cameras. Starting from a glimpse at a given location, our model decides which location to attend to in subsequent frames solely based on previously observed glimpses. Consequently, our model predicts an action using only the

local information and in an online fashion. We choose transformers [57] to learn glimpse-based attention mechanism and action prediction, as they can efficiently encode the relations between spatially and temporally distant glimpses. We thus call our model **Glimpse Transformers (GliTr)**. Following a *factorized encoder* architecture [2], we use a) a spatial encoder that solely models relations between the patches from a single glimpse to predict spatial features, and b) two temporal encoders that model interactions between various glimpse features across time to predict the class label and the next glimpse location, respectively.

Since the ground truth for optimal glimpse locations is unavailable, we propose a novel spatiotemporal teacher-student consistency objective to incentivize GliTr to learn glimpse location in a weakly supervised manner. With only glimpses, GliTr (as the student model) is trained to reproduce the spatial features and class distribution of a teacher model ingesting the complete frames of the video. As the teacher learns to produce predictive features and logits for the downstream task of online action recognition from the full frames, enforcing this consistency loss on the student model implicitly requires focusing attention on the most informative regions, leading to learning a glimpse mechanism. We demonstrate GliTr's effectiveness on Something-Something-v2 [22] and Jester [41] datasets. Our main contributions are as follows.

- We develop GliTr: an online action prediction model that observes only glimpses and predicts ongoing action based on partial spatiotemporal observations. While previous works locate glimpses by first observing full frames, GliTr predicts the next informative glimpse location solely based on the past glimpses.

- We propose a novel spatiotemporal consistency objective to train GliTr without the ground truth for glimpse location. Under this objective, GliTr must select glimpses that summarize features and class distribution predicted from the entire frames. Our proposed consistency yields $\sim 10\%$ gain in accuracy on SSv2 compared to the baseline cross-entropy objective.

- Our model that never observes complete frames and recognizes action solely based on local information gathered through glimpses achieves nearly 53% and 94% accuracy on SSv2 and Jester dataset, respectively, while reducing the total area observed per frame by nearly 67% (with the glimpses of size 128×128 extracted from frames of size 224×224).

2. Related Works

Online Action Recognition. Many state-of-the-art methods perform offline action recognition once the entire video is available [17, 19, 10, 27, 55, 56, 59, 60]. However, these methods are not optimized for the case where the entire video is not yet available, and the model has to predict the action based on a preliminary, incomplete video.

Performing an online or early action recognition based on a partially observed video is a challenging task. A partially observed video may associate with multiple possible actions, leading to the inherent uncertainty in the prediction task. Several methods focus on predicting actions from partial videos. Zhao *et al*. [69], Wu *et al*. [65], and Pang *et al*. [43] anticipate future actions based on the motion and object relations in the past frames. Many analyze micro-motions in the available early frames [52, 34, 30, 28]. Other approaches such as dynamic bag-of-words [49], global-local saliency [32], memorizing hard-to-predict samples [29], soft regression with multiple soft labels [24], and probabilistic modeling [35, 9] are also used. While the existing online action recognition methods focus on partial observation in the temporal dimension, we focus on partial information in the temporal as well as the spatial dimension.

Spatial Selection for Action Recognition. Spatial selection is typically performed using hard attention [42]. As opposed to soft attention models [68] that observe all regions of the scene with varying attention level, hard attention models sequentially attend to the most informative glimpses. Hard attention is widely used for image classification [3, 16, 42, 68, 44, 48, 47, 63].

Recently, hard attention has also been applied to video action recognition. Wang *et al*. propose online action recognition model called Adafocus [62, 64]. Chen *et al*. [11], Huang *et al*. [25] and Wang *et al*. [58] present offline models that first observe the entire video in order to predict attention-worthy glimpse locations. Mac *et al*. [40] and Baradel *et al*. [6] also present offline models but locate and observe multiple informative glimpses per-frame. Another line of approach leverages pose information and focuses only on the relevant body parts [5, 13]. The previous approaches, irrespective of their online or offline nature, access full frames to locate informative glimpses. In contrast, our model never observes complete frames; it only observes a narrow glimpse from each frame.

Consistency Learning. Consistency is widely used for the problem of semi-supervised learning [50, 51, 66, 7, 33]. The idea is to force the output of the model to be invariant to different augmentations of the same input [51, 66, 7, 36], or variations in the internal representations [4, 50], or the model parameters at different training epochs [33]. Another related approach is pseudo-labeling [67, 45], where a separate teacher model generates pseudo-labels for unlabeled samples under no perturbations, and the student model is trained to predict the pseudo-labels under some perturbations. This approach is similar to Knowledge Distillation

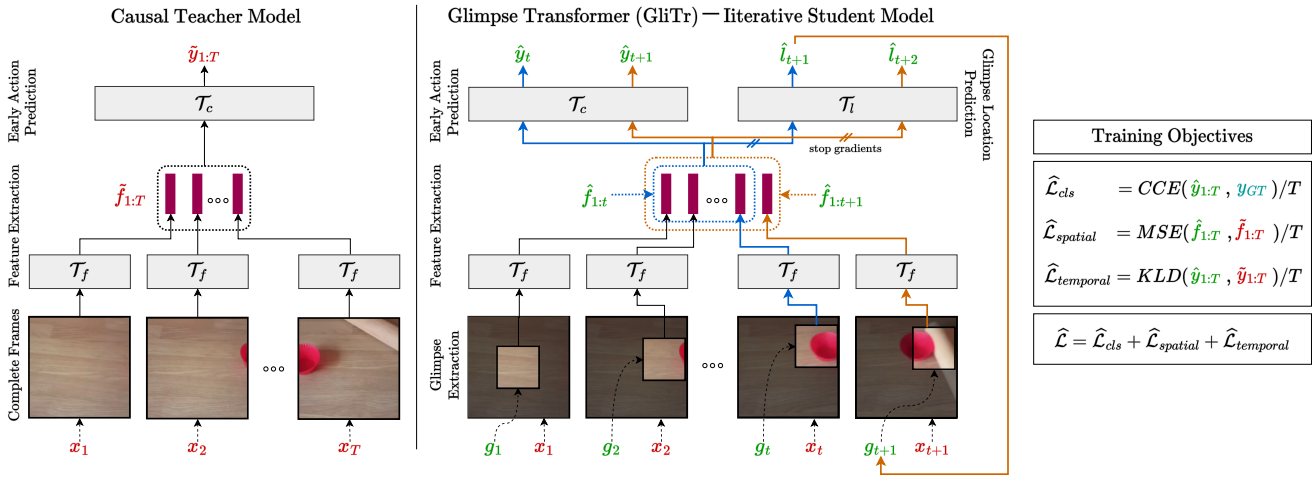


Figure 2: **An overview of our GliTr.** GliTr consists of a frame-level spatial transformer $\mathcal{T}_f$ and causal temporal transformers $\mathcal{T}_c$ and $\mathcal{T}_l$. One training iteration requires T forward passes through our model. Above, we show two consecutive forward passes at time $t \leq T-1$ and $t+1 \leq T$. **Forward pass t** (blue path): Given a new glimpse g_t, $\mathcal{T}_f$ extracts glimpse-features $\hat{f}_t$. We append $\hat{f}_t$ to $\hat{f}_{1:t-1}$, *i.e.* features extracted from $g_{1:t-1}$ during previous passes. Next, $\mathcal{T}_c$ predicts label $\hat{y}_t$ from $\hat{f}_{1:t}$. Simultaneously, $\mathcal{T}_l$ predicts next glimpse location $\hat{l}_{t+1}$ from $\hat{f}_{1:t}$. **Forward pass t+1** (orange path): Given a predicted location $\hat{l}_{t+1}$, we extract a glimpse g_{t+1} at $\hat{l}_{t+1}$ from a frame x_{t+1}. Then, we follow the same steps as the blue path. After T forward passes, we compute the losses shown in the right. To find targets $\tilde{y}_{1:T}$ and $\tilde{f}_{1:T}$ for spatial and temporal consistency, we use a separate pre-trained and fixed teacher model (shown on the left and explained in Figure 3) that observes complete frames $x_{1:T}$. To maintain stability, we stop gradients from $\mathcal{T}_l$ to $\mathcal{T}_f$.

[23], where the student is trained to reconstruct the output or internal representation [1] of the teacher.

Many early action recognition models learn to predict the class distribution consistent with the complete video using only a subset of early frames [8, 20, 31, 46, 61]. Others have also leveraged spatiotemporal consistency for complete frames [53, 18]. Inspired by previous work, we use a teacher model that predicts features from complete frames and predicts class distribution in an online fashion. Our student model observes partial spatiotemporal information and tries to predict features and class distribution consistent with the teacher model.

3. Models

We use a teacher model to i) initialize our GliTr - a student model and ii) compute targets for the spatiotemporal consistency objective used for training GliTr. We discuss our teacher model in Sec 3.1 followed by GliTr in Sec 3.2. We crown the quantities computed by our models using complete frames and glimpses with (˜) and (ˆ), respectively.

3.1. Teacher

Given spatially complete frames $x_{1:t}$ from a preliminary video at time $t \leq T$, our online teacher model predicts $\tilde{y}_t$, an early estimate of true action y_{GT}. We adapt *factorized transformers encoder* architecture [2] for our teacher model, and aggregate spatial and temporal information sequentially. It includes the following components.

Feature Extraction ($\mathcal{T}_f$). We use a spatial transformer $\mathcal{T}_f$ to extract features $\tilde{f}_t$ from each individual frame x_t for all t. We use the ViT architecture [57, 54] without the final classification head and collect features from the output corresponding to the input class token.

Early Action Prediction ($\mathcal{T}_c$). We use a temporal transformer $\mathcal{T}_c$ to aggregate features $\tilde{f}_{1:t}$ and predict label $\tilde{y}_t$. Since transformers are permutation invariant, we enforce order in the input sequence using temporal position embeddings. Moreover, we do not use a separate class token and pass the output corresponding to $\tilde{f}_t$ to the linear classifier to predict $\tilde{y}_t$. Further, to reduce training time, we use causal attention masking [21, 12]. Hence, during training, $\mathcal{T}_c$ observes $\tilde{f}_{1:T}$ and produces $\tilde{y}_{1:T}$ in a single forward pass while aggregating features in an online progressive manner, referencing only $\tilde{f}_{1:t}$ to produce output $\tilde{y}_t$ at index t.

Glimpse Location Prediction ($\mathcal{T}_l$). We include temporal transformer $\mathcal{T}_l$ to predict glimpse location $\tilde{l}_{t+1}$ from $\tilde{f}_{1:t}$. $\mathcal{T}_l$ has the same architecture as $\mathcal{T}_c$, except the final linear classifier is replaced by a linear regression head to predict coordinates $\tilde{l}_{t+1}$. Though not required for online action prediction from full frames, we train $\mathcal{T}_l$ to initialize the corresponding module in our student model. Once the student model is initialized, we discard $\mathcal{T}_l$ from the teacher model.

Algorithm 1 Inference using GliTr

$\hat{l}_1$ is predefined.
for $t \in \{1, \ldots, T\}$ **do**
  Sample g_t at $\hat{l}_t$ from x_t. ▷ **Glimpse Extraction**
  $\hat{f}_t = \mathcal{T}_f(g_t, \hat{l}_t)$ ▷ **Feature Extraction**
  $\hat{y}_t = \mathcal{T}_c(\hat{f}_{1:t})$ ▷ **Early Action Prediction**
  $\hat{l}_{t+1} = \mathcal{T}_l(\hat{f}_{1:t})$ ▷ **Glimpse Location Prediction**
  Save $\hat{f}_t$.
end for

3.2. Glimpse Transformer (GliTr) — Student

Our Glimpse Transformer (GliTr) is derived and initialized from the teacher model discussed in Sec 3.1. It is an iterative model that actively locates and attends to narrow glimpses in a scene and predicts an ongoing action early based on spatially and temporally incomplete observations. At time t, GliTr senses a new glimpse g_t at location $\hat{l}_t$ from frame x_t. Using glimpses $g_{1:t}$, it predicts i) $\hat{y}_t$, an early approximation of label y_{GT} and ii) $\hat{l}_{t+1}$, location of the next glimpse. We display schematics of GliTr in Figure 1. We illustrate GliTr's operation in Algorithm 1 and Figure 2. It consists of the following components.

Glimpse Extraction. Given a location $\hat{l}_t = (i, j)$, we crop a glimpse g_t centered at location $\hat{l}_t$ in frame x_t. To maintain differentiability through the cropping operation, we use a spatial transformer network (STN) [26][1].

Feature Extraction ($\mathcal{T}_f$). Similar to the teacher model, we use $\mathcal{T}_f$ to extract features $\hat{f}_t$ from glimpse g_t. We derive position embeddings for patches in g_t using STN.

Early Action Prediction ($\mathcal{T}_c$). We input glimpse features $\hat{f}_{1:t}$ to $\mathcal{T}_c$ which in turn predicts class label $\hat{y}_t$.

Glimpse Location Prediction ($\mathcal{T}_l$). Similarly, we pass the features $\hat{f}_{1:t}$ to $\mathcal{T}_l$ which predicts next glimpse location $\hat{l}_{t+1}$.

4. Training Objectives

We discuss training objectives for GliTr in Sec 4.1. Considering GliTr as the downstream model, we design training objectives suitable for our teacher model in Sec 4.2. We crown training objectives of GliTr and the teacher model with ($\widehat{\ }$) and ($\widetilde{\ }$), respectively.

4.1. Glimpse Transformer (GliTr) — Student

Classification Loss. Since our goal is to predict action label y_{GT} early using the spatially and temporally incomplete video, we minimize the cross-entropy loss given by

$$\widehat{\mathcal{L}}_{cls} = CCE(\hat{y}_{1:T}, y_{GT})/T. \tag{1}$$

[1]Not to be confused with (spatial) Vision Transformers (ViT) [15].

Spatial Consistency Loss. We require GliTr to attend to the glimpses that produce features as predictive of the action as the ones predicted using complete frames by our teacher model. Hence, we minimize the mean squared error (MSE) between the glimpse features $\hat{f}_t$ predicted by GliTr and the frame features $\tilde{f}_t$ predicted by our teacher model, which is

$$\widehat{\mathcal{L}}_{spatial} = MSE(\hat{f}_{1:T}, \tilde{f}_{1:T})/T. \tag{2}$$

Temporal Consistency Loss. While the teacher model has all instantaneous spatial information available in a complete frame, GliTr must rely on past glimpses to reason about the unobserved yet informative regions in the current frame. To incentivize GliTr to aggregate spatial information from the past to mitigate partial observability, we minimize the KL-divergence between the class logits predicted by GliTr using glimpses ($\hat{y}_t$) and the teacher using complete frames ($\tilde{y}_t$), yielding

$$\widehat{\mathcal{L}}_{temporal} = KLD(\hat{y}_{1:T}, \tilde{y}_{1:T})/T. \tag{3}$$

Our final training objective for GliTr is the following:

$$\widehat{\mathcal{L}} = \widehat{\mathcal{L}}_{cls} + \widehat{\mathcal{L}}_{spatial} + \widehat{\mathcal{L}}_{temporal} \tag{4}$$

4.2. Teacher

Classification loss. For all t, we minimize cross-entropy loss between the prediction $\tilde{y}_t$ and the ground-truth label y_{GT} of the action,

$$\widetilde{\mathcal{L}}_{cls} = CCE(\tilde{y}_{1:T}, y_{GT})/T. \tag{5}$$

Distillation loss. When available, we also use a more powerful offline action recognition model such as VideoMAE [53] to predict action $y_T^{offline}$ from a complete video, *i.e.* $x_{1:T}$. Then, we minimize the KL-divergence between the final prediction $\tilde{y}_T$ and the above $y_T^{offline}$ given by

$$\widetilde{\mathcal{L}}_{dist} = KLD(\tilde{y}_T, y_T^{offline}). \tag{6}$$

Spatiotemporal Consistency losses. Note that the above two losses train only $\mathcal{T}_f$ and $\mathcal{T}_c$. We use the following strategy to train $\mathcal{T}_l$. First, we use the locations $\tilde{l}_1$ (learnable parameter) and $\tilde{l}_{2:T}$ predicted by $\mathcal{T}_l$, to extract glimpses $g_{1:T}$ from frames $x_{1:T}$. Next, we create copies of $\mathcal{T}_f$ and $\mathcal{T}_c$ denoted as $\mathcal{T}_f'$ and $\mathcal{T}_c'$. We input $g_{1:T}$ and the corresponding position embeddings to $\mathcal{T}_f'$ and predict glimpse features $\hat{f}_{1:T}$. Given $\hat{f}_{1:T}$, $\mathcal{T}_c'$ predicts actions $\hat{y}_{1:T}$ in an online fashion. Then we minimize,

$$\widetilde{\mathcal{L}}_{spatial} = MSE(\hat{f}_{1:T}, \tilde{f}_{1:T})/T, \tag{7}$$

$$\widetilde{\mathcal{L}}_{temporal} = KLD(\hat{y}_{1:T}, \tilde{y}_{1:T})/T. \tag{8}$$

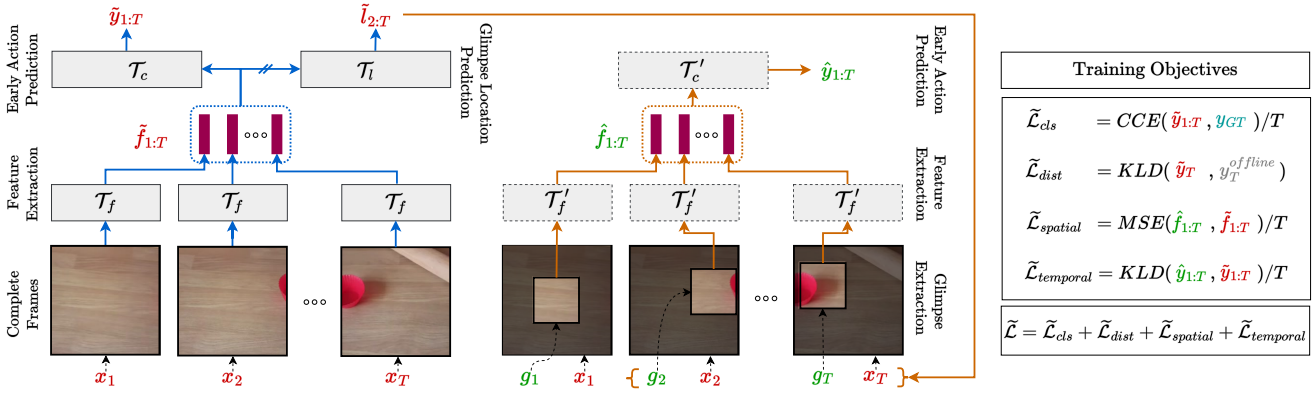


Figure 3: **An overview of our teacher model.** Our teacher model consists of a spatial transformer $\mathcal{T}_f$ and *causal* temporal transformers $\mathcal{T}_c$ and $\mathcal{T}_l$. Each training iteration of the teacher model consists of two steps. **Step 1** (blue path): Given complete video frames $x_{1:T}$, $\mathcal{T}_f$ extracts frame features $\tilde{f}_{1:T}$. Next, $\mathcal{T}_c$ and $\mathcal{T}_l$ predict class labels $\tilde{y}_{1:T}$ and glimpse locations $\tilde{l}_{2:T+1}$ from $\tilde{f}_{1:T}$, respectively. We discard $\tilde{l}_{T+1}$. **Step 2** (orange path): Given $\tilde{l}_1$ (learnable parameter) and $l_{2:T}$ (predicted in step 1), we extract glimpses $g_{1:T}$ from $x_{1:T}$. Then, we create non-learnable copies of $\mathcal{T}_f$ and $\mathcal{T}_c$ denoted as $\mathcal{T}'_f$ and $\mathcal{T}'_c$. $\mathcal{T}'_f$ extracts glimpse-features $\hat{f}_{1:T}$ from $g_{1:T}$ and $\mathcal{T}'_c$ predicts labels $\hat{y}_{1:T}$ from $\hat{f}_{1:T}$. We compute losses shown on the right and update model parameters. To achieve stability during training, we stop gradients from $\mathcal{T}_l$ to $\mathcal{T}_f$.

We use the above two losses to update parameters of $\mathcal{T}_l$ only. We design these consistency objectives based on the spatiotemporal consistency objectives of GliTr (equations 2 and 3). As discussed in Sec 4.1, they encourage $\mathcal{T}_l$ to locate glimpses covering the most useful task-relevant regions in the frames, but based on complete frames observed in the past. We demonstrate the training procedure in Figure 3.

The final objective for our teacher model is as follows.

$$\widetilde{\mathcal{L}} = \widetilde{\mathcal{L}}_{cls} + \widetilde{\mathcal{L}}_{dist} + \widetilde{\mathcal{L}}_{spatial} + \widetilde{\mathcal{L}}_{temporal} \tag{9}$$

5. Experiments

Datasets. We experiment with two publicly available large-scale real-world datasets, namely, Something-Something-v2 (SSv2) [22] and Jester [41]. We adopt the official training-validation splits. SSv2 dataset contains videos recording 174 human actions using everyday objects. There are ∼170K videos for training and ∼25K for validation. Jester dataset is a collection of videos capturing 27 basic hand gestures, consisting of ∼120K videos for training and ∼15K videos for validation.

Implementation. We sample a sequence of 16 and 8 frames per video from SSv2 and Jester, respectively. We resize each frame to size 224×224 and use glimpses of size 96×96, unless stated otherwise. We use ViT-Small [54] architecture for $\mathcal{T}_f$. For $\mathcal{T}_c$ and $\mathcal{T}_l$, we use a custom transformers architecture with 768 embedding dimensions, 6 heads, and a depth of 4.

Optimization. First, we discuss the common setting followed by a model-specific setting. For all models and datasets, we use the same data augmentation scheme as the one used for VideoMAE [53]. Similar to Wang *et al*. [64], we stop gradients from $\mathcal{T}_l$ to $\mathcal{T}_f$ to maintain stability during training. We use AdamW optimizer [39] with weight decay of 5e-2 and cosine learning rate schedule with no warmup unless stated otherwise. We run experiments for SSv2 and Jester on 4 A100 GPUs with 40 GB of memory and 4 V100-SXM2 GPUs with 32 GB of memory, respectively.

To train a teacher model on SSv2 dataset, we initialize $\mathcal{T}_f$ using an open-source ViT-S model [71] pretrained on the ImageNet dataset [14], and initialize $\mathcal{T}_c$ and $\mathcal{T}_l$ randomly. We form a mini-batch using $b = 60$ videos and use an initial learning rate of $\frac{\alpha b}{128}$, with base learning rate α being 1e-5, 1e-4 and 1e-4 for $\mathcal{T}_f$, $\mathcal{T}_c$ and $\mathcal{T}_l$, respectively. We train the teacher model for 40 epochs with a warmup of 15 epochs for $\mathcal{T}_l$. For the Jester dataset, we initialize the teacher model with the teacher model trained on the SSv2 dataset. We do not use distillation loss $\widetilde{\mathcal{L}}_{dist}$ for Jester dataset. We use a batch size b of 100 and α of 1e-5 for all modules. The model is trained for 50 epochs.

Each student model (GliTr) is initialized from a teacher model trained on the corresponding dataset. We use base learning rate α = 1e-5 for all modules and train them for 100 and 150 epochs with a batch-size b of 360 and 800 videos from SSv2 and Jester, respectively.

5.1. Empirical Comparisons

Glimpse Mechanisms Under Partial Observability

We compare the glimpse attention strategy learned by GliTr with four baselines and an approximate upper bound:

- *Uniform random*: Glimpse locations are independently

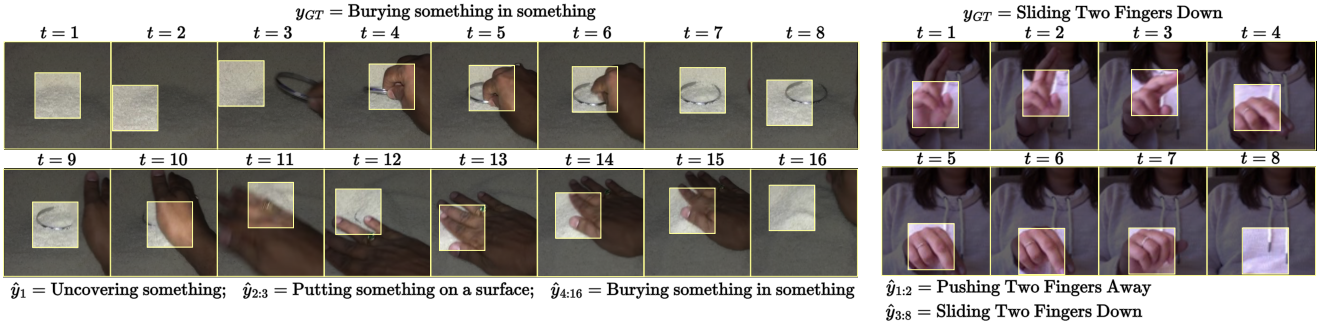


Figure 4: **Glimpses selected by GliTr** on (left) SSv2 and (right) Jester. The complete frames are shown for reference only. GliTr does not observe full frames. It only observes glimpses. We show additional examples in the supplementary material.

drawn from a uniform distribution for each t.

- *Gaussian random*: Similar to *uniform random* but instead, the glimpse locations are sampled from Gaussian distribution with zero mean and unit variance and passed through a $\tanh()$ function to constrain locations to remain within the bounds of the frame.
- *Center*: The model observes glimpses from a constant location at the center of each frame.
- *Bottom Left*: The model attends to the glimpses in the bottom left corner of the frames.
- *Teacher (an upper bound)*: Glimpse locations are chosen as predicted by the teacher model which looks at the full frames. In the absence of ground truth glimpse locations, this provides an approximate upper bound.

To isolate the glimpse strategy's effect on performance, we evaluate the glimpses selected by various strategies using the same model *i.e.* GliTr. While assessing the baselines and the upper bound, we ignore predictions from $\mathcal{T}_l$ and instead use locations given by the specific strategies described above. We show results in Figure 5, plotting online action prediction accuracy after each t. As expected, the prediction accuracy for all strategies increases as the model observes more glimpses. The *Center* and the *Bottom Left* strategies outperform other baselines on SSv2 and Jester datasets, respectively. We suspect this is because the object of interest frequently appears in the center in SSv2; while in most examples from Jester, hand movements begin and end in the region near the bottom left corner of the frames. On the other hand, GliTr outperforms all baselines and achieves performance closest to the upper bound (*i.e.* the *Teacher* strategy). We plot a histogram of glimpse regions selected by GliTr in Figure 6. We observe that not only does GliTr successfully capture different biases (center vs. bottom left) in the two datasets, but it also ignores the bias if necessary. Notice the spread in the histograms for $t > 1$, suggesting GliTr observes various regions in different videos. Consequently, GliTr achieves better accuracy faster than the baselines, and at time T, outperforms the best performing baselines with the respective margins of nearly 5% and 11% on SSv2 and Jester. We visualize glimpses selected by GliTr on example videos from SSv2 and Jester in Figure 4.

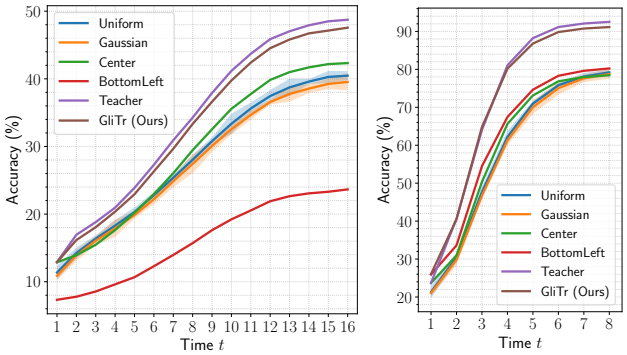


Figure 5: **Comparison of online action prediction accuracy using different glimpse mechanisms.** (a) SSv2 and (b) Jester. The *Uniform* and the *Gaussian* strategies sample locations from the respective distributions. We display mean$\pm 5\times$std computed using five independent runs. The *Center* and the *Bottom Left* strategies always observe glimpses at the constant locations. The *Teacher* (an approximate upper bound) and our GliTr locate informative glimpses based on past frames and glimpses, respectively.

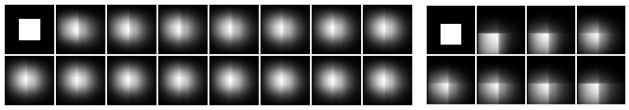


Figure 6: **Histograms of the glimpse regions selected by GliTr** with increasing time (raster scan order) on (a) SSv2 and (b) Jester. Recall that GliTr observes the first glimpse at a predetermined location followed by active selection.

Models with Complete Spatial Observability

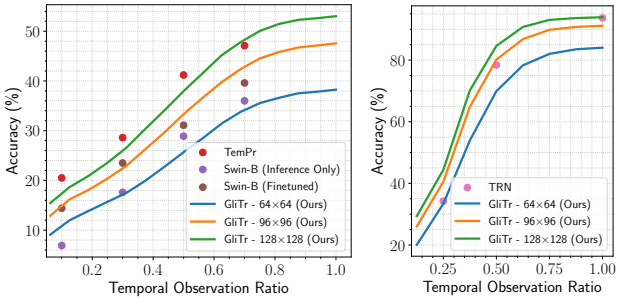
Glimpse-based offline models. We compare our GliTr with previous glimpse-based offline action recognition models in Table 1. We note that a direct comparison between these approaches is unfair since previous models

Method	Online/ Offline?	Observes full frames?	SSv2[22] Glimpse-size	#frames	#pixels	Accuracy (%)	Jester[41] Glimpse-size	#frames	#pixels	Accuracy (%)
AdaFocus [62]◇	Offline	Yes	144×144	8+12	1M	59.70	-	-	-	-
			160×160	8+12	1M	60.20	-	-	-	-
			176×176	8+12	1M	60.70	-	-	-	-
AdaFocusV2 [64]◇	Offline	Yes	128×128	8+12	1M	59.60	128×128	8+12	1M	96.60
			144×144	8+12	1M	60.50	176×176	8+12	1M	96.90
			160×160	8+12	1M	60.80	-	-	-	-
			176×176	8+12	1M	61.30	-	-	-	-
GFNet [25]§	Offline	Yes	96×96 (×2)⋆	8	401K	59.50	80×80 (×2)⋆	8	401K	95.50
			96×96 (×2)⋆	12	602K	61.00	96×96 (×2)⋆	12	602K	95.80
			96×96 (×2)⋆	16	803K	62.00	128×128 (×2)⋆	16	803K	96.10
GliTr (Ours)	**Online**	**No**	64×64	16	**66K**	38.24	64×64	8	**33K**	84.03
			96×96	16	**147K**	47.56	96×96	8	**74K**	91.15
			128×128	16	**262K**	53.02	128×128	8	**131K**	93.91

Table 1: **Comparison with glimpse-based action recognition models.** We count the number of pixels sensed by different approaches to perform recognition. Previous approaches are offline and use complete frames to locate informative glimpses and to recognize actions. GliTr is an online model and only observes glimpses, not complete frames. GliTr achieves competitive performance with a significant saving in the total area observed. ◇AdaFocus [62] and AdaFocusV2 [64] first observe 8 frames to locate useful glimpses and then sample additional 12 frames to extract glimpses, which requires sensing total 20 frames in advance due to their offline nature. §Results are based on Figure 13 from [25]. ⋆GFNet observes two glimpses per frame. For comparison with online methods, refer to Figure 7.

also observe complete frames. Further, unlike offline approaches that initially observe a complete video and select an informative glimpse at t based on the current, past, and future frames, our GliTr - an online model - relies only on the past information to locate glimpses in the current frame. Moreover, previous methods use global information gathered from complete frames to locate glimpses and predict actions; however, GliTr only uses local information. Nevertheless, we include this analysis to highlight the savings achieved by GliTr in terms of the amount of area observed for recognition while still achieving competitive performance with partial observations.

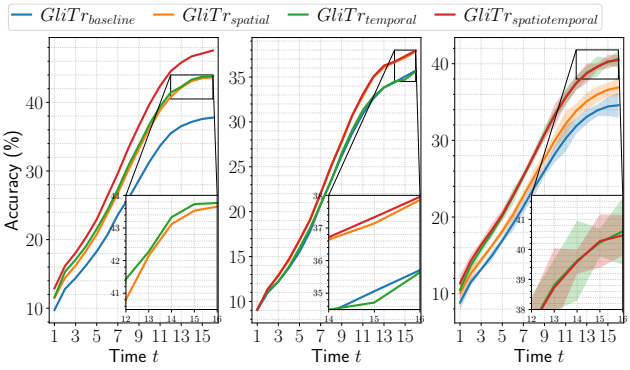
We calculate and compare the number of pixels sensed by various methods to perform action recognition. AdaFocus [62] and AdaFocusV2 [64] uniformly sample 8 frames from a complete video to predict glimpse locations, followed by uniform sampling of another 12 frames to extract glimpses. In total, they require sensing 20 complete frames ($20\times(224\times224)\approx$1M pixels) in advance due to their offline nature. GFNet [25], on the other hand, locates and extracts glimpses from the same set of complete frames. When compared to AdaFocusV2 with glimpses of size 128×128, our GliTr reduces the amount of sensing by nearly 74% and 87% while compromising only around 6% and 3% accuracy on SSv2 and Jester, respectively. Further, while GFNet outperforms GliTr by nearly 14.4% and 4.7% with glimpses of size 96×96 on SSv2 and Jester, GliTr (with 16 and 8 glimpses, respectively) reduces the amount of sensing by nearly 82% and 88% compared to GFNet (with 16 and 12 frames, respectively) on these datasets. We emphasize that GFNet observes full frames and *two* glimpses per frame in an offline manner, while GliTr observes only one glimpse per frame in an online fashion.



(a) (b)

Figure 7: **Comparison with early action prediction models.** (a) SSv2 and (b) Jester. While Swin-B [38], TemPr [52] and TRN [70] predict action early based on complete frames, GliTr predicts action based on early glimpses.

Early action prediction models. We additionally compare GliTr with early action prediction models in Figure 7. We emphasize that these approaches observe entire frames (*i.e.* global information) from a preliminary video; whereas, GliTr observes frames only partially through glimpses (*i.e.* local information). For SSv2 dataset, we consider Swin-B [38] and TemPr [52]. We cite Swin-B results from [52], who evaluate Swin-B for early action prediction before (*i.e.* direct inference with pretrained model) and after finetuning on preliminary videos. Notice that, with glimpses of size 96×96 and higher, GliTr outperforms Swin-B finetuned for early action prediction. Further, GliTr also outperforms TemPr with the glimpses of size 128×128 when both have observed early 70% video. For the Jester dataset, GliTr outperforms TRN [70] for early action prediction with glimpses of size 96×96 and higher. The results demonstrate the efficiency of GliTr for early action prediction using only local information.



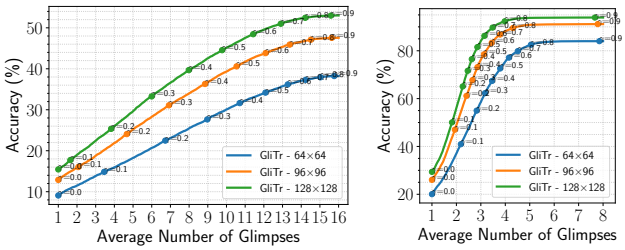
(a) (b) (c)

Figure 8: **Ablation study on the spatiotemporal consistency objective** on SSv2 dataset. (a) accuracy of GliTr when trained using different combinations of the training objectives (b) accuracy of the teacher with the glimpses selected by the above variants. (c) accuracy of the above variants of GliTr when tested with the *Uniform random* strategy. We display mean$\pm 5\times$std from five independent runs.

5.2. Ablation on Spatiotemporal Consistency

To demonstrate the value of the proposed spatiotemporal training objectives, we perform an ablation study for each on the SSv2 dataset. We train four variants of GliTr using the following combinations of the training objectives: i) GliTr$_{\text{baseline}}$ using $\widehat{\mathcal{L}}_{cls}$, ii) GliTr$_{\text{spatial}}$ using $\widehat{\mathcal{L}}_{cls} + \widehat{\mathcal{L}}_{spatial}$, iii) GliTr$_{\text{temporal}}$ using $\widehat{\mathcal{L}}_{cls} + \widehat{\mathcal{L}}_{temporal}$, and iv) our default variant GliTr$_{\text{spatiotemporal}}$ using $\widehat{\mathcal{L}}_{cls} + \widehat{\mathcal{L}}_{spatial} + \widehat{\mathcal{L}}_{temporal}$. Note that the above variants have the same architecture and operation; only their training objectives are different. Figure 8(a) shows results. We observe that including only spatial or only temporal consistency in the training objectives boosts GliTr's accuracy by nearly 6% at t=16. Moreover, including both spatial and temporal consistency provides the highest improvement of around 10%.

To understand the sources of improvements provided by the two consistency losses, we perform two more experiments. First, we evaluate glimpse selection strategies learnt by the above versions of GliTr using an impartial teacher model in Figure 8(b). We observe better performance for GliTr when spatial consistency is included in the training objectives, indicating that spatial consistency helps GliTr learn *better glimpse selection strategy* and in turn improves its performance. Second, we evaluate above four versions of GliTr using an impartial *Uniform random* strategy in Figure 8(c). We observe that GliTr provides the highest performance for the *Uniform random* strategy when we include temporal consistency in the training objective, suggesting that temporal consistency improves GliTr's performance by learning *a better classifier under partial observability*. We experiment with different training procedures for the teacher model in the supplementary material.



(a) (b)

Figure 9: **GliTr with early exit.** We display accuracy vs an average number of glimpses seen by GliTr per video to predict a class with probability $> \gamma$. (a) SSv2 and (b) Jester.

5.3. Early Exit

We extend GliTr for applications that require timely decision-making. We terminate sensing and conclude a class when GliTr makes a sufficiently confident prediction. We evaluate confidence using the maximum value in the predicted class logits, $\mathcal{C}_t = \max(p(\hat{y}_t))$ and exit when GliTr achieves confidence $\mathcal{C}_t > \gamma$. We show the performance of GliTr for varying γ in Figure 9. We observe a trade-off between the glimpse size and the average number of glimpses required for confident prediction. GliTr achieves higher confidence early with larger glimpse sizes and thus requires fewer glimpses to achieve certain performance.While continued sensing improves GliTr's performance on SSv2, the performance saturates on Jester after the initial 50% of the glimpses, rendering further sensing unnecessary.

6. Conclusions

We develop a novel online action prediction model called Glimpse Transformer (GliTr) that observes video frames only partially through glimpses and predicts an ongoing action solely based on spatially and temporally incomplete observations. It predicts an informative glimpse location for a current frame based on the glimpses observed in the past. Without any ground truth for the glimpse locations, we train GliTr using a novel spatiotemporal consistency objective. On the Something-Something-v2 (SSv2) dataset, the proposed consistency objective yields around 10% higher accuracy than the cross-entropy-based baseline objective. Further, we establish that spatial consistency helps GliTr learn a better glimpse selection strategy, whereas temporal consistency improves classification performance under partial observability. While never observing frames completely, GliTr achieves 53.02% and 93.91% accuracy on SSv2 and Jester datasets and reduces the sensing area per frame by $\sim 67\%$. Finally, we also showcase a trade-off between the glimpse size and the number of glimpses required for early action prediction. GliTr is useful for lightweight, low-cost devices with small field-of-view cameras.

References

[1] Gustavo Aguilar, Yuan Ling, Yu Zhang, Benjamin Yao, Xing Fan, and Chenlei Guo. Knowledge distillation from internal representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6836–6846, 2021.

[3] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. In *ICLR*, 2015.

[4] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in Neural Information Processing Systems*, 27:3365–3373, 2014.

[5] Fabien Baradel, Christian Wolf, and Julien Mille. Human action recognition: Pose-based attention draws focus to hands. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 604–613, 2017.

[6] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 469–478, 2018.

[7] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2019.

[8] Yijun Cai, Haoxin Li, Jian-Fang Hu, and Wei-Shi Zheng. Action knowledge transfer for action prediction with partial videos. In *Proceedings of the AAAI conference on artificial intelligence*, 2019.

[9] Yu Cao, Daniel Barrett, Andrei Barbu, Siddharth Narayanaswamy, Haonan Yu, Aaron Michaux, Yuewei Lin, Sven Dickinson, Jeffrey Mark Siskind, and Song Wang. Recognize human activities from partially observed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2658–2665, 2013.

[10] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[11] Guangyi Chen, Jiwen Lu, Ming Yang, and Jie Zhou. Learning recurrent 3d attention for video-based person reidentification. *IEEE Transactions on Image Processing*, 29:6963–6976, 2020.

[12] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32, 2019.

[13] Srijan Das, Arpit Chaudhary, Francois Bremond, and Monique Thonnat. Where to focus on for human action recognition? In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 71–80. IEEE, 2019.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[16] Gamaleldin Elsayed, Simon Kornblith, and Quoc V Le. Saccader: improving accuracy of hard attention models for vision. In *Advances in Neural Information Processing Systems*, pages 702–714, 2019.

[17] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020.

[18] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv preprint arXiv:2205.09113*, 2022.

[19] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[20] Basura Fernando and Samitha Herath. Anticipating human actions by correlating past with the future with jaccard similarity measures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13224–13233, 2021.

[21] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021.

[22] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[23] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[24] Jian-Fang Hu, Wei-Shi Zheng, Lianyang Ma, Gang Wang, Jianhuang Lai, and Jianguo Zhang. Early action prediction by soft regression. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2568–2583, 2018.

[25] Gao Huang, Yulin Wang, Kangchen Lv, Haojun Jiang, Wenhui Huang, Pengfei Qi, and Shiji Song. Glance and focus networks for dynamic visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2022.

[26] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[27] Dotan Kaufman, Gil Levi, Tal Hassner, and Lior Wolf. Temporal tessellation: A unified approach for video analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 94–104, 2017.

[28] Yu Kong and Yun Fu. Max-margin action prediction machine. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1844–1858, 2015.

[29] Yu Kong, Shangqian Gao, Bin Sun, and Yun Fu. Action prediction from videos via memorizing hard-to-predict samples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[30] Yu Kong, Dmitry Kit, and Yun Fu. A discriminative model with multiple temporal scales for action prediction. In *European conference on computer vision*, pages 596–611. Springer, 2014.

[31] Yu Kong, Zhiqiang Tao, and Yun Fu. Deep sequential context networks for action prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1481, 2017.

[32] Shaofan Lai, Wei-Shi Zheng, Jian-Fang Hu, and Jianguo Zhang. Global-local temporal saliency action prediction. *IEEE Transactions on Image Processing*, 27(5):2272–2285, 2017.

[33] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

[34] Tian Lan, Tsung-Chuan Chen, and Silvio Savarese. A hierarchical representation for future action prediction. In *European conference on computer vision*, pages 689–704. Springer, 2014.

[35] Kang Li and Yun Fu. Prediction of human activity by discovering temporal sequence patterns. *IEEE transactions on pattern analysis and machine intelligence*, 36(8):1644–1657, 2014.

[36] Kevin J Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and Lawrence Carin. Mixkd: Towards efficient distillation of large-scale language models. In *International Conference on Learning Representations*, 2021.

[37] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7083–7093, 2019.

[38] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022.

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[40] Khoi-Nguyen C Mac, Minh N Do, and Minh P Vo. Efficient human vision inspired action recognition using adaptive spatiotemporal sampling. *arXiv preprint arXiv:2207.05249*, 2022.

[41] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[42] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*, pages 2204–2212, 2014.

[43] Guoliang Pang, Xionghui Wang, Jianfang Hu, Qing Zhang, and Wei-Shi Zheng. Dbdnet: Learning bi-directional dynamics for early action prediction. In *IJCAI*, pages 897–903, 2019.

[44] Athanasios Papadopoulos, Pawel Korus, and Nasir Memon. Hard-attention for scalable image classification. *Advances in Neural Information Processing Systems*, 34:14694–14707, 2021.

[45] Hieu Pham, Zihang Dai, Qizhe Xie, and Quoc V Le. Meta pseudo labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11557–11568, 2021.

[46] Jie Qin, Li Liu, Ling Shao, Bingbing Ni, Chen Chen, Fumin Shen, and Yunhong Wang. Binary coding for partial action analysis with limited observation ratios. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2017.

[47] Samrudhdhi B. Rangrej and James J. Clark. A probabilistic hard attention model for sequentially observed scenes. *British Machine Vision Conference*, 2021.

[48] Samrudhdhi B Rangrej, Chetan L Srinidhi, and James J Clark. Consistency driven sequential transformers attention model for partially observable scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2518–2527, 2022.

[49] Michael S Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. In *2011 International Conference on Computer Vision*, pages 1036–1043. IEEE, 2011.

[50] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29:1163–1171, 2016.

[51] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

[52] Alexandros Stergiou and Dima Damen. Temporal progressive attention for early action prediction. *arXiv preprint arXiv:2204.13340*, 2022.

[53] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022.

[54] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357, 2021.

[55] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019.

[56] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal

convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[58] Gang Wang, Wenmin Wang, Jingzhuo Wang, and Yaohua Bu. Better deep visual attention with reinforcement learning in action recognition. In *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–4. IEEE, 2017.

[59] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[60] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[61] Xionghui Wang, Jian-Fang Hu, Jian-Huang Lai, Jianguo Zhang, and Wei-Shi Zheng. Progressive teacher-student learning for early action prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3556–3565, 2019.

[62] Yulin Wang, Zhaoxi Chen, Haojun Jiang, Shiji Song, Yizeng Han, and Gao Huang. Adaptive focus for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16249–16258, 2021.

[63] Yulin Wang, Kangchen Lv, Rui Huang, Shiji Song, Le Yang, and Gao Huang. Glance and focus: a dynamic approach to reducing spatial redundancy in image classification. *Advances in Neural Information Processing Systems*, 33, 2020.

[64] Yulin Wang, Yang Yue, Yuanze Lin, Haojun Jiang, Zihang Lai, Victor Kulikov, Nikita Orlov, Humphrey Shi, and Gao Huang. Adafocus v2: End-to-end training of spatial dynamic networks for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20062–20072, 2022.

[65] Xinxiao Wu, Jianwei Zhao, and Ruiqi Wang. Anticipating future relations via graph growing for action prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[66] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020.

[67] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020.

[68] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[69] He Zhao and Richard P Wildes. Spatiotemporal feature residual propagation for action prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7003–7012, 2019.

[70] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 803–818, 2018.

[71] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. Image bert pre-training with online tokenizer. In *International Conference on Learning Representations*, 2021.