SAFCAR: Structured Attention Fusion for Compositional Action Recognition

Tae Soo Kim   Gregpry D. Hager

Johns Hopkins University

3400 N. Charles St, Baltimore, MD

{tkim60,hager}@jhu.edu

Abstract

We present a general framework for compositional action recognition – i.e. action recognition where the labels are composed out of simpler components such as subjects, atomic-actions and objects. The main challenge in compositional action recognition is that there is a combinatorially large set of possible actions that can be composed using basic components. However, compositionality also provides a structure that can be exploited. To do so, we develop and test a novel Structured Attention Fusion (SAF) self-attention mechanism to combine information from object detections, which capture the time-series structure of an action, with visual cues that capture contextual information. We show that our approach recognizes novel verb-noun compositions more effectively than current state of the art systems, and it generalizes to unseen action categories quite efficiently from only a few labeled examples. We validate our approach on the challenging Something-Else tasks from the Something-Something-V2 dataset. We further show that our framework is flexible and can generalize to a new domain by showing competitive results on the Charades-Fewshot dataset.

1. Introduction

Descriptions of human action can be naturally composed from basic components such as subjects (nouns), atomic-actions (verbs) and objects. Automated recognition of human actions in video thus faces the fundamental challenge that the set of labels is combinatorially large. As a result, this makes enumeration of all possible descriptions to train end-to-end 3D convolutional methods [36, 6, 16, 59, 28] impractical in this domain. Instead, a model must have a means to represent and "assemble" the spatial-temporal structure of an action so that it can generalize to novel instances of actions composed with different and possibly previously unseen components.

Many methods have been proposed to model the spatial-temporal structure of actions using objects. In such methods, the structure of an action is approximated using variations of specified space-time graphs [39, 29], learned using a graph convolutional network [57] or manually specified by annotating the relations between objects [30]. However, such methods do not exhibit clear and obvious improvements unless ensembled with end-to-end RGB baselines [39] at test time or ground truth knowledge about actor-object relations is provided by an oracle [30]. Findings suggest that the structured information learned using objects and the visual information are not fused effectively.

In this paper, we build on the idea that action structure learning using object detections can be synergistically coupled with the rich visual cues available in video. This leads
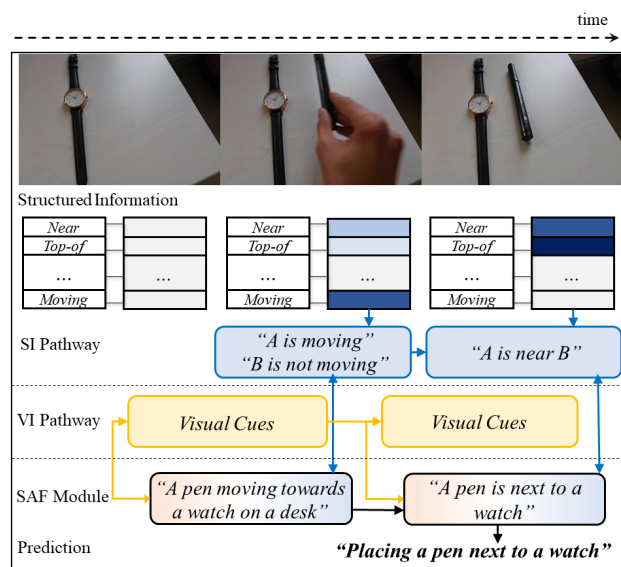

Figure 1: Our framework uses two pathways (Structured and Visual Information pathways) to learn structured components of an action. Structured Information (SI) pathway captures structure of an action using object detection. We show that the Structured Attention Fusion mechanism allows the rich information embedded in a video to refine the structured representation learned in the SI pathway.

to a generic framework for effectively recognizing compositional actions using two pathways. First, the Structured Information (SI) pathway explicitly reasons about the spatial-temporal structure of an action from object detections. Using a temporal model such as a temporal convolutional network [34, 32], the SI pathway learns to encode spatial relations between actors as well as consistent temporal patterns of those relations. Second, the Visual Information (VI) pathway extracts a time-series of visual features from the images comprising the video.

A common way to fuse object-level and visual information is via a naive concatenation or ensembling [39]. Instead, we introduce the Structured Attention Fusion (SAF) mechanism which allows both pathways to attend to each other and fuse the structural and visual information from the two pathways. We show that we can learn structures of an action more effectively by allowing both pathways to attend to each other to refine the fused representation. Figure 1 illustrates the interplay of the two pathways through the SAF module. For example, when only using 2D object detections as inputs to the SI-pathway, fine-grained spatial structure between actors such as whether an object is next to, inside or above another object may be ambiguous. By allowing the structured representation to attend to the visual cues from the VI-pathway, we can eliminate such ambiguities as illustrated in Figure 1.

In the remainder of this paper, we show that our approach significantly improves on current methods for compositional action recognition. We demonstrate that our approach can generalize over object appearance while preserving the core spatial-temporal structure of an action. With this ability, our approach can recognize unseen verb-noun compositions at test time better than other approaches and establishes a new state of the art on the Something-Else task [39] from the Something-Something-V2 dataset [21]. In addition to novel verb-noun compositions, we show that the learned representation generalizes to unseen action categories (verbs) efficiently using only a few labeled examples. We verify this by evaluating on the very challenging few-shot compositional split proposed by [39] where we again establish a new state of the art. Finally, we show that our model generalizes to a different domain of actions found in the Charades dataset [48] where our approach performs competitively on the Charades-Fewshot benchmark [30]. In summary, the main contributions of the paper are:

1. A general framework for compositional action recognition using multiple pathways with the SAF module.

2. A new state-of-the-art for recognizing compositional actions with unseen components.

3. A representation of an action that requires fewer training examples to adapt to unseen actions, validated

across multiple domains including the Something-Else and Charades-Fewshot tasks.

2. Related Work

Action Classification in Videos: With the introduction of large scale video datasets for action classification [14, 22, 31], many deep neural architectures have been proposed to extract powerful representations from videos [4, 6, 16, 17, 28, 36, 52, 59]. Most approaches for video classification pretrain such models on large-scale datasets and later finetune to target domains. However, the findings in [61] suggest that such pretrained models focus more on appearance rather than the temporal structure of actions. To evaluate whether a video model can generalize over appearance, researchers have proposed more fine-grained benchmarks and structured tasks [21, 30, 39, 48] including compositional action recognition. In this work, we focus on compositional and few-shot compositional action recognition tasks where a model must be able to generalize to actions defined by novel verb-noun compositions.

Few-shot Learning: The few-shot setting challenges a model's ability to generalize to unseen examples; we additionally validate our approach in the few-shot setting. Recent work [7] showed that a complex tasks/episodes-based meta-learning approach [3, 54, 49, 51, 44, 13] is not necessary when a good initialization of a model is combined with a distance based classifier [18, 15, 40, 7]. We demonstrate with our few-shot experiments that our approach leads to a good initialization such that it only requires few examples to transfer to novel action categories. We show in our experiments that our approach has learned structural components of an action which can adapt quickly to novel categories using only a handful of labeled instances.

Use of Attention: Attention mechanisms have been widely adopted in many static image-based [26, 27, 41, 58, 62] and video-based [12, 27, 8, 35, 43, 55, 3] classification problems. The role of the attention mechanism can be interpreted as a way to implicitly focus feature extraction on important aspects of an action. In existing methods, this process happens within a single pathway using only RGB input. We use the attention mechanism to fuse the information *between* pathways operating on different input modalities.

In particular, the recently introduced self-attention [53] mechanism captures long-range dependencies within data points; it has quickly become the core building block of state of the art approaches for multiple natural language modeling tasks [5, 11, 37, 60]. In the vision domain, recent papers have investigated how self-attention can generalize to the image domain [2, 45, 10] as well as to video-based applications [56]. The Non-local Neural Network [56] captures long-range dependencies within a video by use of a non-local operator which is a generalization of the

self-attention unit [53]. Our key observation is that the attention operation leads to a fusion of information whether it is within the model's spatial-temporal feature maps [56] or between long-term and short-term features [59] in a feature-bank framework. We investigate its use in fusing structural and visual information of an action in a multimodal framework.

Structured Representations of Videos using Objects: A growing line of work uses structured information extracted from videos, such as object detections and scene graphs, to improve fine-grained analysis of actions. The most common form of structure in action classification is obtained through detectable objects and actors in the scene [1, 29, 19, 59, 57, 38, 50, 30, 39]. Instead of learning features using only video, these approaches often combine features extracted from regions of interest using detections provided by state of the art object detectors [23]. The object centric representations can then be used to learn pairwise relations between objects [39, 38, 1], between objects and global context [59, 50, 19, 39] and within a specified graph structure [57, 30] to improve action classification. The object-centric information can then be used to predict action labels directly [1, 38, 50, 19, 39] via end-to-end training. However, such methods do not exhibit clear and obvious improvements unless ensembled with end-to-end RGB baselines at test time [39] or ground truth knowledge about actor-object relations is provided [30]. Findings suggest that the structured information learned using objects and the visual information are not fused effectively. We address this issue by using an attention based SAF module to fuse structural and visual information. This is similar to the recent Feature-Bank-Operator [59] where long-term ROI-aligned object-centric features [46] are merged with short-term visual features from 3D CNNs. However, the information being fused with attention is fundamentally different between the approaches. In [59], visual features extracted around humans or ROI-aligned object features are used directly without extracting high-level structure from them. We extract high-level action structures from objects using the SI pathway which get fused with the visual information.

Architecture: The well known two-stream architecture [17] has been modified and adopted in state-of-the-art video recognition models such as [6, 16]. Two stream architectures have shown to be effective when learning with optical-flow + RGB inputs [17, 6] and when merging RGB videos with different temporal speeds [16]. We adopt a two stream architecture that leverages object detection in one stream and raw pixels in the other pathway. We empirically show that fusing information between the two pathways through the SAF module greatly improves performance in compositional action recognition settings.

3. SAFCAR: Structured Attention Fusion for Compositional Action Recognition

A successful model for compositional action recognition must abstract out structure of an action from appearance of its components. To that end, we use a dedicated Structured Information (SI) (Sec. 3.2) pathway to explicitly learn the structure of an action using objects and a Visual Information (VI) (Sec. 3.1) pathway to learn useful complementary visual concepts from a video. Structured Attention Fusion (SAF) (Sec. 3.3) mechanism fuses the representations learned from the two pathways by allowing them to attend to each other. Figure 2a illustrates an example instantiation of the framework where object detections are used in the SI pathway.

3.1. Visual Information (VI) Pathway

Let $X \in \mathbb{R}^{T \times H \times W \times C}$ be a video with C channels with T frames with spatial dimension H and W. The VI pathway can be any feed-forward neural architecture V that has the following form:

$$V(X) = v_L(v_{L-1}(\ldots v_2(v_1(X))))\tag{1}$$

where the l-th intermediate representation is computed by sub-modules $v_{1:l}$ of the network and $V(X) \in \mathbb{R}^{T_V^L \times H_V^L \times W_V^L \times C_V^L}$. There are many well-established convolutional neural architectures that satisfy the above conditions [6, 36, 16, 17] and our general framework supports the use of any such models. For classification problems, a prediction $\hat{y}$ is made using a learned classification head $v_{cls} : \mathbb{R}^{C_V^L} \to \mathbb{R}^N$:

$$\hat{y} = \text{softmax}(v_{cls}(\sigma(V(X))))\tag{2}$$

where N is the number of classes and σ is an aggregation function such as a global pooling operator over both spatial and temporal dimensions of $V(X)$. When trained sufficiently, we know that $V(X)$ contains high-level visual features such as movement patterns, context and other visual cues useful for the action classification task.

3.2. Structured Information (SI) Pathway

The goal of the SI pathway is to extract structured information regarding the spatial-temporal components of an action. Let $Z(X) \in \mathbb{R}^{D \times T}$ be some structured representation of the video X. As illustrated in Figure 2a, a concrete example of $Z(X)$ may be a sequence of per-frame object detections provided by some object detector Z. Given a time-series of frame-level structural information, let S be a feed-forward neural architecture such that :

$$S(X) = s_L(s_{L-1}(\ldots s_2(s_1(Z(X)))))\tag{3}$$

where s_l is a submodule of S, $1 \leq l \leq L$ and $S(X) \in \mathbb{R}^{T_S^L \times C_S^L}$. Our framework does not require the depth of the

(a) An example instantiation of our framework. (b) Attention based SAF module.
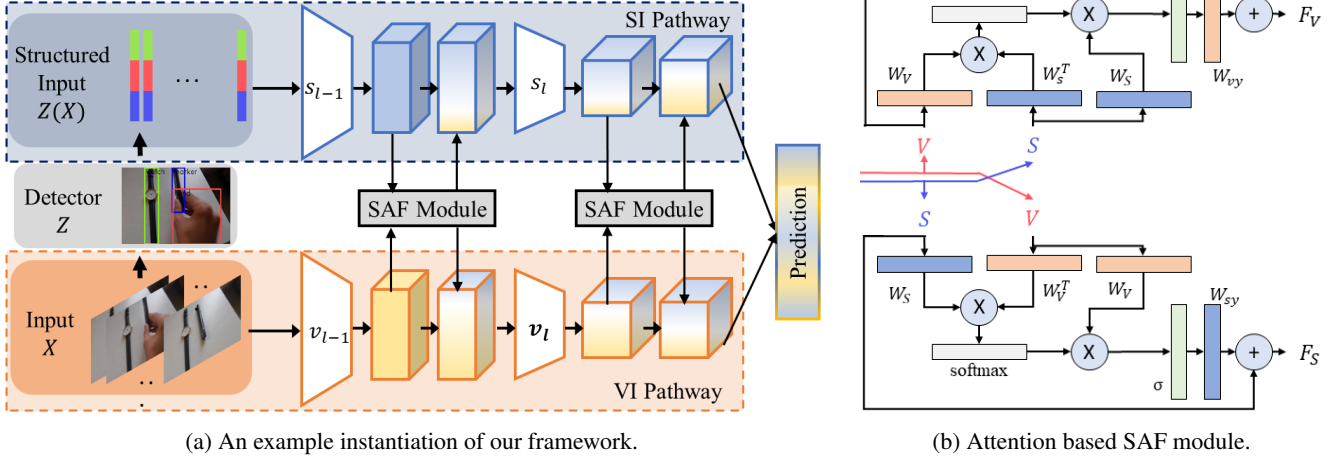
Figure 2: Figure 2a is an example instantiation of our framework which accepts objects detections and a video as inputs to their corresponding pathways. We plot the Structured Attention Fusion module in Figure 2b. The inputs to the module are intermediate representation S (Structured) and V (Visual) from the respective pathways.

V and S to be the same; it only requires that at least one component of s_l contains a neural operator with learnable parameters that performs temporal feature extraction. For example, each s_l can be implemented as a temporal convolution layer with 1D convolutions followed by a non-linear operation [33], a recurrent layer such as [25, 9], a self-attention based transformer encoder [53] or any mixture of such components. We show in our experiments that the SI-pathway alone can outperform complex video models. We describe our approach for fusing the representation learned using both pathways.

3.3. Structured Attention Fusion (SAF) Module

The objective of the SAF module is to fuse time-series structure of an action captured by the SI pathway and the visual cues from the VI pathway. We formulate the SAF module as a self-attention between the representations from the two pathways. The intuition is that the attention operation adds contextual information learned from video to structural information of an action extracted from objects. We also model the opposite direction of information flow such that the SAF module adds structural information of an action to the visual features.

Let the intermediate values $s_L(Z(X)) \in \mathbb{R}^{T_S^L \times C_S^L}$ and $v_L(X) \in \mathbb{R}^{D_v \times C_V^L}$ from both pathways be inputs to a SAF module. Note that $v_L(X)$ can be a tensor resulting from a flattening operation over the spatial-temporal dimensions such that $D_v = H_V^L W_V^L T_V^L$ or an output of some dimension reduction method (eg. pooling). The SAF operator produces two fused outputs F_S, F_V given $s_L(Z(X)), v_L(X)$. For notation clarity, we abbreviate the terms $s_L(Z(X)), v_L(X)$ as S, V from now on. We define

the SAF module as an attention operator between the two pathways. Let us first consider S attending to V denoted as $A_{S \to V}$:

$$A_{S \to V} = \text{softmax}\left(\frac{(SW_S)(W_V^T V^T)}{\sqrt{C}}\right)VW_V \quad (4)$$

where $W_V \in \mathbb{R}^{C_V^L \times C}$ and $W_S \in \mathbb{R}^{C_S^L \times C}$ are learnable parameters of the SAF module and C is a hyperparameter. Then, the output F_S is computed as:

$$F_S = \sigma(A_{S \to V})W_{sy} + S \quad (5)$$

where σ is a normalization operation followed by a non-linearity and $W_{sy} \in \mathbb{R}^{C \times C_S^L}$ is another learnable transformation. By allowing S to attend to V and adding the attended information back to itself, F_S contains information that is fused from both pathways and can be consumed by other operations in the SI pathway. Similarly, we compute F_V by allowing V to attend to S such that:

$$A_{V \to S} = \text{softmax}\left(\frac{(VW_V)(W_S^T S^T)}{\sqrt{C}}\right)SW_S$$
$$F_V = \sigma(A_{V \to S})W_{vy} + V \quad (6)$$

where $W_{vy} \in \mathbb{R}^{C \times C_V^L}$. A visual illustration of the SAF module is shown in Figure 2b. The SAF module can be inserted between the two pathways in different ways to encode different assumptions about the fused representation.

4. Experiments on Something-Else

We use the Something-Else [39] task for validating our approach for recognizing compositional actions. The

Something-Else task is an extension to the Something-Something-V2 [21] dataset with new object annotations and compositional action recognition splits. The dataset contains a total of 174 action categories where a crowd-sourced worker uploads a video capturing an arbitrary composition of an action category (verb) with an object (noun). As a result, the data set contains a very diverse set of verb-noun compositions involving 12,554 different object descriptions. In Sections 4.2 and 4.3, we discuss more in detail the compositional and fewshot compositional action recognition tasks.

We first provide details necessary for reproducing our results in Section 4.1. We then follow with our main experimental results from the compositional split in Section 4.2 and the few-shot split in Section 4.3. Finally, we perform thorough ablations on different components of the framework in Section 4.4.

4.1. Implementation Details

Our framework is general and can thus support most recent state of the art approaches for the VI and SI pathways. We detail our choices below.

VI Pathway. The VI pathways is based on the TSM [36] model because of its efficiency and validated performance on the Something-Something-V2 dataset. We use ResNet-50 [24] for the backbone and initialize with ImageNet [47] pretrained weights. As standard, we use a base learning rate of 0.01 and a batch size of 64 without a separate warm up schedule. We decrease the learning rate by a factor of 10 at epochs 20 and 40. Training is terminated at 50 epochs. We use stochastic gradient descent with a momentum term with weight 0.9 and a weight-decay of 0.0005. All settings are adjusted consistently as proposed in [20] when a change to the batch size was inevitable due to hardware limitations. All our experiments are implemented on the Pytorch framework [42] using 8 Titan Xp GPUs.

SI Pathway. The SI pathway is a flexible concept and we can implement it using any temporal model including a temporal conv-net (TCN) [32, 34], a recurrent architecture [9, 25] or a recent Transformer encoder model [53]. We use a TCN because it is efficient and easy to train. All training conditions are consistent with the VI pathway. We report ablations on the architectural design choices in Section 4.4. The details necessary for reproducing our results including SI pathway implementation can be found in the supplementary material.

SI Pathway Input. We use object detections to form the structured input $Z(X)$ for the SI pathway. Given the t-th frame of video X with N objects, we represent a frame as a concatenation of object coordinates such that $Z(X(t)) = (x_1^1, y_1^1, x_2^1, y_2^1, \ldots, x_1^N, y_2^N, x_2^N, y_2^N,) \in \mathbb{R}^{4N}$ where $(x_1^n, y_1^n, x_2^n, y_2^n)$ are the bounding-box coordinates of the n-th object. Then, the video is repre-

Model	Input		Evaluation	
	RGB	Objects	top-1	top-5
I3D [6]	o		46.8	72.2
TSM [36]	o		52.3	78.0
STIN [39]		o	51.4	79.3
STIN	o	o	54.6	79.4
STIN ensemble	o	o	58.1	83.2
Ours (SI Only)		o	53.8	79.8
Ours	o	o	**60.5**	**84.3**
Ours ensemble	o	o	**67.3**	**89.3**

Table 1: Results on compositional action recognition using **ground-truth** objects.

Model	Input		Evaluation	
	RGB	Objects	top-1	top-5
I3D [6]	o		46.8	72.2
TSM [36]	o		52.3	78.0
STIN [39]	o	o	48.2	72.6
STRG [57]	o	o	52.3	78.3
Ours	o	o	**60.7**	**84.2**
STIN ensemble [39]	o	o	51.5	77.1
STRG ensemble[57]	o	o	56.2	81.3
Ours ensemble	o	o	**65.3**	**87.9**

Table 2: Results on compositional action recognition using **detected** objects.

sented as a temporal concatenation such that $Z(X) = [Z(X(1)), \ldots, Z(X(T)] \in \mathbb{R}^{4N \times T}$

We construct $Z(X)$ using both ground-truth object locations and predictions obtained from a fine-tuned object detector [46]. For fair comparison, we use the same tracking results provided by the authors of [39] to establish temporal correspondence between object predictions.

4.2. Compositional Action Recognition

In the original action recognition split of Something-Something-V2, the same verb-noun composition may appear in both training and validation sets. We first describe the compositional action recognition split [39] that prevents this.

Problem formulation. Let there be two disjoint sets of nouns (objects), $\{\mathcal{A}, \mathcal{B}\}$, and two disjoint sets of verbs (actions) $\{1, 2\}$. The goal of the compositional action recognition task is to recognize novel verb-noun compositions at test time. The model can observe instances from the set $\{1\mathcal{A}+2\mathcal{B}\}$ during training but will be tested using instances from $\{1\mathcal{B} + 2\mathcal{A}\}$. In this setting, there are 174 action categories with 54,919 training and 57,876 validation instances.

Baselines. For RGB-only baselines, we use the popular I3D [6] model and the state-of-the-art (on Something-Something-V2) TSM [36]. For methods that use both

RGB and object bounding boxes, we compare to a very recent state-of-the-art Spatial-Temporal Interaction Network (STIN) [39] that models actions as sparse and semantically-rich object graphs. We compare to the Space-Time Region Graph (STRG) model as well when comparable results are available.

Results. In Table 1, we first report our experiments using ground truth object detections to isolate the effect of object detector performance from the effect of our approach. When accurate object locations are known, we show that simple baselines using only object-level information outperform video models. This shows that popular RGB-only baselines do not generalize over object appearance and/or fail to capture structural information in the scene, and thus does not perform well when presented with novel verb-noun compositions at test time.

When using videos in conjunction with object detections, our approach of fusing information from the SI and VI pathways using the SAF module improves the current state-of-the-art results by a considerable margin. As seen in Table 1, the benefit of using RGB video in addition to object detections is 3.2% for STIN whereas SAFCAR improves 6.7% – more than double the positive gain. Although SAFCAR and STIN perform similarly on the top-5 evaluation (79.8% vs. 79.3%) when only using object detections, there is a significant performance gain of 4.5% when visual information is fused using our approach. This compares to a minimal improvement of 0.1% for STIN. We show that a simple ensemble of predictions (combining RGB-only, an object-only and a RGB+object model, all independently trained) leads to additional performance gains where we establish a strong new state-of-the-art performance of 67.3% top-1 (an increase of 9.2 points) and 89.3% top-5 (an increase of 6.1 points) accuracies.

Table 2 shows action recognition performance on the same compositional split when using detected objects instead of ground truth annotations. While all methods perform worse with object predictions, the overall findings are consistent with the experiments reported in Table 1 – namely, our approach outperforms other methods whether ensemble-ed with baselines or not.

4.3. Few-shot Compositional Action Recognition

Problem formulation. In the compositional split, a common set of all 174 action categories appear during training and testing of the model. In the *few-shot* compositional split, the original 174 action categories are divided into a *base* set with 88 categories and a *novel* set with 86 action categories. The *base* set contains 112,397 labeled training instances where the model is 'pretrained'. Then, we only use few labeled examples (fewshot-train) from each novel category to 'finetune' the model to recognize rest of the videos from the novel action categories (fewshot-

Model	Input		Evaluation	
	RGB	Objects	5-shot	10-shot
I3D [6]	o		21.8	26.7
TSM [36]	o		22.5	27.3
STIN [39]		o	**27.7**	**33.5**
STIN	o	o	28.1	33.6
STIN ensemble	o	o	34.0	40.6
Ours (SI only)		o	27.0	32.8
Ours	o	o	**31.5**	**36.8**
Ours ensemble	o	o	**39.0**	**45.4**

Table 3: Results on few-shot compositional action recognition using **ground-truth** objects

Model	Input		Evaluation	
	RGB	Objects	5-shot	10-shot
I3D [6]	o		21.8	26.7
TSM [36]	o		22.5	27.3
STIN [39]		o	17.7	20.7
Ours (SI only)		o	15.8	19.8
STIN	o	o	23.7	27.0
STRG [57]	o	o	24.8	29.9
Ours	o	o	**28.5**	**32.0**
STIN ensemble	o	o	27.3	32.6
STRG ensemble	o	o	29.1	34.6
Ours ensemble	o	o	**37.0**	**43.4**

Table 4: Results on few-shot compositional action recognition using **detected** objects

validation). For example, when performing 5-shot transfer, only $5 \times 86 = 430$ labeled instances are available during the finetuning stage. As with the compositional split, the objects that appear in fewshot-train are independent from the objects that appear in the fewshot-validation. There are 49,822 and 43,957 instances in the 5-shot and 10-shot validation sets respectively. We pretrain a model using the base categories and freeze the trained parameters when finetuning with the few-shot examples.

Results For a model to perform well under this setting, it must learn re-usable spatial-temporal structures during pretraining with available aciton categories such that it can quickly adapt to unseen actions (verbs) during finetuning. In Table 3, we compare our approach to other state of the art methods for this task that use ground-truth object bounding boxes as part of the input. We observe that both RGB only baselines (I3D,TSM) perform poorly which further underscores that they struggle to generalize over appearance of objects when given novel verb-noun compositions. Compared to the compositional split results in Table 1 where RGB-only baselines performed on par with object-only models, RGB-only baselines actually perform worse than models that only use object level information in the

few-shot setting.

The few-shot experiments clearly highlight the benefit of our approach in making a better use of the visual information when learning with objects. Compared to other state of the art baselines, the performance boost induced by the visual content is large. The object-only baselines of both our approach and that of the STIN model perform similarly: 27.7% vs. 27.0% for the 5-shot and 33.5% vs. 32.8% for the 10-shot splits. When the models are allowed to observe RGB videos, the STIN model gains 0.4 for the 5-shot setting and 0.1 for the 10-shot setting. In contrast, we gain 4.5 and 4.0 points respectively. This is a strong evidence that our approach fuses visual information more effectively.

We compare results when using object detections instead of ground truth object annotations in Table 4. Our SI only model is actually weaker than other state of the art baselines. However, when using video data in addition to object detection, we observe a significant improvements in performance of 12.7%. In comparison, the STIN model gains only 6%.

4.4. Ablations

We perform ablation studies on the few-shot compositional split of the Something-Else task.

SAF architectures: Table 5a shows various ways of using the SAF module. A simple concatenation of the outputs from both pathways improves the performance of the VI-only baseline. However, naive concatenation actually performs worse than the model using only object detections (SI-only baseline). We ablate our approach with three different ways of using the SAF module as illustrated in Figure 3. The Late-SAF version inserts a SAF module between the final feature extraction layers of both pathways to fuse high-level information. The Cascaded-SAF version allows the high-level structural information to affect the lower level visual feature learning process. Finally, Sequential-SAF fuses intermediate layers sequentially. Table 5a shows that all variations that use the SAF module perform better than when the information from both pathways are naively merged. We also observe that fusing high-level information (Late-SAF) is more effective than other variants.

SI pathway architectures: Table 5b shows different SI pathway implementations. All versions are based on 1D convolution layers with filter length 3 and a stride of 1. Models v0 and v1 share identical temporal convolutional layers but v1 contains extra multi-headed attention transformer encoder [53] layers before the prediction head. Model v2 is a purely convolutional architecture with residual connections [24, 32]. The architecture of v3 is the same as v2 but with more convolutional filters per layer to contain almost double the number of trainable parameters. We see that adding parameters improves performance when only using objects as input. However, in the next ablation study,

	5-shot	10-shot
Baseline (VI only)	22.5	27.3
Baseline (SI only)	27.0	32.8
Concat(SI,VI) (No-SAF)	25.6	32.1
Late-SAF	**31.5**	**36.8**
Cascaded-SAF	29.9	34.8
Sequential-SAF	30.4	35.6

(a) Different ways to use the SAF module.

SI Pathway Model	FS-5	FS-10	# Params
v0	26.4	31.2	5.19M
v1	27.0	32.8	5.79M
v2	29.3	36.0	5.27M
v3	29.7	36.9	9.23M

(b) Different SI pathway instantiations using only ground truth detections.

Combination	FS-5	FS-10
v0 + Late-SAF	28.2	33.7
v1 + Late-SAF	**31.5**	**36.8**
v2 + Late-SAF	31.1	36.1
v3 + Late-SAF	29.4	34.3

(c) Different SI pathway instantiations combined with the the same VI pathway.

Table 5: Ablations using the few-shot compositional split.

we show that increasing the complexity of the SI pathway does not necessarily lead to improved few-shot performance when combined with the VI pathway.

SI+VI pathway combinations: Table 5c shows the performance of different SI pathway implementations when combined with a constant VI implementation. We find that adding the transformer encoder layers (v0 vs. v1) in the SI pathway improves performance when fused with the VI pathway. From Table 5b, we saw that v2 outperforms v1 when only object input is used. However, we find that when combined with the VI pathway, both models perform similarly. Lastly, we begin to see signs of overfitting with v3.

5. Experiments on Charades-Fewshot

To demonstrate that our framework for recognizing compositional actions is flexible, we evaluate our method on the Charades [48] dataset which contains 9,848 videos and 157 action categories with an average length of 30 seconds which may contain one or more action instances. Many action categories found in the Charades dataset are defined as a verb+noun composition (ie. Opening a book, Opening a door, Throwing a book). The task is to predict all the actions contained in the video without having to predict the
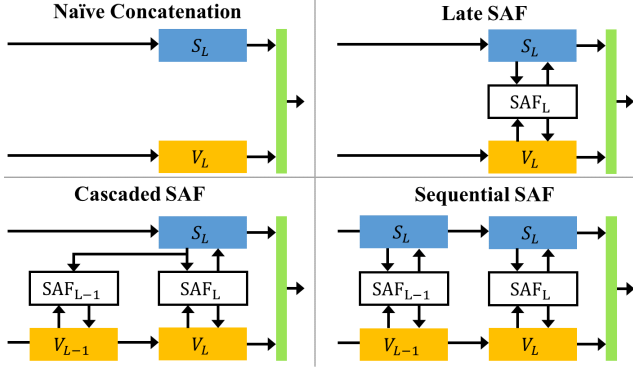
Figure 3: A visual comparison of different ways the SAF module can be inserted between the SI and VI pathways.

	1-shot	5-shot	10shot
LFB [59]	28.3	36.3	39.6
SGFB [30]	28.8	37.9	42.7
SGFB-oracle [30]	**30.4**	40.2	50.5
Ours, VI-only (SlowFast [16])	18.9	31.1	36.4
Ours	21.8	**41.8**	**50.9**

Table 6: Results on the Charades-Fewshot task.

start and end frames of the instances. Despite the different set of activities and objects, we show that our approach generalizes effectively across domains and show convincing few-shot compositional action recognition capabilities on the Charades-Fewshot task [30].

5.1. Implementation Details

Our goal with the Charades dataset is to demonstrate the flexibility of our framework when applied to a different problem domain. Hence, we do not perform any architecture search and simply use the same SI-pathway architecture (v1) used in Section 4.1. For the VI pathway, we use the SlowFast [16] model given its competitive performance on the Charades dataset. We finetune the model with ResNet-50 backbone pretrained on Kinetics-400 [31]. Similar to the Something-Else experiments, the input to the SI pathway is a time-series of ground-truth object locations. There are 36 annotated objects such that $Z(X) \in \mathbb{R}^{36*4 \times T}$. Detailed parameters necessary for reproducing our results are included in the supplementary material.

5.2. Few-shot Action Recognition

Problem Formulation The Charades dataset contains 37 annotated objects. Following the Charades-Fewshot setup [30], six 'novel' objects (pillow, broom, television, bed, vacuum, refrigerator) are excluded during training. There are 20 action categories that involve an interaction with an object in the novel-object set. Therefore, there are 137 base

action categories where we pretrain our model and 20 novel actions that are held out. From the 20 novel categories, a fixed set of few (k=1,5,10) examples are sampled per category to form the fewshot-train set where our model is finetuned. The remaining instances are used for few-shot validation.

Baselines. We compare our approach to recent state-of-the-art methods that reported results on the Charades-Fewshot task, including the Long-term Feature Banks (LFB) [59] and Scene-Graph Feature Banks (SGFB) [30]. LFB extracts ROI-aligned features using object detection as the feature bank representation which gets merged with a video feature for action prediction. SGFB uses additional manual annotations collected using the Action Genome [30] framework which contains ground truth relations between objects and attribute level information. The annotations include relational attributes such as 'in-front-of', 'is-looked-at', 'being-touched' and more. The annotated spatio-temporal scene graphs are provided to the SGFB model as an additional supervision during training. The SGFB-oracle model assumes correct spatial-temporal scene graphs are available during test time.

Results. We note that all state of the art models for this task use additional information such as objects (LFB) or annotated scene-graphs (SGFB). As expected, our baseline approach using only RGB videos trails other methods in all few-shot settings. Consistent with our findings in the Something-Else task in the previous sections, we show strong results when RGB information is combined with the SI-pathway using the SAF module. We show large improvements over our baseline: we gain 2.9, 10.7 and 14.5 points in 1-shot, 5-shot and 10-shot experiments respectively. We observe that our approach does not sufficiently model the high-level structure of an action using only one example per category when compared to other state of the art baselines. In the extreme 1-shot case, the SGFB-oracle model which assumes knowledge of manually annotated object-level attributes outperforms other methods. When enough examples are available (K ≥ 5), we empirically show our approach can even outperform the SGFB-oracle model. This finding suggests that the SI-pathway of our approach implicitly learns similar high-level concepts that are defined and annotated in the scene-graphs representation of the SGFB-oracle model.

6. Conclusion

We have presented a general framework that uses Structured and Visual pathways with the SAF module. Using object detections as input to the SI pathway, our approach modeled structures of actions better when attending to visual features in the VI pathway via the SAF mechanism. We validate our approach on multiple tasks including the Something-Else (compositional, few-shot compo-

sitional splits) and the Charades-Fewshot tasks where we significantly improved over the current state of the art.

Acknowledgement

References

[1] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *ECCV*, 2018. 4323

[2] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens. Attention augmented convolutional networks. In *ICCV*, pages 3285–3294, 2019. 4322

[3] Mina Bishay, Georgios Zoumpourlis, and I. Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. In *BMVC*, 2019. 4322

[4] Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *International Conference on Communication, Information and Computing Technology (ICCICT)*, pages 4489–4497, 12 2015. 4322

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arxiv*, 2020. 4322

[6] J. Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 4321, 4322, 4323, 4325, 4326

[7] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019. 4322

[8] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. Aˆ2-nets: Double attention networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 352–361. Curran Associates, Inc., 2018. 4322

[9] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014. 4324, 4325

[10] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *International Conference on Learning Representations*, 2020. 4322

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 4322

[12] Yang Du, Chunfeng Yuan, Bing Li, Lili Zhao, Yangxi Li, and Weiming Hu. Interaction-aware spatio-temporal pyramid attention networks for action classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 4322

[13] S. K. Dwivedi, V. Gupta, Rudrajit Mitra, S. Ahmed, and Arjun Jain. Protogan: Towards few shot learning for action recognition. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1308–1316, 2019. 4322

[14] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. 4322

[15] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611, Apr. 2006. 4322

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *International Conference on Computer Vision*, pages 6202–6211, 2018. 4321, 4322, 4323, 4328

[17] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016*, pages 1933–1941, 6 2016. 4322, 4323

[18] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In Doina Precup and Yee Whye Teh, editors, *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. 4322

[19] Rohit Girdhar, João Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *CVPR*, 2019. 4323

[20] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arxiv*, 06 2017. 4325

[21] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic. The "something something" video database for learning and evaluating visual common sense. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5843–5851, 2017. 4322, 4325

[22] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijaya-

narasimhan, George Toderici, Susanna Ricco, Rahul Suk-thankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4322

[23] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 4323

[24] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4325, 4327

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 4324, 4325

[26] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Andrea Vedaldi. Gather-excite: Exploiting feature context in convolutional neural networks. In *NeurIPS*, 2018. 4322

[27] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):2011–2023, 2020. 4322

[28] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *CVPR*, 2019. 4321, 4322

[29] Ashesh Jain, A. Zamir, S. Savarese, and A. Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5308–5317, 2016. 4321, 4323

[30] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4321, 4322, 4323, 4328

[31] W. Kay, J. Carreira, K. Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, F. Viola, T. Green, T. Back, A. Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *ArXiv*, abs/1705.06950, 2017. 4322, 4328

[32] T. S. Kim and A. Reiter. Interpretable 3d human action analysis with temporal convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1623–1631, 2017. 4322, 4325, 4327

[33] C. Lea, M. Flynn, R. Vidal, A. Reiter, and Gregory Hager. Temporal convolutional networks for action segmentation and detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1003–1012, 2017. 4324

[34] Colin Lea, René Vidal, Austin Reiter, and Gregory Hager. Temporal convolutional networks: A unified approach to action segmentation. In *Computer Vision – ECCV 2016 Workshops, Proceedings*, volume 9915 LNCS of *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 47–54. Springer Verlag, 2016. 14th European Conference on Computer Vision, ECCV 2016 ; Conference date: 08-10-2016 Through 16-10-2016. 4322, 4325

[35] Jun Li, Xianglong Liu, Wenxuan Zhang, Mingyuan Zhang, Jingkuan Song, and Nicu Sebe. Spatio-temporal attention networks for action recognition and detection. *IEEE Transactions on Multimedia*, PP:1–1, 01 2020. 4322

[36] Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. *CoRR*, abs/1811.08383, 2018. 4321, 4322, 4323, 4325, 4326

[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. 4322

[38] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan Alregib, and H.P. Graf. Attend and interact: Higher-order object interactions for video understanding. In *CVPR*, 2018. 4323

[39] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4321, 4322, 4323, 4324, 4325, 4326

[40] A. Mishra, V. K. Verma, M. S. K. Reddy, A. S., P. Rai, and A. Mittal. A generative approach to zero-shot and few-shot action recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 372–380, 2018. 4322

[41] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 4322

[42] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 4325

[43] Juan-Manuel Perez-Rua, Brais Martinez, Xiatian Zhu, Antoine Toisoul, Victor Escorcia, and Tao Xiang. Knowing what, where and when to look: Efficient video action modeling with attention, 2020. 4322

[44] Hang Qi, David Lowe, and Matthew Brown. Low-shot learning with imprinted weights. In *CVPR*, 2018. 4322

[45] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jon Shlens. Stand-alone self-attention in vision models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 68–80. Curran Associates, Inc., 2019. 4322

[46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M.

Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 91–99. Curran Associates, Inc., 2015. 4323, 4325

[47] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 4325

[48] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, 2016. 4322, 4327

[49] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4077–4087. Curran Associates, Inc., 2017. 4322

[50] C. Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, R. Sukthankar, and C. Schmid. Actor-centric relation network. In *ECCV*, 2018. 4323

[51] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. S. Torr, and T. M. Hospedales. Learning to compare: Relation network for few-shot learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1199–1208, 2018. 4322

[52] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 International Conference on Computer Vision, ICCV 2015*, Proceedings of the IEEE International Conference on Computer Vision, pages 4489–4497. Institute of Electrical and Electronics Engineers Inc., Feb. 2015. 15th IEEE International Conference on Computer Vision, ICCV 2015 ; Conference date: 11-12-2015 Through 18-12-2015. 4322

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc., 2017. 4322, 4323, 4324, 4325, 4327

[54] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3630–3638. Curran Associates, Inc., 2016. 4322

[55] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4322

[56] 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 (CVPR)*, June 2018. 4322, 4323

[57] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 4321, 4323, 4325, 4326

[58] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 4322

[59] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross Girshick. Long-Term Feature Banks for Detailed Video Understanding. In *CVPR*, 2019. 4321, 4322, 4323, 4328

[60] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 5753–5763. Curran Associates, Inc., 2019. 4322

[61] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *ECCV, Germany, September 8-14, 2018, Proceedings, Part I*, volume 11205 of *Lecture Notes in Computer Science*, pages 831–846. Springer, 2018. 4322

[62] Gao Zilin, Xie Jiangtao, Wang Qilong, and Li Peihua. Global second-order pooling convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4322