Revisiting spatio-temporal layouts for compositional action recognition

Gorjan Radevski[12]
gorjan.radevski@esat.kuleuven.be
Marie-Francine Moens[2]
sien.moens@cs.kuleuven.be
Tinne Tuytelaars[1]
tinne.tuytelaars@esat.kuleuven.be

[1] PSI-ESAT,
KU Leuven,
Leuven, Belgium

[2] LIIR-CS Department,
KU Leuven,
Leuven, Belgium

Abstract

Recognizing human actions is fundamentally a spatio-temporal reasoning problem, and should be, at least to some extent, invariant to the appearance of the human and the objects involved. Motivated by this hypothesis, in this work, we take an object-centric approach to action recognition. Multiple works have studied this setting before, yet it remains unclear (i) how well a carefully crafted, spatio-temporal layout-based method can recognize human actions, and (ii) how, and when, to fuse the information from layout- and appearance-based models. The main focus of this paper is compositional/few-shot action recognition, where we advocate the usage of multi-head attention (proven to be effective for spatial reasoning) over spatio-temporal layouts, i.e., configurations of object bounding boxes. We evaluate different schemes to inject video appearance information to the system, and benchmark our approach on background cluttered action recognition. On the Something-Else and Action Genome datasets, we demonstrate (i) how to extend multi-head attention for spatio-temporal layout-based action recognition, (ii) how to improve the performance of appearance-based models by fusion with layout-based models, (iii) that even on non-compositional background-cluttered video datasets, a fusion between layout- and appearance-based models improves the performance.

1 Introduction

Whether a person is "taking an apple out of a box" or "taking a screwdriver out of a box", we can recognize the action performed with ease. In fact, even if we have never seen the object before, we are still able to recognize the action that occurred. Moreover, for us, it makes no difference where the action takes place (indoors, outdoors, etc.), as long as the objects involved in the action are visible. This suggests that action recognition should be, to a degree, invariant to the appearance of the objects, as well as the environment where the action takes place. Yet, most state-of-the-art action recognition methods are appearance-based 3D CNNs [7, 26, 52, 55]. These methods are indeed powerful, albeit heavily reliant on large-scale (pre)training datasets [21]. Unfortunately, in spite of the great pre-training efforts taken, their performance rapidly deteriorates on compositional action recognition, i.e., when the objects encountered at test time are novel [30].

For this reason, multiple other works [4, 15, 41, 49, 57] have advocated an object-centric approach for video action recognition, reporting an improved robustness, interpretability, and overall performance. To achieve object-centric reasoning, on top of the video appearance (RGB frames), these works either use a region proposal network [4, 15, 41, 49], or leverage an independently trained object detector, e.g., Faster R-CNN [37], to obtain object detections for each video frame [30, 54]. Within these models, the *layout* module operates on object detections, while the *appearance* module operates on RGB frames. In a *two-branch model* the layout and appearance input follow separate pathways and are fused late, while in a *one-branch model* they follow a single pathway (fused early in the model). Some of the limitations include: (i) With a two-branch model, fusion is performed by concatenation, not fully exploiting the complementarity of the spatio-temporal layouts and the video appearance [4, 30, 49], and (ii) the layout module is treated as a peripheral component [20, 49], so it remains unclear to what extent in different evaluation settings (compositional, few-shot, background cluttered videos), a well assembled layout-based model can recognize human actions. At the same time, a multi-head attention model [45] has been demonstrated to be a powerful common-sense reasoning tool over sets of spatially distributed objects in images for visual question-answering [28, 43], layout generation [36], etc. By applying multiple heads of beyond-pairwise spatial reasoning, it encapsulates the scene's global spatial context, which is indicative of its semantics, to a certain extent. Just as importantly, a variety of works specifically examine the problem of multimodal fusion [33, 34, 46], attempting to determine how and where to fuse the different modalities.

Contributions. The main focus of this paper is compositional and few-shot action recognition, where we hold on to the object-level video reasoning and (i) reveal how a multi-head attention based method, applied purely over highly abstract concepts (no appearance information), i.e., spatio-temporal layouts, can be extended for action recognition, (ii) investigate how to fuse the information from the layout- and appearance-based branch for improved action recognition, (iii) find that, even on non-compostional, background cluttered video dataset such as Action Genome [20], reasoning over the spatio-temporal layouts significantly improves the performance. The codebase and trained models are released here[1].

2 Related work

Action recognition methods are mostly 3D CNN based [7, 14, 21, 26, 39, 39, 48, 52, 55]. These methods often use a 2D CNN pre-trained on ImageNet [12], subsequently inflated to 3D [7]. Other works explore (pre)training 3D CNNs [21] on large-scale curated datasets [8, 32], as well as the best practices for doing so [48], reducing the computational complexity [26, 44, 52], or propose plug-in components to improve the temporal reasoning [58].

Multi-head attention (MHA) in computer vision. The applications of MHA in computer vision are rapidly expanding [22]. So far, MHA has been applied in conjunction with a CNN [6, 15, 31, 42], as a stand-alone MHA over low level, raw image pixels [5, 10, 13, 47], for vision + text tasks [28, 43], or tasks involving spatial reasoning [36] to name a few. In contrast, we apply MHA: (i) over high level, abstract, spatio-temporal layouts, and (ii) to fuse the features of two distinct modalities (layout and appearance).

Object level reasoning for action recognition (with attention). The issue with appearance-based action recognition methods is their inherent tendency to overfit on the appearance

[1] https://github.com/gorjanradevski/revisiting-spatial-temporal-layouts

of the environment and of the objects, deteriorating the performance on fine-grained [4] or compositional datasets [30]. Recently, multiple works that address this issue emerged [9, 15, 20, 30, 30, 35, 41, 49, 53, 54]. Object Relation Network [9] performs spatio-temporal reasoning over detected video objects with a GRU [9]. STAR [53] regresses the bounding box coordinates where the action occurs and classifies the action. STRG [49] uses a graph CNN [23] and a region proposal network (RPN) [37], applied over I3D [7] features, with a non-local neural network (NL) [50] temporal module. Actor Centric Relation Network [41] fuses the cropped feature map of the actors' regions and the global video feature map. In parallel, multiple works leverage attention, to augment existing methods or propose new ones. LFB [51] uses attention as a non-local block [50] to accumulate video features. SINet's [29] coarse- and fine-grained branch are attention-based, subsequently fused for action recognition. Compared to us, SINet's fine-grained branch applies attention over the region of interest (RoI) pooled features from an RPN for each frame, subsequently fed to LSTM [19], while we apply MHA over the object detections (category + bounding box) to encode the videos' spatio-temporal context. W3 [35] is an attention based plug-in module on top of appearance models, while SGFB [20] utilizes attention through LFB [51] over per-frame scene graphs, combined with I3D [7] and NL [50]. The Video Action Transformer (VAT) [15] uses I3D [7] in conjunction with RPN [37] and a transformer [45]. It (i) obtains an I3D feature map around a center frame, (ii) generates region proposals for the center frame, (iii) applies MHA where the I3D feature map is the memory and the center frame RoI pooled features are the query. In our work, we also benchmark a VAT inspired fusion scheme between the appearance and layout branch. Lastly, STIN [30] and SFI [54] demonstrate that a graph neural network [23] layout model can surpass I3D's [7] performance for compositional/few-shot action recognition with ground truth object detections, and fusion with I3D improves performance. Unlike these works, inspired by MHA-based methods for spatial reasoning, we (i) develop a specifically tailored model for layout-based action recognition which applies attention over high-level, spatio-temporal layouts, (ii) empirically evaluate different state-of-the-art MHA-based fusion methods, to uncover how and when the appearance- and layout-based models should be fused.

Multimodal fusion attempts to extract the relevant, complementary information from multimodal input, resulting in a better joint model, compared to training separate models on the individual modalities. The literature is extensive [2, 24, 33, 46], without a universal approach that generalizes across different modalities and tasks. Many works [28, 43] that use MHA have demonstrated remarkable results on multimodal tasks, e.g., VQA [1], when fusing image and text features. In action recognition, late fusion by concatenation works well [30], also confirmed in our work. We demonstrate that multimodal fusion with cross-attention [43], based on the CentralNet approach [46], further improves the performance.

3 Methodology

Next, we introduce the necessary multi-head attention background (Sec. 3.1), describe how it is extended for modelling spatio-temporal layouts, i.e., object detections (Sec. 3.2), and discuss the appearance models and schemes to fuse them with the layout model (Sec. 3.3).

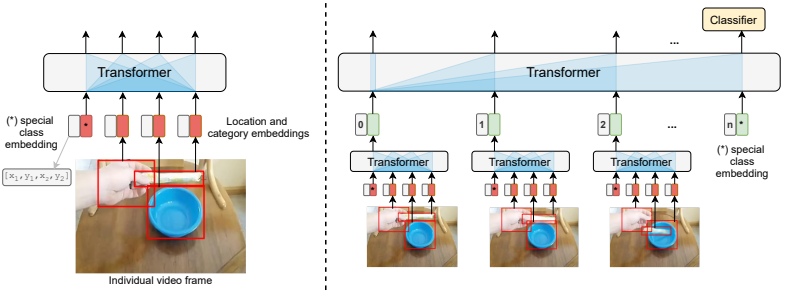

Figure 1: STLT overview. **Left: Spatial Transformer**. The inputs are the object categories, and their $[x_1, y_1, x_2, y_2]$ frame location normalized by the frame size. We select the special `class` embedding as the module output. **Right: Temporal Transformer**. The inputs are the Spatial Transformer outputs, summed with trainable position embeddings. We select the temporal `class` embedding as the module output, and add a classifier for action recognition.

3.1 Multi-head attention revisited

The core Transformer model component [45] is the multi-head attention module, defined as: $\text{Attention}(Q,K,V) = \text{Softmax}(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V$, where Q, K, V are the queries, keys and values respectively, and d_k is the per-attention head hidden size. With self-attention, Q, K and V come from the same and only modality (in our case, the layout modality), and with cross-attention, Q originates from the target, while K and V from the source modality we attend on. Due to MHA's permutation invariance, the input should include positional information. In this work, we rely on (i) bidirectional and causal attention, (ii) self- and cross-attention, and (iii) different variants of position embeddings. Refer to [45] for details.

3.2 Layout branch: Spatial-Temporal Transformer

The input to the layout model (Fig. 1) is a frame sequence $S = (f_0, f_1, ..., f_{n-1})$ of length n. Each frame f_i is composed of m objects, $f_i = \{o_0, o_1, ..., o_{m-1}\}$, where o_j consists of the object category c_j and location in the frame $l_j = [x_1, y_1, x_2, y_2]$. *Note that this is all the information required and used by the layout branch: no appearance information, just bounding boxes and category labels.* Two separate fully-connected layers yield the category embedding $\hat{c}_j$ and the frame location embedding $\hat{l}_j$, which we subsequently sum together and apply layer-normalization [3] and dropout [40] to obtain the final object embedding: $\hat{o}_j = \text{Dropout}(\text{LayerNorm}(\hat{c}_j + \hat{l}_j))$.

With the layout model, dubbed as Spatial-Temporal Layout Transformer, henceforth abbreviated as STLT, we decouple the spatial (per-frame) reasoning from the temporal (across-video) reasoning. To that end, to model the per-frame spatial relations, given a set of frame objects $f_i = \{o_0, o_1, ..., o_{m-1}\}$, we firstly prepend an $o_{\texttt{class}}$ object, with $c_{\texttt{class}}$ (special `class` category) and $l_{\texttt{class}}$ equal to the frame size. Then, we obtain the embedding $\hat{o}_j$ for each frame object, $\hat{f}_i = \{\hat{o}_{\texttt{class}}, \hat{o}_0, \hat{o}_1, ..., \hat{o}_{m-1}\}$, and use a bidirectional transformer (each object embedding can attend on all others). We denote this module as Spatial-Transformer (Fig. 1, left), which we apply on the set of object embeddings for each frame $\hat{f}_i$ separately. Subsequently we select the output hidden state corresponding to the `class` category as a global representation of a single frame: $\hat{s}_i = \text{Spatial-Transformer}(\hat{f}_i)$.

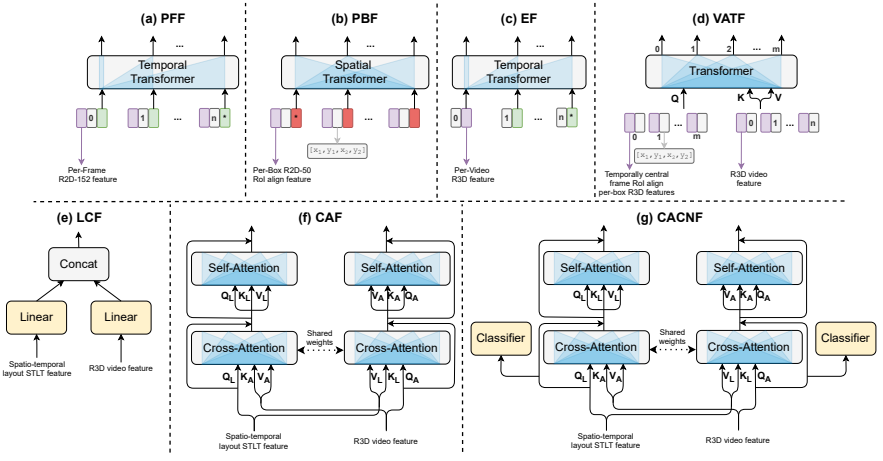

Figure 2: Fusion schemes. **Top:** One-branch. **Bottom:** Two-branch fusion methods.

To model the temporal evolution of the spatial relations, we use a causal transformer (each frame embedding can attend on the past ones), denoted as Temporal-Transformer (Fig. 1, right). We firstly append another special `class` embedding $\hat{s}_{\texttt{class}}$ to the outputs of the Spatial-Transformer. Then, for each frame, with a fully-connected layer, we obtain frame-position-in-the-video embedding $\hat{p}_i$. This is summed with the frame spatial embedding $\hat{s}_i$, followed by layer-normalization and dropout: $\hat{t}_i = \text{Dropout}(\text{LayerNorm}(\hat{\text{s}}_\text{i} + \hat{\text{p}}_\text{i}))$. Finally, we forward propagate the sequence of frame embeddings $\hat{T} = (\hat{t}_0, \hat{t}_1, ...\hat{t}_{n-1}, \hat{t}_{\texttt{class}})$ through the Temporal Transformer: $\hat{H} = \text{Temporal-Transformer}(\hat{T})$, where $\hat{H}$ are the output hidden states. If we use STLT as a standalone action recognizer, i.e., given spatio-temporal layouts as input we want to infer the action, we select the hidden state corresponding to the `class` embedding, and add a classifier on top: $\hat{y} = \text{Linear}(\hat{h}_{\texttt{class}})$, where $\hat{y}$ are the logits.

3.3 Appearance branch and multimodal fusion

As an appearance model, we deem a neural network, usually a CNN [25], applied over pixels of the video frame(s). In this work, we use four types of appearance features, each tightly coupled with the corresponding fusion approaches: (i) 2D Resnet152 (R2D-152) [17], pre-trained on ImageNet [12], applied over individual frames; (ii) 2D Resnet50 (R2D-50) backbone, from a COCO [27] pre-trained Faster R-CNN [37]. Given class-agnostic bounding boxes from a video frame, we extract RoI align [18] features from each; (iii) An inflated 3D Resnet50 (R3D) [21], pre-trained on [8, 32, 56]; (iv) R3D, same as (iii), with a Transformer encoder to enable multimodal fusion with cross-attention.

Due to the specific nature of the appearance features, we devise different ways to fuse them with the layout model (Fig. 2). We evaluate different configurations where each has its weaknesses, while we gradually build the ultimate, empirically superior fusion approach (for details see supplementary): (a) **Per-Frame Fusion (PFF) –** We sum the R2D-152 and the Spatial Transformer features for each frame individually, and feed the obtained embedding as input to the Temporal Transformer; (b) **Per-Box Fusion (PBF) –** We sum the RoI aligned R2D-50 features with the bounding box and the category embeddings, and feed the fused embedding as input to the Spatial Transformer; (c) **Early Fusion (EF) –** We feed the

R3D video features as a first token to the Temporal Transformer, essentially allowing each spatial frame embedding to attend on the entire video; (d) **Video Action Transformer Fusion (VATF) –** A fusion approach which is an adapted Video Action Transformer (VAT) model [15] for our task. We use the R3D as a trunk, and leverage the externally obtained bounding boxes to extract RoI align features from the temporally central frame as a query, and the trunk features as a memory to a transformer model which predicts the action (for details refer to [15]); (e) **Late Concatenation Fusion (LCF) –** The R3D video embedding is concatenated with the STLT embedding right before the classifier, commonly used as a standard baseline for multimodal fusion; (f) **Cross-Attention Fusion (CAF) –** A multimodal fusion with cross-attention [43] to fuse the layout (STLT) and appearance (R3D) branch embeddings; (g) **Cross-Attention CentralNet Fusion (CACNF) –** A CentralNet [46] based CAF implementation. Despite performing end-to-end training by minimizing the loss from the cross-attention fusion module (CAF) output, we additionally minimize the loss for each branch (layout and appearance) independently. To that end, by using the CentralNet fusion approach, we achieve multimodal fusion on a two-branch model, where the individual branches are enforced to preserve their individual abilities.

4 Evaluation and discussion

We perform experiments on the Something-Something [16] / Else [30] and the Action Genome datasets [20]. During training, we randomly sample 16 frames (each represented as spatio-temporal layouts) for STLT, and uniformly sample 32 RGB frames for models using R3D, subsequently rescaled to 112 × 112 (complete experimental setup in supplementary).

	Something-Else: Compositional setting				Something-Something	
	Obj. predictions		Oracle		Oracle	
Method	Top 1 acc.	Top 5 acc.	Top 1. acc.	Top 5 acc.	Top 1 acc.	Top 5 acc.
GNN-NL	33.3	58.9	50.7	78.6	47.1	76.3
S&TLT	40.6	66.9	57.7	84.7	55.6	84.3
STLT	**41.6**	**67.7**	**59.4**	**85.8**	**57.0**	**85.2**
R3D	51.3	78.6	51.3	78.6	52.2	80.7
PFF	47.3	73.7	62.5	87.5	62.9	88.0
PBF	48.5	73.2	62.9	86.5	**64.5**	89.1
EF	**52.8**	**79.3**	**63.8**	**88.1**	64.4	**89.4**
VATF	49.1	78.0	53.0	79.6	54.9	82.8
LCF	54.1	79.8	66.1	88.8	64.4	89.3
CAF	52.3	78.9	64.4	88.6	64.5	89.1
CACNF	**56.9**	**82.5**	**67.1**	**90.4**	**66.8**	**90.6**

Table 1: Comparison between the different model configurations. **From top to bottom:** Layout-based methods, R3D, one-branch fusion methods, two branch fusion methods. Best method within group in bold, overall best method in red.

Something-Something V2 [16] consists of egocentric videos of people performing actions with their hands, with 174 unique actions. To deal with the environment bias, videos recorded by the same person can be in either the training or validation set. Nevertheless, the objects the person interacts with might still overlap between training and test time, indicating that appearance-based models can overfit on the objects' appearance.

Something-Else [30] proposes two data splits according to the objects' distribution at training and test time. In the compositional split, to validate the compositional generalization, the data is divided such that the models encounter distinct objects during training and testing. The training and validation set contain ∼55k and ∼58k videos respectively, with 174 actions. In the few-shot split, there are ∼112k pre-training videos (with 88 base actions), 5 × 86 and 10 × 86 videos in the 5-shot and 10-shot setup respectively for fine-tuning, and ∼49k testing videos (with 86 novel actions)[2]. On the compositional and few-shot splits, we

[2]When fine-tuning, we freeze the backbone's weights and only train the action classifier following [30].

perform experiments with Faster R-CNN [37] object detections (object predictions setting), and with ground truth object detections (oracle setting), both released by [30]. The input object categories in STLT are either "hand" or "object". We measure performance using top-1 and top-5 accuracy (acc.), and perform training with cross-entropy loss.

Action Genome [20], built on top of Charades [38], has ∼10k videos of people doing daily activities. Multiple object-specific actions simultaneously occur in each video out of 157 unique ones. The frames where the action occurs, i.e., the person interacts with the objects, are annotated with bounding boxes and categories. We train a Faster R-CNN [37] on these frames and obtain object detections (for details see supplementary). We perform experiments on the Charades train/validation split with our object detections (obj. predictions setting), as well as the ground truth object detections (oracle setting) released by [20]. We measure performance using mean average precision (mAP), and perform training with binary cross-entropy loss.

4.1 Ablation studies and main findings

We ablate our models to gain insights in *why* one should leverage spatio-temporal layouts for action recognition, and *how* to come up with the best approach for it. We do an ablation study on the Something-Else compositional dataset, while we also verify our findings on the Something-Something (non-compositional) validation set, albeit only in an oracle setting.

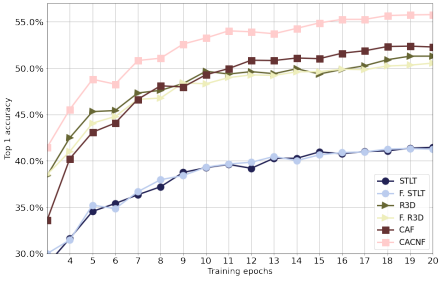

Figure 3: Top-1 validation acc. (epoch 3 to 20) of STLT and R3D trained individually, trained in conjunction with CACNF, CAF and CACNF.

Layout branch: How to model the spatio-temporal layouts? In Table 1 (Top), we measure STLT's performance against: (i) A baseline model [30] with a spatial reasoning graph neural network (GNN) and temporal reasoning non-local block (NL); (ii) An STLT variant performing joint spatial-temporal reasoning (S&TLT) on unrolled frames' bounding boxes (for details see supplementary). Across different settings and layout types (obj. predictions or oracle), we observe that a decoupled spatio-temporal reasoning is preferable. Furthermore, STLT and S&TLT significantly outperform GNN-NL, suggesting the appropriateness of MHA-based methods for modelling spatio-temporal layouts.

Multimodal fusion: How and where to fuse? In Table 1 (middle, bottom) we report action recognition results with the fusion methods we consider in this work. We compare among the fusion methods plus a fine-tuned R3D [21], and succinctly summarize our empirical findings as: (i) 2D fusion methods (PFF, PBF) exhibit good performance on non-compositional datasets (Something-Something), where object appearance matters, while their performance deteriorates on compositional datasets; (ii) Early fusion (EF) yields a competitive performance across different datasets and is superior to the other one-branch fusion methods; (iii) The Video Action Transformer [15] fusion type (VATF), does not fully exploit the spatial-temporal layout and video-context specific to the action (it only performs RoI align on the temporally central frame), thus it yields consistently lower results on these types of data; (iv) One-branch methods only marginally outperform R3D without oracle layouts; (v) Late fusion by concatenation (LCF) remains a strong baseline, as reported by others

[33, 46]; (vi) Cross-Attention Fusion (CAF) is consistently weaker than LCF and CACNF, while CACNF (CAF with CentralNet [46]) outperforms other methods regardless of the data type (compositional, non-compositional) and type of layouts (obj. predictions or oracle).

Why is CACNF superior to CAF despite the conceptual similarity? In Fig. 3, we observe the top 1 acc. on the Something-Else compositional split, in the obj. predictions setting, of: (i) STLT trained individually; (ii) STLT trained within the CACNF model (F. STLT); (iii) R3D trained individually; (iv) R3D trained within the CACNF model (F. R3D); (v) CAF; (vi) CACNF. We see that the performance of STLT and F. STLT, and, R3D and F. R3D, is remarkably similar, indicating that with CACNF the layout and appearance branch preserve their individual capabilities. What is interesting is that this phenomenon results in better overall performance of the cross-attention fusion module (CACNF), compared to training it without CentralNet [46] – CAF.

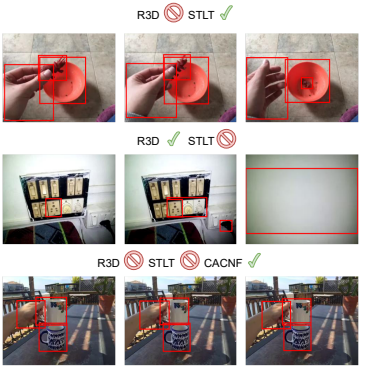

Figure 4: **Top:** R3D mispredicts, STLT predicts correctly. **Middle:** STLT mispredicts, R3D predicts correctly. **Bottom:** STLT and R3D mispredict, CACNF predicts correctly.

How does it look visually? We are interested in visually inspecting three error types: (i) R3D predicts wrong, STLT predicts correct action; (ii) STLT predicts wrong, R3D predicts correct action; (iii) STLT and R3D predict wrong, CACNF predicts correct action. In Fig. 4 (top), we observe the action "Dropping smth. into smth.", suitable for layout-based methods, e.g., STLT, as they directly model the spatial properties of the objects (location, movement, size, etc.). On the contrary, STLT is unable to recognize the action "Turning the camera upwards while filming smth.", Fig. 4 (middle), due its spatial ambiguity, while R3D recognizes the change in appearance, indicative of the action. Lastly, the action "Holding smth. over smth." in Fig. 4 (bottom), requires modelling both the temporal consistency of the layout and appearance, which STLT and R3D individually fail, while CACNF recognizes the correct action. Furthermore, in Fig. 5, we compare the performance of R3D and STLT with CACNF, on five actions from Something-Else, where the difference between the averaged R3D and STLT accuracy, and CACNF accuracy for each action is most prominent. We observe a consistent pattern across all five actions, i.e., CACNF successfully fuses the appearance (R3D) and layout branch (STLT), and it yields superior performance compared to the unimodal (layout or appearance) methods.

4.2 Something-Else: State-of-the-art comparisons

We compare against the following methods: (i) **I3D** [7]: An inflated Resnet50 [17] based 3D CNN as in [49], pre-trained on ImageNet [12], subsequently fine-tuned on the Something-Else dataset; (ii) **STRG** [49]: A multimodal method, with a GNN [23] applied over region proposals, combined with I3D using late fusion; (iii) **STIN** [30]: A GNN [23] for spatial, and a non-local neural network [50] for temporal reasoning; (iv) **STIN + I3D**: STIN combined with I3D in a late-fusion manner; (v) **SFI** [54]: A layout-appearance fusion method, trained with an auxiliary task of predicting the future state of the video objects.

	Compositional setting				Few-shot setting			
	Obj. predictions		Oracle		Obj. predictions		Oracle	
Method	Top 1 acc.	Top 5 acc.	Top 1. acc.	Top 5 acc.	Top 1 acc. (5-shot)	Top 1. acc. (10-shot)	Top 1 acc. (5-shot)	Top 1 acc. (10-shot)
STIN [22]	37.2	62.4	51.4	79.3	17.7	20.8	27.7	33.5
SFI [23]	—	—	44.1	74.0	—	—	24.3	29.8
STLT (Ours)	**41.6**	**67.9**	**59.0**	**86.0**	**18.8**	**24.8**	**31.4**	**38.6**
I3D [4]	46.8	72.2	46.8	72.2	21.8	26.7	21.8	26.7
STIN [22] + I3D [4]	48.2	72.6	54.6	79.4	23.7	27.0	28.1	33.6
STRG [18]	52.3	78.3	—	—	24.8	29.9	—	—
SFI [23]	—	—	59.6	85.8	—	—	30.7	36.2
CACNF (Ours)	**56.9**	**82.5**	**67.1**	**90.4**	**27.1**	**33.9**	**37.1**	**45.5**

Table 2: Something-Else SOTA comparisons. **Left:** Compositional setting, **Right:** Few-shot setting. **Top:** Layout-based methods, **Bottom:** I3D and Multimodal methods.

Compositional action recognition. We report results in Table 2 (Left). In both the obj. predictions and oracle setting, we observe that STLT outperforms STIN and the other methods, with a more prominent difference in performance in the oracle setting. We also observe that STLT's performance is remarkably close to the best multimodal concurrent method in the oracle setting – SFI, an indication that, if perfect object detections are available, MHA captures finer interactions between the objects compared to a (convolutional) GNN, 1D convolution, etc. In the multimodal section – Table 2 (bottom), obj. predictions setting, CACNF outperforms the other methods significantly, suggesting a noteable improvement in robustness w.r.t. compositional data (a likely real-life scenario).

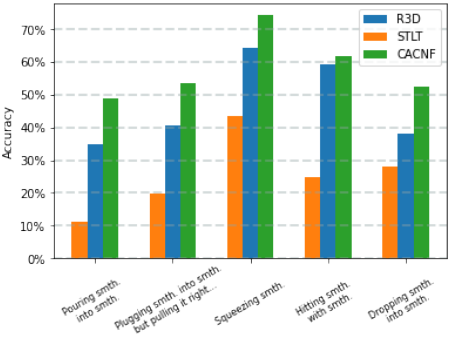

Figure 5: Five Something-Else actions where the performance difference between R3D and STLT (averaged) with CACNF is most prominent.

Few-shot action recognition. We report results in Table 2 (Right). We observe that in the obj. predictions setting, STLT outperforms STIN in both the 5-shot and 10-shot setup. In the oracle setting, STLT significantly outperforms the other methods as per the top-1 accuracy, allowing a significant gap for improvement as object detectors continue to improve [11]. Furthermore, CACNF outperforms the other multimodal methods in the obj. predictions setting, with a bigger difference in top-1 acc. in the 10-shot setup. We interpret the performance improvement as evidence that STLT and CACNF can successfully generalize in a low-data regime, a valuable observation considering the cost of acquiring curated video data.

4.3 Action Genome: Coping with background-cluttered videos

Object-centric layout-based models appear to be unsuited for dealing with background-cluttered videos. To address such concerns, we use Action Genome [20], as it consists of videos (i) of people performing actions at home, where it is hard to isolate the objects specific to the action, (ii) with objects presence overlapping between training and testing. We measure action recognition mAP, as well as mAP relative improvement on top of trained I3D [4], by ensembling it with STLT. We conduct experiments in a setup where we replace all

specific categories, e.g., book, broom, phone, etc., except for person, with a generic "object" category, therefore having only 2 object categories (person, object), and the default setup, where we utilize all 38 categories as input to STLT. We compare against: (i) **LFB** [51]: Long-term feature bank model, which performs well when video features are aggregated over time; (ii) **SGFB** [20]: Method which predicts (or uses ground truth in oracle setting) symbolic scene graph, further encoded and combined with LFB [51].

Method	Input modalities	Num. categories	Obj. predictions mAP	Oracle mAP
LFB	Video Appearance	—	42.5	42.5
SGFB	Scene Graphs & Video Appearance	38	44.3 (1.8)	60.3 (17.8)
I3D (Ours)	Video Appearance	—	33.5	33.5
STLT (Ours)	Obj. detections	2	16.1	19.9
STLT + I3D (Ours)	Obj. detections & Video Appearance	2	33.8 (0.3)	36.5 (3.0)
STLT (Ours)	Obj. detections	38	30.2	60.6
STLT + I3D (Ours)	Obj. detections & Video Appearance	38	38.5 (5.0)	61.63 (28.1)

Table 3: Action Genome results. **From top to bottom:** Baselines, I3D, STLT and STLT + I3D ensemble with 2 generic obj. categories (person, obj.), STLT and STLT + I3D ensemble with all 38 obj. categories (person, book, phone, etc.). Relative mAP improvement over appearance method in parenthesis.

In Table 3, when only 2 object categories (person, object) are registered in STLT, we observe that STLT yields significantly weaker performance compared to a standard appearance method, e.g., I3D. Interestingly, despite the negatively biased setup, i.e., the actions in Action Genome are object-specific, e.g., opening a book, while the input categories are object-agnostic – person/object, STLT still boosts I3D's performance by 3.0 mAP points in the oracle setting. When all 38 object categories are registered in STLT, we observe that STLT performs well even in the obj. predictions setting, considering the Faster R-CNN's ~ 11.5 average precision (AP) on the validation set. In the oracle setting the performance increases drastically, being on par with SGFB (which relies on ground truth scene graph), indicating a high upper bound, considering that object detection is merely a subset of scene graph generation. In the obj. predictions setting, we observe a solid relative improvement over I3D by ensembling STLT with I3D, even larger compared to SGFB and LFB[3], concluding that STLT reasonably copes with background clutter, and successfully boosts the performance of an appearance model – I3D.

5 Conclusion

In this paper we shed light on the problem of compositional and few-show action recognition. We advocated the use of multi-head attention over spatio-temporal layouts, and attempted to reach a conclusion how layout- and appearance-based models should be fused. Our main empirical findings suggest that (i) a layout-based model is robust w.r.t. compositional data, and generalizes from a few samples, (ii) when fusing a layout- and appearance-based model, it is crucial for the individual models to preserve their capabilities, (iii) even on non-compositional, background cluttered video datasets, a layout-based model can reasonably recognize human actions, and boosts the performance of appearance-based models.

A limitation which remains is that layout-based models are highly dependent on the object detections quality. Notice, however, the high upper bound (oracle setting), combined with the fact that for compositional action recognition, the requirement is class-agnostic object detections. Lastly, by relying on an object detector the overall model complexity increases, which is detrimental to the speed. Ideally, an off-the-shelf appearance based model should exhibit object-centric reasoning abilities, which we leave for future work.

[3]The appearance model performance and the relative improvement are most likely inversely proportional.

Acknowledgements

We acknowledge funding from the Flemish Government under the Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen programme. We also thank Dina Trajkovska for the help with the figures.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[2] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 105–121, 2018.

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[7] 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.

[8] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[9] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL https://www.aclweb.org/anthology/D14-1179.

[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. URL https://openreview.net/forum?id=HJlnC1rKPB.

[11] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7373–7382, 2021.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[14] 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.

[15] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019.

[16] 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.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

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

[20] 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*, pages 10236–10247, 2020.

[21] Hirokatsu Kataoka, Tenga Wakamiya, Kensho Hara, and Yutaka Satoh. Would mega-scale datasets further enhance spatiotemporal 3d cnns? *arXiv preprint arXiv:2004.04968*, 2020.

[22] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*, 2021.

[23] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

[24] Dana Lahat, Tülay Adali, and Christian Jutten. Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, 103(9):1449–1477, 2015.

[25] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[26] 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.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf.

[29] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2018.

[30] 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*, pages 1049–1059, 2020.

[31] Lili Meng, Bo Zhao, Bo Chang, Gao Huang, Wei Sun, Frederick Tung, and Leonid Sigal. Interpretable spatio-temporal attention for video action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[32] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019.

[33] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015.

[34] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6966–6975, 2019.

[35] 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. *arXiv preprint arXiv:2004.01278*, 2020.

[36] Gorjan Radevski, Guillem Collell, Marie Francine Moens, and Tinne Tuytelaars. Decoding language spatial relations to 2d spatial arrangements. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4549–4560, 2020.

[37] 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. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

[38] 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*, pages 510–526. Springer, 2016.

[39] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/00ec53c4682d36f5c4359f4ae7bd7ba1-Paper.pdf.

[40] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[41] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Kevin Murphy, Rahul Sukthankar, and Cordelia Schmid. Actor-centric relation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–334, 2018.

[42] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[43] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[44] 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.

[45] 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*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[46] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[47] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020.

[48] 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.

[49] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 399–417, 2018.

[50] 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.

[51] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 284–293, 2019.

[52] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018.

[53] Huijuan Xu, Lizhi Yang, Stan Sclaroff, Kate Saenko, and Trevor Darrell. Spatio-temporal action detection with multi-object interaction. *arXiv preprint arXiv:2004.00180*, 2020.

[54] Rui Yan, Lingxi Xie, Xiangbo Shu, and Jinhui Tang. Interactive fusion of multi-level features for compositional activity recognition. *arXiv preprint arXiv:2012.05689*, 2020.

[55] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 591–600, 2020.

[56] Yuya Yoshikawa, Jiaqing Lin, and Akikazu Takeuchi. Stair actions: A video dataset of everyday home actions. *arXiv preprint arXiv:1804.04326*, 2018.

[57] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. A structured model for action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9975–9984, 2019.

[58] 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.