Reasoning About Human-Object Interactions Through Dual Attention Networks

Tete Xiao[1,2*] Quanfu Fan[2] Dan Gutfreund[2] Mathew Monfort[3] Aude Oliva[3] Bolei Zhou[4]
[1]University of California, Berkeley [2]MIT-IBM Watson AI Lab, IBM Research
[3]Massachusetts Institute of Technology [4]The Chinese University of Hong Kong

Abstract

Objects are entities we act upon, where the functionality of an object is determined by how we interact with it. In this work we propose a Dual Attention Network model which reasons about human-object interactions. The dual-attentional framework weights the important features for objects and actions respectively. As a result, the recognition of objects and actions mutually benefit each other. The proposed model shows competitive classification performance on the human-object interaction dataset Something-Something. Besides, it can perform weak spatiotemporal localization and affordance segmentation, despite being trained only with video-level labels. The model not only finds when an action is happening and which object is being manipulated, but also identifies which part of the object is being interacted with.

1. Introduction

Affordance, introduced by James Gibson [9], refers to the properties of an object, often its shape and material, that dictate how the object should be manipulated or interacted with. The possible set of actions that an object can afford is constrained. For instance, we can drink from a plastic bottle, pour water into it, squeeze it, or spin it, but we cannot tear it easily into two pieces (see Figure 1). Similarly, for a given action, the possible objects which it can apply to are also limited. For example, we can fold a paper but not a bottle.

A handful of works have exploited object information for the recognition of Human-Object Interactions (HOIs) and more general action recognition [3, 10, 31, 19, 44]. However, understanding HOIs goes beyond the perception of objects and actions: it involves reasoning about the relationships between how the action is portrayed and the consequence on the object (*i.e.*, whether the shape or location of the object is changed by the action upon it). Most of the previous works pre-define human-object or action-object pairs for HOI [3, 10, 31]. The classification is done by either a graphical model [12], a classifier based on the appearance features [10], or a graph parsing model [31]. A potential issue with the previous approaches is that the complexity of an HOI model grows quickly as the number of objects and actions increase. The reasoning capability of these approaches is also limited due to the action-object pairs being preset for modeling. In addition, full annotations including action labels and object bounding-boxes are often required by these approaches for the effective modeling of HOIs.

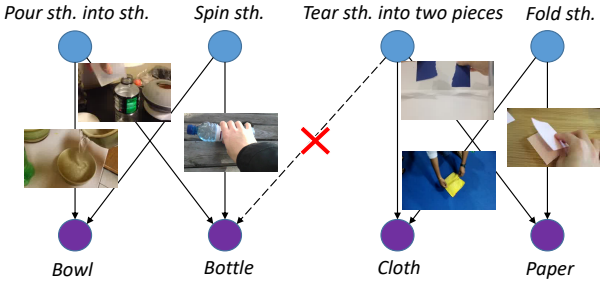

Figure 1: **Object and action co-dependence**. The action *tearing something into two pieces* can be performed on a piece of *paper* but not a *bottle*. Given the object *bottle*, we can pour water into it or spin it, but cannot fold it or tear it.

Here we propose a Dual Attention Network model that leverages object priors as the guidance to where actions are likely to be performed in a video stream and *vice versa*. The focus of attention is represented by a heatmap indicating the likelihood of where an action is taking place or where an object is being manipulated in each frame. These attention maps can enhance video representation and improve both action and object recognition, yielding very competitive performance on Something-something [11] dataset. We show that the attention maps are intuitive and interpretable, enabling better video understanding and model diagnosis. Such attention maps also facilitate weakly-supervised spatiotemporal localization of objects and actions.

1.1. Related work

Action recognition. Deep convolutional neural networks have been used with success for action recognition [23, 38, 45, 13, 16]. For instance, we can exploit the suc-

*The work was done when Tete Xiao was an intern at IBM Research, Cambridge.

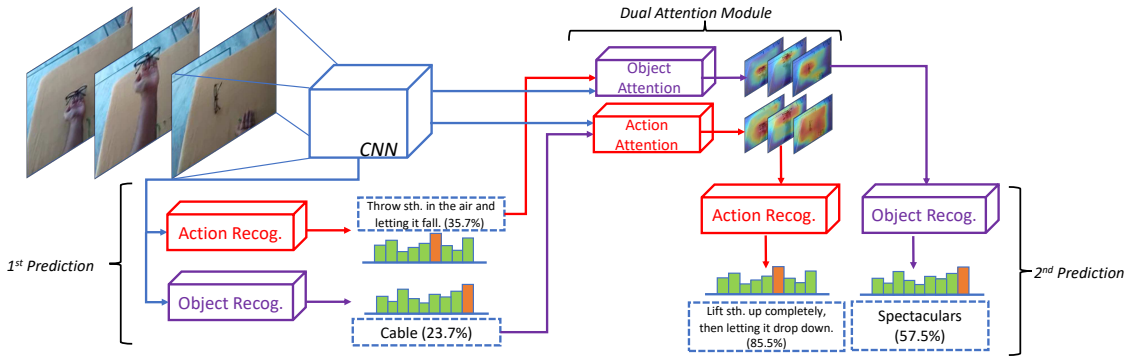

Figure 2: **Framework overview**. Our approach exploits the role of human action and object in human-object interactions via the dual attention module. The Dual Attention Network first predicts plausible action and object labels independently as the priors (1^{st} prediction). Then the priors are used to generate attention maps that weight the features of object and action for the 2^{nd} prediction. Action Recog.: action recognition head. Object Recog.: object recognition head.

cess of CNNs for static images and RNNs for temporal relations by feeding CNN-based features from single frame into an RNN model [49, 5, 20]. An alternative approach is to extend 2D CNNs by applying 3D convolutional filters (C3D) on raw videos to directly capture the spatiotemporal information [39]. The 3D filters can be "inflated" from 2D filters (I3D) [2] and can be initialized with an ImageNet [4] pre-trained model. Recent works involve Non-local Networks [43], which uses space-time non-local operations to capture long-range dependencies; and Temporal Relation Network [50], which sparsely samples frames from different time segments and learns their causal relations. In addition to the end-to-end frameworks on raw video inputs, optical flow [15] has also proven to be useful [35, 2, 50] when combined with features extracted from raw RGB images.

Human-object interactions and visual affordance. Several works have exploited human-object interactions and affordance for action recognition. Gupta *et al*. [12] integrate perceptual tasks to exploit the spatial and functional constrains for understanding human-object interactions. Koppula *et al*. [22] frame the problem as a graph, where the nodes represent objects and sub-activities while the edges represent the affordance and relations between human actions and objects. The graph model can be optimized using structural Support Vector Machine (SVM) [22] or Conditional Random Field (CRF) [24]. Jain *et al*. [18] merge spatiotemporal graph with an RNN to model different kinds of spatial-temporal problems such as motion, action prediction and anticipation. Gkioxari *et al*. [10] propose InteractNet to detect ⟨ human, verb, object ⟩ triplets by exploiting the appearance features from detected persons. Dutta and Zielinska [7] employ a probabilistic method to predict the next action in human-object interactions [41]. Fang *et al*. [8] propose a model to learn the interactive region and action label of an object via watching demonstration videos.

Attention models. Attention mechanism has been

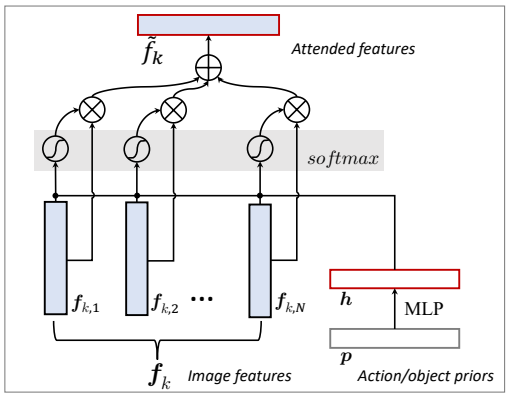

Figure 3: **Illustration of the attention module** for the k^{th} frame. It encodes action (object) priors and attends image regions accordingly, yielding the representation for object (action) recognition.

adopted for action recognition. Sharma *et al*. [33] use a soft attention module to re-weight CNN features spatially. Ramanathan *et al*. [32] propose to attend people involved in specific events for event detection in multi-person videos. Song *et al*. [36] exploit skeleton data for attention module to extract more discriminative features in human-centered actions. Du *et al*. [6] propose to incorporate a spatial-temporal attention module into a classical CNN-RNN video recognition model.

Co-attention models [34, 46, 47, 48, 27, 26] are widely adopted in tasks relating to language and vision such as image captioning [40], visual question answering (VQA) [1] and visual question generation (VQG) [29]. Lu *et al*. [27] propose a hierarchical co-attention model for VQA, in which image representation is used to guide the question attention and *vice versa*, exploiting the relation between the two modalities, image and text.

Comparison to our approach. In contrast to the self-attention and human-attention models for action recognition, and the co-attention models for multi-modal (text and vision) tasks, our framework applies dual attention in the context of multi-task learning on a single input modality, namely the raw video. Our novel iterative model exploits the action/object relations to simultaneously learn cross-task object/action attention maps, which significantly differs from previous works that use self-guided attention [33, 6]. Our model is able to not only outperform the previous state-of-the-art on a human-object interaction dataset [11] but also yield interpretable attention maps (see Section 4).

2. Dual Attention Network for Human-Object Interactions

The dual attention network is designed in such a way that the streams of human activity and objects interact with each other by cross-weighting the intermediate features of action and object for recognition. Our attention module is general and can be plugged into any CNN-based action recognition models for feature enhancement. We first describe CNN-based feature representations for video understanding in Sections 2.1 and 2.2. We then introduce the dual attention model in Section 2.3, the building block for reasoning about actions and objects. Finally we detail the full framework in Section 2.4.

2.1. Representing videos with neural networks

There are two *de facto* paradigms to extract video representations: 1) *Image-based* models which use spatial convolutional kernels to process frames independently, and later perform temporal feature aggregation by another model such as a Long Short-Term Memory network (LSTM) [14] or a Temporal Relation Network (TRN) [50]; 2) *Video-based* models which apply convolutional kernels across frames to process a video with spatial and temporal dimensions directly.[1]

Image-based models. Given a video V with T frames, CNN features from each frame are extracted *independently*, resulting in a set of T raw features $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_T\}$, where $\boldsymbol{f}_k \in \mathbb{R}^{d \times N}$, d is the feature dimension and $N = HW$ is the vectorized spatial dimension of the feature map. The CNN features are then averaged by global pooling over the spatial dimension, *i.e.*,

$$\bar{\boldsymbol{f}}_k = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{f}_k [., i] \tag{1}$$

After that, various modules that process and fuse information across temporal domain can be applied on top of extracted features. For example, all frames can be modeled by an LSTM, resulting in the final representation of a video $\hat{\boldsymbol{v}}$ as:

$$\hat{\boldsymbol{v}} = \text{LSTM}\left(\bar{\boldsymbol{f}}_1, \bar{\boldsymbol{f}}_2, \cdots, \bar{\boldsymbol{f}}_T\right) \tag{2}$$

Alternatively, TRN [50] is a simple yet effective network module recently proposed to explicitly learn and model temporal dependencies across sparsely sampled frames at different temporal scales. TRN can be applied on top of any 2D CNN architecture. More specifically, an n-order relation, for a given number n, is modeled as:

$$R_n(V) = h_\phi \left(\sum_{k_1 < k_2 < \cdots < k_n} g_\theta(\bar{\boldsymbol{f}}_{k_1}, \bar{\boldsymbol{f}}_{k_2}, \cdots, \bar{\boldsymbol{f}}_{k_n}) \right) \tag{3}$$

Here h_ϕ and g_θ are both multi-layer perceptrons (MLPs) fusing features of different frames. For the sake of efficiency, rather than summing over all possible choices of n ordered frames, a small number of tuples uniformly sampled are chosen. The model can be extended to capture relations at multiple temporal scales by considering different values of n. The final representation of a video is an aggregation of a 2-order TRN up to an n-order TRN:

$$\hat{\boldsymbol{v}} = R_2(V) + R_3(V) + \cdots + R_n(V) \tag{4}$$

where n is a hyperparameter of the model.

Video-based models. A video-based model operates on multiple frames within a video. As a result, given a video with T frames, features of each frame are not independent anymore, so that cross-time dynamics may be learned in this way. Besides, temporal down-sampling is often adopted to form a sufficiently large receptive field over temporal domain, so that the number of remaining frames T' is less or equal than T. Denote a set of T' features as $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_{T'}\}$, where $\boldsymbol{f}_k$ can be a super frame if $T' < T$, and like in image-based models each frame is then averaged by spatial global pooling. Since dynamics are expected to be learned implicitly within convolutional neural networks, the final representation of a video is usually acquired by averaging across all (super) frames:

$$\hat{\boldsymbol{v}} = \frac{1}{T'} \sum_{k=1}^{T'} \bar{\boldsymbol{f}}_k \tag{5}$$

2.2. Object recognition in videos

Given a video of an HOI, we want to recognize the action and the object associated with the interaction. For example, for a "*playing*" action, we want our model to recognize that it is "*playing a violin*" rather than "*playing a piano*". For training we assume that labels are provided at the video level without bounding-boxes. A straight-forward method for joint action-object recognition is to add a separate classification head for object recognition alongside the head for

[1] We group the models based on the *domain* that they use for extracting features of each frame, *i.e.*, whether cross-time dynamics are exploited, rather than types of convolutional kernels.

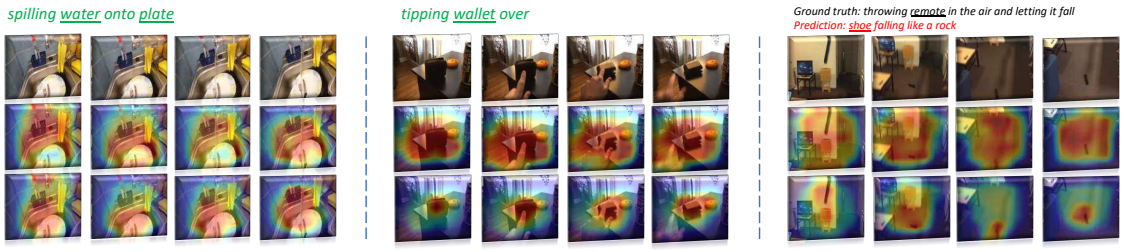

Figure 4: **Examples of attention maps** yielded by the Dual Attention Network with their predicted labels above. For each clip four frames are shown out of eight frames for TRN with a stride of two. The first row is the input frames while the second and third ones are attention maps for recognizing *action* and *object* respectively. The model accurately learns the alignment between actions and objects, even when the background is complicated (*e.g.*, 1^{st} clip), or the predicted labels are wrong (*e.g.*, 3^{rd} clip). These examples are drawn from validation subsets.

action recognition. Note that the task differs from the standard object recognition in static images, because the model should look for the objects being manipulated by the actor instead of those in the background. As a result, the object head should also utilize feature representations containing temporal information, *i.e.*, for image-based models such as TRN, another multi-scale TRN module is used for object recognition, whereas for video-based models we simply use another MLP.

2.3. Dual attention module

We propose a dual attention model for action and object recognition, as illustrated in Figure 3. The model is not dependent on a specific CNN architecture thus it is general and extensible. The dual attention uses action priors to attend image features for objects, and object priors for actions. Suppose that we have the probabilities $\boldsymbol{p}^a$ and $\boldsymbol{p}^o$ over actions and objects respectively of their likelihood to appear in the video. First, we apply two MLPs to encode these probability vectors into two intermediate feature representations $\boldsymbol{h}^a, \boldsymbol{h}^o \in \mathbb{R}^d$. The dual attention module takes input of the visual features at each frame and generates the object and action attention distributions over N regions of each frame:

$$\boldsymbol{z}_k^a = \boldsymbol{w}_a^T \tanh\left(\boldsymbol{W}_a \boldsymbol{f}_k + \boldsymbol{W}_{ota} \boldsymbol{h}^o \mathbb{1}^T\right) \quad (6)$$

$$\boldsymbol{z}_k^o = \boldsymbol{w}_o^T \tanh\left(\boldsymbol{W}_o \boldsymbol{f}_k + \boldsymbol{W}_{ato} \boldsymbol{h}^a \mathbb{1}^T\right) \quad (7)$$

$$\boldsymbol{\alpha}_k = \mathrm{softmax}(\boldsymbol{z}_k^a) \quad (8)$$

$$\boldsymbol{\beta}_k = \mathrm{softmax}(\boldsymbol{z}_k^o) \quad (9)$$

where $\mathbb{1} \in \mathbb{R}^N$ is a vector whose elements are all equal to 1. $\boldsymbol{W}_a, \boldsymbol{W}_o \in \mathbb{R}^{N \times d}$ and $\boldsymbol{w}_a, \boldsymbol{w}_o \in \mathbb{R}^N$ are the weights to be learned. $\boldsymbol{W}_{ota}, \boldsymbol{W}_{ato} \in \mathbb{R}^{N \times d}$ are parameters for object-to-action attention and action-to-object attention, respectively. $\boldsymbol{\alpha}_k, \boldsymbol{\beta}_k \in \mathbb{R}^N$ are the attention weights over spatial features in $\boldsymbol{f}_k$. The representation of each frame is obtained by a weighted-average over its spatial domain:

$$\tilde{\boldsymbol{f}}_k^a = \sum_{i=1}^{N} \alpha_{k,i} \boldsymbol{f}_k\,[,i] \quad (10)$$

$$\tilde{\boldsymbol{f}}_k^o = \sum_{i=1}^{N} \beta_{k,i} \boldsymbol{f}_k\,[,i] \quad (11)$$

Finally, for x in $\{a, o\}$, we obtain representations of a video for action and object respectively by substituting $\bar{\boldsymbol{f}}_k^x$ with $\tilde{\boldsymbol{f}}_k^x$ in Equation 3 or 5.

2.4. Full architecture

The full architecture is illustrated in Figure 2. Given a video, the network first predicts the plausible action and object labels using two separate heads. The prediction results serve as the priors of actions and objects, which are subsequently used to produce the attention maps for objects and actions, via the dual attention module. A second prediction is performed with the attention-based enhanced features. This two-step scheme expresses the interaction between human and objects. The two prediction modules along with the attention module are integrated into one network for end-to-end learning. Some actions may involve multiple entities, for instance, "*put something on something*". We therefore use two softmax classifiers to predict the objects. If the order of the objects is exchangeable, *e.g.*, in the category "*move something and something closer*", the classifiers will learn to predict the objects in the order as they appear in the ground truth to avoid ambiguity. We use a *null* label as a placeholder for those action classes with only a single object.

3. Experiments

We conduct comprehensive experiments below to validate the efficacy of our proposed Dual Attention Network.

3.1. Implementation details

We choose Temporal Segment Networks (TSN) [42] and TRN [50] as the backbones among image-based models, and Temporal Shift Module (TSM) [25] among video-based models, given their demonstrated superior performance. We did not choose I3D [2] because the temporal down-sampling

method	top-1(A)	top-5(A)	top-1(O)
baseline	44.6	73.9	58.2
multi-tasking	45.7	75.0	59.9
dual attention	**46.6**	**75.6**	**60.1**

(a) Joint learning of two tasks. Dual attention is better than multi-task learning at exploiting action and object information.

method	top-1(A)	top-5(A)	top-1(O)
baseline	44.6	73.9	-
	-	-	58.2
GT-object att.	**50.2**	**79.7**	-
GT-action att.	-	-	**67.0**

(b) Attention guided by ground-truth labels. The significant improvements indicate that actions and objects are indeed closely intertwined.

method	top-1(A)	top-5(A)	top-1(O)
baseline	44.6	73.9	58.2
self attention	45.3	74.4	58.3
dual attention	**46.6**	**75.6**	**60.1**

(c) Self attention vs. dual attention. Action and object priors offer a better attention mechanism for recognition.

Table 1: Ablation study. A: action recognition; O: object recognition. The baseline is a TRN-4 network.

model	backbone	domain	modality	frames	top-1 val	top-5 val	top-1 test	top-5 test
TSN† [42]	BN-Inception	2D	RGB	8	41.1	69.3	-	-
TSN Dual Attention [ours]	BN-Inception	2D	RGB	8	42.1	71.2	-	-
I3D† [2]	ResNet-50	3D	RGB	16	43.8	73.2	-	-
2D-CNN w/ LSTM [28]	VGG-like	2D	RGB	48	40.2	-	38.8	-
3D-CNN w/ LSTM [28]	VGG-like	3D	RGB	48	51.9	-	51.1	-
2D-3D-CNN w/ LSTM [28]	VGG-like	2D + 3D	RGB	48+48	51.6	-	50.4	-
TSM‡ [25]	ResNet-50	3D	RGB	8	56.7	83.7	-	-
TSM†	ResNet-50	3D	RGB	8	54.0	81.3	-	-
TSM Dual Attention [ours]	ResNet-50	3D	RGB	8	55.0	82.0	-	-
TRN [50]	BN-Inception	2D	RGB	8	48.8	77.6	50.8	79.3
	BN-Inception	2D	RGB + Flow	8+8	55.5	83.0	56.2	83.1
TRN Dual Attention [ours]	BN-Inception	2D	RGB	8	51.6	80.3	54.0	81.9
	BN-Inception	2D	RGB + Flow	8+8	58.4	85.2	60.1	86.1

Table 2: Comparisons to state-of-the-art methods on Something-V2, with results on both the validation and test subsets. †: Our re-implemented model. ‡: From original paper, pre-trained on Kinetics [21] asides from ImageNet [4].

rate is overly large (*e.g.*, 16 frames input and 2 super frames output) so that it is naturally inproper to use spatial attention.

Base networks. Following [50] and [25], we adopt Inception with Batch Normalization [17] (BN-Inception) as our base models of TSN and TRN, while ResNet-50 [13] pre-trained on ImageNet [4] as it of TSM for fair comparisons. The input size is set to 224×224. The spatial size of output features is 7×7 with 1024 and 2048 channels for BN-Inception and ResNet-50, respectively. We append dropout [37] after the extracted features, with a ratio of 0.5.

Dual attention module. The dual attention module generates distributions over the spatial grids of feature maps for each frame. To embed the probabilities of action or object labels, we use a two-layer MLP with ReLU activations [30]. Both layers of the MLP have 512 channels. We project image features into 512 channels by a single-layer perceptron before feeding them into the attention module.

Recognition heads. We use the same sampling strategy as in [50] for multi-scale TRNs. g_ϕ is a two-layer MLP with 256 units per layer, while h_ϕ is a two-layer MLP whose output channels match the number of classes. We do not use dropout within classification heads. For both TSN and TSM the recognition head is a two-layer MLP. The classification heads do not share weights between the first (pre-attention) prediction and the second (post-attention) prediction. This design does not introduce computational overhead as the CNN feature extraction, which dominates the computation, is shared.

3.2. Setup

Dataset. We use Something-something dataset V2 [11], a video action dataset for human-object interactions, with $220,847$ videos from 174 classes. Those classes are fine-grained so a model needs to distinguish actions such as "lifting up *one end of* something then letting it drop down" from actions such as "lifting something up *completely* then letting it drop down". This requires the model to look into details of different actions. Note that object labels (*nouns*), provided in V2 by the workers, may result in some inconsistencies: a mobile phone can be depicted as *"phone"*, *"mobile phone"*, *"a phone"*, *"a black phone"* or even *"iPhone"*. We therefore merge nouns describing the same or similar objects, for a total of 307 object clusters (see supplementary material for details). We conduct our study on this dataset because it is among the very few ones containing videos of diverse human-object interactions instead of a few pre-defined relations between actions and objects. Also it is one of a few large-scale video dataset which provides object labels.

Training details. We use a multi-scale 4-frame TRN (TRN-4) in all our ablation study for efficiency. Results of an 8-frame TSN (TSN-8), an 8-frame TRN (TRN-8) and an 8-frame TSM (TSM-8) are included in the final experiment. The networks are trained end-to-end. We augment the data during training by scale and aspect-ratio jittering. The batch

size is set to 32 for TRN-4, 16 for TSN-8 and TRN-8, and 8 for TSM-8 due to GPU memory limitation. We train all models on a server with 8 GPUs for 70 epochs. It starts with a learning rate of 0.01, and is reduced by a factor of 10 at epoch 50 and 65. We use a momentum of 0.9. The weight decay of models with BN-Inception is set as 0.0001, whereas 0.0005 for models with ResNet. We train our models with unfrozen Batch Normalization, which effectively stabilizes the training procedure.

3.3. Main results

Results on the validation subset are in Table 1a. Our dual attention model attached on a TRN-4 network yields accuracies of $46.6/75.6$ (top1/top-5) on action recognition and of 60.1 (top-1) on object recognition, a $2.0/1.7$ and 1.9-point boost over the baseline. Compared to a separately trained model of joint learning of actions and objects (multi-tasking), our approach achieves superior performance, indicating that dual attention is a better approach to utilize the action and object information interchangeably.

Figure 4 visualizes attention maps learned by our model. For each clip, the first row contains four out of the eight frames chosen by the TRN module. The second and third rows are attention maps for *actions* and *objects*. We see that our model learns meaningful alignment between actions and objects. For action recognition, the attention map generally covers a larger space capturing the global information of the entire action series; for object recognition, the attention map is sharp and neat, mostly on the object being manipulated by the actor. Surprisingly, the model can attend to the relevant region and predict correct classes even when the background is complex, *e.g.*, the first example, in which the model finds the dishes in the sink as well as the water being spilled but ignores the background. In cases where the model produces inaccurate predictions, *e.g.*, the third example, our model still looks at reasonable regions across frames, although it seems unable to recognize the fine-grained categories.

Cohesion of actions and objects. In order to understand the potential maximum performance gain that we can expect from our approach, we experiment with using ground-truth annotations as action and object priors instead of predicting them. As shown in Table 1b, ground-truth guided attentions show a remarkable improvement for both action recognition ($5.6\%/5.8\%$) and object recognition (8.8%). This demonstrates that action and object recognition are closely intertwined, and that improving the first prediction in our approach can lead to an even bigger boost to performance.

Figure 5 shows the class-wise improvements over baseline with dual attention. We can see that the performance of action categories which are closely associated with certain types of objects is boosted. For example, a *liquid* prior is helpful for recognizing *"spill something"*. Similarly, *something untwistable* is helpful for predicting *"pretend or try and fail to twist something"*. Meanwhile, the performance of actions related to the physical localization of objects, such as *"turn the camera upwards while film something"* and *"move something away"* also gets better. The improvement on these action categories strongly indicates that our dual attention mechanism facilitate the model to trace the object manipulated by the actor.

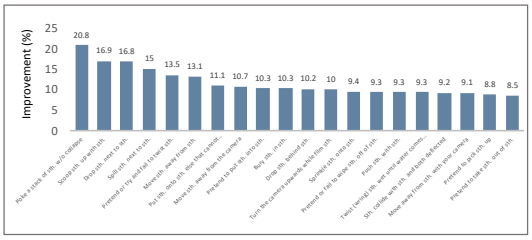

Figure 5: **Class-wise improvements** of the Dual Attention Network with respect to the baseline model. Action classes closely associated with certain objects are improved.

Self attention vs. Dual attention. We train a model by generating attention maps from image features only w/o the guidance of priors, termed self-attention model. Table 1c compares our approach with the self-attention one. The inferior performance of self attention suggests that actions and objects priors indeed provide useful information for objects and actions recognition respectively.

3.4. Comparisons with state-of-the-art methods

Table 2 summarizes results on Something-Something V2. We do not use the earlier version as object annotations are missing in V1 and label noise is greatly reduced in the latest release. We compare our approach with TSN [42], I3D [2], 2D and 3D CNNs with LSTM from Something-Something [28], and previous state-of-the-art TRN and TSM. These approaches differ from each other in many aspects such as backbones, temporal feature fusion techniques, training schemes, number of input frames, model domains and modalities. Still, models with the dual attention module surpass all their counterparts. Specifically, TSN dual attention is better than original TSN by $1.0/1.9\%$ points; TSM dual attention is better than original TSM by $1.0/0.7\%$ points; TRN dual attention achieves top-1 accuracy of 51.6% and top-5 accuracy of 80.3% on the validation subset, which is better than any previous 2D model. We conjecture that performance boost of TRN being larger than it of TSN and TSM is because TRN has more complex recognition heads so that it might be able to better exploit attended features. When TRN dual attention is turned into a two-stream models by adding an optical flow branch in TRN, our approach further boosts the performance to $58.4/85.2\%$.

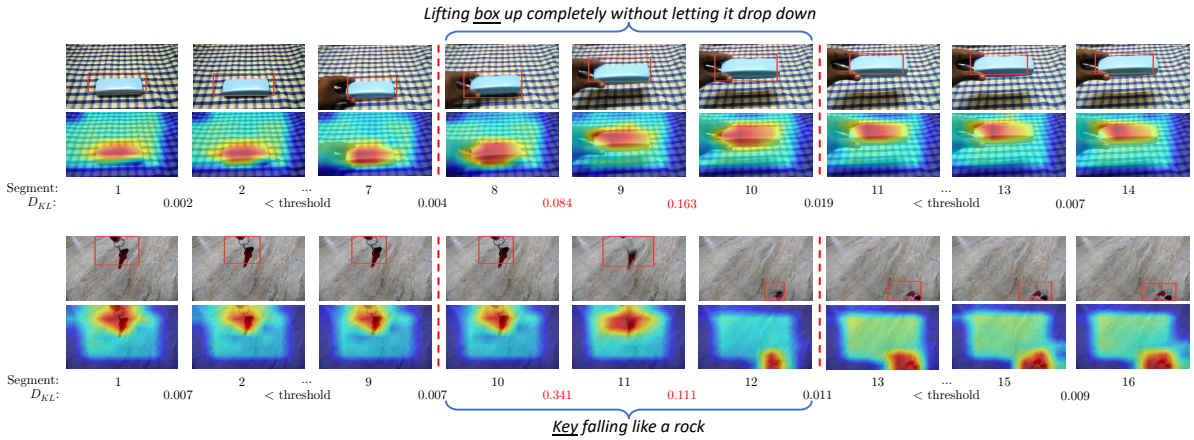

Figure 6: **Visualization of spatial and temporal localization**. We visualize one frame out of each segment (four frames). Our method find the object being manipulated, as well as the segments in which the action actually happens.

4. Weakly-supervised Localization

We can reason about human-object interactions by inspecting attention maps yielded by the model. Here, we apply the model to two weakly-supervised localization tasks: *spatiotemporal localization* and *object affordance segmentation*.

4.1. Spatiotemporal localization

The attention maps, learned from video-level action and object labels only, can accurately localize objects in the spatial domain, and actions in the temporal domain. The localization task requires the object attention map of *all* frames available. This is achieved by running the dual attention network alone with predictions of the object and action from our model, and the CNN features extracted from every frame, as input. A single forward pass with a batch of n frames can generate n attention maps.

The "where": spatial object localization. We generate object bounding boxes by thresholding the object attention map. We set the threshold to 60% of the maximum weight in the map. We then apply the flood-fill algorithm to find the connected regions. Bounding-boxes are generated by calculating the minimum and maximum coordinates of each region. We always take the largest bounding-box as a prediction, while the second largest one (if available) is optionally taken based on its size and the number of predicted objects.

The "when": temporal action localization. We observe that a large amount of human-object interactions take place once the object starts to move. Thus, we can associate the start and end of an action via the alteration of the attention maps. We divide a video into segments covering $1/3$ second each, *i.e.*, four frames in one segment for videos from the Something-Something dataset. We average the attention maps within each segment to reduce the margin of error. We measure the difference between two object attention maps

Action Category	**A.D.**
Piling something up	49.1
Stacking number of something	48.4
Pouring sth. into sth. until it overflows	47.9
Pouring something into something	43.7
Digging something out of something	43.5
Pretending to put something on a surface	35.4
Spinning something so it continues spinning	34.8
Putting something on a surface	29.8
Spinning something that quickly stops spinning	28.5
Tipping something over	27.7
Uncovering something	26.8
Something falling like a rock	25.7
Throwing something onto a surface	22.7

Table 3: The average duration (A.D.) of *trimmed* videos in each action category. The A.D. is measured by frames and the fps rate is 12. The results are in accordance with our human knowledge.

Model	IoU=0.3	IoU=0.4	IoU=0.5
Dual Attention	72.5	56.0	33.7
Self Attention	62.2	40.4	26.4

Table 4: Object localization results (in Average Precision) of dual attention and self attention on the *validation* subset.

P and Q (as they are two discrete probability distributions) via the Kullback–Leibler divergence, *i.e.*:

$$D_{\mathrm{KL}}(P||Q) = -\sum_i P(i) \log \frac{Q(i)}{P(i)} \tag{12}$$

where the sum is over the discrete points in the domain of the distribution. We filter out the leading (trailing) segments if the difference to its preceding (succeeding) segment is below

a threshold. The remaining segments are considered as the interval in which an action happens. We set the threshold to 0.06. We consider an action spanning over the entire video if the filtered video is shorter than one second to avoid actions such as *"holding something"* or *"showing something"*.

Results. We perform temporal and spatial localization on videos from the validation subset (see Figure 6). We can see that the object being interacted is highlighted and a reasonable bounding-box associated is generated accordingly. Due to the stable and accurate attention map, we are able to eliminate leading and trailing frames irrelevant to the action and find the segments wherein the box is being lifted and the key is falling. We note that compared to using optical flow our approach has more advantages that it can be performed with sparsely sampled segments if the video is very long, and is more robust to camera shake.

We conduct quantitative evaluation of weakly-supervised spatial localization by dual attention model and self attention model on the validation subset, as shown in Table 4. We randomly sample 100 videos from validation subset and annotate 2 random frames in each video. We report the average precision (AP) under various intersection-of-union (IoU) criteria. As can be seen from Table 4, our dual attention model yields much better localization accuracy than the self-attention model, indicating that action priors help the model better localize the object being manipulated. We further analyze the statistics of the trimmed videos. Out of total 24,777 videos, 16,592 ($\sim$67%) are trimmed by our temporal localization technique. The average trimmed length is 13 frames ($\sim$1 second), which, compared to the average length of 3.1 seconds, accounts for $1/3$ of overall frames. After performing temporal localization, we also analyze the average length of videos in each action category and summarize the results in Table 3. The longest actions involve *"piling something up"* and *"stacking number of something"* whereas the shortest ones involve *"throwing something onto a surface"* and *"something falling like a rock"*. It is also interesting to see that *pretending* to do something is longer than actually doing it, something *continues* spinning is longer than it quickly stops spinning, and pouring something until *overflowing* is longer than pouring something.

4.2. Object affordance segmentation

Humans can learn the roles of different parts of an object by observing how the object is being used, *i.e.*, by watching examples of "pouring water into a bottle" we can infer not only that water can be poured, but also through which part of the bottle the water can be poured. Our model can learn that detailed information. Given a question such as *"Where to plug cables?"*, we find the videos with related labels, *e.g.*, videos labeled with *"plugging a cable into a computer"*. Note that since we would have acquired the ground-truth object label when retrieving videos, the ground-truth object-guided attention is used to create attention maps; meanwhile, this model is trained with ground-truth objects to perform action recognition thus it mixes actions and objects for affordance discovery. After acquiring the attention maps, we then segment by a threshold of 60% of the maximum attention weight. The results are in Figure 7: the model focuses on the object parts associated with the action instead of the whole object. For example, the model focuses on the brim of a cup for videos involving *pouring something into a cup* or *putting something into a cup*. Importantly, the model knows to focus on the handle of a shelf even in a still image. This enables us to effectively parse object parts and infer their affordance even when the labels used for training are at the video-level.

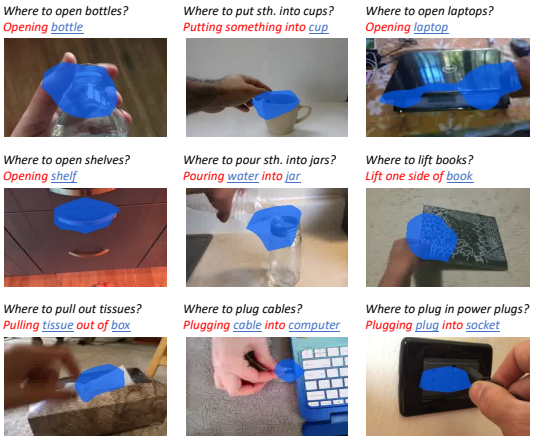

Figure 7: **Examples of object-affordance segmentation**. The model trained with video-level annotations can find object parts associated with possible ongoing actions.

5. Conclusion

Dual Attention Networks is proposed to recognize human-object interactions. It achieves very competitive performance on Something-Something V2 dataset. Based on actions/objects priors, The model is able to produce intuitive and interpretable attention maps which can enhance video feature representations for improving the recognition of both objects and actions and enable better video understanding.

Acknowledgement: This work was supported by the MIT-IBM Watson AI Lab, as well as the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00341. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

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

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017. 2, 4, 5, 6

[3] Vincent Delaitre, Josef Sivic, and Ivan Laptev. Learning person-object interactions for action recognition in still images. In *Advances in neural information processing systems*, pages 1503–1511, 2011. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009. 2, 5

[5] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[6] Wenbin Du, Yali Wang, and Yu Qiao. Recurrent spatial-temporal attention network for action recognition in videos. *IEEE Transactions on Image Processing*, 27:1347–1360, 2018. 2, 3

[7] Vibekananda Dutta and Teresa Zielinska. Action prediction based on physically grounded object affordances in human-object interactions. In *Robot Motion and Control (RoMoCo), 2017 11th International Workshop on*, pages 47–52. IEEE, 2017. 2

[8] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J. Lim. Demo2vec: Reasoning object affordances from online videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[9] James Jerome Gibson. The senses considered as perceptual systems. 1966. 1

[10] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions, 2017. 1, 2

[11] 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 *The IEEE International Conference on Computer Vision (ICCV)*, volume 1, page 3, 2017. 1, 3, 5

[12] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009. 1, 2

[13] 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. 1, 5

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[15] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 2

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*. 5

[18] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5308–5317, 2016. 2

[19] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. Joint learning of object and action detectors. In *ICCV 2017-IEEE International Conference on Computer Vision*, 2017. 1

[20] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2

[21] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 5

[22] Hema S Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):14–29, 2016. 2

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[24] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 2

[25] Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. *arXiv preprint arXiv:1811.08383*, 2018. 4, 5

[26] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6, page 2, 2017. 2

[27] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016. 2

[28] Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. Fine-grained video classification and captioning. *arXiv preprint arXiv:1804.09235*, 2018. 5, 6

[29] Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*, 2016. 2

[30] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. 5

[31] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. *arXiv preprint arXiv:1808.07962*, 2018. 1

[32] Vignesh Ramanathan, Jonathan Huang, Sami Abu-El-Haija, Alexander Gorban, Kevin Murphy, and Li Fei-Fei. Detecting events and key actors in multi-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3043–3053, 2016. 2

[33] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015. 2, 3

[34] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621, 2016. 2

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 2

[36] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[37] 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. 5

[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[39] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[40] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[41] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011. 2

[42] 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. 4, 5, 6

[43] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[44] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1

[45] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 5987–5995. IEEE, 2017. 1

[46] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406, 2016. 2

[47] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016. 2

[48] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016. 2

[49] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 2

[50] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 3, 4, 5