IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 19, 2021, accepted May 31, 2021, date of publication June 18, 2021, date of current version June 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3090471

Joint Object Affordance Reasoning and Segmentation in RGB-D Videos

SPYRIDON THERMOS[1], GERASIMOS POTAMIANOS[2], (Member, IEEE), AND PETROS DARAS[3], (Senior Member, IEEE)

[1]School of Engineering, The University of Edinburgh, Edinburgh EH9 3JL, U.K.
[2]Department of Electrical and Computer Engineering, University of Thessaly, 38221 Volos, Greece
[3]Visual Computing Lab, Centre for Research and Technology Hellas, Information Technologies Institute, 57001 Thessaloniki, Greece

Corresponding author: Spyridon Thermos (sthermos@ed.ac.uk)

This work was supported in part by the European Commission through the H2020 Project HR-Recycler under Contract 820742, and in part by Nvidia Corporation through the Titan X GPU Donation.

ABSTRACT Understanding human-object interaction is a fundamental challenge in computer vision and robotics. Crucial to it is the ability to infer ''object affordances'' from visual data, namely the types of interaction supported by an object of interest and the object parts involved. Such inference can be approached as an ''affordance reasoning'' task, where object affordances are recognized and localized as image heatmaps, and as an ''affordance segmentation'' task, where affordance labels are obtained at a more detailed, image pixel level. To tackle the two tasks, existing methods typically: (i) treat them independently; (ii) adopt static image-based models, ignoring the temporal aspect of human-object interaction; and / or (iii) require additional strong supervision concerning object class and location. In this paper, we focus on both tasks, while addressing all three aforementioned shortcomings. For this purpose, we propose a deep-learning based dual encoder-decoder model for joint affordance reasoning and segmentation, which learns from our recently introduced SOR3D-AFF corpus of RGB-D human-object interaction videos, without relying on object localization and classification. The basic components of the model comprise: (i) two parallel encoders that capture spatio-temporal interaction information; (ii) a reasoning decoder that predicts affordance heatmaps, assisted by an affordance classifier and an attention mechanism; and (iii) a segmentation decoder that exploits the predicted heatmap to yield pixel-level affordance segmentation. All modules are jointly trained, while the system can operate on both static images and videos. The approach is evaluated on four datasets, surpassing the current state-of-the-art in both affordance reasoning and segmentation.

INDEX TERMS Object affordances, human-object interaction, reasoning, semantic segmentation, deep learning, encoder-decoder model, attention mechanism, RGB-D video.

I. INTRODUCTION

Despite significant advances in the tasks of object detection [1]–[7], tracking [8]–[11], segmentation [12]–[16], and classification [17]–[20], the goal of understanding the utility of perceived objects remains elusive. To accomplish this objective, a computer vision system should be able to predict the so-called ''object affordances'', namely the set of actions that humans can perform while interacting with an object [21], and, further, to locate which object parts support interaction. Our paper aims to advance this challenging topic, focusing on the problem of pinpointing object affordances in visual data. Specifically, it addresses this at both a coarse information level, in the form of localizing object affordances as image heatmaps, as well as at a finer, more detailed extent, seeking the object affordance segmentation at pixel level.

In the literature, object affordances have been traditionally considered as auxiliary information in computer vision and robotics applications, focusing on the effect that applied actions have on object appearance [22]–[25]. More recently though, there has been increasing interest in affordance localization and segmentation, since these tasks lead to more detailed representations of affordance information that can, for example, be exploited in scene understanding and action recognition. Regarding localization, existing approaches [26]–[29] rely mostly on predicting saliency-based heatmaps in static images, without associating them with specific affordance classes though. Similarly,

The associate editor coordinating the review of this manuscript and approving it for publication was Xiaohui Yuan.

existing affordance segmentation methods [30]–[34] predict affordance classes at the pixel level on objects detected in static images.

However, such image-based approaches to affordance localization and segmentation base their learning on static representations that lack temporal information. Hence, they are limited by the sheer fact that affordances are spatio-temporal by nature, since they involve interaction. It is thus not surprising that very recent works on affordance reasoning (i.e., affordance localization and recognition, combined), have successfully demonstrated the advantages of learning spatio-temporal features, compared to predicting salient hotspots using static information alone [35], [36]. These works adopt the so-called ''learning from observation'' paradigm, processing human-object interaction videos to learn the object part that affords specific actions to be carried out. Interestingly, these methods do not depend on object details.

Inspired by the above, in this paper we proceed further and argue that affordance reasoning can be exploited to improve affordance segmentation by a joint modeling approach within the aforementioned ''learning from observation'' paradigm, aiming to achieve robust affordance understanding. In particular, we advocate that it is possible to localize and segment object affordances based on hand-object interaction information, without requiring strong object-related supervision (i.e., object labels and bounding boxes) or an intermediate step of object detection. We provide a high-level visualization of our strategy in Fig. 1.

We develop our approach starting from our recent preliminary work [37], where we proposed an end-to-end encoder-decoder model for affordance segmentation. As depicted in Fig. 2(a), that model was based on spatio-temporal information encoded from human-object interaction, and it contained a single decoder that relied solely on a soft-attention mechanism to focus on the interaction location and guide affordance segmentation. Here, we extend this early architecture to serve our joint affordance reasoning and segmentation strategy, by introducing a model with two decoders, one for each task, that can be trained jointly for both. To demonstrate the advantages of having a dedicated decoder for affordance localization, we first investigate the reasoning task independently via a single decoder, as shown in Fig. 2(b), and we compare the resulting model to state-of-the-art affordance reasoning approaches. We then introduce the second decoder targeting semantic segmentation to form our proposed model, as depicted in Fig. 2(c), and train it jointly without any object-related supervision. To operate, our model encodes at its front-end both appearance and motion information of human-object interaction through two respective encoders, each learning from both RGB and depthmap visual streams.

Thus, our main contribution lies on the exploration of both affordance reasoning and segmentation tasks within a joint spatio-temporal approach with no need of strong object-related supervision, introducing an end-to-end

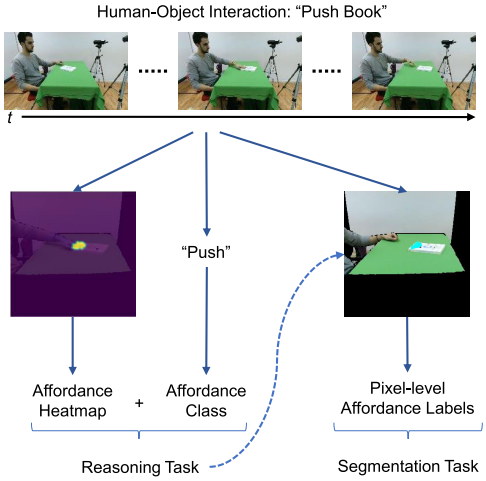

FIGURE 1. High-level overview of our learning approach to joint object affordance reasoning (left) and segmentation (right) based on human-object interaction (top). The latter is used to learn the heatmap of the interaction spot and the corresponding affordance class, with the predicted heatmap further utilized to improve pixel-level affordance segmentation. Notably, our approach does not require strong object-related supervision (i.e., object class and bounding box) and can operate on both static images and videos.

convolutional dual encoder-decoder model that encodes color, depth, and motion information from human-object interaction data and predicts affordance heatmaps, classes, and segmentations. To our knowledge, this constitutes the first such work in the affordance understanding literature.

In order to investigate the suitability of our proposed approach, we conduct a large number of experiments on our publicly released SOR3D-AFF corpus,[1] which we introduced recently [37] to support affordance reasoning and segmentation research with its numerous RGB-D human-object interaction videos and corresponding annotations. In addition, whenever suitable, we consider alternative datasets available in the literature [30]–demo2vec2018cvpr. As we report in our experiments, we significantly outperform the state-of-the-art in both affordance reasoning and segmentation tasks.

The remainder of the paper is organized as follows: Section II overviews related work on affordance reasoning and segmentation; Section III details the architecture of the proposed encoder-decoder model; Section IV presents the experimental framework and results; and, finally, Section V summarizes the work.

II. RELATED WORK

Object affordance information has attracted significant interest in the literature, as it can be exploited in a wide spectrum of computer vision and robotics tasks. Here, related work is summarized in three categories.

A. AFFORDANCE AS AUXILIARY INFORMATION

It is well-established by cognitive neuroscience that the human brain leverages object affordance information to identify objects or to interact with them. This fact has

[1] Available at http://sor3d.vcl.iti.gr/

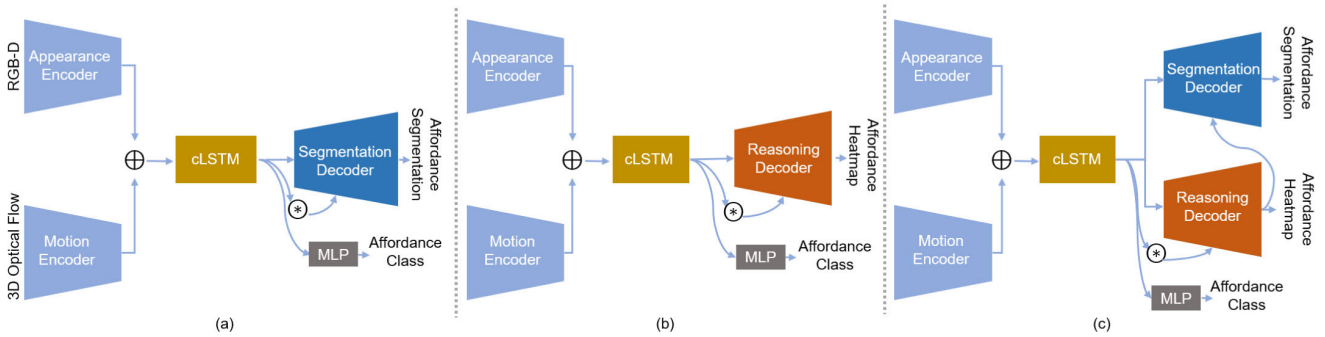

FIGURE 2. Overview of the proposed model and its two variants considered. In all cases, appearance and motion information of human-object interaction is encoded, producing spatio-temporal embeddings through a convolutional LSTM (cLSTM) module. These are fed to the soft-attention mechanism (⊛) focusing on the human-object interaction spot, to the MLP-based affordance recognition branch, and to three different decoding schemes, namely: (a) the segmentation-only decoder of our preliminary work [37], predicting pixel-level affordance labels; (b) the reasoning-only decoder, predicting affordance heatmaps (detailed in Fig. 3); and (c) the proposed two-decoder model, jointly trained for both affordance reasoning and segmentation (detailed in Fig. 5).

motivated early research on the potential of affordance information in computer vision [25], [38], [39]. For example, affordance information has been exploited through active embodiment and interaction observation to improve object recognition [40]–[42], or used to learn semantics and boost object localization for improved scene understanding [43], [44]. In action understanding, object affordances have been utilized for action anticipation [45]–[49], hand grasp generation [50], [51], and used as context information to improve action recognition [52], [53].

B. AFFORDANCE REASONING

The affordance reasoning task is realized as the combination of affordance localization and recognition. Early studies on the topic focus solely on its localization aspect, exploring saliency-based methods to predict affordance heatmaps on static images [26]-salgan. In contrast, more recent works adopt the ''learning from observation'' perspective by processing human-object interaction videos, and reason about object affordances by associating each predicted heatmap with the corresponding affordance class. Specifically, Fang *et al.* [35] present ''Demo2Vec'' that learns spatio-temporal embeddings from product demonstrations and predicts keypoints on the object affordance part, while Nagarajan *et al.* [36] propose a model that infers heatmaps based on gradient-weighted attention maps for pre-defined actions.

C. AFFORDANCE SEGMENTATION

Object affordance segmentation, i.e., the pixel-wise identification of the object part that enables a specific interaction, is a challenging task that has been mostly treated in the literature as a static semantic segmentation problem, in most cases in a strongly supervised fashion and usually coupled with object detection. For example, Myers *et al.* [30] used hierarchical matching pursuit, as well as normal and curvature features derived from RGB-D data, to learn pixel-wise labels of affordances for common household objects, while Nguyen *et al.* [31] proposed an encoder-decoder architecture to predict pixel-wise affordances based on depthmaps. Similarly, Do *et al.* [32] expanded the architecture of [31] by adding a region proposal network [54] to predict the bounding box of the target object and also investigated the joint learning of detecting and segmenting the object affordance part. Further, Sawatzky *et al.* [34] proposed a weakly-supervised setting using convolutional neural networks (CNNs) and keypoints annotation to predict reasonable but not precise pixel-level affordances, which they subsequently refined by the GrabCut algorithm [55]. Finally, deviating from the above, in our recent preliminary work [37] we proposed a spatio-temporal model with a single decoder to learn affordance segmentations from human-object interaction videos. Its architecture constitutes the basis of this paper, and it is suitably extended to jointly address object affordance reasoning and segmentation, as discussed in the Introduction and detailed below.

III. MODEL ARCHITECTURE AND LEARNING

The architecture of the proposed encoder-decoder model for joint object affordance reasoning and segmentation is depicted in Fig. 2(c) and is inspired by the so-called U-Net model [56]. More specifically, it consists of two encoders (one capturing appearance and the other motion information), a bottleneck with a residual block and a convolutional long short-term memory (cLSTM) module [57], the soft-attention mechanism of our earlier work [37], a multi-layer perceptron (MLP) for affordance class recognition, and two decoders (one for affordance hotspot localization and one for its segmentation at the pixel level). The architecture includes skip connections between the layers of the appearance encoder and the two decoders, as such are known to help recover the full spatial resolution and improve gradient flow [56]-shelhamer, he2016cvpr. In the following, prior to the description of all model components, we present the supported input representations.

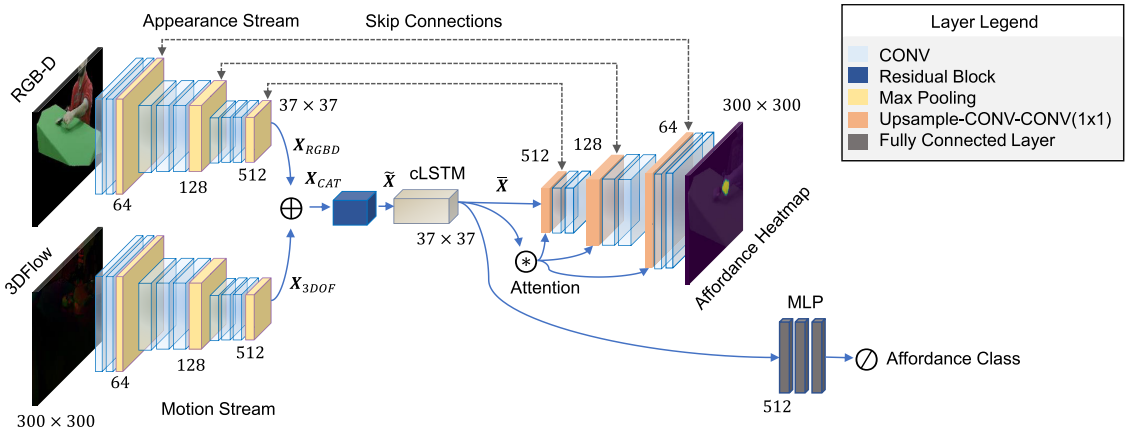

FIGURE 3. Detailed architecture of the affordance reasoning model of Fig. 2(b). From left to right: (i) RGB-D and 3D flow input streams pass through respective convolutional encoders, and the resulting encoded features are fused; (ii) the latent space consists of one residual block and a cLSTM module, followed by a soft-attention mechanism (⊛) that is further detailed in Fig. 4; (iii) the convolutional decoder predicts the affordance heatmap; and (iv) the fully-connected network (MLP) predicts the affordance class following a softmax operator (⊘). Skip connections between the appearance encoder and the reasoning decoder are also used. In the diagram, single numbers below or above layers denote the number of channels used, whereas $H \times W$ numbers indicate spatial resolution.

A. INPUT REPRESENTATIONS

Localizing and segmenting object parts based on human-object interaction can benefit from both appearance and motion information. Thus, in our approach we use color and depth information to accurately represent appearance, while for motion we employ 3D optical flow that can efficiently encode the temporal dynamics of hand movement [60]. The resulting representations are then fed to corresponding encoders, as described in Section III-B.

In more detail, we choose to combine RGB and depth information, by first mapping each RGB frame to the corresponding depthmap resolution (employing the RGB-D alignment process described in [42]) and subsequently appending the resulting color frame to the depthmap along the channel dimension. This yields a $4 \times H \times W$ input, where $H = W = 300$ represent the height and width of the input image and depthmap after center-cropping (see also Section IV).

Regarding 3D optical flow, we utilize the algorithm of [61] that computes the 3D motion vectors between consecutive RGB frames and their corresponding aligned depth images. We then colorize these vectors by normalizing their values along each axis to the $[0, 255]$ range, thus transforming them to a three-channel image of size $3 \times H \times W$. Such colorization enables us to exploit transfer learning by using deep-learning models that are pre-trained on large-scale image datasets.

Examples of these input representations can be found in Section IV-A, where the datasets of our experiments are described. Note also that in case of static images, the motion input representation is set to all zeros. Further, if depth information is unavailable, RGB-only appearance information and 2D optical flow are used instead. Such modifications are necessary to allow experiments on traditional image and video databases, as described in Section IV-B.

B. APPEARANCE AND MOTION FEATURE ENCODERS

In order to exploit the appearance and motion features of the human-object interaction, we encode the extracted RGB-D and the 3D flow information using two encoders, as depicted in Fig. 3 where the affordance reasoning model variant of Fig. 2(b) is detailed. Both encoders share the typical structure of a VGG CNN [62], i.e., 11 convolutional (CONV) layers each followed by a rectified linear unit (ReLU) activation function. Three max pooling (POOL) layers with 2×2 kernel size and stride 2 are used to downsample the RGB-D and 3D flow input representations, while the downsampled feature maps are concatenated at the model bottleneck.

In particular,[2] let $\boldsymbol{X}_{RGBD}^{d \times h \times w}$ and $\boldsymbol{X}_{3DOF}^{d \times h \times w}$ denote the output features of the RGB-D encoder and the 3D optical flow one, respectively, with $d = 512$, $h = w = 37$ representing the number of channels, height, and width of both features. Then, the two convolutional features are concatenated along the channel dimension and convolved with d kernels of size 1×1, producing the activation map $\boldsymbol{X}_{CAT}^{d \times h \times w}$ (see also Fig. 3).

C. BOTTLENECK AND AFFORDANCE RECOGNITION

The bottleneck of the model, also visible in Fig. 3, consists of a residual block and two cLSTM layers, aiming to capture temporal dependencies of the human-object interaction. The residual block follows the ReLU-CONV-ReLU-CONV structure, adopting the pre-activation method and the identity mapping proposed in [63] for performance improvement. Both the residual block and cLSTM use CONV layers with 3×3 kernel size and stride equal to 1. The activation maps after the residual block and the last cLSTM layer, respectively denoted by $\widetilde{\boldsymbol{X}}^{d \times h \times w}$ and $\overline{\boldsymbol{X}}^{d \times h \times w}$, have the same dimensionality.

[2] In our notation, we use bold-italic capitals for vectors and matrices, with optional superscripts denoting their dimensionality, and plain italic capitals for their elements, with subscripts denoting their position.

The cLSTM layers are followed by a soft-attention mechanism that is detailed in Section III-D.

Besides the attention module, the cLSTM output is also processed by an average pooling and three 512-dimensional MLP layers, respectively. The MLP is followed by a softmax classifier for affordance recognition. We use this to further regularize the model parameters during training, as well as to associate the predicted affordance label to the corresponding affordance heatmap of the reasoning decoder (detailed in Section III-E). Note that we choose to place the affordance recognition branch after the cLSTM module, inspired by the 2D/3D action recognition literature, where both appearance and motion features are encoded in the context of various CNN-LSTM model architectures [64], [65].

D. SOFT-ATTENTION MECHANISM

Depending on object detection before predicting its affordance requires extra supervision of the object class and bounding box, while also adding significant complexity to the model architecture. Since the affordance part of the object is ''exposed'' during the interaction with the human, we instead chose to utilize the soft-attention mechanism of [37], which forces the model to focus on that specific object part based on both spatial and temporal information [66].

The aforementioned mechanism is object-agnostic, and its structure is summarized in three steps, as depicted in Fig. 4: First, the $\widetilde{\boldsymbol{X}}$ and $\overline{\boldsymbol{X}}$ activation maps are concatenated at the channel dimension, and the resulting feature is convolved using a kernel of size 1×1. Then, the softmax function is employed to normalize the activation values to the $[0, 1]$ range, forming the ''excitation'' (attention) mask $\boldsymbol{M}^{1\times h\times w}$. Finally, the mask is multiplied with each channel of $\overline{\boldsymbol{X}}$ in an element-wise manner. The result of this mechanism is upsampled by nearest-neighbor interpolation and is re-applied to the activation maps of the reasoning decoder after each upsampling layer, as also shown in Figs. 3 and 5.

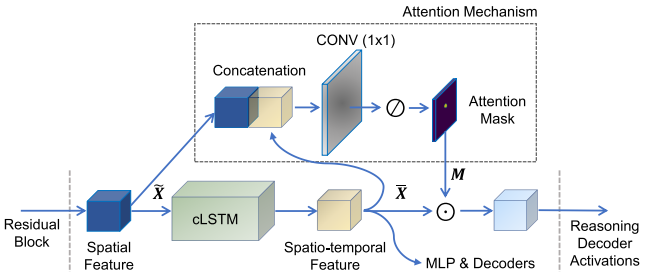

FIGURE 4. Details of the employed spatio-temporal soft-attention mechanism: The spatial feature of the last residual block of the network bottleneck is first concatenated with the cLSTM output. The result is processed by a CONV layer with kernel size 1 × 1 and a softmax operator (⊘) to yield the attention mask. The attention is applied to the latent pipeline by multiplying the mask with the spatio-temporal feature in an element-wise manner (⊙).

E. REASONING AND SEGMENTATION DECODERS

We use separate decoders for the tasks of affordance reasoning and segmentation. These have similar structure, however the segmentation decoder is deeper, since more detailed spatial information is required for semantic segmentation at the pixel level, as compared to the coarser heatmap prediction.

In more detail, the reasoning decoder is a combination of 6 CONV, 6 ReLU, and 3 upsampling layers, and predicts an $H \times W$ heatmap, as also depicted in Fig. 3 (or in the middle part of Fig. 5). More specifically, after each upsampling layer a CONV layer follows, and its output feature is concatenated with the corresponding one from the appearance and motion encoders through skip connections. A CONV layer with 1×1 kernel size follows, forcing intra-channel correlation learning. Note that each channel of the produced activation map is multiplied with the attention mask $\boldsymbol{M}$ in an element-wise manner, as described in Section III-D. The last CONV layer results in a $1 \times H \times W$ feature, where a softmax function is applied over all its values to yield the final heatmap $\boldsymbol{D}^{H\times W}$ that can be viewed as a probability mass function. Note that the ReLU activation function is employed after each CONV layer, and that nearest-neighbor interpolation is used for upsampling.

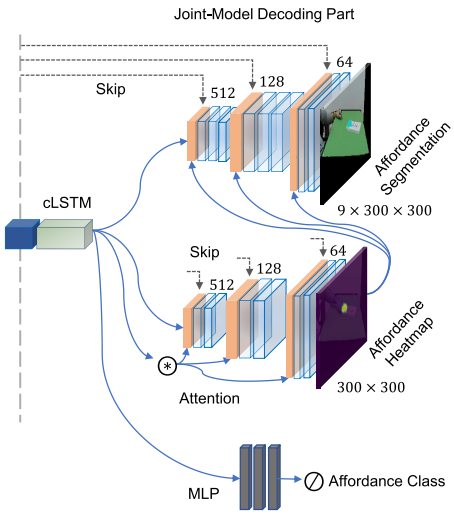

FIGURE 5. Details of the decoding part of the proposed model. The MLP and the two decoders receive the spatio-temporal cLSTM feature as input. The soft-attention mechanism (⊛) guides the reasoning decoder at different levels of granularity, while the predicted interaction hotspot operates as an additional attention mechanism, masking the segmentation decoder activations to improve the pixel-wise prediction of nine affordance classes.

As mentioned, the segmentation decoder follows a similar architecture with the reasoning one, but employs a larger number of CONV and ReLU layers (14 each) in order to better preserve spatial information details. The main difference between the two decoders is that, instead of using the output of the soft-attention mechanism to mask the activations after each upsampling module, the segmentation decoder exploits the predicted affordance heatmap $\boldsymbol{D}$. For this purpose, $\boldsymbol{D}$ is multiplied in an element-wise manner with each channel of the activation map after each upsampling layer. Note that the heatmap is downsampled to two different spatial resolutions, namely 75×75 and 150×150, in order to match the

height and width of the activation map after each upsampling layer. The segmentation decoder finally produces a 3D feature $\boldsymbol{U}^{C\times H\times W}$ using a softmax function, or equivalently a total of C predicted 2D affordance maps, where C denotes the number of affordance classes (nine, in this paper). Its architecture is also depicted in Fig. 5 (upper part).

F. JOINT-TASK LEARNING

We argue that the affordance reasoning and segmentation tasks are complementary to each other as: (i) their predictions are based on the same spatio-temporal embedding that is designed to focus on the human-object interaction hotspot; and (ii) the segmentation task can benefit from the localization of this hotspot, as the affordance heatmap and segmentation mask should overlap.

To take advantage of this complementarity, we train our model jointly for the two tasks, by minimizing the following loss function:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{seg}} + \lambda_2 \mathcal{L}_{\text{heat}} + \lambda_3 \mathcal{L}_{\text{aff}}, \tag{1}$$

where $\lambda_1, \lambda_2, \lambda_3 \in [0, 1]$ are hyper-parameters that add to 1. In (1), we compute $\mathcal{L}_{\text{heat}}$ as the Kullback-Leibler divergence (KLD) between the predicted $\widehat{\boldsymbol{D}}$ and the ground-truth $\boldsymbol{D}$ heatmaps (probability mass functions) as follows:

$$\mathcal{L}_{\text{heat}} = \sum_{i=1}^{H}\sum_{j=1}^{W} \widehat{D}_{i,j} \log \frac{\widehat{D}_{i,j}}{D_{i,j}}, \tag{2}$$

with label smoothing of $\boldsymbol{D}$ to avoid zeros. Further, we define $\mathcal{L}_{\text{seg}}$ as the per-pixel cross-entropy of the predicted and ground-truth affordance labels (normalized over the total number of pixels), given by:

$$\mathcal{L}_{\text{seg}} = -\frac{1}{HW}\sum_{c=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W} U_{c,i,j} \log(\widehat{U}_{c,i,j}), \tag{3}$$

where $\widehat{\boldsymbol{U}}, \boldsymbol{U}$ are the predicted and the ground-truth affordance maps, respectively. Finally, we define $\mathcal{L}_{\text{aff}}$ as:

$$\mathcal{L}_{\text{aff}} = -\sum_{c=1}^{C} A_c \log(\widehat{A}_c), \tag{4}$$

where $\widehat{\boldsymbol{A}}, \boldsymbol{A}$ are the C-dimensional vectors of the predicted and ground-truth affordance labels, respectively.

We provide additional training setup details in Sections IV-B and IV-C. Further, in order to enrich the comparison of our model with an alternative approach in the literature that depends on strong object supervision, we also develop a suitable model variant incorporating additional terms to the training loss of (1), as we discuss in Section IV-C.

IV. EXPERIMENTAL FRAMEWORK AND RESULTS

We next proceed with evaluating our proposed approach. For this purpose, we first introduce our publicly available SOR3D-AFF corpus and overview three additional datasets, all used in our experiments. We then provide implementation details of our models, review the evaluation metrics used, and report a large number of experiments. In these, we compare our approach to alternative methods for the affordance reasoning and segmentation tasks, report both quantitative and qualitative results, and conduct suitable ablation studies.

A. DATASETS

To address the challenging tasks of affordance reasoning and segmentation based on spatio-temporal human-object interaction information, we have created the "Sensorimotor Object Recognition 3D AFFordance" dataset (SOR3D-AFF). This constitutes a subset of the SOR3D database that was first introduced in [24] solely for sensorimotor object recognition. The database has been subsequently augmented with suitable affordance annotations and also used in our recent preliminary work [37]. It is the first dataset to contain human-object interaction videos coupled with pixel-level affordance annotations, while also enabling the investigation of the depthmap added value to affordance reasoning and segmentation.

Besides SOR3D-AFF, we also employ the OPRA video database [35] to evaluate our reasoning model alone, since it does not have pixel-level affordance annotations to allow segmentation evaluation. Note that for both sets, we predict affordance heatmaps and/or segmentations on the so-called "target" frame of the given video sequence (as defined in their descriptions below). In addition, we use the UMD [30] and IIT-AFF [33] databases to qualitatively evaluate our segmentation model on objects unseen during training. The latter two sets contain static images only, coupled with pixel-level segmentation annotations. We provide an overview of all four databases in Table 1 and more details next.

TABLE 1. Overview of the video and static image datasets used in our experiments for affordance reasoning (R) and / or segmentation (S) evaluation with reported quantitative (N), qualitative (L), and / or ablation study (A) results.

Dataset	Data Type	# Afford. Classes	Examples Shown in	Conducted Evaluation(s)
SOR3D-AFF	video, RGB-D	9	Figs. 6, 7	**R**(N,L,A), **S**(N,A)
OPRA	video, RGB	7	Fig. 7	**R**(N,L)
UMD	image, RGB-D	7	Fig. 8	**S**(L)
IIT-AFF	image, RGB	9	Fig. 8	**S**(L)

1) THE SOR3D-AFF VIDEO CORPUS

This dataset consists of 1201 RGB-D human-object interaction videos from the SOR3D database [24], captured at 1920×1080 and 512×424-pixel resolution for the RGB and depth streams, respectively. The data are split into 962 videos for training and 239 videos for testing. SOR3D-AFF supports nine affordance classes, namely "grasp", "cut", "lift", "push", "rotate", "hammer", "squeeze", "paint", and "type", of ten household object types, namely "ball", "book", "brush", "can", "cup", "hammer", "knife", "pitcher", "smartphone", and "sponge".

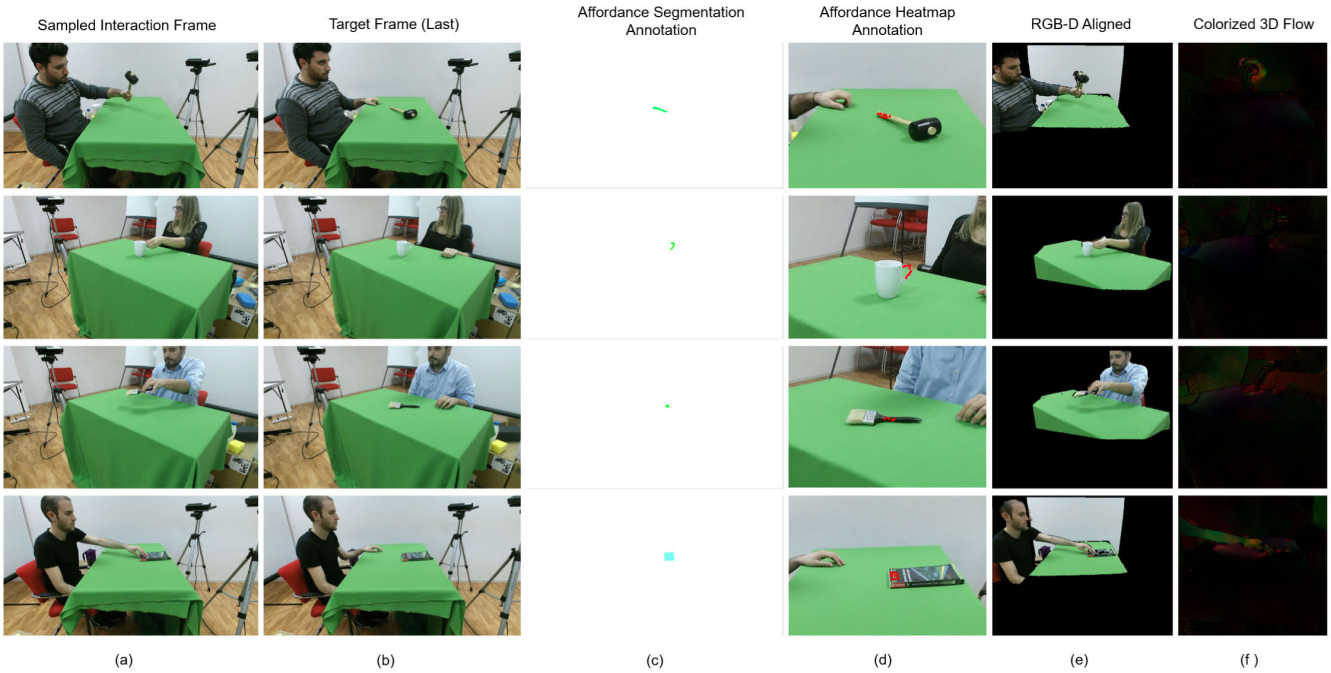

FIGURE 6. SOR3D-AFF data, annotation, and pre-processing examples. Four interaction sequences are considered (top-to-bottom rows): "hammer with hammer", "grasp cup", "lift brush", and "push book". Depicted are (left-to-right columns): (a) a sample sequence frame; (b) the last ("target") sequence frame; (c) its affordance segmentation annotation; (d) its affordance heatmap annotation (cropped for clarity); (e) the sample frame after RGB-D alignment, with the color image mapped to the depthmap 512 × 424-pixel resolution; and (f) its colorized 3D optical flow, center-cropped to 300 × 300 pixels. The resolution in (a)-(c) is 1920 × 1080 pixels.

In creating this dataset, we chose to omit some object categories from the original SOR3D set and their corresponding affordances, as their annotation turned out problematic (e.g., typically, object "pen" was mostly occluded during interaction and had very noisy depthmap, while object "box" had a removable cover, thus its shape ended up different at the interaction conclusion). Further, we only considered SOR3D videos with the interaction spot visible in all frames.

The developed dataset is endowed with multiple annotations, but only at the last frame of each video sequence ("target" frame), namely: (i) pixel-level affordance labels; (ii) the affordance heatmap, based on Gaussian blurring of 10-15 marked pixels that indicate the human-object interaction hotspot; (iii) the affordance label; (iv) the object bounding box; and (v) the object label. Indicative SOR3D-AFF examples are depicted in Fig. 6, including some of the aforementioned annotations, as well as the input representations detailed in Section III-A.

2) THE OPRA VIDEO CORPUS

This dataset consists of 20,774 RGB video clips at various resolutions, split into 16,976 clips for training and 3,798 clips for testing. Each video contains the demonstration of an appliance feature that involves human-object interaction (e.g., "scoop food from the pan"), and it is paired with a static image that depicts the corresponding object (e.g., "pan") without any background or occlusion ("target" image). Each such image is annotated with an affordance heatmap, which is the result of Gaussian blurring applied on ten marked pixels that indicate the human-object interaction spot. In total, OPRA supports seven affordance classes, namely "hold", "touch", "rotate", "push", "pull", "pick up", and "put down".

3) THE UMD IMAGE CORPUS

This dataset contains color images and aligned depthmaps (both at 640×480-pixel resolution) of 105 kitchen, workshop, and garden tools that belong to 17 object categories and are captured at various view-points. These static images are accompanied by pixel-level affordance labels, corresponding to seven affordance classes, namely "grasp", "cut", "scoop", "contain", "pound", "support", and "wrap-grasp".

4) THE IIT-AFF IMAGE CORPUS

This dataset contains images from ImageNet [67] and additional ones at various resolutions. All images depict cluttered scenes that include multiple objects, and they are annotated with pixel-level affordance labels and object bounding boxes. The IIT-AFF dataset supports nine affordance classes, namely "contain", "cut", "display", "engine", "grasp", "hit", "pound", "support", and "w-grasp".

B. AFFORDANCE REASONING EVALUATION

We first evaluate the affordance reasoning branch of our proposed architecture, i.e., the model depicted in Figs. 2(b) and 3. Specifically, we provide implementation details of the model (including a variation of it), alternative algorithms to compare

it against, the evaluation metrics used, our quantitative and qualitative results, as well as an ablation study to justify its chosen architecture. This evaluation is conducted on the SOR3D-AFF and OPRA video datasets, since the UMD and IIT-AFF sets lack heatmap annotations.

1) IMPLEMENTATION DETAILS AND A PROPOSED MODEL VARIANT

We temporally sub-sample the videos of SOR3D-AFF and OPRA to 10 frames per second (fps) and center-crop all frames to 300×300 pixels. We pre-train the two encoders of our model on separate datasets. Specifically, we train the RGB-D encoder followed by the cLSTM for 50 epochs on the UTKinect action recognition dataset [68], while for the colorized 3D flow encoder we utilize the weights of a VGG16 model pre-trained on ImageNet [67]. For the decoder and the MLP layer weights initialization, we employ the Xavier method [69]. We fine-tune the model in an end-to-end fashion for 80 epochs, using batch size equal to 6, Adam optimization [70], and learning rate set to 2×10^{-5}. Due to the small batch size, we choose to use group normalization [71] between each CONV and ReLU layer. Further, in (1), we set $\lambda_1 = 0$, since there is no segmentation here. We set the remaining loss-function hyper-parameters to $\lambda_2 = 0.3$ and $\lambda_3 = 0.7$ for the first 50 epochs, since recognition of the affordance class is a critical step towards the prediction of the affordance heatmap (and thus should initially guide the total loss), while for the last 30 epochs we set both weights to 0.5. We implement the model[3] in PyTorch [72] and run all computations on two Nvidia Titan X GPUs.

To be fair in our comparisons to alternative literature models that employ color information alone (e.g., without the depth stream), we consider a color-only variant of our proposed model, as already indicated in Section III-A. This variant uses 2D optical flow instead of 3D in the motion encoder. In fact, we compute the 2D displacement vector fields between sequential frames following the optical flow stacking approach of [73]. We refer to this model as "RGB-only" to easily distinguish it from our RGB-D system.

2) ALTERNATIVE MODELS FROM THE LITERATURE

We evaluate our reasoning model against the considered as the state-of-the-art in the affordance reasoning literature:

- *Demo2Vec* [35], a model designed to predict affordance heatmaps on target object images based on demonstration videos and trained with heatmap annotations. Here, we re-implement it to support the input resolutions of both SOR3D-AFF and OPRA data. Compared to our proposed model, the Demo2Vec architecture differs in the following aspects: it uses deeper encoders; instead of one cLSTM, it contains two (one for each stream), with the resulting spatio-temporal representations fused by an attention mechanism; its affordance classifier involves an LSTM; its decoder is shallow, based on transpose convolutions and tiled features from the encoder pooling layers (as in [74]); and, finally, it lacks encoder-decoder residual connections.

In addition, we compare our model against the following:

- *Grounded human-object interactions (GHOI)* [36], a model designed to predict affordance heatmaps based on human-object interaction videos. GHOI is trained in a weakly supervised manner, employing affordance class annotations alone. Here, we use the original implementation of [36].
- *Img2Heat*, adopting the term and goal from [36] and defining it as a static Demo2Vec variant. The model follows the Demo2Vec architecture, however it is trained without the video context, i.e., using only static images.
- *Saliency generative adversarial network (SalGAN)* [28], which estimates the most salient regions in an image by predicting heatmaps. The model is trained in a supervised manner using saliency annotations. Here, we consider its original implementation [28] and pre-train it on the SALICON dataset [75].

Among these models, we consider SalGAN as weakly supervised for affordance class prediction, since it is trained using saliency annotations that do not correspond to specific affordance classes. Similarly, GHOI is weakly supervised for affordance heatmap prediction, as it is trained using affordance class labels only. In contrast, Demo2Vec, Img2Heat, and our proposed model are strongly supervised, as they are all trained employing affordance heatmap annotations. Note also that SalGAN and Img2Heat can only operate on static images, in contrast to GHOI, Demo2Vec, and our model.

3) EVALUATION METRICS

To quantitatively evaluate our reasoning model against the aforementioned alternative models, we consider three metrics: (i) the KLD, described in (2), with lower values signifying better results; (ii) the similarity or histogram intersection (SIM), with higher values being better, and (iii) a variant of the area under the receiver operating characteristic curve, termed AUC-J, with higher values indicating better results.

SIM is a popular metric in the saliency literature [76], measuring the similarity between two heatmaps, $\widehat{\boldsymbol{D}}$ (predicted) and $\boldsymbol{D}$ (ground truth), that assume values within $[0\,,\ 1]$ and sum to 1 over the $H \times W$-pixel image. It is defined as:

$$\mathrm{SIM}(\widehat{\boldsymbol{D}}, \boldsymbol{D}) = \sum_{i=1}^{H} \sum_{j=1}^{W} \min(\widehat{D}_{i,j}\,, D_{i,j}). \tag{5}$$

Since both KLD and SIM are distribution-based metrics, to diversify our evaluation we also consider the location-based AUC-J metric [77], [78]. AUC-J views heatmap prediction as a binary classification problem (object affordance vs. background regions) and is computed based on the plot of the true positive vs. false positive classification rates, when comparing appropriately binarized versions of $\widehat{\boldsymbol{D}}$ and $\boldsymbol{D}$.

[3]Code is made publicly available at https://github.com/spthermo/STCAE

TABLE 2. Evaluation of our proposed affordance reasoning model against alternatives on the SOR3D-AFF and OPRA test sets, in terms of the KLD, SIM, and AUC-J metrics. Arrow ↑ indicates that higher values of the corresponding metric are better, while arrow ↓ implies the opposite.

	Dataset:	SOR3D-AFF			OPRA		
Inference	**Method**	KLD (↓)	SIM (↑)	AUC-J (↑)	KLD (↓)	SIM (↑)	AUC-J (↑)
Image-based	SalGAN	2.452	0.289	0.614	2.121	0.308	0.744
	Img2Heat	2.026	0.312	0.656	1.498	0.352	0.790
Video-based	GHOI	1.992	0.319	0.667	1.425	0.363	0.795
	Demo2Vec	1.961	0.322	0.711	1.203	0.486	0.856
	Ours (RGB-only)	1.818	0.332	0.728	**1.189**	**0.488**	**0.862**
	Ours (RGB-D)	**1.439**	**0.412**	**0.762**	n/a	n/a	n/a

4) QUANTITATIVE EVALUATION RESULTS

In Table 2, we report the performance of the aforementioned models on the SOR3D-AFF and OPRA datasets (on their test-set "target" frames), in terms of the KLD, SIM, and AUC-J metrics. To be fair in the comparisons and to accommodate OPRA that lacks depth data, we consider the RGB-only variant of our RGB-D reasoning model, which ignores depth as discussed earlier. In addition, to benchmark the depth stream contribution, we also evaluate our RGB-D model on SOR3D-AFF, where this stream is available.

Based on Table 2, we deduce that video-based models (lower part of table) are superior to ones operating on static images (upper part). Further, we readily observe that our proposed reasoning approach achieves the best results on both datasets and for all three metrics. In more detail, on the SOR3D-AFF database, our RGB-D model yields the best KLD, SIM, and AUC-J (1.44, 0.41, and 0.76, respectively), surpassing its RGB-only variant and thus demonstrating the depth value to affordance reasoning. Notably, our RGB-only model also prevails over the remaining alternatives, slightly outperforming Demo2Vec. We believe this is due to our reasoning decoder design, namely the upsampling CONV layers used instead of just transposed ones. This way, our decoder is able to preserve more fine-grained spatial information up to the heatmap prediction [79]. We observe similar trends on the OPRA dataset, where again our RGB-only model outperforms Demo2Vec slightly and all other alternatives by larger margins. It should be noted that the above comparisons (i.e., of our RGB-D model vs. its RGB-only variant, as well as of our models against their alternatives) hold with statistical significance for all metrics, based on the Wilcoxon non-parametric test at $p < 0.05$ [80].

Regarding computational resources, our model is again superior to its closest performing competitor (Demo2Vec). Indeed, it is slightly faster at inference, processing one video frame in 24 ms against 27 ms of Demo2Vec. In addition, it is leaner, having 10.7 M parameters vs. 15.8 M of the latter. These differences can be attributed to our lighter encoder architecture and the additional cLSTM of Demo2Vec, although are somewhat mitigated by its shallower decoder. Further, the pre-activation and identity mapping in the residual block of our model bottleneck contribute to its faster runtime [63].

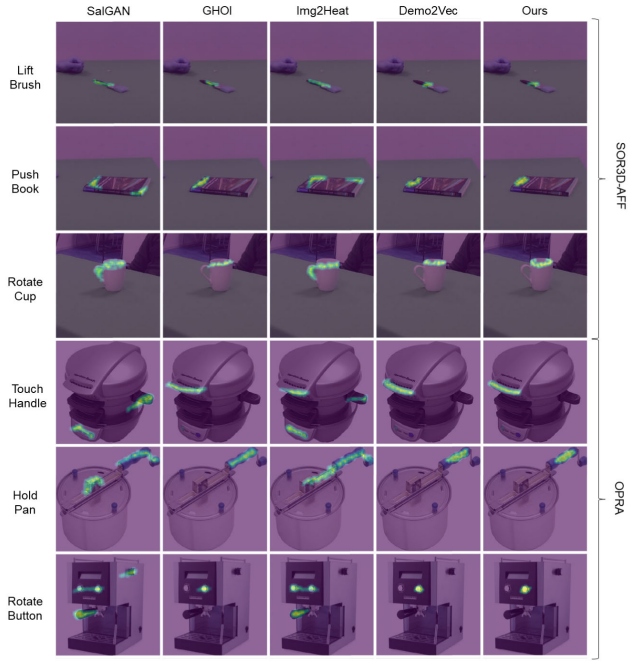

FIGURE 7. Affordance heatmap prediction on "target" images of the SOR3D-AFF (upper three rows) and OPRA (lower three rows) datasets, using different models per column (model names are shown at the top). Each heatmap is associated with the corresponding affordance class predicted by each model (left). The heatmaps are overlaid on the target images for better visualization, while the object classes are also shown for clarity (left).

5) QUALITATIVE EVALUATION RESULTS

In Fig. 7, we depict some indicative affordance heatmap predictions on target frames of the SOR3D-AFF and OPRA test sets. From the visualized examples, we observe that our model is able to predict accurate heatmaps, associated with the affordance label of the interaction (e.g., the predicted heatmap highlights the pan handle in the target image of the "hold pan" OPRA video sample). In contrast, the static approaches Img2Heat and SalGAN, both trained on images, highlight the most salient regions of the objects regardless of the affordance class (e.g., the predicted heatmap of the SOR3D-AFF "rotate cup" video sample highlights both the handle and the top area of the cup). Finally, by observing the performance of the GHOI model, which is able to effectively predict heatmaps associated with specific affordances

despite being trained without heatmap ground truth, we conclude that temporal information of human-object interaction is more valuable to affordance reasoning than strong supervision.

6) ABLATION STUDY

To demonstrate the contribution of each individual component of the proposed affordance reasoning architecture, we conduct an ablation study considering three variations for each of our RGB-only and RGB-D models. Specifically, we evaluate: (i) the single appearance-only encoder model; (ii) the single appearance-only encoder with the soft-attention mechanism; and (iii) the two-encoder model (both appearance and motion) with the soft-attention mechanism present.

In Table 3, we report performance of the six resulting variants on the SOR3D-AFF test set in terms of the KLD, SIM, and AUC-J metrics. Evidently, RGB-D models are superior to their RGB counterparts in all three cases examined. Integration of the soft-attention mechanism to the appearance-only encoder improves results, and incorporation of optical flow (2D or 3D) through the motion encoder further boosts performance of both RGB-only and RGB-D models.

TABLE 3. Ablation study on our affordance reasoning model architecture. Various model variants (with or without depth information, soft-attention mechanism ("α"), and optical flow) are evaluated for affordance heatmap prediction in terms of KLD, SIM, and AUC-J on the SOR3D-AFF test set.

Model Variations	KLD (↓)	SIM (↑)	AUC-J (↑)
RGB	2.209	0.231	0.651
RGB + α	2.031	0.294	0.697
RGB + α + 2D flow	1.818	0.332	0.728
RGB-D	2.189	0.292	0.673
RGB-D + α	1.914	0.368	0.733
RGB-D + α + 3D flow	**1.439**	**0.412**	**0.762**

In addition to heatmap prediction, the reasoning task includes affordance class inference, which is achieved by the MLP classifier in our proposed architecture (see Fig. 3). To investigate its performance, we conduct an ablation study similar in spirit to the one just described, varying the encoding scheme (note that the attention mechanism does not relate to the MLP). Specifically, for each of our RGB-only and RGB-D systems, we consider: (i) the single appearance-only encoder; and (ii) the two-encoder model.

In Table 4, we report performance of the four resulting systems on the SOR3D-AFF test set in terms of classification accuracy for each of the nine affordance classes and overall. As expected, the RGB-D models turn out superior to their RGB counterparts, showcasing the contribution of the depth stream, while the two-encoder models outperform the corresponding single-encoder ones, demonstrating the importance of temporal information. Interestingly, the above remarks hold per affordance class as well. Finally, among all classes, the "grasp" affordance reaches the highest accuracy, followed by the "hammer" one, while the more complex affordance "squeeze" achieves the lowest accuracy. These observations hold for all model variations considered.

C. AFFORDANCE SEGMENTATION EVALUATION

We finally evaluate the affordance segmentation performance of our proposed architecture, paralleling in our presentation the structure of Section IV-B. We report results on the datasets of Section IV-A with available segmentation ground truth, namely on SOR3D-AFF, UMD, and IIT-AFF.

1) IMPLEMENTATION DETAILS AND A PROPOSED MODEL VARIANT

Similarly to the reasoning setup, all data from the aforementioned corpora are resized to a 300 × 300-pixel resolution, while each video from SOR3D-AFF is decimated to 10 fps. Focusing on the joint model, we pre-train both encoders as reported in the reasoning setup and fine-tune the model for 200 epochs. We set the batch size equal to 4 and apply group normalization to the activations after each CONV layer, while we use a learning of 2×10^{-5}. Following (1), we set $\lambda_1 = 0.3$, $\lambda_2 = 0.1$, $\lambda_3 = 0.6$, and we optimize the model by the Adam algorithm [70].

In addition, since the SOR3D-AFF database also provides labels of object bounding boxes and object classes, we implement a model variant with "extra supervision" to exploit such annotations, by introducing two extra terms to training loss (1). This allows us additional comparisons with an alternative affordance segmentation model in the literature, which is based on strong object supervision and is discussed below. For this purpose, we employ the L_2-norm to quantify the object bounding-box error, defined as:

$$\mathcal{L}_{\text{bbox}} = \| \widehat{\boldsymbol{R}} - \boldsymbol{R} \|_2, \tag{6}$$

where $\widehat{\boldsymbol{R}}$ and $\boldsymbol{R}$ are 4-dimensional vectors containing respectively the predicted and ground-truth object bounding-box information (top left corner coordinates, width, and height). Further, we modify (4) to measure the object recognition loss that we denote as $\mathcal{L}_{\text{obj}}$ (the vectors of (4) now correspond to the predicted and ground-truth object classes, with the number of classes being $C = 10$). The total loss for joint model training then becomes:

$$\mathcal{L}_{\text{total}}^{(\text{upd})} = \lambda_1 \mathcal{L}_{\text{seg}} + \lambda_2 \mathcal{L}_{\text{heat}} + \lambda_3 \mathcal{L}_{\text{aff}} + \lambda_4 \mathcal{L}_{\text{bbox}} + \lambda_5 \mathcal{L}_{\text{obj}}, \tag{7}$$

where we set $\lambda_1 = 0.2$, $\lambda_2 = \lambda_4 = 0.1$, and $\lambda_3 = \lambda_5 = 0.3$. We also modify the model architecture suitably, by introducing a region proposal network (RPN) [54] and a region-of-interest (RoI) align layer [2] at the end of the model bottleneck. These are followed by a regressor and a classifier MLP, positioned parallel to the two decoders, to predict the object bounding-box coordinates and class, respectively.

2) ALTERNATIVE MODEL FROM THE LITERATURE

We evaluate our model against the state-of-the-art in the affordance segmentation literature:

TABLE 4. Ablation study on our reasoning model architecture concerning the MLP affordance classifier performance on the SOR3D-AFF test set. Four model variations are evaluated (with or without depth information and optical flow), with classification accuracy (%) reported per affordance class and overall.

Model Variations	"cut"	"grasp"	"hammer"	"lift"	"paint"	"push"	"rotate"	"squeeze"	"type"	all classes
RGB	78.72	88.87	85.76	81.58	74.81	81.05	78.25	68.51	80.73	79.81
RGB + 2D flow	83.12	91.27	87.91	84.14	77.52	84.41	80.88	72.74	85.32	83.03
RGB-D	80.87	90.04	86.88	82.74	75.66	83.28	79.34	69.68	82.51	81.22
RGB-D + 3D flow	86.71	**92.57**	90.34	86.23	80.11	86.52	83.98	76.22	88.39	**85.67**

- *AffordanceNet* [32], a convolutional encoder-decoder that operates on static images and relies on strong object supervision. The model comprises a single appearance encoder, a convolutional bottleneck, and a single decoder dedicated to affordance segmentation. Such segmentation is restricted within the bounding box of the detected object, thus, following the bottleneck, the model also contains an RPN module, a RoI Align layer, and MLPs for object classification and bounding-box prediction. Here, we re-implement AffordanceNet to allow 300 × 300-pixel input, training it for 50 epochs with batch size 8 and learning rate equal to 2×10^{-5}.

Note that we do not consider other alternatives in our comparisons, since AffordanceNet is the only model in the literature designed explicitly for affordance segmentation, being the most recent evolution of earlier versions by its authors that are now considered obsolete.

3) EVALUATION METRICS

We use two metrics to evaluate model performance for affordance segmentation. The first is the intersection-over-union (IoU), originally proposed in [81]. Adapted to the problem of interest, IoU quantifies the pixel-level overlap between the predicted affordance segmentation[4] $\widehat{\boldsymbol{U}}$ and the corresponding ground truth $\boldsymbol{U}$, and it is defined as:

$$\mathrm{IoU}(\widehat{\boldsymbol{U}}, \boldsymbol{U}) = \frac{|\widehat{\boldsymbol{U}} \odot \boldsymbol{U}|}{|\widehat{\boldsymbol{U}}| + |\boldsymbol{U}| - |\widehat{\boldsymbol{U}} \odot \boldsymbol{U}|}, \tag{8}$$

where $\odot$ denotes element-wise matrix multiplication and $|\bullet|$ is the number of ones in its matrix argument (equivalently here, the sum of the matrix elements).

The second metric used is the F-score and its weighted version, since some affordances are associated with more objects than others in our data. The weighted F-score has been proposed in [82] and is defined as:

$$\mathrm{F}_w = \frac{2\,\mathrm{P}_w\,\mathrm{R}_w}{\mathrm{P}_w + \mathrm{R}_w}, \tag{9}$$

where P_w and R_w are the weighted versions of the standard precision and recall metrics, respectively. In particular, for our experiments on SOR3D-AFF, we set the weights of the most dominant affordance classes "grasp" and "lift" to 0.2, that of next dominant class "push" to 0.1, and for the remaining six affordances we set their weights to 0.0833, so that they all add to 1.

[4]Compared to its use in (3), $\widehat{\boldsymbol{U}}$ is binarized following the softmax.

4) QUANTITATIVE EVALUATION RESULTS

In Table 5, we report the affordance segmentation evaluation of our proposed joint model on the SOR3D-AFF test set, when applied on human-object interaction videos, or just on the target (last) frame of the dataset sequences. In the first case, since there exist no alternative models in the literature for inferring pixel-level affordance labels from videos, we only provide results of our model. In the static image inference scenario though, we also evaluate AffordanceNet.

From the table, we easily deduce that video-based affordance segmentation performs much better than image-based one, with our model achieving an IoU of 0.731 against 0.559 of the static run and an F-score of 0.820 vs. 0.617. Such improvements demonstrate the capacity of our proposed approach to exploit the spatio-temporal nature of human-object interaction for object affordance segmentation, supporting our argument that a model can be trained on interaction sequences to infer affordance labels in both videos and images.

TABLE 5. Affordance segmentation evaluation of various models on the SOR3D-AFF test set, in terms of IoU, F-score, and weighted F-score.

Inference	Method	IoU (↑)	F (↑)	F_w (↑)
Video-based	Ours	**0.731**	**0.820**	**0.821**
Image-based	AffordanceNet	0.561	0.618	0.621
	Ours	0.559	0.617	0.622
	Ours (extra supervision)	**0.575**	**0.625**	**0.638**

In the static case, we observe from the bottom part of Table 5 that our model performs comparably to AffordanceNet. Note, however, that the latter depends on strong object-related supervision (e.g., object class and bounding-box information) and requires object detection. In contrast, we have designed our model with no need of such annotations or object detection, limiting its supervision to only affordance-related annotations (and only at the last frame of the interaction video). However, if we choose to exploit object-related annotations by developing a suitable variant of our model (with "extra supervision", as discussed earlier), our approach ends up outperforming AffordanceNet. Indeed, this model variant

achieves an IoU of 0.575 vs. 0.561 of AffordanceNet, an F-score of 0.625 against 0.618, and a weighted F-score of 0.638 vs. 0.621. All three gains are statistically significant, based on the Wilcoxon non-parametric test at $p < 0.05$ [80].

Concerning speed, our proposed model is significantly faster than AffordanceNet, due to the expensive object detection components of the latter. As a result, AffordanceNet requires 113 ms to infer a segmentation, assuming a single detected object. In comparison, our model performs the task in only 37 ms, adding approximately 13 ms of overhead to our reasoning-only model runtime, due to the second decoder and the heatmap-guided segmentation process. As expected, ''extra supervision'', when incorporated to our model, slows it down considerably to 97 ms, due to the region proposals and RoI alignment overhead. Still, this model variant remains slightly faster than AffordanceNet.

Regarding model size, AffordanceNet has 226.7 M trainable parameters, 186 M of which belong to the object class and bounding-box prediction branch, and the remaining 40.7 M to the encoder-decoder part. In comparison, our model is significantly lighter, having only 13.2 M parameters. This is due to our shallower encoder and decoder architectures, as well as the upsampling-CONV layers that are used instead of deconvolution layers with larger filter size.

Completing our quantitative evaluation, in Table 6, we present object affordance segmentation results of our proposed model per affordance class, based on video or static image inference. There, we can observe the superiority of the dominant affordances (i.e., those associated with most of the dataset objects), such as ''grasp'' and ''lift'', as well as the satisfactory performance of complex affordances that modify the object visual appearance, such as ''squeeze''. Note that affordance label weighting leads to slightly better F-score overall, which is expected given the very confident predictions for the dominant affordances.

5) QUALITATIVE EVALUATION RESULTS

Besides the quantitative evaluation, we use samples from the image-only UMD and IIT-AFF datasets to qualitatively evaluate our proposed model performance and judge its generalization to unseen data during its training. We depict predicted affordance segmentations in Fig. 8, color-coding the pixel-level predictions and drawing them if $\widehat{U}_{c,i,j} > 0.75$ (the images are also center-cropped to 300×300 pixels).

In the upper two rows of Fig. 8, we can observe that our model is able to confidently predict affordances on objects of the UMD dataset that are similar to those in SOR3D-AFF (e.g., ''cup'', ''hammer'', and ''knife''), although it has never seen the exact same objects during its training. Further, in the lower two rows of the figure, our model is able to infer reasonable segmentations of the dominant affordances on samples from the challenging IIT-AFF dataset, although it has never been trained on multi-object cluttered data scenes. Notably, our model has never seen object class ''cable'', but manages to provide ''grasp'' or ''rotate'' pixel-level affordance predictions for it (right-most examples).

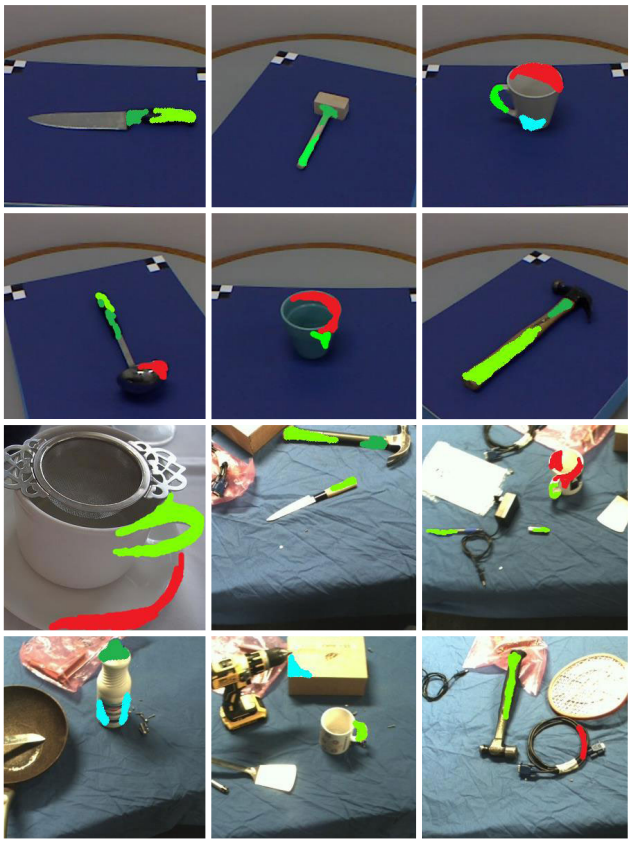

FIGURE 8. Predicted affordance segmentations by our proposed model on objects of the UMD (upper two rows) and IIT-AFF (lower two rows) datasets, unseen during training. Affordances are color-coded (''grasp'': light green, ''lift'': green, ''rotate'': red, ''push'': cyan) and are shown when $\widehat{U}_{c,i,j} > 0.75$.

6) ABLATION STUDY

Similarly to Section IV-B, we report an ablation study in Table 7, evaluating different variants of our affordance segmentation model, in order to demonstrate the contribution of each individual architecture component. In this study we focus on the segmentation-only model of our recent work [37] (also depicted in Fig. 2(a)) – denoted as ''Single'' in the table, as it employs one decoder only. This allows us to also demonstrate the improvement achieved by our proposed joint model training approach (see Fig. 2(c)) – denoted as ''Joint''.

We follow the same logic as in Table 3 for video-based inference (upper part of Table 7), considering: (i) the single appearance-only encoder; (ii) the single appearance-only encoder with the soft-attention mechanism; and (iii) the two-encoder model (both appearance and motion) with the soft-attention mechanism present, in all cases in conjunction with RGB-only or RGB-D input encoding and the single-decoder setup. Similarly to the reasoning task, we observe that adding the attention mechanism and then the motion encoder both improve results, with all three RGB-D variants outperforming their RGB-only counterparts. In addition, we can see that incorporating the reasoning decoder in the model architecture and jointly training the system improves results further, for example increasing the F-score

TABLE 6. Affordance segmentation performance of our model on the SOR3D-AFF test set, shown per class and overall, based on video or static image inference.

Inference	Metric	"cut"	"grasp"	"hammer"	"lift"	"paint"	"push"	"rotate"	"squeeze"	"type"	all classes
Video-based	IoU (↑)	0.477	0.878	0.761	0.920	0.713	0.777	0.754	0.671	0.633	0.731
	F (↑)	0.612	0.929	0.859	0.952	0.790	0.892	0.847	0.769	0.731	0.820
	F_w (↑)	0.613	0.931	0.860	0.952	0.791	0.894	0.859	0.772	0.734	0.821
Image-based	IoU (↑)	0.381	0.673	0.594	0.709	0.532	0.634	0.579	0.533	0.408	0.559
	F (↑)	0.447	0.707	0.652	0.761	0.590	0.678	0.641	0.592	0.479	0.617
	F_w (↑)	0.468	0.716	0.661	0.772	0.592	0.681	0.643	0.594	0.481	0.622

TABLE 7. Ablation study on affordance segmentation model architectures, based on video or static image inference on the SOR3D-AFF test set. Various model variants are evaluated in terms of IoU and F-score: single decoder vs. two jointly trained decoders, as well as with or without depth information, soft attention mechanism ("α"), and optical flow.

Inference	Model Variations	Decoder	IoU (↑)	F (↑)
Video-based	RGB	Single	0.619	0.733
	RGB + α		0.652	0.769
	RGB + α + 2D flow		0.663	0.784
	RGB-D		0.640	0.771
	RGB-D + α		0.703	0.797
	RGB-D + α + 3D flow		0.718	0.801
	RGB-D + α + 3D flow	Joint	**0.731**	**0.820**
Image-based	RGB + α + 2D flow	Single	0.471	0.529
	RGB-D + α + 3D flow		0.537	0.582
	RGB-D + α + 3D flow	Joint	**0.559**	**0.617**

from 0.80 to 0.82. We observe similar trends in static image inference (lower part of Table 7) with RGB-D encoding outperforming RGB-only, and the joint two-decoder system outperforming the single-decoder one, for example in the latter case improving F-score from 0.58 to 0.62.

V. CONCLUSION

In this paper, the related tasks of object affordance reasoning and segmentation are jointly investigated, following the "learning from observation" paradigm based on human-object interaction videos. In particular, an end-to-end deep-learning based dual encoder-decoder model is introduced for this purpose, without the need of strong object-related supervision. The proposed model encodes color, depth, and motion information from human-object interaction videos and predicts the desired affordance heatmaps and segmentation maps through two dedicated decoders. The model also employs a spatio-temporal soft-attention mechanism that enforces implicit localization of the interaction hotspot to improve both tasks. An extensive evaluation of the proposed approach is reported on the recently introduced SOR3D-AFF corpus, which consists of RGB-D interaction videos coupled with affordance heatmaps and pixel-wise affordance class annotations, as well as on three other state-of-the-art datasets. The experiments demonstrate that the presented model predicts more accurate affordance heatmaps compared to alternative state-of-the-art methods in the literature, while achieving better performance in segmentation prediction when operating either on videos or static images. Finally, its generalization ability is illustrated qualitatively on one video and two static image datasets.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[2] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," in *IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[3] Z. Jie, W. F. Lu, S. Sakhavi, Y. Wei, E. H. F. Tay, and S. Yan, "Object proposal generation with fully convolutional networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 62–75, Jan. 2018.

[4] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, Nov. 2019.

[5] B. Xue and N. Tong, "DIOD: Fast and efficient weakly semi-supervised deep complex ISAR object detection," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 3991–4003, Nov. 2019.

[6] B. Bamne, N. Shrivastava, L. Parashar, and U. Singh, "Transfer learning-based object detection by using convolutional neural networks," in *Proc. Int. Conf. Electron. Sustain. Commun. Syst. (ICESC)*, Jul. 2020, pp. 328–332.

[7] Q. Fan, W. Zhuo, C.-K. Tang, and Y.-W. Tai, "Few-shot object detection with attention-RPN and multi-relation detector," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4012–4021.

[8] Z. Soleimanitaleb, M. A. Keyvanrad, and A. Jafari, "Object tracking methods: A review," in *Proc. Int. Conf. Comput. Knowl. Eng. (ICCKE)*, Oct. 2019, pp. 282–288.

[9] Z. He and X. Chen, "Object tracking based on channel attention," *IEEE Access*, vol. 8, pp. 17824–17832, 2020.

[10] D. Yuan, W. Kang, and Z. He, "Robust visual tracking with correlation filters and metric learning," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105697.

[11] J. Jin, X. Li, X. Li, and S. Guan, "Online multi-object tracking with Siamese network and optical flow," in *Proc. IEEE Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2020, pp. 193–198.

[12] S. Hao, Y. Zhou, and Y. Guo, "A brief survey on semantic segmentation with deep learning," *Neurocomputing*, vol. 406, pp. 302–321, Sep. 2020.

[13] Y. Yang, X. Shu, R. Wang, C. Feng, and W. Jia, "Parallelizable and robust image segmentation model based on the shape prior information," *Appl. Math. Model.*, vol. 83, pp. 357–370, Jul. 2020.

[14] J. H. Giraldozuluaga, S. Javed, and T. Bouwmans, "Graph moving object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 12, 2020, doi: 10.1109/ICIVC50857.2020.9177480.

[15] J. Luiten, I. E. Zulfikar, and B. Leibe, "UnOVOST: Unsupervised offline video object segmentation and tracking," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1989–1998.

[16] X. Shu, Y. Yang, and B. Wu, "Adaptive segmentation model for liver CT images based on neural network and level set method," *Neurocomputing*, vol. 453, pp. 438–452, Sep. 2021.

[17] A. Andreopoulos and J. K. Tsotsos, "50 years of object recognition: Directions forward," *Comput. Vis. Image Understand.*, vol. 117, no. 8, pp. 827–891, Aug. 2013.

[18] A. Kanezaki, Y. Matsushita, and Y. Nishida, "RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5010–5019.

[19] Z. Yang and L. Wang, "Learning relationships for multi-view 3D object recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7504–7513.
[20] M. Mandal, L. K. Kumar, M. Singh Saran, and S. K. Vipparthi, "MotionRec: A unified deep framework for moving object recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2723–2732.
[21] J. J. Gibson, "The theory of affordances," in *Perceiving, Acting, and Knowing: Toward an Ecological Psychology*, R. Shaw and J. Bransford, Eds. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 1977, pp. 67–82.
[22] Y. Zhu, Y. Zhao, and S.-C. Zhu, "Understanding tools: Task-oriented object modeling, learning and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2855–2864.
[23] N. Lyubova, S. Ivaldi, and D. Filliat, "From passive to interactive object learning and recognition through self-identification on a humanoid robot," *Auto. Robots*, vol. 40, no. 1, pp. 33–57, Jan. 2016.
[24] S. Thermos, G. T. Papadopoulos, P. Daras, and G. Potamianos, "Deep affordance-grounded sensorimotor object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 49–57.
[25] M. Hassanin, S. Khan, and M. Tahtali, "Visual affordance and function understanding: A survey," 2018, *arXiv:1807.06775*. [Online]. Available: http://arxiv.org/abs/1807.06775
[26] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A deep multi-level network for saliency prediction," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Dec. 2016, pp. 3488–3493.
[27] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "DeepGaze II: Reading fixations from deep features trained on object recognition," 2016, *arXiv:1610.01563*. [Online]. Available: https://arxiv.org/abs/1610.01563
[28] J. Pan, C. Canton-Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i-Nieto, "SalGAN: Visual saliency prediction with generative adversarial networks," 2017, *arXiv:1701.01081*. [Online]. Available: https://arxiv.org/abs/1701.01081
[29] Y. Huang, M. Cai, Z. Li, and Y. Sato, "Predicting gaze in egocentric video by learning task-dependent attention transition," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 11208, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. New York, NY, USA: Springer, 2018, pp. 789–804.
[30] A. Myers, C. L. Teo, C. Fermuller, and Y. Aloimonos, "Affordance detection of tool parts from geometric features," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 1374–1381.
[31] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Detecting object affordances with convolutional neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2016, pp. 2765–2770.
[32] T.-T. Do, A. Nguyen, and I. Reid, "AffordanceNet: An end-to-end deep learning approach for object affordance detection," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 5882–5889.
[33] A. Nguyen, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, "Object-based affordances detection with convolutional neural networks and dense conditional random fields," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5908–5915.
[34] J. Sawatzky, A. Srikantha, and J. Gall, "Weakly supervised affordance detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5197–5206.
[35] K. Fang, T.-L. Wu, D. Yang, S. Savarese, and J. J. Lim, "Demo2Vec: Reasoning object affordances from online videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 2139–2147.
[36] T. Nagarajan, C. Feichtenhofer, and K. Grauman, "Grounded human-object interaction hotspots from video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8687–8696.
[37] S. Thermos, P. Daras, and G. Potamianos, "A deep learning approach to object affordance segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2358–2362.
[38] E. Rivlin, S. J. Dickinson, and A. Rosenfeld, "Recognition by functional parts," *Comput. Vis. Image Understand.*, vol. 62, no. 2, pp. 164–176, Sep. 1995.
[39] M. Sutton, L. Stark, and K. Bowyer, "Function from visual analysis and physical interaction: A methodology for recognition of generic classes of objects," *Image Vis. Comput.*, vol. 16, no. 11, pp. 745–763, Aug. 1998.
[40] V. Hogman, M. Bjorkman, A. Maki, and D. Kragic, "A sensorimotor learning framework for object categorization," *IEEE Trans. Cognit. Develop. Syst.*, vol. 8, no. 1, pp. 15–25, Mar. 2016.
[41] D. Jayaraman and K. Grauman, "End-to-end policy learning for active visual categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1601–1614, Jul. 2019.
[42] S. Thermos, G. T. Papadopoulos, P. Daras, and G. Potamianos, "Deep sensorimotor learning for RGB-D object recognition," *Comput. Vis. Image Understand.*, vol. 190, Jan. 2020, Art. no. 102844.
[43] V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. A. Efros, "Scene semantics from long-term observation of people," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 7577, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. New York, NY, USA: Springer, 2012, pp. 284–298.
[44] X. Wang, R. Girdhar, and A. Gupta, "Binge watching: Scaling affordance learning from sitcoms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3366–3375.
[45] H. S. Koppula and A. Saxena, "Physically grounded spatio-temporal object affordances," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 8691, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. New York, NY, USA: Springer, 2014, pp. 831–847.
[46] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 14–29, Jan. 2016.
[47] N. Rhinehart and K. M. Kitani, "Learning action maps of large environments via first-person vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 580–588.
[48] Y. Zhou, B. Ni, R. Hong, X. Yang, and Q. Tian, "Cascaded interactional targeting network for egocentric video analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1904–1913.
[49] T. Nagarajan, Y. Li, C. Feichtenhofer, and K. Grauman, "Ego-topo: Environment affordances from egocentric video," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 160–169.
[50] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez, "GanHand: Predicting human grasp affordances in multi-object scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5030–5040.
[51] P. Ardon, E. Pairet, Y. Petillot, R. P. A. Petrick, S. Ramamoorthy, and K. S. Lohan, "Self-assessment of grasp affordance transfer," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 9385–9392.
[52] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, Jul. 2013.
[53] T. Xiao, Q. Fan, D. Gutfreund, M. Monfort, A. Oliva, and B. Zhou, "Reasoning about human-object interactions through dual attention networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3918–3927.
[54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[55] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004.
[56] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. New York, NY, USA: Springer, 2015, pp. 234–241.
[57] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 802–810.
[58] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.
[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[60] G. T. Papadopoulos and P. Daras, "Human action recognition using 3D reconstruction data," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1807–1823, Aug. 2018.
[61] M. Jaimez, M. Souiai, J. Gonzalez-Jimenez, and D. Cremers, "A primal-dual framework for real-time dense RGB-D scene flow," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 98–104.
[62] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 9908, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. New York, NY, USA: Springer, 2016, pp. 630–645.
[64] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 677–691, Apr. 2017.
[65] D. S. Alexiadis and P. Daras, "Quaternionic signal processing techniques for automatic evaluation of dance performances from MoCap data," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1391–1406, Aug. 2014.
[66] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
[67] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[68] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 20–27.
[69] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, vol. 9, 2010, pp. 249–256.
[70] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.
[71] Y. Wu and K. He, "Group normalization," in *Computer Vision—ECCV* (Lecture Notes in Computer Science), vol. 11217, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. New York, NY, USA: Springer, 2018, pp. 3–19.
[72] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst*, vol. 32. Red Hook, NY, USA: Curran Associates, 2019, pp. 8024–8035.
[73] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst., Annu. Conf. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 568–576.
[74] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
[75] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1072–1080.
[76] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Mar. 2019.
[77] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.
[78] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1153–1160.
[79] Z. Wojna, V. Ferrari, S. Guadarrama, N. Silberman, L.-C. Chen, A. Fathi, and J. Uijlings, "The devil is in the decoder: Classification, regression and GANs," *Int. J. Comput. Vis.*, vol. 127, nos. 11–12, pp. 1694–1706, Dec. 2019.
[80] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in Statistics: Methodology and Distribution*, vol. 2, S. Kotz and N. L. Johnson, Eds. New York, NY, USA: Springer, 1992, pp. 196–202.
[81] P. Jaccard, "The distribution of the flora in the alpine zone," *New Phytologist*, vol. 11, no. 2, pp. 37–50, Feb. 1912.
[82] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.

SPYRIDON THERMOS was born in Thessaloniki, Greece, in 1989. He received the Diploma, M.Sc., and Ph.D. degrees in electrical and computer engineering from the University of Thessaly, Volos, Greece, in 2013, 2015, and 2020, respectively. Since 2019, he has been a Postdoctoral Researcher with The University of Edinburgh, U.K. From 2015 to 2019, he was a Research Assistant with the Visual Computing Lab, Centre for Research and Technology Hellas. He has coauthored 15 peer-reviewed conference papers and journal articles and one book chapter. His research interests include object affordance understanding, human-object interaction modeling, more recently disentangled representation learning, and controllable image synthesis.

GERASIMOS POTAMIANOS (Member, IEEE) received the Diploma degree from the National Technical University of Athens, Greece, in 1988, and the M.S.E. and Ph.D. degrees in electrical and computer engineering from Johns Hopkins University, Baltimore, MD, USA, in 1990 and 1994, respectively. He is currently an Associate Professor with the Department of Electrical and Computer Engineering, University of Thessaly, Volos, Greece. Previously, he has worked with the Center of Speech and Language Processing, Johns Hopkins at AT&T Labs-Research, Murray Hill and Florham Park, NJ, at IBM Research, Yorktown Heights, NY, and at the FORTH and NCSR Demokritos Research Centers, Greece. He has published over 150 articles in these areas that have received around 6k citations and he holds seven U.S. patents. His research interests include the fields of audio-visual speech processing, automatic speech recognition, sign language recognition, multimedia signal processing, and fusion, as well as multimodal scene analysis. In addition to IEEE, he is a member of EURASIP, ISCA, and the National Technical Chamber of Greece.

PETROS DARAS (Senior Member, IEEE) received the Diploma, M.Sc., and Ph.D. degrees in electrical and computer engineering from the Aristotle University of Thessaloniki, Greece, in 1999, 2002, and 2005, respectively. He is currently the Research Director and the Chair of the Visual Computing Lab, Centre for Research and Technology Hellas. He has been involved in over 50 European and Greek-funded projects and has published over 300 articles in refereed journals and international conferences. His main research interests include visual content processing, multimedia indexing, search engines, recommendation algorithms, and relevance feedback.