Generating Videos of Zero-Shot Compositions of Actions and Objects

Megha Nawhal[1,2], Mengyao Zhai[2], Andreas Lehrmann[1], Leonid Sigal[1,3,4,5], and Greg Mori[1,2]

[1] Borealis AI, Vancouver, Canada
[2] Simon Fraser University, Burnaby, Canada
[3] University of British Columbia, Vancouver, Canada
[4] Vector Institute for AI [5] CIFAR AI Chair

Abstract. Human activity videos involve rich, varied interactions between people and objects. In this paper we develop methods for generating such videos – making progress toward addressing the important, open problem of video generation in complex scenes. In particular, we introduce the task of generating human-object interaction videos in a zero-shot compositional setting, *i.e.*, generating videos for action-object compositions that are unseen during training, having seen the target action and target object separately. This setting is particularly important for generalization in human activity video generation, obviating the need to observe every possible action-object combination in training and thus avoiding the combinatorial explosion involved in modeling complex scenes. To generate human-object interaction videos, we propose a novel adversarial framework HOI-GAN which includes multiple discriminators focusing on different aspects of a video. To demonstrate the effectiveness of our proposed framework, we perform extensive quantitative and qualitative evaluation on two challenging datasets: EPIC-Kitchens and 20BN-Something-Something v2.

Keywords: Video Generation; Compositionality in Videos

1 Introduction

Visual imagination and prediction are fundamental components of human intelligence. Arguably, the ability to create realistic renderings from symbolic representations are considered prerequisite for broad visual understanding. Computer vision has seen rapid advances in the field of image generation over the past few years. Existing models are capable of generating impressive results in this static scenario, ranging from hand-written digits [3, 11, 19] to realistic scenes [5, 29, 34, 53, 78]. Progress on *video generation* [4, 25, 57, 64, 66, 69, 70], on the other hand, has been relatively moderate and remains an open and challenging problem. While most approaches focus on the expressivity and controllability of the underlying generative models, their ability to generalize to unseen scene compositions has not received as much attention. However, such generalizability

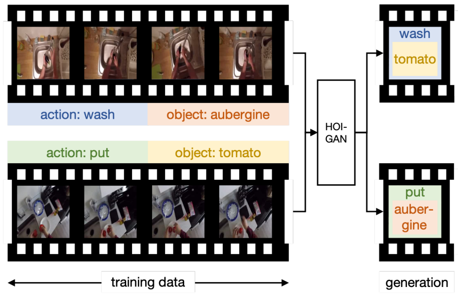


Fig. 1. Generation of Zero-Shot Human-Object Interactions. Given training examples *"wash aubergine"* and *"put tomato"*, an intelligent agent should be able to imagine action sequences for unseen action-object compositions, *i.e.*, *"wash tomato"* and *"put aubergine"*.

is an important cornerstone of robust visual imagination as it demonstrates the capacity to reason over elements of a scene.

We posit that the domain of human activities constitutes a rich realistic testbed for video generation models. Human activities involve people interacting with objects in complex ways, presenting numerous challenges for generation – the need to (1) render a variety of objects; (2) model the temporal evolution of the effect of actions on objects; (3) understand spatial relations and interactions; and (4) overcome the paucity of data for the complete set of action-object pairings. The last, in particular, is a critical challenge that also serves as an opportunity for designing and evaluating generative models that can generalize to myriad, possibly unseen, action-object compositions. For example, consider Figure 1. The activity sequences for *"wash aubergine"* (action a_1: "wash"; object o_1: "aubergine") and *"put tomato"*(action a_2: "put"; object o_2: "tomato") are observed in the training data. A robust visual imagination would then allow an agent to imagine videos for *"wash tomato"* (a_1, o_2) and *"put aubergine"* (a_2, o_1).

We propose a novel framework for generating human-object interaction (HOI) videos for unseen action-object compositions. We refer to this task as *zero-shot HOI video generation*. To the best of our knowledge, our work is the first to propose and address this problem. In doing so, we push the envelope on conditional (or controllable) video generation and focus squarely on the model's ability to generalize to unseen action-object compositions. This zero-shot compositional setting verifies that the model is capable of semantic disentanglement of the action and objects in a given context and recreating them separately in other contexts.

The desiderata for performing zero-shot HOI video generation include: (1) mapping the content in the video to the right semantic category, (2) ensuring spatial and temporal consistency across the frames of a video, and (3) producing interactions with the right object in the presence of multiple objects. Based on these observations, we introduce a novel multi-adversarial learning scheme involving multiple discriminators, each focusing on different aspects of an HOI video. Our framework *HOI-GAN* generates a fixed length video clip given an action, an object, and a target scene serving as the context.

Concretely, the conditional inputs to our framework are semantic labels of action and object, and a single start frame with a mask providing the background and location for the object. Then, the model has to create the object, reason over the action, and enact the action on the object (leading to object translation and/or transformation) over the background, thus generating the whole interaction video. During training of the generator, our framework utilizes four discriminators – three pixel-centric discriminators, namely, *frame* discriminator, *gradient* discriminator, *video* discriminator; and one object-centric *relational* discriminator. The three pixel-centric discriminators ensure spatial and temporal consistency across the frames. The novel relational discriminator leverages spatio-temporal scene graphs to reason over the object layouts in videos ensuring the right interactions among objects. Through experiments, we show that our HOI-GAN framework is able to disentangle objects and actions and learns to generate videos with unseen compositions.

In summary, our contributions are as follows:

- We introduce the task of zero-shot HOI video generation. Specifically, given a training set of videos depicting certain action-object compositions, we propose to generate unseen compositions having seen the target action and target object individually, *i.e.*, the target action was paired with a different object and the target object was involved in a different action.
- We propose a novel adversarial learning scheme and introduce our HOI-GAN framework to generate HOI videos in a zero-shot compositional setting.
- We demonstrate the effectiveness of HOI-GAN through empirical evaluation on two challenging HOI video datasets: *20BN-something-something v2* [20] and *EPIC-Kitchens* [9]. We perform both quantitative and qualitative evaluation of the proposed approach and compare with state-of-the-art approaches.

Overall, our work facilitates research in the direction of enhancing generalizability of generative models for complex videos.

2 Related Work

Our paper builds on prior work in: (1) modeling of human-object interactions and (2) GAN-based video generation. In addition, we also discuss literature relevant to HOI video generation in a zero-shot compositional setting.

Modeling Human-Object Interactions. Earlier research attempts to study human-object interactions (HOIs) aimed at studying object affordances [21, 38] and semantic-driven understanding of object functionalities [24, 62]. Recent work on modeling HOIs in images range from studying semantics and spatial features of interactions between humans and objects [10, 18, 77] to action information [13, 17, 76]. Furthermore, there have been attempts to create large scale image and video datasets to study HOI [7, 8, 20, 39]. To model dynamics in HOIs, recent works have proposed methods that jointly model actions and objects in videos [33, 35, 60]. Inspired by these approaches, we model HOI videos as compositions of actions and objects.

GAN-based Image & Video Generation. Generative Adversarial Network (GAN) [19] and its variants [3, 11, 79] have shown tremendous progress in high quality image generation. Built over these techniques, conditional image generation using various forms of inputs to the generator such as textual information [55,75,78], category labels [48,52], and images [29,36,43,80] have been widely studied. This class of GANs allows the generator network to learn a mapping between conditioning variables and the real data distribution, thereby allowing control over the generation process. Extending these efforts to conditional video generation is not straightforward as generating a video involves modeling of both spatial and temporal variations. Vondrick et al. [66] proposed the Video GAN (VGAN) framework to generate videos using a two-stream generator network that decouples foreground and background of a scene. Temporal GAN (TGAN) [57] employs a separate generator for each frame in a video and an additional generator to model temporal variations across these frames. MoCoGAN [64] disentangles the latent space representations of motion and content in a video to perform controllable video generation using seen compositions of motion and content as conditional inputs. In our paper, we evaluate the extent to which these video generation methods generalize when provided with unseen scene compositions as conditioning variables. Furthermore, promising success has been achieved by recent video-to-video translation methods [4, 69, 70] wherein video generation is conditioned on a corresponding semantic video. In contrast, our task does not require semantic videos as conditional input.

Video Prediction. Video prediction approaches predict future frames of a video given one or a few observed frames using RNNs [61], variational auto-encoders [67,68], adversarial training [42,46], or auto-regressive methods [32]. While video prediction is typically posed as an image-conditioned (past frame) image generation (future frame) problem, it is substantially different from video generation where the goal is to generate a video clip given a stochastic latent space.

Video Inpainting. Video inpainting/completion refers to the problem of correctly filling up the missing pixels given a video with arbitrary spatio-temporal pixels missing [14,22,50,51,59]. In our setting, however, the model only receives a single static image as input and not a video. Our model is required to go beyond merely filling in pixel values and has to produce an output video with the right visual content depicting the prescribed action upon a synthesized object. In doing so, the background may, and in certain cases should, evolve as well.

Zero-Shot Learning. Zero-shot learning (ZSL) aims to solve the problem of recognizing classes whose instances are not seen during training. In ZSL, external information of a certain form is required to share information between classes to transfer knowledge from seen to unseen classes. A variety of techniques have been used for ZSL ranging from usage of attribute-based information [16, 40], word embeddings [71] to WordNet hierarchy [1] and text-based descriptions [15, 23, 41, 81]. [72] provides a thorough overview of zero-shot learning techniques.

Similar to these works, we leverage word embeddings to reason over the unseen compositions of actions and objects in the context of video generation.

Learning Visual Relationships. Visual relationships in the form of scene graphs, *i.e.*, directed graphs representing relationships (edges) between the objects (nodes) have been used for image caption evaluation [2], image retrieval [31] and predicting scene compositions for images [44, 49, 74]. Furthermore, in a generative setting, [30] aims to synthesize an image from a given scene graph and evaluate the generalizability of an adversarial network to create images with unseen relationships between objects. Similarly, we leverage spatio-temporal scene graphs to learn relevant relations among the objects and focus on the generalizability of video generation models to unseen compositions of actions and objects. However, our task of zero-shot HOI video generation is more difficult as it requires learning to map the inputs to spatio-temporal variations in a video.

Learning Disentangled Representations for Videos. Various methods have been proposed to learn disentangled representations in videos [12,27,64], such as, learning representations by decoupling the content and pose [12], or separating motion from content using image differences [65]. Similarly, our model implicitly learns to disentangle the action and object information of an HOI video.

3 HOI-GAN

Intuitively, for a generated human-object interaction (HOI) video to be realistic, it must: (1) contain the object designated by a semantic label; (2) exhibit the prescribed interaction with that object; (3) be temporally consistent; and (4 – optional) occur in a specified scene. Based on this intuition, we propose an adversarial learning scheme in which we train a generator network $\mathbf{G}$ with a set of 4 discriminators: (1) a frame discriminator $\mathbf{D}_f$, which encourages the generator to learn spatially coherent visual content; (2) a gradient discriminator $\mathbf{D}_g$, which incentivizes $\mathbf{G}$ to produce temporally consistent frames; (3) a video discriminator $\mathbf{D}_v$, which provides the generator with global spatio-temporal context; and (4) a relational discriminator $\mathbf{D}_r$, which assists the generator in producing correct object layouts in a video. We use pretrained word embeddings [54] for semantic representations of actions and objects. All discriminators are conditioned on word embeddings of the action ($\mathbf{s}_a$) and object ($\mathbf{s}_o$) and trained simultaneously in an end-to-end manner. Figure 2 shows an overview of our proposed framework *HOI-GAN*. We now formalize our task and describe each module in detail.

3.1 Task Formulation

Let $\mathbf{s}_a$ and $\mathbf{s}_o$ be word embeddings of an action a and an object o, respectively. Furthermore, let I be an image provided as context to the generator. We encode I using an encoder $\mathbf{E}_v$ to obtain a visual embedding $\mathbf{s}_I$, which we refer to as a context vector. Our goal is to generate a video $V = (V^{(i)})_{i=1}^{T}$ of length T depicting the action a performed on the object o with context image I as the background of V. To this end, we learn a function $\mathbf{G} : (\mathbf{z}, \mathbf{s}_a, \mathbf{s}_o, \mathbf{s}_I) \mapsto V$, where $\mathbf{z}$ is a noise vector sampled from a distribution $p_{\mathbf{z}}$, such as a Gaussian distribution.

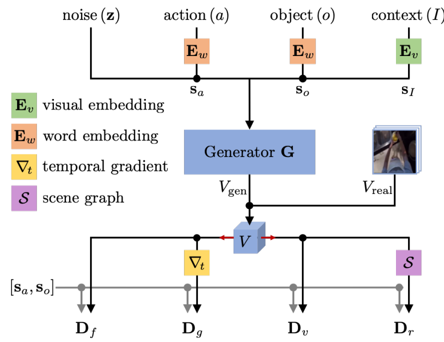


Fig. 2. Architecture Overview. The generator network $\mathbf{G}$ is trained using 4 discriminators simultaneously: a frame discriminator $\mathbf{D}_f$, a gradient discriminator $\mathbf{D}_g$, a video discriminator $\mathbf{D}_v$, and a relational discriminator $\mathbf{D}_r$. Given the word embeddings of an action $\mathbf{s}_a$, an object $\mathbf{s}_o$, and a context image $\mathbf{s}_I$, the generator learns to synthesize a video with background I in which the action a is performed on the object o.

3.2 Model Description

We describe the elements of our framework below. Overall, the four discriminator networks, *i.e.*, frame discriminator $\mathbf{D}_f$, gradient discriminator $\mathbf{D}_g$, video discriminator $\mathbf{D}_v$, and relational discriminator $\mathbf{D}_r$ are all involved in a zero-sum game with the generator network $\mathbf{G}$. Refer to the supplementary for implementation details.

Frame Discriminator. The frame discriminator network $\mathbf{D}_f$ learns to distinguish between real and generated frames corresponding to the real video V_{real} and generated video $V_{\text{gen}} = \mathbf{G}(\mathbf{z}, \mathbf{s}_a, \mathbf{s}_o, \mathbf{s}_I)$ respectively. Each frame in V_{gen} and V_{real} is processed independently using a network consisting of stacked `conv2d` layers, *i.e.*, 2D convolutional layers followed by spectral normalization [47] and leaky ReLU layers [45] with $a = 0.2$. We obtain a tensor of size $N^{(t)} \times w_0^{(t)} \times h_0^{(t)}$ ($t = 1, 2, \ldots, T$), where $N^{(t)}$, $w_0^{(t)}$, and $h_0^{(t)}$ are the channel length, width and height of the activation of the last `conv2d` layer respectively. We concatenate this tensor with spatially replicated copies of $\mathbf{s}_a$ and $\mathbf{s}_o$, which results in a tensor of size $(\dim(\mathbf{s}_a) + \dim(\mathbf{s}_o) + N^{(t)}) \times w_0^{(t)} \times h_0^{(t)}$. We then apply another `conv2d` layer to obtain a $N \times w_0^{(t)} \times h_0^{(t)}$ tensor. We now perform 1×1 convolutions followed by $w_0^{(t)} \times h_0^{(t)}$ convolutions and a sigmoid to obtain a T-dimensional vector corresponding to the T frames of the video V. The i-th element of the output denotes the probability that the frame $V^{(i)}$ is real. The objective function of the network $\mathbf{D}_f$ is the loss function:

$$L_f = \frac{1}{2T} \sum_{i=1}^{T} [\log(\mathbf{D}_f^{(i)}(V_{\text{real}}; \mathbf{s}_a, \mathbf{s}_o)) + \log(1 - \mathbf{D}_f^{(i)}(V_{\text{gen}}; \mathbf{s}_a, \mathbf{s}_o))], \quad (1)$$

where $\mathbf{D}_f^{(i)}$ is the i-th element of the output of $\mathbf{D}_f$.

Gradient Discriminator. The gradient discriminator network $\mathbf{D}_g$ enforces temporal smoothness by learning to differentiate between the temporal gradient of a real video V_{real} and a generated video V_{gen}. We define the temporal gradient $\nabla_t V$ of a video V with T frames $V^{(1)}, \dots, V^{(T)}$ as pixel-wise differences between two consecutive frames of the video. The i-th element of $\nabla_t V$ is defined as:

$$[\nabla_t V]_i = V^{(i+1)} - V^{(i)}, \quad i = 1, 2, \dots, (T-1). \tag{2}$$

The architecture of the gradient discriminator $\mathbf{D}_g$ is similar to that of the frame discriminator $\mathbf{D}_f$. The output of $\mathbf{D}_g$ is a $(T-1)$-dimensional vector corresponding to the $(T-1)$ values in gradient $\nabla_t V$. The objective function of $\mathbf{D}_g$ is

$$L_g = \frac{1}{2(T-1)} \sum_{i=1}^{T-1} [\log(\mathbf{D}_g^{(i)}(\nabla_t V_{\text{real}}; \mathbf{s}_a, \mathbf{s}_o)) + \log(1 - \mathbf{D}_g^{(i)}(\nabla_t V_{\text{gen}}; \mathbf{s}_a, \mathbf{s}_o))], \tag{3}$$

where $\mathbf{D}_g^{(i)}$ is the i-th element of the output of $\mathbf{D}_g$.

Video Discriminator. The video discriminator network $\mathbf{D}_v$ learns to distinguish between real videos V_{real} and generated videos V_{gen} by comparing their global spatio-temporal contexts. The architecture consists of stacked `conv3d` layers, *i.e.*, 3D convolutional layers followed by spectral normalization [47] and leaky ReLU layers [45] with $a = 0.2$. We obtain a $N \times d_0 \times w_0 \times h_0$ tensor, where N, d_0, w_0, and h_0 are the channel length, depth, width, and height of the activation of the last `conv3d` layer respectively. We concatenate this tensor with spatially replicated copies of $\mathbf{s}_a$ and $\mathbf{s}_o$, which results in a tensor of size $(\dim(\mathbf{s}_a) + \dim(\mathbf{s}_o) + N) \times d_0 \times w_0 \times h_0$, where $\dim(\cdot)$ returns the dimensionality of a vector. We then apply another `conv3d` layer to obtain a $N \times d_0 \times w_0 \times h_0$ tensor. Finally, we apply a $1 \times 1 \times 1$ convolution followed by a $d_0 \times w_0 \times h_0$ convolution and a sigmoid to obtain the output, which represents the probability that the video V is real. The objective function of the network $\mathbf{D}_v$ is the following loss function:

$$L_v = \frac{1}{2} [\log(\mathbf{D}_v(V_{\text{real}}; \mathbf{s}_a, \mathbf{s}_o)) + \log(1 - \mathbf{D}_v(V_{\text{gen}}; \mathbf{s}_a, \mathbf{s}_o))]. \tag{4}$$

Relational Discriminator. In addition to the three pixel-centric discriminators above, we also propose a novel object-centric discriminator $\mathbf{D}_r$. Driven by a spatio-temporal scene graph, this relational discriminator learns to distinguish between scene layouts of real videos V_{real} and generated videos V_{gen} (Figure 3).

Specifically, we build a spatio-temporal scene graph $\mathcal{S} = (\mathcal{N}, \mathcal{E})$ from V, where the nodes and edges are represented by $\mathcal{N}$ and $\mathcal{E}$ respectively. We assume one node per object per frame. Each node is connected to all other nodes in the same frame, referred to as spatial edges. In addition, to represent temporal evolution of objects, each node is connected to the corresponding nodes in the adjacent frames that also depict the same object, referred to as temporal edges. To obtain the node representations, we crop the objects in V using Mask-RCNN [26], compute a convolutional embedding for them, and augment the resulting vectors with the aspect ratio (AR) and position of the corresponding

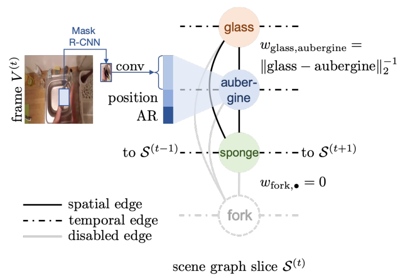


Fig. 3. Relational Discriminator. The relational discriminator $\mathbf{D}_r$ leverages a spatio-temporal scene graph to distinguish between object layouts in videos. Each node contains convolutional embedding, position and aspect ratio (AR) of the object crop obtained from MaskRCNN. The nodes are connected in space and time and edges are weighted based on their inverse distance. Edge weights of (dis)appearing objects are 0.

bounding boxes. The weights of spatial edges in $\mathcal{E}$ are given by inverse Euclidean distances between the centers of these bounding boxes corresponding to the object appearing in the frame. The weights of the temporal edges is set to 1 by default. When an object is not present in a frame (but appears in the overall video), spatial edges connecting to the object will be absent by design. This is implemented by setting the weights to 0 depicting distance between the objects as ∞. Similarly, if an object does not appear in the adjacent frame, the temporal edge is set to 0. In case of multiple objects of the same category, the correspondence is established based on the location in the adjacent frames using nearest neighbour data association.

The relational discriminator $\mathbf{D}_r$ operates on this scene graph $\mathcal{S}$ by virtue of a graph convolutional network (GCN) [37] followed by stacking and average-pooling of the resulting node representations along the time axis. We then concatenate this tensor with spatially replicated copies of $\mathbf{s}_a$ and $\mathbf{s}_o$ to result in a tensor of size $(\dim(\mathbf{s}_a) + \dim(\mathbf{s}_o) + N^{(t)}) \times w_0^{(t)} \times h_0^{(t)}$. As before, we then apply convolutions and sigmoid to obtain the final output which denotes the probability of the scene graph belonging to a real video. The objective function of the network $\mathbf{D}_r$ is given by

$$L_r = \frac{1}{2}\left[\log(\mathbf{D}_r(\mathcal{S}_{\text{real}}; \mathbf{s}_a, \mathbf{s}_o)) + \log(1 - \mathbf{D}_r(\mathcal{S}_{\text{gen}}; \mathbf{s}_a, \mathbf{s}_o))\right]. \tag{5}$$

Generator. Given the semantic embeddings $\mathbf{s}_a$, $\mathbf{s}_o$ of action and object labels respectively, and context vector $\mathbf{s}_I$, the generator network $\mathbf{G}$ learns to generate video V_{gen} consisting of T frames (RGB) of height H and width W. We concatenate noise $\mathbf{z}$ with the conditions, namely, $\mathbf{s}_a$, $\mathbf{s}_o$, and $\mathbf{s}_I$. We provide this concatenated vector as the input to the network $\mathbf{G}$. The network comprises stacked `deconv3d` layers, *i.e.*, 3D transposed convolution layers each followed by Batch Normalization [28] and leaky ReLU layers [45] with $a = 0.2$ except the last convolutional layer which is followed by a Batch Normalization layer [28] and a `tanh` activation layer. The network is optimized according to the following objective function:

$$
\begin{aligned}
L_{gan} = \frac{1}{T}\sum_{i=1}^{T}&[\log(1-\mathbf{D}_f^{(i)}(V_{\text{gen}};\mathbf{s}_a,\mathbf{s}_o))]+\\
&\frac{1}{(T-1)}\sum_{i=1}^{T-1}[\log(1-\mathbf{D}_g^{(i)}(\nabla_t V_{\text{gen}};\mathbf{s}_a,\mathbf{s}_o))]+\\
&\log(1-\mathbf{D}_v(V_{\text{gen}};\mathbf{s}_a,\mathbf{s}_o))+\log(1-\mathbf{D}_r(\mathcal{S}_{\text{gen}};\mathbf{s}_a,\mathbf{s}_o)).
\end{aligned}
\tag{6}
$$

4 Experiments

We conduct quantitative and qualitative analysis to demonstrate the effectiveness of the proposed framework HOI-GAN for the task of zero-shot generation of human-object interaction (HOI) videos.

4.1 Datasets and Data Splits

We use two datasets for our experiments: EPIC-Kitchens [9] and 20BN-Something-Something V2 [20]. Both of these datasets comprise a diverse set of HOI videos ranging from simple translational motion of objects (*e.g.* push, move) and rotation (*e.g.* open) to transformations in state of objects (*e.g.* cut, fold). Therefore, these datasets, with their wide ranging variety and complexity, provide a challenging setup for evaluating HOI video generation models.

EPIC-Kitchens [9] contains egocentric videos of activities in several kitchens. A video clip V is annotated with action label a and object label o (*e.g.* open microwave, cut apple, move pan) along with a set of bounding boxes $\mathcal{B}$ (one per frame) for objects that the human interacts with while performing the action. There are around 40k instances in the form of $(V, a, o, \mathcal{B})$ across 352 objects and 125 actions. We refer to this dataset as EPIC hereafter.

20BN-Something-Something V2 [20] contains videos of daily activities performed by humans. A video clip V is annotated with a label l, an action template and object(s) on which the action is applied (*e.g.* 'hitting ball with racket' has action template 'hitting something with something'). There are 220,847 training instances of the form (V, l) spanning 30,408 objects and 174 action templates. To transform l to action-object label pair (a, o), we use NLTK POS-tagger. We consider the verb tag (after stemming) in l as action label a. We observe that all instances of l begin with the present continuous form of a which is acting upon the subsequent noun. Therefore, we use the noun that appears immediately after the verb as object o. Hereafter, we refer to the transformed dataset in the form of (V, a, o) as SS.

Splitting by Compositions. We believe it is reasonable to only generate combinations that are semantically feasible, and do so by only using action-object pairs seen in the original datasets. We use a subset of action-object pairs as testing pairs – these pairs are not seen during training but are present in the original dataset, hence are semantically feasible. To make the dataset training /

Table 1. Generation Scenarios. Description of the conditional inputs for the two generation scenarios GS1 & GS2 used for evaluation. ✓ denotes 'Yes', ✗ denotes 'No'.

Target Conditions	GS1	GS2
Target action a seen during training	✓	✓
Target object o seen during training	✓	✓
Background of target context I seen during training	✗	✓
Object mask in target context I corresponds to target object o	✓	✗
Target action a seen with target context I during training	✗	✓/ ✗
Target object o seen with target context I during training	✗	✗
Target action-object composition (a-o) seen during training	✗	✗

testing splits suitable for our zero-shot compositional setting, we first merge the data samples present in the default train and validation sets of the dataset. We then split the combined dataset into training set and test set based on the condition that all the unique object and action labels in appear in the training set, however, any composition of action and object present in the test set is absent in training set and vice versa. We provide further details of the splits for both datasets EPIC and SS in the supplementary.

Generation Scenarios. Recall that the generator network in the HOI-GAN framework (Fig. 2) has 3 conditional inputs, namely, action embedding, object embedding, and context frame I. The context frame serves as the background in the scene. Thus, to provide this context frame during training, we apply a binary mask $M^{(1)}$ corresponding to the first frame $V^{(1)}$ of a real video as $I = (\mathbb{1} - M^{(1)}) \odot V^{(1)}$, where $\mathbb{1}$ represents a matrix of size $M^{(1)}$ containing all ones and $\odot$ denotes elementwise multiplication. This mask $M^{(1)}$ contains ones in regions (either rectangular bounding boxes or segmentation masks) corresponding to the objects (non-*person* classes) detected using MaskRCNN [26] and zeros for other regions. Intuitively, this helps ensure the generator learns to map the action and object embeddings to relevant visual content in the HOI video.

During testing, to evaluate the generator's capability to synthesize the right human-object interactions, we provide a background frame as described above. This background frame can be selected from either the test set or training set, and can be suitable or unsuitable for the target action-object composition. To capture these possibilities, we design two different generation scenarios. Specifically, in *Generation Scenario 1 (GS1)*, the input context frame I is the masked first frame of a video from the test set corresponding to the target action-object composition (unseen during training). In *Generation Scenario 2 (GS2)*, I is the masked first frame of a video from the training set which depicts an object other than the target object. The original action in this video could be same or different than the target action. See Table 1 for the contrast between the scenarios.

As such, in GS1, the generator receives a context that it has not seen during training but the context (including object mask) is consistent with the target action-object composition it is being asked to generate. In contrast, in GS2, the generator receives a context frame that it has seen during training but is

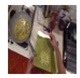
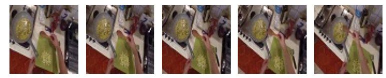
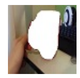
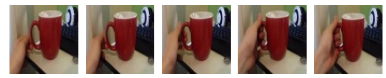
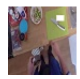
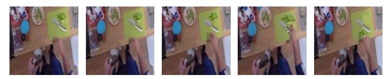
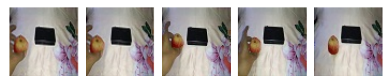
Scenario	(a, o)	Context	Generated Output
GS1	take spoon (EPIC)		
GS1	hold cup (SS)		
GS2	move broccoli (EPIC)		
GS2	put apple (SS)		

Fig. 4. Qualitative Results: Videos generated using our best version of HOI-GAN using embeddings for action (a)-object (o) composition and the context frame. We show 5 frames of the video clip generated for both generation scenarios GS1 and GS2. The context frame in GS1 is obtained from a video in the test set depicting an action-object composition same as the target one. The context frame for GS2 scenarios shown here are from videos depicting *"take carrot"* (for row 3) and *"put bowl"* (for row 4). Refer to supplementary section for additional videos generated using HOI-GAN.

not consistent with the action-object composition it is being asked to generate. Particularly, the object mask in the context does not correspond to the target object. Although the background is seen, the model has to evolve the background in ways different from training samples to make it suitable for the target composition. Thus, these generation scenarios help illustrate that the generator indeed generalizes over compositions.

4.2 Evaluation Setup

Evaluation of image/video quality is inherently challenging, thus, we use both quantitative and qualitative metrics.

Quantitative Metrics. Inception Score **(I-score)** [58] is a widely used metric for evaluating image generation models. For images x with labels y, I-score is defined as $\exp(\mathbf{KL}(\rho(y|x)||\rho(y)))$ where $\rho(y|x)$ is the conditional label distribution of an ImageNet [56] -pretrained Inception model [63]. We adopted this metric for video quality evaluation. We fine-tune a Kinetics [6]-pretrained video classifier ResNeXt-101 [73] for each of our source datasets and use it for calculating I-score (higher is better). It is based on one of the state-of-the-art video classification architectures. We used the same evaluation setup for the baselines and our model to ensure a fair comparison.

In addition, we believe that measuring realism explicitly is more relevant for our task as the generation process can be conditioned on any context frame arbitrarily to obtain diverse samples. Therefore, in addition to *I-score*, we also analyze the first and second terms of the KL divergence separately. We refer to these terms as: (1) Saliency score or **S-score** (lower is better) to specifically measure realism, and (2) Diversity score or **D-score** (higher is better) to indicate the diversity in generated samples. A smaller value of S-score implies that the generated videos are more realistic as the classifier is very confident in classifying the generated videos. Specifically, the saliency score will have a low value (low is good) only when the classifier is confidently able to classify the generated videos into action-object categories matching the conditional input composition (action-object), thus indicating realistic instances of the required target interaction. In fact, even if a model generates realistic-looking videos but depicts an action-object composition not corresponding to the conditional action-object input, the saliency score will have high values. Finally, a larger value of D-score implies the model generates diverse samples.

Human Preference Score. We conduct a user study for evaluating the quality of generated videos. In each test, we present the participants with two videos generated by two different algorithms and ask which among the two better depicts the given activity, *i.e.*, action-object composition (*e.g.* lift fork). We evaluate the performance of an algorithm as the overall percentage of tests in which that algorithm's outputs are preferred. This is an aggregate measure over all the test instances across all participants.

Baselines. We compare HOI-GAN with three state-of-the-art video generation approaches: (1) VGAN [66], (2) TGAN, [57] and (3) MoCoGAN [64]. We develop the conditional variants of VGAN and TGAN from the descriptions provided in their papers. We refer to the conditional variants as C-VGAN and C-TGAN respectively. We observed that these two models saturated easily in the initial iterations, thus, we added dropout in the last layer of the discriminator network in both models. MoCoGAN focuses on disentangling motion and content in the latent space and is the closest baseline. We use the code provided by the authors.

4.3 Results

Next, we discuss the results of our qualitative and quantitative evaluation.

Comparison with Baselines. As shown in Table 2, HOI-GAN with different conditional inputs outperforms C-VGAN and C-TGAN by a wide margin in both generation scenarios. In addition, our overall model shows considerable improvement over MoCoGAN, while MoCoGAN has comparable scores to some ablated versions of our models (where gradient discriminator and/or relational discriminator is missing). Furthermore, we varied the richness of the masks in the conditional input context frame ranging from bounding boxes to segmentation masks obtained corresponding to non-*person* classes using MaskRCNN framework [26]. We observe that providing masks during training leads to slight improvements in both scenarios as compared to using bounding boxes (refer to Table 2). We also show the samples generated using the best version of HOI-GAN for the two generation scenarios (Figure 4). See supplementary for more generated samples and detailed qualitative analysis.

Ablation Study. To illustrate the impact of each discriminator in generating HOI videos, we conduct ablation experiments (refer to Table 3). We observe

Table 2. Quantitative Evaluation. Comparison of HOI-GAN with C-VGAN, C-TGAN, and MoCoGAN baselines. We distinguish training of HOI-GAN with bounding boxes (*bboxes*) and segmentation masks (*masks*). Arrows indicate whether lower (↓) or higher (↑) is better. [I: inception score; S: saliency score; D: diversity score]

	Model	EPIC						SS					
		GS1			GS2			GS1			GS2		
		I↑	S↓	D↑	I↑	S↓	D↑	I↑	S↓	D↑	I↑	S↓	D↑
	C-VGAN [66]	1.8	30.9	0.2	1.4	44.9	0.3	2.1	25.4	0.4	1.8	40.5	0.3
	C-TGAN [57]	2.0	30.4	0.6	1.5	35.9	0.4	2.2	28.9	0.6	1.6	39.7	0.5
	MoCoGAN [64]	2.4	30.7	0.5	2.2	31.4	1.2	2.8	17.5	1.0	2.4	33.7	1.4
(ours)	HOI-GAN (bboxes)	6.0	14.0	3.4	5.7	20.8	4.0	6.6	12.7	3.5	6.0	15.2	2.9
(ours)	HOI-GAN (masks)	**6.2**	**13.2**	**3.7**	**5.2**	**18.3**	**3.5**	**8.6**	**11.4**	**4.4**	**7.1**	**14.7**	**4.0**

Table 3. Ablation Study. We evaluate the contributions of our pixel-centric losses (F,G,V) and relational losses (first block vs. second block) by conducting ablation study on HOI-GAN (masks). The last row corresponds to the overall proposed model.[F: frame discriminator $\mathbf{D}_f$; G: gradient discriminator $\mathbf{D}_g$; V: video discriminator $\mathbf{D}_v$; R: relational discriminator $\mathbf{D}_r$]

	Model	EPIC						SS					
		GS1			GS2			GS1			GS2		
		I↑	S↓	D↑	I↑	S↓	D↑	I↑	S↓	D↑	I↑	S↓	D↑
−R	HOI-GAN (F)	1.4	44.2	0.2	1.1	47.2	0.3	1.8	34.7	0.4	1.5	39.5	0.3
−R	HOI-GAN (F+G)	2.3	25.6	0.7	1.9	30.7	0.5	3.0	24.5	0.9	2.7	28.8	0.7
−R	HOI-GAN (F+G+V)	2.8	21.2	1.3	2.6	29.7	1.7	3.3	18.6	1.2	3.0	20.7	1.0
+R	HOI-GAN (F)	2.4	24.9	0.8	2.2	26.0	0.7	3.1	20.3	1.0	2.9	27.7	0.9
+R	HOI-GAN (F+G)	5.9	15.4	3.5	4.8	21.3	3.3	7.4	12.1	3.5	5.4	19.2	3.4
+R	HOI-GAN (F+G+V)	6.2	13.2	3.7	5.2	18.3	3.5	8.6	11.4	4.4	7.1	14.7	4.0

that the addition of temporal information using the gradient discriminator and spatio-temporal information using the video discriminator lead to improvement in generation quality. In particular, the addition of our scene graph based relational discriminator leads to considerable improvement in generation quality resulting in more realistic videos (refer to second block in Table 3). Additional quantitative studies and results are in the supplementary.

Human Evaluation. We recruited 15 sequestered participants for our user study. We randomly chose 50 unique categories and chose generated videos for half of them from generation scenario GS1 and the other half from GS2. For each category, we provided three instances, each containing a pair of videos; one generated using a baseline model and the other using HOI-GAN. For each instance, at least 3 participants (ensuring inter-rater reliability) are asked to choose the video that best depicts the given category. The (aggregate) human preference scores for our model versus the baselines range between 69-84% for both generation scenarios (refer Table 4). These results indicate that HOI-GAN generates more realistic videos than the baselines.

Table 4. Human Evaluation. Human Preference Score (%) for scenarios GS1 and GS2. All the results have p-value less than 0.05 implying statistical significance.

Ours / Baseline	GS1	GS2
HOI-GAN / MoCoGAN	**71.7**/28.3	**69.2**/30.8
HOI-GAN / C-TGAN	**75.4**/34.9	**79.3**/30.7
HOI-GAN / C-VGAN	**83.6**/16.4	**80.4**/19.6

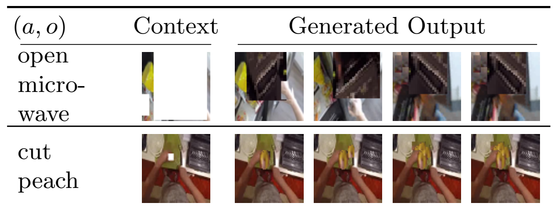


Fig. 5. Failure Cases. Videos generated using HOI-GAN corresponding to the given action-object composition (a, o) and the context frame. We show 4 frames of the videos.

Failure Cases. We discuss the limitations of our framework using qualitative examples shown in Figure 5. For *"open microwave"*, we observe that although HOI-GAN is able to generate conventional colors for a microwave, it shows limited capability to hallucinate such large objects. For *"cut peach"* (Figure 5), the generated sample shows that our model can learn the increase in count of partial objects corresponding to the action cut and yellow-green color of a peach. However, as the model has not observed the interior of a peach during training (as *cut peach* was not in training set), it is unable to create realistic transformations in the state of *peach* that show the interior clearly. We provide additional discussion on the failure cases in the supplementary.

5 Conclusion

In this paper, we introduced the task of zero-shot HOI video generation, *i.e.*, generating human-object interaction (HOI) videos corresponding to unseen action-object compositions, having seen the target action and target object independently. Towards this goal, we proposed the HOI-GAN framework that uses a novel multi-adversarial learning scheme and demonstrated its effectiveness on challenging HOI datasets. We show that an object-level relational discriminator is an effective means for GAN-based generation of interaction videos. Future work can benefit from our idea of using relational adversaries to synthesize more realistic videos. We believe relational adversaries to be relevant beyond video generation in tasks such as layout-to-image translation.

Acknowledgements. This work was done when Megha Nawhal was an intern at Borealis AI. We would like to thank the Borealis AI team for participating in our user study.

References

1. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: European Conference on Computer Vision (ECCV) (2016)
3. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning (ICML) (2017)
4. Bansal, A., Ma, S., Ramanan, D., Sheikh, Y.: Recycle-gan: Unsupervised video retargeting. In: European Conference on Computer Vision (ECCV) (2018)
5. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. In: International Conference on Learning Representations (ICLR) (2019)
6. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
7. Chao, Y.W., Liu, Y., Liu, X., Zeng, H., Deng, J.: Learning to detect human-object interactions. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2018)
8. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: Hico: A benchmark for recognizing human-object interactions in images. In: IEEE International Conference on Computer Vision (ICCV) (2015)
9. Damen, D., Doughty, H., Farinella, G.M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., Wray, M.: Scaling egocentric vision: The epic-kitchens dataset. In: European Conference on Computer Vision (ECCV) (2018)
10. Delaitre, V., Fouhey, D.F., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: European Conference on Computer Vision (ECCV) (2012)
11. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
12. Denton, E.L., et al.: Unsupervised learning of disentangled representations from video. In: Advances in neural information processing systems (NIPS) (2017)
13. Desai, C., Ramanan, D.: Detecting actions, poses, and objects with relational phraselets. In: European Conference on Computer Vision (ECCV) (2012)
14. Ebdelli, M., Le Meur, O., Guillemot, C.: Video inpainting with short-term windows: application to object removal and error concealment. In: IEEE Transactions on Image Processing (2015)
15. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: Zero-shot learning using purely textual descriptions. In: IEEE International Conference on Computer Vision (ICCV) (2013)
16. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
17. Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People watching: Human actions as a cue for single view geometry. In: International Journal of Computer Vision (IJCV) (2014)

18. Gkioxari, G., Girshick, R., Dollár, P., He, K.: Detecting and recognizing human-object interactions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
19. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS) (2014)
20. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The "something something" video database for learning and evaluating visual common sense. In: IEEE International Conference on Computer Vision (ICCV) (2017)
21. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
22. Granados, M., Kim, K.I., Tompkin, J., Kautz, J., Theobalt, C.: Background inpainting for videos with dynamic objects and a free-moving camera. In: European Conference on Computer Vision (ECCV) (2012)
23. Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., Saenko, K.: Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: IEEE International Conference on Computer Vision (ICCV) (2013)
24. Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object perception. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2007)
25. He, J., Lehrmann, A., Marino, J., Mori, G., Sigal, L.: Probabilistic video generation using holistic attribute control. In: European Conference on Computer Vision (ECCV) (2018)
26. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: IEEE International Conference on Computer Vision (ICCV) (2017)
27. Hsieh, J.T., Liu, B., Huang, D.A., Fei-Fei, L.F., Niebles, J.C.: Learning to decompose and disentangle representations for video prediction. In: Advances in neural information processing systems (NIPS) (2018)
28. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning (ICML) (2015)
29. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
30. Johnson, J., Gupta, A., Fei-Fei, L.: Image generation from scene graphs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
31. Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D., Bernstein, M., Fei-Fei, L.: Image retrieval using scene graphs. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
32. Kalchbrenner, N., van den Oord, A., Simonyan, K., Danihelka, I., Vinyals, O., Graves, A., Kavukcuoglu, K.: Video pixel networks. In: International Conference on Machine Learning (ICML) (2017)
33. Kalogeiton, V., Weinzaepfel, P., Ferrari, V., Schmid, C.: Joint learning of object and action detectors. In: IEEE International Conference on Computer Vision (ICCV). IEEE (2017)
34. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: International Conference on Learning Representations (ICLR) (2018)

35. Kato, K., Li, Y., Gupta, A.: Compositional learning for human object interaction. In: European Conference on Computer Vision (ECCV) (2018)
36. Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks (2017)
37. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (ICLR) (2017)
38. Kjellström, H., Romero, J., Kragić, D.: Visual object-action recognition: Inferring object affordances from human demonstration. In: Computer Vision and Image Understanding (CVIU) (2011)
39. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. In: International Journal of Computer Vision (IJCV) (2017)
40. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
41. Lei Ba, J., Swersky, K., Fidler, S., et al.: Predicting deep zero-shot convolutional neural networks using textual descriptions. In: IEEE International Conference on Computer Vision (2015)
42. Liang, X., Lee, L., Dai, W., Xing, E.P.: Dual motion gan for future-flow embedded video prediction. In: IEEE International Conference on Computer Vision (ICCV) (2017)
43. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in neural information processing systems (NIPS) (2017)
44. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European Conference on Computer Vision (ECCV) (2016)
45. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: International Conference on Machine Learning (ICML) (2013)
46. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. In: International Conference on Learning Representations (ICLR) (2016)
47. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations (ICLR) (2018)
48. Miyato, T., Koyama, M.: cgans with projection discriminator. In: International Conference on Learning Representations (ICLR) (2018)
49. Newell, A., Deng, J.: Pixels to graphs by associative embedding. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
50. Newson, A., Almansa, A., Fradet, M., Gousseau, Y., Pérez, P.: Video inpainting of complex scenes. In: SIAM Journal on Imaging Sciences (2014)
51. Niklaus, S., Mai, L., Liu, F.: Video frame interpolation via adaptive separable convolution. In: IEEE International Conference on Computer Vision (ICCV) (2017)
52. Odena, A., Olah, C., Shlens, J.: Conditional image synthesis with auxiliary classifier gans. In: International Conference on Machine Learning (ICML) (2017)
53. van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. In: Advances in neural information processing systems (NIPS) (2016)
54. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)

55. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text-to-image synthesis. In: International Conference on Machine Learning (ICML) (2016)
56. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge (2015)
57. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: IEEE International Conference on Computer Vision (ICCV) (2017)
58. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. In: Advances in neural information processing systems (NIPS) (2016)
59. Shen, Y., Lu, F., Cao, X., Foroosh, H.: Video completion for perspective camera under constrained motion. In: International Conference on Pattern Recognition (ICPR) (2006)
60. Sigurdsson, G.A., Russakovsky, O., Gupta, A.: What actions are needed for understanding human actions in videos? In: IEEE International Conference on Computer Vision (ICCV) (2017)
61. Srivastava, N., Mansimov, E., Salakhudinov, R.: Unsupervised learning of video representations using lstms. In: International conference on Machine Learning (ICML) (2015)
62. Stark, L., Bowyer, K.: Achieving generalized object recognition through reasoning about association of function to structure. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (1991)
63. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: IEEE conference on computer vision and pattern recognition (CVPR) (2016)
64. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
65. Villegas, R., Yang, J., Hong, S., Lin, X., Lee, H.: Decomposing motion and content for natural video sequence prediction. In: International Conference on Learning Representations (ICLR) (2017)
66. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. In: Advances in neural information processing systems (NIPS) (2016)
67. Walker, J., Doersch, C., Gupta, A., Hebert, M.: An uncertain future: Forecasting from static images using variational autoencoders. In: European Conference on Computer Vision (ECCV) (2016)
68. Walker, J., Marino, K., Gupta, A., Hebert, M.: The pose knows: Video forecasting by generating pose futures. In: IEEE International Conference on Computer Vision (ICCV) (2017)
69. Wang, T.C., Liu, M.Y., Tao, A., Liu, G., Kautz, J., Catanzaro, B.: Few-shot video-to-video synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2019)
70. Wang, T.C., Liu, M.Y., Zhu, J.Y., Liu, G., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. In: Advances in neural information processing systems (NeurIPS) (2018)
71. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

72. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
73. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
74. Xu, D., Zhu, Y., Choy, C.B., Fei-Fei, L.: Scene graph generation by iterative message passing. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
75. Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., He, X.: Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
76. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
77. Zellers, R., Yatskar, M., Thomson, S., Choi, Y.: Neural motifs: Scene graph parsing with global context. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
78. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: IEEE International Conference on Computer Vision (ICCV) (2017)
79. Zhao, J., Mathieu, M., LeCun, Y.: Energy-based generative adversarial network. In: International Conference on Learning Representations (ICLR) (2017)
80. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE International Conference on Computer Vision (ICCV) (2017)
81. Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., Elgammal, A.: A generative adversarial approach for zero-shot learning from noisy texts. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)