Concept Vectors for Zero-Shot Video Generation

Riya J. Dani

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Masters of Science
in
Computer Science and Application

Ismini Lourentzou, Chair
Hoda M. Eldardiry
Dawei Zhou

April 25, 2022
Blacksburg, Virginia

Keywords: Computing methodologies, Computer vision tasks, Vision for robotics, Neural networks, Information systems, Multimedia content creation

Concept Vectors for Zero-Shot Video Generation

Riya J. Dani

(ABSTRACT)

Zero-shot video generation involves generating videos of concepts (action classes) that are not seen in the training phase. Even though the research community has explored conditional video generation for long high-resolution videos, zero-shot video remains a fairly unexplored and challenging task. Most recent works can generate videos for action-object or motion-content pairs, where both the object (content) and action (motion) are observed separately during training, yet results often lack spatial consistency between foreground and background and cannot generalize to complex scenes with multiple objects or actions. In this work, we propose Concept2Vid that generates zero-shot videos for classes that are completely unseen during training. In contrast to prior work, our model is not limited to a predefined fixed set of class-level attributes, but rather utilizes semantic information from multiple videos of the same topic to generate samples from novel classes. We evaluate qualitatively and quantitatively on the Kinetics400 and UCF101 datasets, demonstrating the effectiveness of our proposed model.

Concept Vectors for Zero-Shot Video Generation

Riya J. Dani

(GENERAL AUDIENCE ABSTRACT)

Humans are able to generalize unseen scenarios without explicit feedback. They can be thought of as self-learning Artificial Intelligence agents that can collect data from various modalities (video, audio, text) found in surrounding environments, to develop new knowledge and acclimate to unseen situations without explicit feedback. Many recent studies have learned how to perform this process for images, but very few have been able to extend it to videos. Videos provide rich multi-modal data, such as text, audio, and images, and hence are composed of multifaceted knowledge that can introduce more complex temporal and spatial constraints. Leveraging videos in combination with text and audio data can assist intelligent systems to learn similar to how humans do. Zero-shot video generation (ZSVG) involves generating videos of concepts that are not seen in the training phase of a machine learning model. Generating a zero-shot video requires a multitude of temporal and spatial dependencies. In generating a video, not only does the model need temporal coherence but also the understanding of object properties. Current approaches for ZSVG are not well suited due to these challenges. We propose Concept2Vid which generates zero-shot videos for classes that are completely unseen during training. In contrast to prior work, our model is not limited to a predefined fixed set of class descriptions, but rather utilizes semantic information from multiple videos of the same topic to generate samples from novel classes. We evaluate qualitatively and quantitatively on the Kinetics400 and UCF101 datasets, demonstrating the effectiveness of our proposed model.

Dedication

Dedicated to my family and close friends

Acknowledgments

First and foremost, I would like to sincerely thank my advisor, Dr. Ismini Lourentzou, for providing me with constant support and guidance during my time as a graduate student. She has inspired me that women too can be at the forefront of spearheading technology. I am so honored to be the first student to defend his/her thesis and graduate under her advising.

I am very grateful for Dr. Hoda Eldardiry and Dr. Dawei Zhou to serve as my committee members.

I am incredibly appreciative for my labmate, mentor, friend, and Ph.D. candidate Afrina Tabassum for continuously guiding me by answering my countless late-night questions. She has helped me become a stronger and more confident female engineer.

I would like to show my gratitude towards the Department of Computer Science for accepting me as part of the Accelerated BS/MS program, allowing me to complete both my B.S and M.S in four years. VT CS has allowed me to meet the most kind, hard-working individuals with brilliant minds, some of whom I call my closest friends today. I am thankful for their endless love and support.

Finally, I express my deepest appreciation to my family. I am eternally thankful for my dad for introducing me to the Accelerated Masters program and my mom for constantly checking on me and supporting me during my toughest times. I would also like to thank my brother for instilling a deep appreciation and sparking my curiosity for technology.

I would like to acknowledge all of those who believed in me and supported me during some of the most challenging times.

Contents

List of Figures

List of Tables

List of Abbreviations

GAN Generative Adversarial Network

HOI Human Object Interaction

ZSL Zero Shot Learning

Chapter 1

Introduction

1.1 Motivation

The advancements in video-capturing technologies and the increase of social media video platforms have led to exponentially growing volumes of available video data covering a wide range of application domains. While image generation models exhibit impressively realistic results [17, 36, 68], video generation and understanding remain fairly challenging. Many of the existing works propose variants of conditional generative models to address video generation [15, 64, 79, 96, 97]. Yet, videos of new concepts are being created daily and the ability to generalize to new complex scenes heavily depends on the availability of video annotations, e.g., objects detected and actions recognized. Obtaining annotations that capture the spatio-temporal dependencies found in video data can quickly become time-consuming, expensive, and inefficient. To effectively address these issues, a couple of recent methods have also targeted zero-shot conditional video generation [41, 55].

Zero-shot conditional video generation aims to generate videos for a specified class or concept that is not seen during training. Methods typically involve class-conditional deep generative models such as Generative Adversarial Networks [41, 55], with varying inputs as starting points for initiating the video generation process, e.g., synthesizing videos from action-object pairs, each encoded with real-valued predefined vector representations, alongside a starting image frame [55] or from encoded motion-content pairs [41]. Hence, conditional classes are

frequently decomposed into motion and content or action and object [41, 55]. However, motion, content, action, or object labels are not easily available for all existing videos, and learning such disentangled representations may not be possible for all concepts and scenes. Moreover, the proposed zero-shot methods can only generate videos when both objects and actions or motions and contents are seen during training (but not necessarily their combination, e.g., for an unseen object-action pair "cutting tomato", there should be an example of "cutting" an object and an example with an action on a "tomato"). In practice, the set of possible visual contents such as actions, objects, etc. is huge, and therefore such methods lack the required flexibility for realistic zero-shot settings, as they cannot properly handle multiple objects simultaneously or complex combinations of actions.

On a related video understanding subfield, multi-video summarization targets the task of summarizing multiple videos by selecting representative frames from the input videos. Almost all the previous multi-video summarization techniques are extractive [50, 58], i.e., they select a subset of frames from multiple videos to form a summary. As a result, the summaries generated by these methods lack temporal consistency. Recent works in abstractive single-video summarization generate textual summaries of open-source videos by utilizing either only the video modality [18] or transcripts along with video [57] or audio and text (transcript) [46, 47]. However, these methods solve the inverse problem and are not able to generate videos. Moreover, these video2text summarization methods rely on rich information from additional modalities, such as lengthy documents and transcripts, which may not always be available and could result in ambiguous outcomes, either due to misalignment between modalities [25, 51, 92], missing information such as described entities not seen in video [70, 74, 98], or the inherent challenges in fusing modalities, which remains an open problem [5, 45]. Overall, abstractive text summarization methods for videos are not extended to either summarization from multiple videos or zero-shot settings. Generating videos for

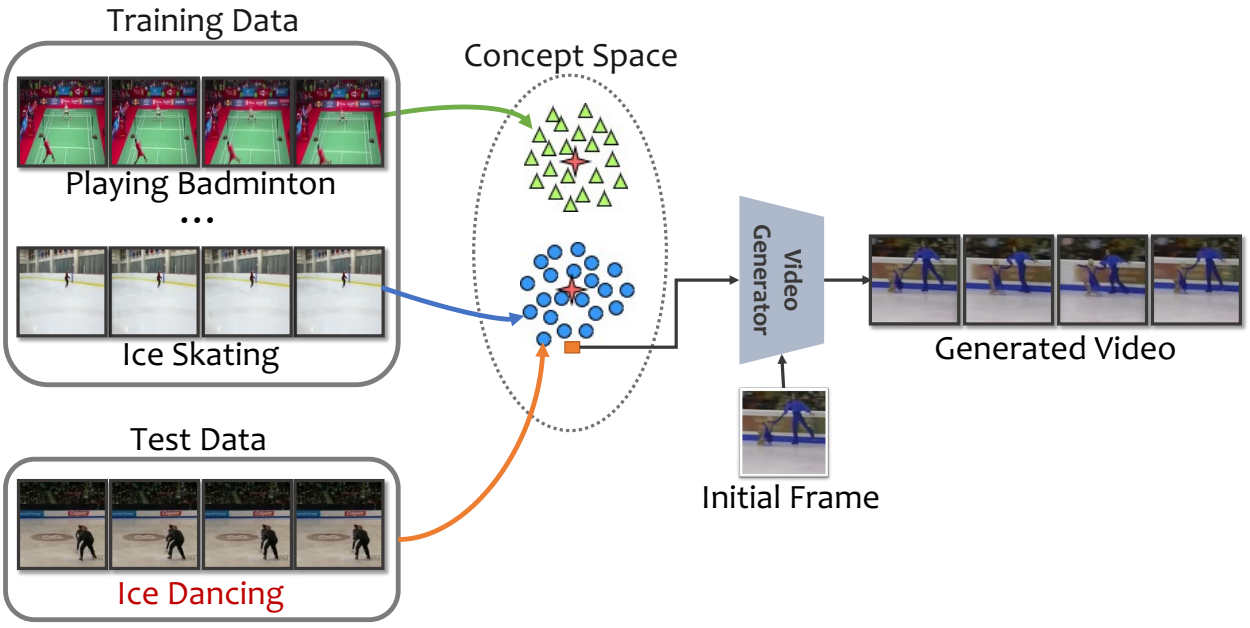


Figure 1.1: Overview of Concept2Vid. Given a set of training video examples, e.g., "playing badminton", "ice skating", etc., a zero-shot video generation process can utilize semantic class-level information to generalize to unseen concepts, e.g., "ice dancing".

novel classes and complex scenes remains a fairly challenging and open problem with limited prior work.

In this work, we introduce a more flexible zero-shot video generation framework. To address the aforementioned challenges, we propose Concept2Vid, a conditional generative model that can generate videos of concepts (topics) that are not observed during training. In contrast to prior work in zero-shot video generation, our model is not limited to a predefined fixed set of class-level attributes, but rather utilizes semantic information from multiple videos of the same topic to generate samples from novel classes. To the best of our knowledge, we are the first to introduce a zero-shot video generation method that can generalize to an open-set scenario for novel out-of-distribution classes. Starting by mapping the videos in a latent semantic space, Concept2Vid creates concept vectors that are utilized as class-conditional input when generating a video. We present two variations for concept vector generation,

a simple centroid-based approach and a set-based method that employs a permutation-equivariant Transformer network to embed multiple input videos to concept vectors. By mapping video concepts into a continuous space, Concept2Vid is capable of generating videos of new topics, actions or objects, even if these were not observed in the training phase. To the best of our knowledge, this is the first work that utilizes continuous conditional labels for video generation.

1.2 Contributions

The contributions of our work can be summarized as follows:

- We propose Concept2Vid, a conditional generative model that can generate videos in zero-shot settings, i.e., videos of novel concepts that are not seen during training.
- In contrast to prior works that condition the generation process with one-hot or word embedding vectors for a pre-defined set of attributes, we enhance the generalizability to novel classes and complex topics, by conditioning on per-video latent features. More specifically, our method utilizes continuous concept (class) representations extracted from multiple input videos.
- We empirically demonstrate the effectiveness of the proposed Concept2Vid model via a series of qualitative and quantitative experiments, showing that the proposed method results in videos that are more realistic.

Chapter 2

Related Work

2.1 Zero-Shot Learning

Zero-Shot Learning (ZSL) is a recent research area. In ZSL methods, the learner observes samples from classes that have not been seen during training. A ZSL approach is able to recognize a category of an object without it being seen before. ZSL is dependent on a labeled training set of seen and unseen classes. Both seen and unseen classes are related in a semantic space where the knowledge from seen classes can be applied to those that are unseen. ZSL techniques aim to learn semantic knowledge, attributes, and apply this to predict a new unseen class.

Essentially for a concrete example, let us say a model that has never seen a horse before is shown a picture of a horse. However, the model has seen a zebra before and is taught that a zebra looks like a horse but possesses black and white stripes. Then, the model can most likely identify it. Similarly, in ZSL methods, models are able to predict the label of an image without it being seen before in the training data. For more information on zero-shot learning, we refer the reader to a recent survey [86].

Narayan et al. [53] suggests enforcing semantics consistency in all the steps of ZSL, including training, feature synthesis, and classification, by leveraging a feedback loop from a Semantic Embedding Decoder (SED) that repeatedly looks over and updates the generated features

during both the training and feature synthesis step. Their method leverages a VAE-GAN architecture to propose a feedback module for zero-shot recognition. They evaluate their method on two generalized zero-shot action recognition in video datasets, HMDB51 [87] and UCF101 [71].

Xian et al. [93] provides a comprehensive evaluation of ZSL by providing a ZSL benchmark. This is done by evaluating three major aspects: methods, datasets, and evaluation protocols. These were evaluated on several datasets and the authors found that unlabeled data of unseen classes can accelerate zero-shot learning results. They also show that generative models and compatibility learning frameworks produce better results than attribute classifiers, learning independent objects, and hybrid models for zero-shot learning.

Changpinyo et al. [14] tackles the ZSL problem of classifying images from unseen classes into a label space via manifold learning. They align the two spaces: semantic and model. The model space coordinates should be the projection of the graph vertices from the semantic space to the model space. This would ensure the relatedness of similar classes encoded in the graph. Then, they leverage adaptable phantom classes to connect the seen and unseen classes together.

There has been a lot of work recently on ZSL [59, 61, 69, 100], but traditional ZSL methods are designed with one modality in mind. However, there is an added complexity of multi-modal data. Multi-modal data requires the fusing of modalities, re-aligning information between sources, and capturing spatio-temporal information between modalities. The following section describes multi-modal ZSL problems.

2.2 Multi-modal Zero-Shot Learning

Liu and Tuytelaars [48] provides a Deep Multi-modal Explanation (DME) model for zero-shot learning, that generates visuals and text to help with classification. By building a DME model, which incorporates a visual-attribute embedding and multi-channel explanation module from end-to-end, they are able to generate visual and textual explanations for classification tasks.

Felix et al. [22] proposes a technique for Generative Adversarial Network (GAN) training that makes the visual features produced to reconstruct to their original semantic features. Their model is trained with a multi-modal cycle-consistent semantic compatibility that allows the creation of more representative visual representations for the seen and unseen classes.

Bendre et al. [8] offers a Multimodal Variational Auto-Encoder (M-VAE) that can learn about the semantic latent space of image features. They use a multi-modal loss when they reconstruct the feature embeddings via the decoder. The reconstructed feature embeddings and the original labels are leveraged to train a Multi-Layer Perceptron (MLP) Classifier.

Several additional studies for multi-modal zero shot learning exist [21, 23, 33, 49]. In the subsequent sections, we will review zero-shot object recognition models that can be generally useful components for visual tasks. Finally, we will focus on zero-shot related works designed specifically for video.

2.3 Object-Recognition

Cacheux et al. [12] studies zero-shot learning with deep neural networks for object recognition. The general approach was to learn a mapping from visual data to semantic prototypes. Then, this is used to classify objects from the class prototypes. This paper presents a review

of a few different approaches using deep neural networks that work towards solving zero-shot learning. The approaches are under three different categories: regression methods, ranking methods, and generative methods.

The Ridge Regression approach views zero-shot learning as a regression problem. This means predicting the continuous attributes from a visual instance. Provided an image that is unseen, its corresponding semantic representation is estimated and the class with the nearest semantic prototype is predicted. The Embarrassingly Simple approach to Zero-Shot Learning (ESZSL) [62] follows similar ideas but instead of predicting the class prototypes from the visual features, it predicts the visual features from the class prototypes' features.

Triplet-loss approaches are ranking methods that learn a compatibility function. The compatibility of matching pairs is expected to be higher than the compatibility of non-matching pairs. In triplet loss approaches for zero-shot learning, the compatibility function is a bilinear mapping between the visual and semantic spaces, parameterized by a matrix. For example, the Deep Visual-Semantic Embedding model (DeViSE) [24] is a direct example of a triplet loss with a linear compatibility function for zero-shot learning, because the total loss is the sum of the triple loss over all training triplets. The Structure Joint Embedding (SJE) approach [4] is similar to DeVISE but when applied to zero-shot learning, only the class that violates the triple-loss constraint the most is used for each sample. The Attribute Label Embedding (ALE) [2, 3] approach views the zero-shot learning task as a ranking problem. The objective is to rank the correct class as high as possible on the list of unseen classes.

Generative approaches solving zero-shot learning create visual samples for unseen classes based on their semantic description, and then these samples can be used to train classifiers. The Generative Framework for Zero-Shot Learning(GFZSL) [84] makes the assumption that visual features are normally distributed given their class. The parameters of the distribution of a class depend on the class prototype. The models' parameters can be retrieved with

ride regression. The Synthesized Samples for Zero-Shot Learning (SSZSL) approach [27] estimates parameters for seen classes similar to GFZSL and predicts parameters for unseen classes. For a non-parametric approach, assumptions of the shape of the distribution of visual features are not made, and instead, generative methods such as Variational Auto-Encoders (VAEs) or Generative Adversarial Networks (GANs) are used to directly synthesize samples. The Synthesized Examples for GSL (SE-GZSL) approach [85] utilizes a conditional VAE. The VAE encoder maps an input to an $\mathbb{R}$-dimensional internal representation or latent code, while the VAE decoder tries to rebuild the input from the internal representation. The feature-GAN (f-GAN) approach [94] also uses conditional GANs to generate visual features.

2.4 Human-Object Recognition

There has been recent work on scaling Human-Object Interaction (HOI) recognition to categories using zero-shot learning. Shen et al. [66] create a factorized model for HOI detection that can produce detection for verb-object pairs after disentangling reasoning on verbs and objects. The factorized model is compromised of visual feature extraction layers, disentangled verb detection, and object detection networks. They model and learn verb and object representations to create combinations for zero-shot learning.

Ollah Maraghi and Faez [56] in their study offers an approach for human-object interaction recognition in videos via zero-shot learning, by recognizing a verb and object from a video to make a human-object interaction (HOI) class. Their model can detect unseen HOI classes.

2.5 Zero-Shot Video Classification

There are few existing zero-shot learning video classification frameworks currently existing. As mentioned, zero-shot classifiers work towards making generalizations to unseen test classes. Brattoli et al. [10] in their recent study explore zero-shot learning for video action recognition. Video action recognition is an area for which data sourcing and annotating are expensive. The paper contributes to a few important areas of zero-shot learning for video classification: novel modeling, evaluation protocol, and in-depth analysis. Novel modeling refers to the first end-to-end trained model for zero-shot action recognition. The evaluation protocol is the first training and evaluation protocol that targets a zero-shot setting for this task. This paper essentially defines zero-shot learning in the context of video classification.

The authors leverage a trainable 3D CNN to learn the visual features of videos. Evaluation is performed on the UCF101 [71], HMDB51 [87], ActivityNet [11], and SUN397 [95] datasets. For training, the authors use 2 protocols. In the initial training method, Kinetics700 [13] classes are filtered out after comparing the distance to classes in UCF101 or HMDB51. This results in a subset of Kinetics with 664 classes. In the second training protocol, certain classes are removed after comparing whether the distance to any class in UCF101 or HMDB51, or ActivityNet is smaller than the same value in the initial protocol. This results in Kinetics with 605 classes.

In the proposed evaluation framework, there exist two testing protocols. In the first protocol, the authors randomly choose half of the test dataset's classes from UCF101 and HMDB51. Then evaluate on this test set, and the results are averaged over 10 independent runs. In the second protocol, they evaluate on all UCF101 classes and HMDB51 classes.

In the pretrained setting, an R(2+1)D-18 model [77] is pre-trained on Kinetics400 [37] and C3D model [75] is pretrained on Sport-1M [35]. Word2Vec encodes every word and then

the multi-world class names are averaged. The proposed zero-shot classification method outperforms previous works, even after using stricter evaluation protocol.

Hong et al. [29] is another study on generalized zero-shot video classification via Generative Adversarial Networks (GAN). This was the first study to add class descriptions to zero-shot video classification. It does so by creating a loss function that combines both visual and text features. They extract text features from a text data set, assuming that videos with similar text dataset types should be considered generally similar.

The aforementioned works are designed for classification tasks and have a simpler output space than video generation tasks. In the next section, we briefly describe works that target zero-shot video generation specifically.

2.6 Zero-Shot Video Generation

The unprecedented success of GANs in generating images inspired video generation research that introduced a plethora of generative models such as VGAN [99], TGAN [63], MoCoGAN [79], DVD-GAN [15], VideoGPT [96], TGAN-v2 [64], DIGAN [97], etc.. Most recently, DIGAN [97] utilized implicit neural representations to generate longer high-resolution videos. However, these methods are not capable of generating videos for unseen classes that are not observed during training.

There has been limited work on zero-shot or few-shot video generation [41, 55]. For example, Nawhal et al. [55] consider action-object pairs as well as a starting image frame as inputs and utilize graph neural networks to map the relation between objects in the same frame and subsequent frames. However, this technique is restricted to a pre-defined fixed set of objects and actions and can generate videos only when both objects and actions are seen separately

during training, e.g., for an unseen object-action pair "cutting tomato", there should be an example of "cutting" an object and an example with an action on a "tomato". Moreover, the generated videos lack consistency between foreground and background.

A Motion and Content decomposed GAN (MoCoGAN) [80] model decomposes a video into two features capturing motion and content. It then generates a video by sampling every input noise vector from a different latent feature space.Kimura and Kawamoto [40] proposes a conditional-MoCoGAN that is able to learn a conditional generative model which divides the latent feature space into different classes via detailed class information on the two features' motion and content. The model is trained on two datasets, the Weizmann action video database [26] & the MUG facial expression video database [1]. Evaluation performed is both quantitative and qualitative.

Kimura and Kawamoto [40] develop a conditional generative adversarial network (GAN) model for zero-shot video generation that is able to generate unseen videos from training samples with missing classes. The proposed model is a conditional-MoCoGAN. A GAN model architecture is comprised of a generator and a discriminator. A generator inputs a noise vector and then generates data pertaining to the training data. A discriminator inputs data and tries to distinguish between fake and real data outputted by the generator. A conditional GAN (cGAN) is similar to a GAN structure, but in a cGAN both the noise vector and a condition vector are passed as input to the generator and discriminator. Kimura and Kawamoto [41] utilize conditional MoCoGAN [79], a video generation framework, to generate zero-shot videos given motion-content pairs as conditional class inputs. However, this technique also suffers from the same limitations and is restricted to generating videos with a static background. The challenge lies in utilizing class-level information that is encoded with one-hot or embedding vectors learned for a fixed set of attributes, i.e., actions, objects, motion or content classes. Such conditional inputs cannot generalize to complex

scenes with varying combinations of attributes or an open set of unseen classes. To alleviate the challenges with class-conditional video generation, we propose a zero-shot video generation method that relies on semantic continuous latent representations of concepts, learned from multiple videos capturing the same concept. Our experiments show that the proposed continuous conditioning can generate more realistic videos for classes that are unseen during training.

Nawhal et al. [54] is one of the first to generate human-object interaction videos for unseen compositions. They use an adversarial framework termed HOI-GAN, which leverages many discriminators to focus on different parts of a video. Their framework produces a fixed-length video clip when provided an action, object, and target scene for a context. For the human-object interaction (HOI) video to be realistic, it needs to contain an object with a semantic label, show the interaction with the object, be temporally consistent, and may occur in a specific scene. To this end, the authors train a generator network with four discriminators: frame, gradient, video, and relational. This will help the generator create accurate object layouts in a video. They also use pre-trained word embeddings that the discriminators are conditioned with.

From this literature review, it can be observed that there exists limited work on zero-shot video generation. There are other recent studies regarding video generation.

Bar et al. [7] is a recent work from 2020 that proposes to represent actions in a graph structure named Action Graph (AG) and shares a new synthesis task "Action Graph To Video." AG creates a generative model for "Action Graph To Video." which decomposes into motion and appearance features as well as includes a scheduling mechanism for actions to create a timely and coordinated video generation. A graph structure is used to represent coordinated and timed actions.

Huang et al. [30] introduces an unsupervised layered controllable video generation method that decomposes the initial frame of a video into foreground and background layers. The user can have input in the video generation process by manipulating the foreground mask. The paper describes a two-stage learning process. In the first stage, the model is learned how to separate the frame into the foreground and background layers to generate the next frame using a VQ-VAE [82] generator. In the second stage, the network is fine-tuned to foresee the edits to the mask.

2.7 Multi-Video Summarization

There have been several recent works in Multi-Video Summarization (MVS) [34, 50, 58, 90, 91]. Messaoud et al. [50] introduce a hierarchical attention network, trained with reinforcement learning, that can create a query-aware chronologically-sound summary from multiple videos. Ji et al. [34] utilize additional side information (title, description) while training a sparse auto-encoder for selecting keyframes from multiple videos. Panda et al. [58] propose a diversity-aware MVS method that selects complementary frames from each video to generate a diverse summary. Wu et al. [91] propose a dynamic graph convolutional network to select representative keyframes from multiple videos. All aforementioned MVS methods are extractive, i.e., they select a subset of frames from the input videos. However, given the lack of ground-truth summaries and the visual differences between videos (colors, audio, etc.), the final summary produced is often temporally and spatially incoherent. In addition, these methods cannot extend to video generation.

Another line of work deals with the task of abstractive single-video text summarization, where the goal is to generate a textual summary of one single video, by utilizing LSTM-based models [18] or hierarchical attention mechanism to integrate transcrips [57] or transformer-

based models to integrate multi-source data (audio, text) [6, 46, 47]. Nevertheless, these works rely on audio and text transcripts and generate abstractive textual summaries rather than video-based summaries. Additionally, such works are not extended to zero-shot or multi-video settings.

2.8 Latent class embeddings for conditional Generative Models

There have been a few works with conditional GANs (cGANs) that utilize latent class embeddings, mostly for image generation. Ding et al. [19] propose a Continuous conditional Generative Adversarial Network (CcGAN), the first generative model for image generation that is conditioned on continuous regression-based labels. Ditria et al. [20] present an open-set GAN architecture (OpenGAN) that is conditioned on a feature embedding drawn from a metric space. This metric space is defined by using a metric learning model that encodes class-level and scrupulous semantic information. Using this information, the generator produces images with features similar to the metric features extracted from real source images. Conditioning the generator on these semantically rich features allows for the generation of novel images.

Triess et al. [78] introduces the generation of 3D point cloud shapes conditioned on a continuous parameter, which is used to guide the generation process to create custom-fit 3D objects. Huh et al. [31] present a method that projects an input image into the space of a class-conditional generative neural network. Given an input target image, the method searches for a transformation to apply. Then, the latent vector that is most similar to the object in the target image is recovered via projection. The generative model can then be fine-tuned to reconstruct missing details and the image can be edited by altering the latent code or the class vector. This image is then inverted and blended back into the original

image. In terms of class-conditional models for video generation, Bar et al. [7] describe a method to represent actions in a graph structure (Action Graph) and an Action Graph to Video (AG2Vid) synthesis task. AG2Vid disentangles motion and appearance features and generates video using a scheduling mechanism. To the best of our knowledge, there has not been prior work on learning latent class embeddings for zero-shot conditional video generation.

Chapter 3

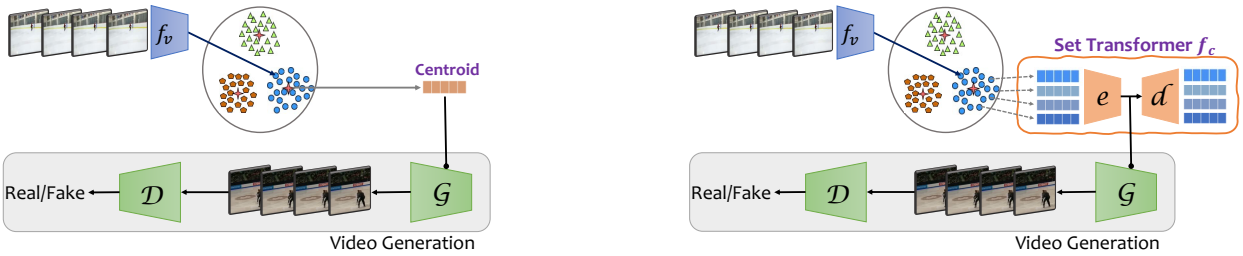
Methodology

3.1 Problem Definition

Let $\mathcal{V}_s = \{(x_i, y_i)\}_{i=1}^{N}$ set of N videos, where each video $x_i = [x_i^{(0)}, \ldots, x_i^{(L)}]$ is a sequence of L frames $x_i^{(l)} \in \mathbb{R}^{H \times W \times 3}$ that is associated with a class label $y_i \in \mathcal{Y}$. Zero-shot video generation aims at learning a function $G\colon \mathcal{Z} \rightarrow \mathcal{V}$ to synthesize videos from latent representations in $\mathcal{Z}$, such that G generalizes to unseen test classes. Previous works decompose $\mathcal{Z}$ to subspaces of content and motion, or object and action. Yet, this limits the flexibility of the framework to sampling based on one attribute from each of these predefined sets. Concept2Vid aims at first extracting semantically rich vector representations that can generalize well to novel classes and then generating videos with a generative adversarial network that is conditioned on embeddings drawn from the learned concept space. Concept2Vid is therefore composed of three sequential parts: i) visual feature extraction, ii) concept vector generation, and ii) zero-shot video generation.

3.2 Visual Feature Extraction

We first extract visual features from the video frames. We utilize a feature extractor that has achieved state-of-the-art performance on zero-shot action recognition [10]. More specifically, and for a given video x_i, a label encoder embeds label into a semantic embedding



(a) Concept2Vid $_C$ (b) Concept2Vid $_T$

Figure 3.1: Pictorial overview of our model, Concept2Vid illustrating two types of input to the video generation model. First, videos are mapped to an embedding space with a visual encoder model f_v. Then, Concept2Vid $_C$ (left) treats class centroids as concept vectors passed to the video generation process. In contrast, Concept2Vid $_T$ (right) utilizes a permutation-equivariant Transformer to directly learn representations for semantic classes.

$f_s(y_i) \in R^{d_z}$ and a separate 3D CNN encoder embeds the video into a visual embedding $f_v(x_i) \in R^{d_v}$, where d_z, d_v denote the dimensionality of the learned semantic and visual vector representations, respectively. The model is trained end-to-end by minimizing the MSE loss $\mathcal{L}_{su}(x_i) = \|f_s(y_i) - f_v(x_i)\|^2$, and the final classification utilizes a nearest-neighbor approach $f_c(x_i) = \arg\min_{y \in \mathcal{Y}} cos\left(f_v(x_i), f_s(y_i)\right)$, where cos denotes the cosine distance [10].

3.3 Concept Vector Generation

After mapping each video to a feature representation, we leverage representations from multiple videos to create semantic concept (class) vectors. We experiment with two variants, a simple centroid-based and a set-based approach.

3.3.1 Centroid-based Concept Vectors

The centroid method aims to analyze the effect of simpler concept vector creation. More specifically, let $\mathcal{V}_c = \{(x_i, y_i) \colon y_i = c \in \mathcal{Y}, i \in \mathcal{I}_c \subset \{1, \ldots, N\}\} \subset \mathcal{V}$ be the subset of

videos with the same class label c. Similarly, we create $|\mathcal{Y}|$ subsets for all unique classes, namely $\mathcal{V}_1, \ldots, \mathcal{V}_c, \ldots, \mathcal{V}_{|\mathcal{Y}|}$. To create a concept vector v_c for class $c \in \mathcal{Y}$, we employ simple averaging of the visual embeddings extracted from the feature extractor $f_v(x_i)$ for all videos originating from the same class

$$v_c = \frac{1}{|\mathcal{V}_c|} \sum_{x_i \in \mathcal{V}_c} f_v(xi), \tag{3.1}$$

and utilize this averaged representation as the conditional concept vector input for the video generation process. Note that compared to a one-hot vector embedding, the feature extractor can create such concept vectors for classes that are unseen during training. We term this model Concept2Vid $_C$.

3.3.2 Set-based Concept Vectors

While centroid-based vectors offer simplicity and no overhead, averaging is sensitive to outliers and may not work well for low-tail classes with fewer samples. To create a more robust method, we can directly learn vector representations for semantic classes. A target semantic class vector v_c can be extracted from autoencoding the visual embeddings originating from multiple videos $x_i \in \mathcal{V}_c$. To this end, we design a Set Transformer model for learning concept representations.

Specifically, we train a permutation-equivariant Transformer model to learn semantic concepts shared by videos of the same classes. Prior work has shown that Transformers generalize first-order permutation models such as DeepSets, and has proposed modifications to generalize to higher-order invariance [39, 42, 44]. Hence, a similar Transformer model that can learn concept representations from multiple videos of the same concept (class) is the most suitable for the proposed Concept2Vid approach.

To train a permutation-equivariant Transformer, we create uniformly sampled batches of b visual embeddings for videos in $\mathcal{V}_c$, i.e., with slight abuse of notation $\mathcal{B}_c = \{f_v(x_i) \colon x_i \in \mathcal{V}_c, i \in \mathcal{I}_B \subset \mathcal{I}_c\}$, with cardinality $|\mathcal{B}| = b << \mathcal{V}_c$. We perform the same operation to create several batches conditioned on class information, i.e., $\mathcal{B} = \{\mathcal{B}_1^k, \ldots, \mathcal{B}_c^k, \ldots, \mathcal{B}_{|\mathcal{Y}|}^k\}_{k=1}^K$. Then, we employ a permutation-equivariant Transformer encoder-decoder f_c as a learnable set transformation that learns to recreate the input set and can account for interactions between the set elements. Similarly to Lee et al. [44], we make two key changes to the original Transformer architecture [83]:

1. We exclude the positional embeddings as input to enforce the model to act similarly to different permutations of the input visual representations $f_v(x_i)$.

2. We exclude the masks that prevent the decoder 'peaking' ahead at the rest of the feature representation.

The loss function for training the Transformer model f_c aims to reconstruct the set of visual embeddings for batch $\mathcal{B}_c$ and is defined as the cross-entropy between the predicted set of features $f_c(f_v(x_i))$ and the original input set $f_v(x_i)$, computed for all batches as follows:

$$\mathcal{L}_c = -\sum_{\mathcal{B}_c \in \mathcal{B}} \sum_{x_i \in \mathcal{B}_c} f_v(x_i) \log \Big(f_c(f_v(x_i)) \Big). \tag{3.2}$$

After training, we can utilize the Transformer encoder to generate concept vectors $v_c = f_c(f_v(x_i))$ for all classes $c \in \mathcal{Y}$. We term this model Concept2Vid $_T$.

3.4 Zero-shot Video Generation

We built Concept2Vid on the top of DIGAN [97]. Specifically, we use the concept vector v_c as input to the DIGAN Generator so that it can learn to generate videos from a continuous space. DIGAN is selected due to its superior video generation performance as compared to other video generation models and the capability of generating high-quality videos in a class-conditional setup.

DIGAN is a state-of-the-art video generation network that is capable of generating long videos of high resolution and is an extension of INR-GAN [68] for image synthesis. The method utilizes implicit neural representations (INR) [67, 73] to map input coordinates of the corresponding 3D video coordinates to RGB values, where a video v is a continuous function of images I_t for time $t \in \mathbb{R}$, and an image I_t is a function of spatial coordinates$(x, y) \in \mathbb{R}^2$. The objective of INR is to model the signal with a neural network $(r, g, b) = v(x, y, t; \phi)$ parameterized by ϕ using a multi-layer perceptron (MLP).

The DIGAN generator $\mathcal{G}$ aims to generate the function of coordinates ϕ i.e., maps a latent vector $z \sim p(z)$ from a given prior distribution $p(z)$ as well as concept vector v_c to an INR parameter ϕ. To improve the generation quality the latent vector z and parameter ϕ are decomposed into motion components (z_M, ϕ_M) and content components (z_I, ϕ_I), respectively and the models learns functions $\phi_I = \mathcal{G}_I(z_I, v_c)$ and $\phi_M = \mathcal{G}_M(z_I, z_M, v_c)$ The DIGAN discriminator $\mathcal{D}$ is an efficient 2D Convolutional Neural Network(CNN) that verifies that video INRs are efficiently generating two frames of arbitrary times t_1, t_2. More specifically, $\mathcal{D}$ distinguishes the triplet consisting of a pair of images and their time difference (I_{t_1}, I_{t_2}, $|t_1 - t_2|$).

Chapter 4

Experiments

4.1 Experimental Details

4.1.1 Datasets

We use the Kinetics-400 [38] and UCF101 [72] datasets for training and testing, respectively.

UCF101 contains approximately $13,000$ YouTube video clips from 101 human action classes. Action classes involve person-object interactions, person-person interactions, body motion, playing sports, and musical instruments. This is a common benchmark for video generation, as the dataset contains variability both in motion and background.

Kinetics-400 is a large-scale video action recognition dataset that contains YouTube video clips from 400 human action classes. Each clip is 10 seconds long and action classes involve person-object interactions, person-person interactions, and single-person actions. There also exist actions operating on the same object, such as "shooting basketball", "dribbling basketball" and "'playing basketball' as well as the same actions on different objects, such as "playing guitar", "playing violin" and "playing cello". To ensure zero-shot settings, we remove from Kinetics-400 all the classes whose distance to any class in UCF101 is smaller than a 0.05 threshold, similarly to Brattoli et al. [10]. After removing overlapping classes, there remain video clips from 372 Kinetics classes for training. In 4.1, sample videos of zero-shot

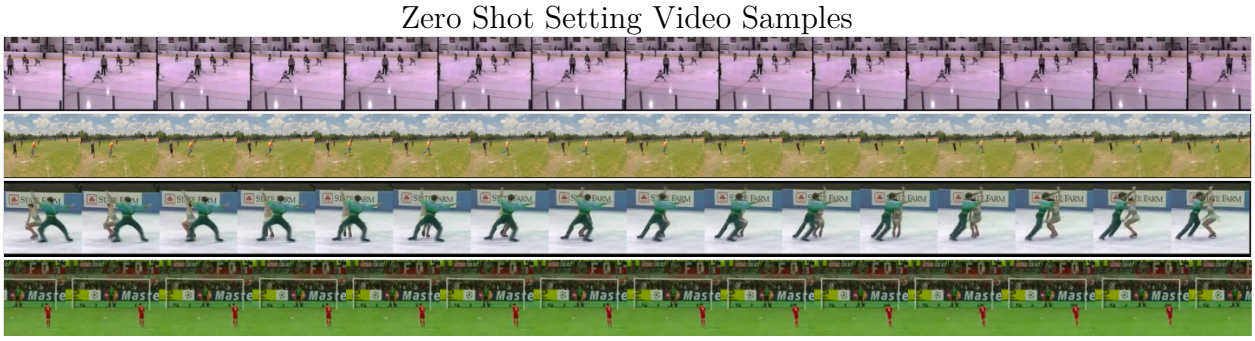


Figure 4.1: Examples from the benchmark datasets. The video samples presented are Kinetics-400 ice skating (first row), Kinetics-400 playing kickball (second row), UCF-101 ice dancing (third row), UCF-101 soccer penalty (fourth row).

settings are portrayed.

4.1.2 Hyper-parameter Details

Concept Vector Generation: The resolution of the input videos is 128×128. For training the feature extractor, we adopt the training settings and implementation details of Brattoli et al. [10]. More specifically, we make use of an R(2+1)D network [77] as the backbone of the extractor and use the semantic embeddings of individual words generated by Word2Vec [52] (Gesim Python implementation [60]). In the case of a multi-word class name, we use the average of the word embeddings for all words in the phrase. We train with Adam optimizer, 4×10^{-3} learning rate and a batch size of 88. We extract 512-dimensional visual embeddings from the last hidden encoder layer, i.e., before the final linear layer. For the permutation-equivariant Transformer, we use a transformer network with 8 multi-attention heads, an output dimension of 512, 4 hidden layers, and a dropout rate of 0.5. We train with Adam optimizer, batch size of 1500 and learning rate of 1×10^{-4}.

Video Generation: For training the generative adversarial network, we make use of videos from the following sports classes from Kinetics-400: ‘ice skating’, ‘ice climbing’, ‘jet skiing’,

'kicking soccer ball', 'playing badminton', 'playing ice hockey, 'playing kickball', 'playing tennis', 'playing cricket', 'playing basketball'. We follow the hyper-parameter settings of DIGAN [97]. We train the generator, image discriminator, and video discriminator models using ADAM optimizer with a batch size of 32 and learning rate of 2.5×10^{-3}. The feature dimensions for one-hot encoding, word2vec embedding, and concept vectors are 10, 300 and 512, respectively.

4.1.3 Baselines

We train all models on a single A100 Nvidia GPU for 3 days. We compare Concept2Vid in both traditional supervised video generation and zero-shot video generation settings, with the following baselines:

DIGAN [97]: The latest video-generation network which utilizes implicit neural representations to generate high-resolution long videos both unconditioned and conditioned on one-hot representations of action classes.

DIGAN$_{w2v}$: a zero-shot variation of the conditional DIGAN, where word2vec representations [52] of the class labels are used instead of one-hot representations.

4.1.4 Evaluation Metrics

For a fair comparison, we follow the evaluation setup of DIGAN [97] across all experiments. Furthermore, we use the following commonly used metrics:

Inception Score (IS) [65] is a popular metric for evaluating image generation methods, and is calculated as $exp(KL(p(y|x)||p(y)))$, where $p(y|x)$ is the conditional label distribution of a pre-trained Inception model trained on the ImageNet dataset [16]. For computing the

inception score for generated videos, we follow the setting of Yu et al. [97] and use a C3D network [76] pretrained on Sports-1M dataset [35] and fine-tuned on UCF101 dataset [72]. Similar to Yu et al. [97], we evaluate all methods over $10,000$ generated videos.

Frèechet Video Distance (FVD) [81] and Kernel Video Distance (KVD) [81] metrics are built based on Frèechet Image Distance (FID) [28], commonly used for image generation. FVD and KVD take into consideration a distribution over the whole video and avoid any drawbacks of previous video generation metrics [32]. Similar to [97], we utilize the I3D network trained on Kinetics-400. We present results by averaging 10 runs of scores computed from $2,054$ generated and real videos conditioned on sampled concepts.

We present extensive quantitative and qualitative comparisons of Concept2Vid with baseline methods under several experimental settings: traditional supervised video generation, zero-shot video generation, and zero-shot video generation given an initial frame as input.

4.2 Experimental Results

4.2.1 Supervised Video Generation

We first report FVD, KVD and Inception (IS) scores for supervised video generation, i.e., training and evaluating on the same dataset. In Table 4.1, we present results for conditional generative models trained with one-hot vectors (DIGAN), word2vec embeddings (DIGAN$_{w2v}$), set-based concept vectors (Concept2Vid $_T$), and centroid-based concept vectors (Concept2Vid $_C$) learned on the Kinetics-400 dataset. Concept2Vid $_T$ performs better than all baselines by 7.4% in terms of FVD and 16.4% in terms KVD distance (lower is better). Moreover, in terms of inception score, Concept2Vid $_C$ performs the best, which shows that while Concept2Vid $_T$ produces temporally-coherent videos, Concept2Vid $_C$ generates

Table 4.1: FVD, KVD and IS for video generation models trained and evaluated on Kinetics-400, i.e., traditional supervised video generation. For FVD and KVD lower values are better (↓), while for IS higher values are better (↑). Mean and standard deviation over 10 trials.

Method	FVD (↓)	KVD (↓)	IS(↑)
DIGAN [97]	564.41±29.60	85.16±6.80	8.10±0.12
DIGAN$_{w2v}$	614.20±41.52	83.29±8.18	8.02±0.10
Concept2Vid $_T$ (Ours)	522.88±25.94	69.60±5.33	8.14±0.11
Concept2Vid $_C$ (Ours)	566.08±22.93	83.02±5.87	8.39±0.16

more diverse distinct videos. Qualitative video generation results for all variations are shown in Figures 4.2, 4.3, and 4.4. Our proposed Concept2Vid models can generate fairly good videos for challenging and diverse scenarios.

4.2.2 Zero-shot Video Generation

We also report FVD, KVD and Inception (IS) scores of zero-shot video generation in Table 4.2. Note that unlike continuous vectors, such as word or latent embeddings, one-hot vectors are not available in zero-shot settings. Due to their discrete nature, one-hot encodings do not allow producing vectors for novel classes that are unobserved during training. We compare Concept2Vid $_C$ and Concept2Vid $_T$ with DIGAN$_{w2v}$ in zero-shot video generation (Table 4.2), i.e., for models trained on the Kinetics-400 dataset and evaluated on novel classes from the UCF101 dataset. We can observe that our proposed models perform better in terms of FVD, KVD and IS. Additionally, Figures 4.5 and 4.6 present qualitative results. Both Concept2Vid $_T$ and Concept2Vid $_C$ generate understandable good videos for novel classes originating from UCF101, which is a fairly challenging dataset that contains a large set of diverse concept classes. In Table 4.2, we can also observe that Concept2Vid $_C$ performs better than Concept2Vid $_T$ in zero-shot video generation, and outperforms the next best model by 1.54% in terms of FVD, 5.6% in terms of KVD and 7.6% in terms of IS

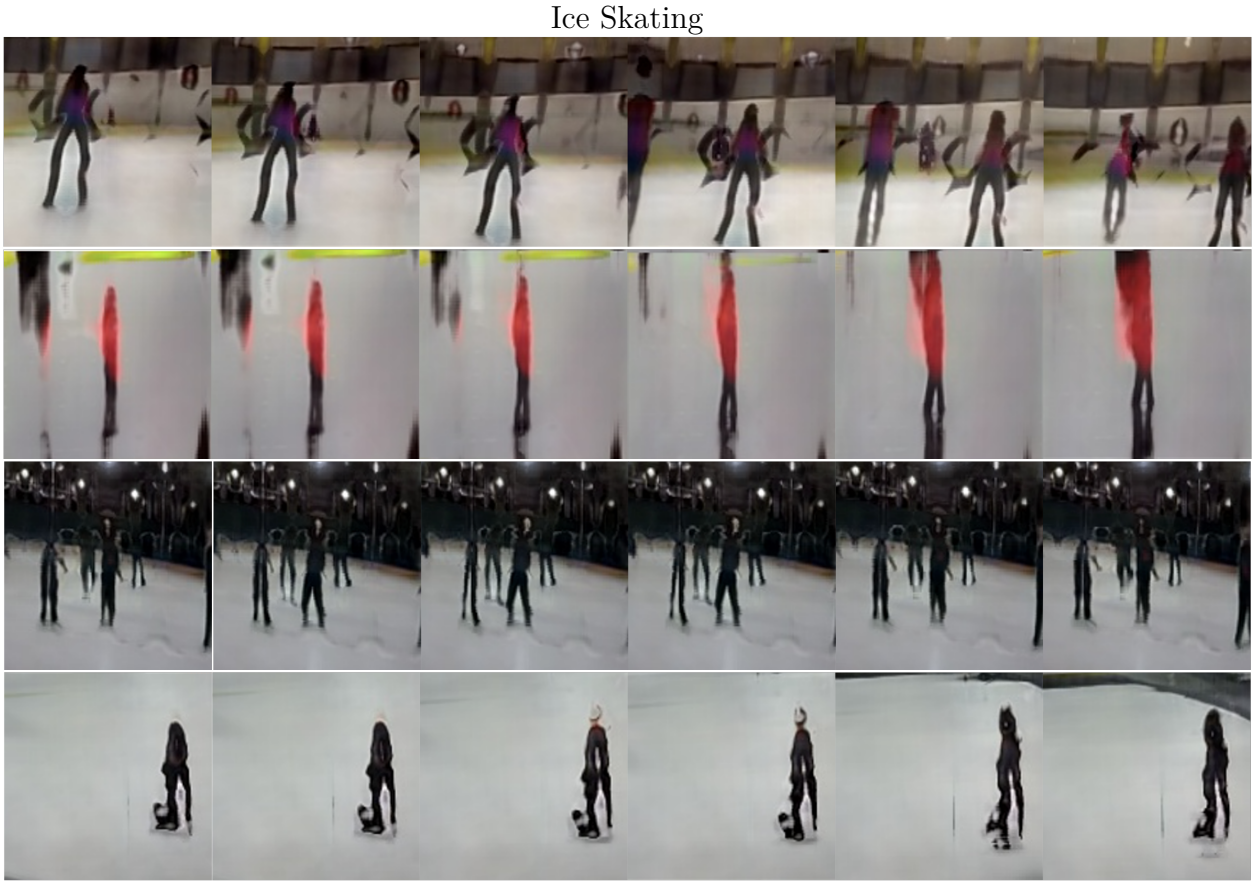


Figure 4.2: Generated video for the Kinetics-400 category 'ice skating'. Results for DIGAN (first row), DIGAN$_{w2v}$ (second row), set transformer based Concept2Vid $_T$ (third row) and centroid based Concept2Vid $_C$ (fourth row).

score. This comparison shows that, albeit being a fairly simple and straightforward method for concept vectors, Concept2Vid $_C$ can generalize better in zero-shot settings.

Moreover, we perform per-class analysis for FVD and KVD to check whether results vary drastically for specific classes, e.g., for semantically distant classes from intricate concepts such as 'field hockey penalty' or 'tennis swing'. Our results reported in Table 4.3 are stable, with Concept2Vid $_T$ being the best performing method across all evaluated 8 classes.

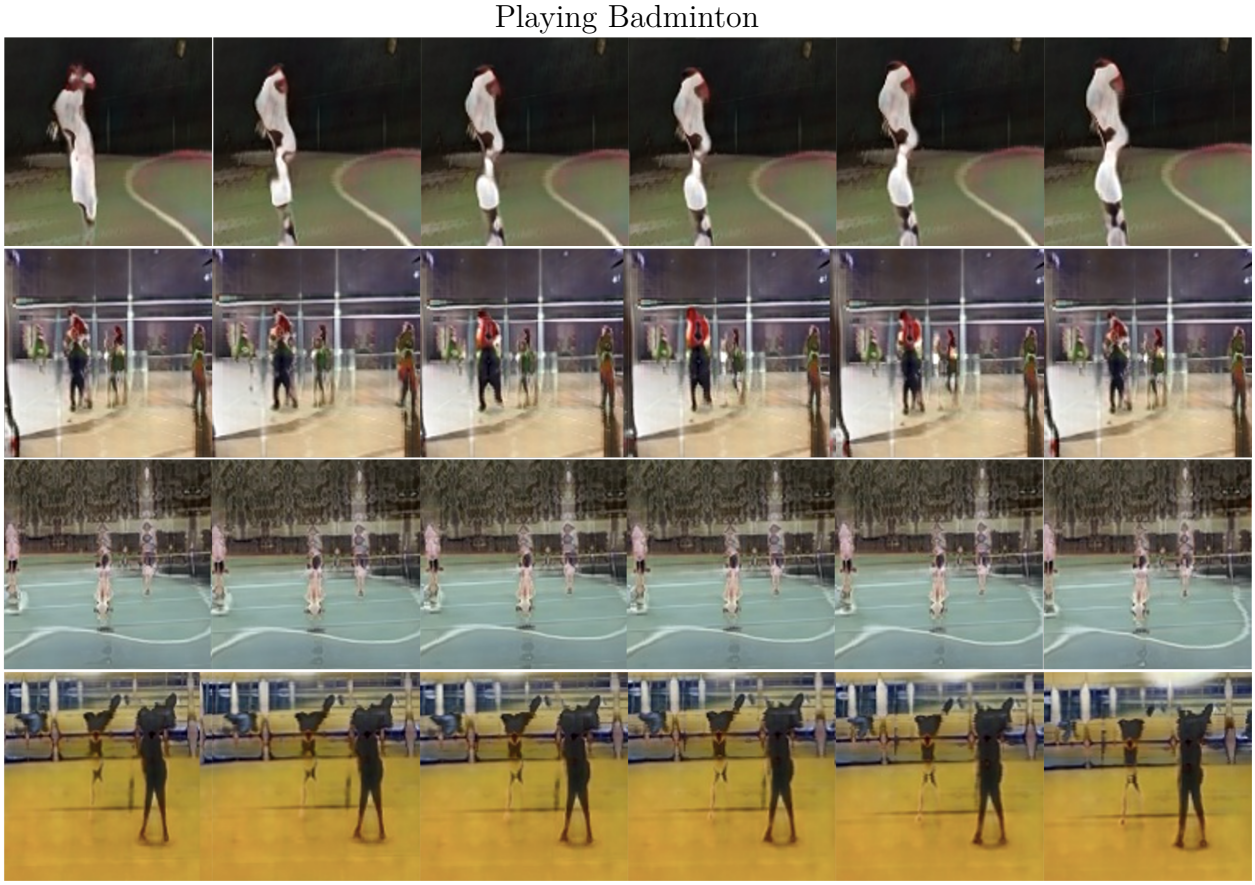


Figure 4.3: Generated video for the Kinetics-400 category 'playing badminton'. Results for DIGAN (first row), DIGAN$_{w2v}$ (second row), set transformer based Concept2Vid $_T$ (third row) and centroid based Concept2Vid $_C$ (fourth row).

Table 4.2: FVD, KVD and IS for video generation models trained on Kinetics-400 and evaluated on UCF101, i.e., zero-shot video generation settings. FVD and KVD lower values are better (↓), while IS higher values are better (↑). Mean and standard deviation reported over 10 trials.

Method	FVD (↓)	KVD (↓)	IS (↑)
DIGAN [97]	-	-	-
DIGAN$_{w2v}$	1312.52±27.93	152.10±6.97	8.12±0.10
Concept2Vid $_T$ (Ours)	1148.01±21.90	132.30±5.01	7.55±0.14
Concept2Vid $_C$ (Ours)	1130.24±16.04	124.85±4.36	8.79±0.17

Figure 4.4: Generated video for the Kinetics-400 category 'playing kickball'. Results for DIGAN (first row), DIGAN$_{w2v}$ (second row), set transformer based Concept2Vid $_T$ (third row) and centroid based Concept2Vid $_C$ (fourth row).

Table 4.3: Per-class analysis of FVD and KVD scores for zero-shot video generation (UCF101). Lower values are better. Mean and standard deviation over 10 trials.

	Method	IceDancing	SoccerPenalty	FieldHockeyPenalty	TennisSwing	TaiChi	Biking	Skiing	HorseRiding
FVD	DIGAN$_{w2v}$	1595.07±128.31	1687.98±229.26	1556.82 ±104.39	1628.17 ±252.92	1647.28 ±254.69	1525.58±98.16	1698.30 ± 237.43	1571.58 ±126.74
	Concept2Vid $_C$	1736.58±222.37	1795.68±410.45	1683.19±260.31	1761.62±397.86	1753.96±413.60	1846.24±136.50	1829.10 ±419.35	1715.03 ±233.10
	Concept2Vid $_T$	1326.83±116.69	1343.81±106.72	1289.38±82.37	1343.50±102.25	1340.43±105.51	1287.91±89.94	1350.85±109.33	1286.07 ±75.00
KVD	DIGAN$_{w2v}$	166.53 ±19.44	174.90 ±24.70	163.18 ±16.70	167.14 ±29.22	169.94 ±28.59	159.15 ±15.87	174.79 ±25.12	163.53 ±18.84
	Concept2Vid $_C$	184.52 ±41.14	182.82 ±40.49	190.91 ±53.91	179.81 ±41.01	177.86 ±41.84	219.96 ±41.51	186.41 ±41.06	185.02 ±44.15
	Concept2Vid $_T$	134.78 ±19.95	139.51 ±19.25	128.21 ±17.70	138.91 ±18.71	139.29 ±19.27	130.00 ±19.12	139.69 ±20.06	129.56 ±16.59

4.2.3 Zero-shot with Initial Frame as Input

We present evaluation results for zero-shot video generation with an initial frame passed as additional input to the generator. In Table 4.4 we can observe that our models perform

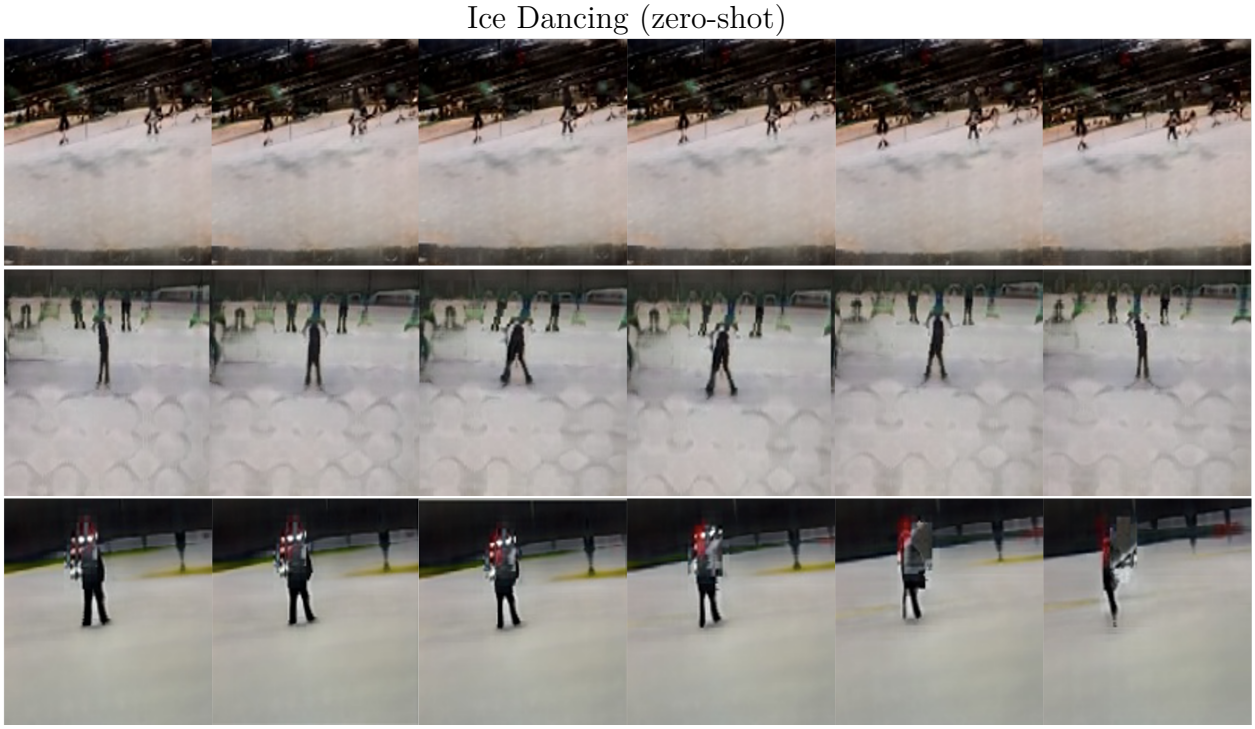


Figure 4.5: Generated video for the UCF101 category 'ice dancing' (zero-shot video generation). Results shown for DIGAN$_{word2vec}$ (first row), set-based Concept2Vid $_T$ (second row) and centroid-based Concept2Vid $_C$ (third row).

Table 4.4: FVD, KVD and IS metrics for video generation models trained on Kinetics-400 and evaluated on UCF101, i.e., zero-shot video generation settings, with an initial frame as additional input to the conditional generation. FVD and KVD lower values are better (↓), while IS higher values are better (↑). Mean and standard deviation reported over 10 trials.

Method	FVD (↓)	KVD (↓)	IS (↑)
DIGAN$_{w2v}$	1333.96±32.04	155.34±5.82	7.00±0.13
Concept2Vid $_T$ (Ours)	1421.73±41.65	164.71±8.42	6.78±0.09
Concept2Vid $_C$ (Ours)	1188.13±33.65	131.46 ±7.99	7.53±0.11

outperform DIGAN$_{w2v}$. We also present qualitative video generation for Zero-shot with Initial Frame as Input as shown in 4.8. In this setting, Concept2Vid $_C$ performs better than all methods across all metrics, surpassing DIGAN$_{w2v}$ by 10%, 15.4% , and 7.5% in terms of FVD, KVD and IS scores.

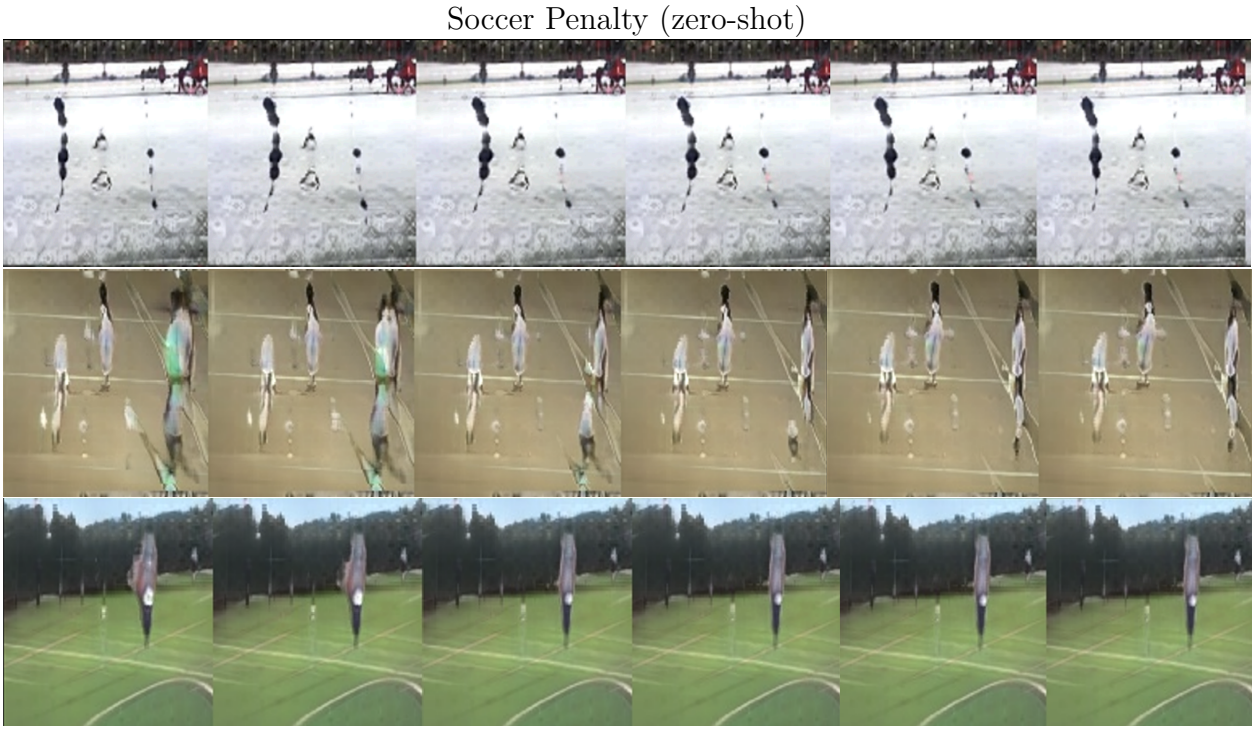


Figure 4.6: Generated video for the UCF101 category 'soccer penalty' (zero-shot video generation). Results shown for DIGAN$_{word2vec}$ (first row), set-based Concept2Vid $_T$ (second row) and centroid-based Concept2Vid $_C$ (third row).

4.2.4 Qualitative Analysis

We also validate whether Concept2Vid extracts semantically rich concept vectors that encode intra-class and inter-class information. We visualize clusters of video representations for classes that are observed during training (Kinetics-400) as well as novel classes (UCF101). The t-SNE visualization in Figure 4.11 shows the joint embedding space for 7 Kinetics-400 and 10 UCF101 classes, represented as triangles and circles, respectively, and with each class being visualized with a unique color. The plot shows that examples from novel classes form semantically coherent clusters and that similar classes are located nearby in the embedding space, which further validates the generalization capabilities of the proposed method.

Figure 4.7: Generated video for the UCF101 category 'tennis swing' (zero-shot video generation). Results shown for DIGAN$_{word2vec}$ (first row), set-based Concept2Vid $_T$ (second row) and centroid-based Concept2Vid $_C$ (third row).

Baseball Pitch (zero-shot with Initial Frame as Input)

Figure 4.8: Generated video for the UCF101 category 'baseball pitch' (Zero-shot with Initial Frame as Input). Results shown for DIGAN$_{word2vec}$ (first row), set-based Concept2Vid $_{T}$ (second row) and centroid-based Concept2Vid $_{C}$ (third row).

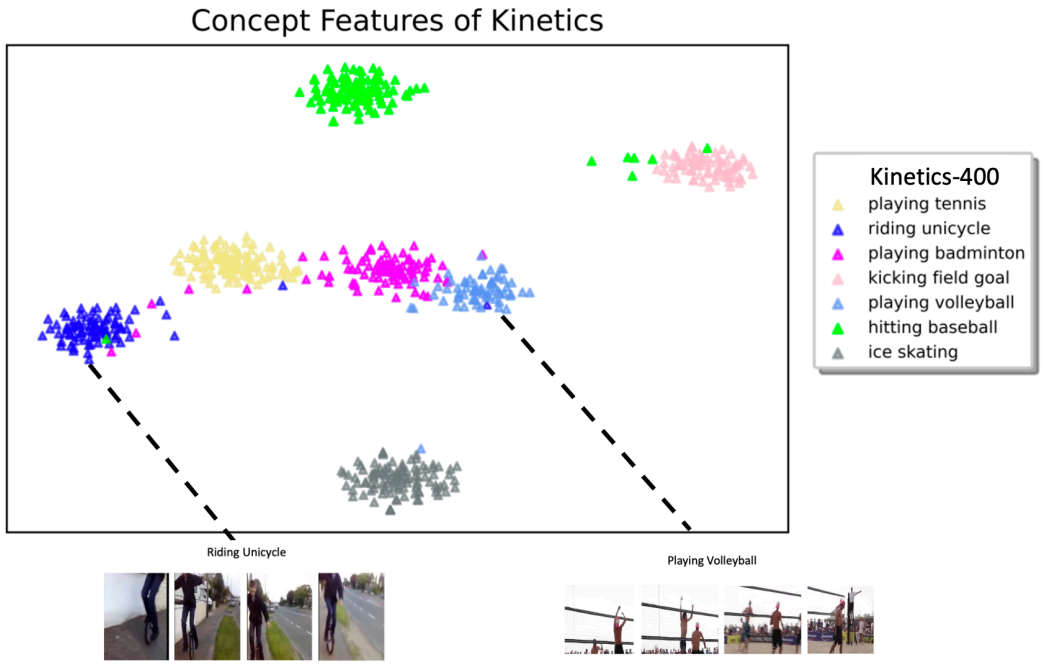


Figure 4.9: Visualisation of the learned concept embedding space for 7 Kinetics-400 classes (seen during training). Triangles correspond to Kinetics-400 data points.

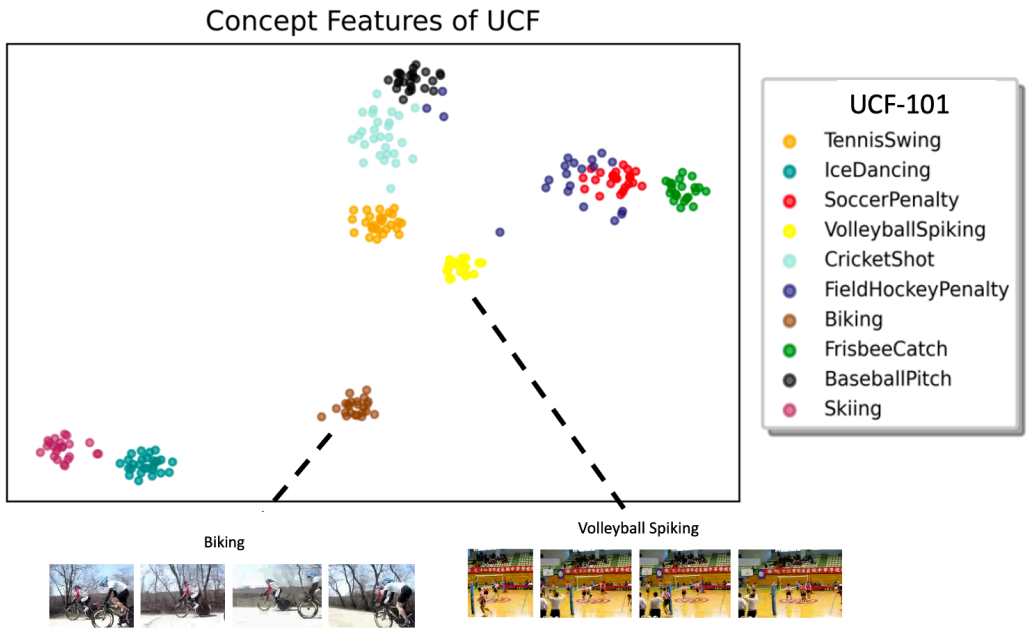


Figure 4.10: Visualisation of the learned concept embedding space for 10 UCF101 classes, represented as circles. The plot shows that examples from novel classes form semantically coherent clusters. which further validates the generalization capabilities of the proposed method.

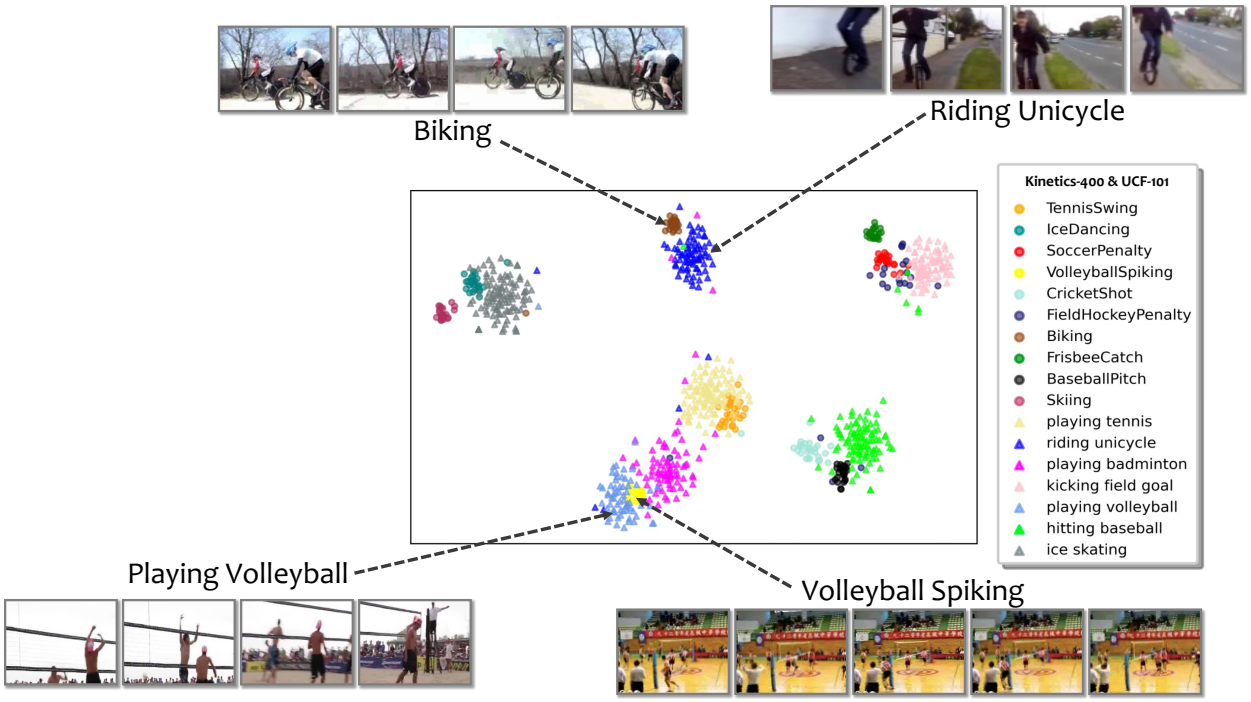


Figure 4.11: Visualisation of the learned concept embedding space for 7 Kinetics-400 classes (seen during training) and 10 UCF-101 classes (novel classes, unseen during training). Triangles correspond to Kinetics-400 and circles correspond to UCF101. Data points for different classes are represented by unique colors. Concept2Vid is not trained on UCF101 but is able to learn semantically meaningful concept embeddings, for which similar classes from both datasets are located nearby in the embedding space.

Chapter 5

Discussion and Limitations

5.1 Zero-Shot Video Generation Motivation

Humans are able to generalize unseen scenarios without explicit feedback or supervision signals. Humans can be thought of as self-learning AI-agents that can collect data from various modalities (video, audio, text) found in surrounding environments, to develop new knowledge and acclimate to unseen situations without explicit feedback. For instance, once someone has learned how to cut a tomato, cutting an apple is a similar process. Humans do this by inherently generalizing knowledge. Many recent studies have learned how to perform this process for images, but very few have been able to extend to videos. Videos provide rich multi-modal data, such as text, audio, images, and hence are composed of multifaceted knowledge that can introduce more complex temporal and spatial constraints. Leveraging videos in combination with text and audio data can assist intelligent systems to learn similar to how humans do. Video understanding methods can learn from video data to represent, summarize and generate videos under specific parameters.

As mentioned in the introduction, there exist limited works that propose zero-shot video generation methods, but can only generate videos when both objects and actions are seen during training, but not necessarily their combination.

In this work, we propose a more flexible method for conditional zero-shot video generation

and demonstrate quantitative and qualitative improvements. However, there is still work to be done and several limitations to tackle in the future. Below, we describe opportunities for improvement and open future directions.

5.2 Additional Experiments

5.2.1 Conditional MoCoGAN

In our preliminary experiments, we have tried conditional MoCoGAN (cMoCoGAN) as our video generation component [41, 79]. cMoCoGAN is trained and evaluated on the Weizmann action database [9] and on the MUG facial expression database [1], where the class of a video can be decomposed into two types of one-hot labels, 1) motion label i.e., the action label such as clapping, waving etc. 2) content label i.e., the participant who is performing the action. cMoCoGAN trains and tests on different combinations of these two types of labels, i.e., even though the motion-content combination is not present in the training phase, the motion and the content are present separately in the training phase.

To adopt the model to zero-shot settings and datasets that cannot be categorized as these two types of labels, we utilize the word2vec class mapping as input, similar to DIGAN$_{w2v}$. We additionally compare against a one-hot vector representation of class labels, similar to the original Conditional MoCoGAN [79]. Results in Table 5.1 show large score variations between the two models. Despite our best efforts, we were not able to produce good results with cMoCoGAN, not even in traditional video generation settings (training and testing on the same dataset).

Table 5.1: FVD scores for cMocoGAN and DIGAN traditional supervised and zero-shot video generation. FVD lower values are better (↓). Mean and standard deviation reported over 10 trials.

Method	Supervised	Zero-shot
cMoCoGAN	2120.93±213	-
DIGAN	564.41±29.60	-
cMoCoGAN$_{w2v}$	2174.84±243	3201.57±442
DIGAN$_{w2v}$	614.2±41.52	1312.52±27.93

Table 5.2: FVD scores for DIGAN traditional supervised and zero-shot video generation for open-set scenario. FVD lower values are better (↓). Mean and standard deviation reported over 10 trials.

Method (open-set)	Supervised	Zero-shot
DIGAN$_{w2v}$	239.35 ±14.71	1024.59±17.37
Concept2Vid $_T$ (Ours)	217.19±17.68	1003.92±16.50

5.2.2 Open-Set Experiments

In our preliminary experiments, we also experimented by training our Concept2Vid $_T$ on open-ended classes that range from a variety of actions. Specifically, we trained Concept2Vid $_T$ on the following classes from Kinetics-400: 'abseiling', 'air drumming', 'answering questions', 'applauding', 'applying cream', 'arm wrestling', 'arranging flowers', 'assembling computer', 'baby waking up'. We evaluated with zero-shot settings on the following classes: 'ApplyEyeMakeup', 'BabyCrawling', 'PlayingDaf', 'PlayingDhol', 'RopeClimbing'. Our results from this experiment, are shown in Table 5.2. Video generation results of traditional supervised video generation and zero-shot settings are, respectively, shown in Figure 5.1 and Figure 5.2. These results show that open-set scenario zero shot video generation performs poorly due to train and test classes being less interconnected and from a diverse set of categories. We discuss many of the challenges and limitations in the next section.

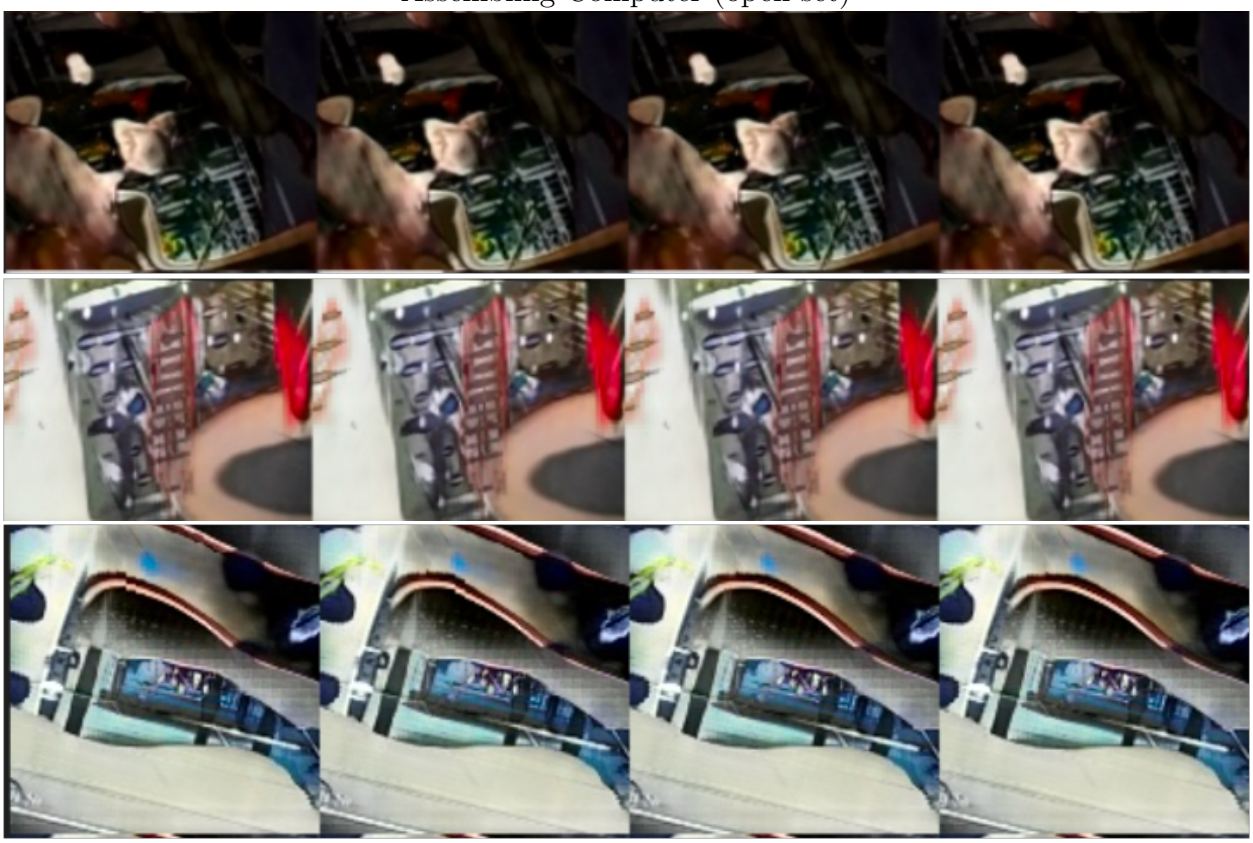


Figure 5.1: Generated video for the Kinetics-400 category 'assembling computer'. Results for DIGAN (first row), DIGAN$_{w2v}$ (second row), set transformer based Concept2Vid $_T$ (third row)

5.3 Why is Zero-Shot Video Generation a Challenging Task?

There are a variety of challenges in the task of ZSVG. In a zero-shot setting, test classes are not seen during training. This makes it difficult for a model to understand what an object looks like without having ever seen it, let alone learn to generate unseen objects and actions and combinations thereof. In order to successfully train ZSVG systems, it is necessary to understand and model objects, object relations, and spatiotemporal dependencies during training. The understanding of objects is vital, specifically capturing the attributes of color, texture, and style. For instance, if we want to generate a video of a Coca-Cola bottle, there

Figure 5.2: Generated video for the UCF101 category 'baby crawling' (zero-shot video generation) in an open-set scenario. Results shown for DIGAN$_{word2vec}$ (first row), set-based Concept2Vid $_T$ (second row)

are specific colors that are expected for the bottle to have. These attributes are difficult to capture in a video and cannot easily be done in a zero-shot setting because a model would simply not know what the object's properties are. When constructing a complex scene with multiple individuals, it is important to consider how these interact with each other as well (e.g., to generate a video of playing football). Actions have long-range spatio-temporal dependencies between interactions of objects and people. For instance, when a football player throws the ball, all of the other players must be coordinated and carefully timed [7]. In generating a video, not only does the model need temporal coherence but also the understanding of object properties. For a ZSVG approach, it is necessary to understand the foreground features, background features, and what the main properties an object should possess to generate a comprehensible video, especially if it is an unseen object. Current approaches for ZSVG are not well suited due to these challenges.

5.4 Future Work in Video Generation

Overall, video generation remains a challenging task, let alone zero-shot video generation. Most prior work on supervised video generation evaluates and trains on constraints settings with datasets for facial expressions, specific categories of objects or motions such as food, tai-chi, sky time-lapse, tabletop scenes with a small predefined set of objects, etc. [7, 97] Other works present relatively good results for video prediction [43], i.e., predicting the next frames given the first few frames, a task that also is challenging; extending such methods to (zero-shot) open-set video generation from scratch remains unclear.

The best models on large and diverse datasets such as Kinetics-400 are typically pretrained on additional data containing millions of videos [15, 88] and often also pre-train on a broad spectrum of related tasks, e.g., text-to-image generation or video prediction [89]. As a result, these models are computationally expensive, requiring more than half a hundred GPUs (e.g., 64 V100 GPUs) and weeks of training time [88, 89] or up to 512 TPU pods to reduce training time [15]. As such, improvements often come from the availability of large computational resources to train models, and even then the video quality remains generally low. Methods that are more efficient and environmentally friendly are necessary for expanding video generation usability and practicality in realistic applications. In addition, due to the increased difficulty, zero-shot video generation remains a fairly unexplored task. We hope our work will fill in this gap and will help spur the development of efficient zero-shot video generation methods.

Chapter 6

Conclusion

In this work, we introduce Concept2Vid, a conditional video generation model that can generate videos for concepts that are unobserved in the training phase. To enable flexible zero-shot video generation for novel classes, Concept2Vid maps videos into a visual embedding space and generates latent concept representations from multiple videos covering the same concept, either by utilizing a set transformer or by simply computing the centroids of the visual embeddings. Subsequently, Concept2Vid utilizes these concept vectors as conditional input to a video generation model.

To illustrate the effectiveness of the proposed method, we present qualitative and quantitative results on two benchmark datasets in traditional supervised and zero-shot video generation settings. Our experimental results show that Concept2Vid surpasses baselines and is able to generate more realistic videos. In the future, we plan to improve the video quality and evaluate our model on other datasets, such as Sports-1M and Howto100M. We also hope to extend our work on latent concept representations to related video understanding tasks, e.g., video prediction, text-to-video generation, etc.

Bibliography

[1] Niki Aifanti, Christos Papachristou, and Anastasios Delopoulos. The mug facial expression database. In 11th International Workshop on Image Analysis for Multimedia Interactive Services, pages 1–4. IEEE, 2010.

[2] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 819–826, 2013.

[3] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. IEEE transactions on pattern analysis and machine intelligence, 38(7):1425–1438, 2015.

[4] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2927–2936, 2015.

[5] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems, 34, 2021.

[6] Yash Kumar Atri, Shraman Pramanick, Vikram Goyal, and Tanmoy Chakraborty. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. Knowledge-Based Systems, 227:107152, 2021.

[7] Amir Bar, Roei Herzig, Xiaolong Wang, Anna Rohrbach, Gal Chechik, Trevor

Darrell, and Amir Globerson. Compositional video synthesis with action graphs. arXiv:2006.15327, 2020.

[8] Nihar Bendre, Kevin Desai, and Peyman Najafirad. Generalized zero-shot learning using multimodal variational auto-encoder with semantic concepts. CoRR, abs/2106.14082, 2021. URL https://arxiv.org/abs/2106.14082.

[9] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In Tenth IEEE International Conference on Computer Vision (ICCV), volume 2, pages 1395–1402. IEEE, 2005.

[10] Biagio Brattoli, Joseph Tighe, Fedor Zhdanov, Pietro Perona, and Krzysztof Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4613–4623, 2020.

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In Proceedings of the ieee conference on computer vision and pattern recognition, pages 961–970, 2015.

[12] Yannick Le Cacheux, Hervé Le Borgne, and Michel Crucianu. Zero-shot learning with deep neural networks for object recognition, 2021.

[13] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724–4733, 2017. doi: 10.1109/CVPR.2017.502.

[14] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning, 2016.

[15] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. arXiv:1907.06571, 2019.

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

[17] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. Advances in Neural Information Processing Systems, 34, 2021.

[18] Aniqa Dilawari and Muhammad Usman Ghani Khan. Asovs: abstractive summarization of video sequences. IEEE Access, 7:29253–29263, 2019.

[19] Xin Ding, Yongwei Wang, Zuheng Xu, William J Welch, and Z Jane Wang. Ccgan: Continuous conditional generative adversarial networks for image generation. In International Conference on Learning Representations, 2020.

[20] Luke Ditria, Benjamin J Meyer, and Tom Drummond. Opengan: Open set generative adversarial networks. In Proceedings of the Asian Conference on Computer Vision, 2020.

[21] Mohamed Elhoseiny, Jingen Liu, Hui Cheng, Harpreet Sawhney, and Ahmed Elgammal. Zero-shot event detection by multimodal distributional semantic embedding of videos. In Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[22] Rafael Felix, B. G. Vijay Kumar, Ian D. Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. CoRR, abs/1808.00136, 2018. URL http://arxiv.org/abs/1808.00136.

[23] Rafael Felix, Ben Harwood, Michele Sasdelli, and Gustavo Carneiro. Generalised zero-

shot learning with a classifier ensemble over multi-modal embedding spaces. CoRR, abs/1908.02013, 2019. URL http://arxiv.org/abs/1908.02013.

[24] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. Advances in neural information processing systems, 26, 2013.

[25] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In European Conference on Computer Vision, pages 214–229. Springer, 2020.

[26] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. IEEE transactions on pattern analysis and machine intelligence, 29(12):2247–2253, 2007.

[27] Yuchen Guo, Guiguang Ding, Jungong Han, and Yue Gao. Synthesizing samples fro zero-shot learning. IJCAI, 2017.

[28] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017.

[29] Mingyao Hong, Guorong Li, Xinfeng Zhang, and Qingming Huang. Generalized zero-shot video classification via generative adversarial networks. pages 2419–2426, 10 2020. doi: 10.1145/3394171.3413517.

[30] Jiahui Huang, Yuhe Jin, Kwang Moo Yi, and Leonid Sigal. Layered controllable video generation. arXiv preprint arXiv:2111.12747, 2021.

[31] Minyoung Huh, Richard Zhang, Jun-Yan Zhu, Sylvain Paris, and Aaron Hertzmann.

Transforming and projecting images into class-conditional generative networks. In European Conference on Computer Vision, pages 17–34. Springer, 2020.

[32] Quan Huynh-Thu and Mohammed Ghanbari. The accuracy of psnr in predicting video quality for different video scenes and frame rates. Telecommunication Systems, 49(1): 35–48, 2012.

[33] Zhong Ji, Kexin Chen, Junyue Wang, Yunlong Yu, and Zhongfei Zhang. Multi-modal generative adversarial network for zero-shot learning. Knowledge-Based Systems, 197: 105847, 2020. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2020.105847. URL https://www.sciencedirect.com/science/article/pii/S095070512030215X.

[34] Zhong Ji, Yuxiao Zhao, Yanwei Pang, and Xuelong Li. Cross-modal guidance based auto-encoder for multi-video summarization. Pattern Recognition Letters, 135:131–137, 2020.

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

[36] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In International Conference on Learning Representations, 2018.

[37] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. CoRR, abs/1705.06950, 2017. URL http://arxiv.org/abs/1705.06950.

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

[39] Jinwoo Kim, Saeyoon Oh, and Seunghoon Hong. Transformers generalize deepsets and can be extended to graphs & hypergraphs. Advances in Neural Information Processing Systems, 34, 2021.

[40] Shun Kimura and Kazuhiko Kawamoto. Conditional mocogan for zero-shot video generation. CoRR, abs/2109.05864, 2021. URL https://arxiv.org/abs/2109.05864.

[41] Shun Kimura and Kazuhiko Kawamoto. Conditional mocogan for zero-shot video generation. arXiv:2109.05864, 2021.

[42] Adam R Kosiorek, Hyunjik Kim, and Danilo J Rezende. Conditional set generation with transformers. In ICML 2020 Workshop on Object-Oriented Learning, 2020.

[43] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Ccvs: Context-aware controllable video synthesis. Advances in Neural Information Processing Systems, 34, 2021.

[44] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In International Conference on Machine Learning, pages 3744–3753. PMLR, 2019.

[45] Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:2476–2483, 2021.

[46] Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. D-mmt: A concise decoder-only multi-modal transformer for abstractive summarization in videos. Neurocomputing, 456:179–189, 2021.

[47] Nayu Liu, Xian Sun, Hongfeng Yu, Fanglong Yao, Guangluan Xu, and Kun Fu. Abstractive summarization for video: A revisit in multistage fusion network with forget gate. IEEE Transactions on Multimedia, 2022.

[48] Yu Liu and Tinne Tuytelaars. A deep multi-modal explanation model for zero-shot learning. IEEE Transactions on Image Processing, 29:4788–4803, 2020.

[49] Dwarikanath Mahapatra. Multimodal generalized zero shot learning for gleason grading using self-supervised learning. arXiv preprint arXiv:2111.07646, 2021.

[50] Safa Messaoud, Ismini Lourentzou, Assma Boughoula, Mona Zehni, Zhizhen Zhao, Chengxiang Zhai, and Alexander G Schwing. Deepqamvs: Query-aware hierarchical pointer networks for multi-video summarization. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1389–1399, 2021.

[51] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9879–9889, 2020.

[52] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv:1301.3781, 2013.

[53] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees G. M. Snoek, and Ling

Shao. Latent embedding feedback and discriminative features for zero-shot classification. CoRR, abs/2003.07833, 2020. URL https://arxiv.org/abs/2003.07833.

[54] Megha Nawhal, Mengyao Zhai, Andreas Lehrmann, Leonid Sigal, and Greg Mori. Generating videos of zero-shot compositions of actions and objects. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, Computer Vision – ECCV 2020, pages 382–401, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58610-2.

[55] Megha Nawhal, Mengyao Zhai, Andreas Lehrmann, Leonid Sigal, and Greg Mori. Generating videos of zero-shot compositions of actions and objects. In European Conference on Computer Vision, pages 382–401. Springer, 2020.

[56] Vali Ollah Maraghi and Karim Faez. Zero-shot learning on human-object interaction recognition in video. In 2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS), pages 1–7, 2019. doi: 10.1109/ICSPIS48872.2019.9066160.

[57] Shruti Palaskar, Jindřich Libovickỳ, Spandana Gella, and Florian Metze. Multimodal abstractive summarization for how2 videos. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6587–6596, 2019.

[58] Rameswar Panda, Niluthpol Chowdhury Mithun, and Amit K. Roy-Chowdhury. Diversity-aware multi-video summarization. IEEE Transactions on Image Processing, 26(10):4712–4724, 2017. ISSN 1941-0042. doi: 10.1109/tip.2017.2708902. URL http://dx.doi.org/10.1109/TIP.2017.2708902.

[59] Shafin Rahman, Salman Khan, and Fatih Porikli. A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning. IEEE Transactions on Image Processing, 27(11):5652–5667, 2018.

[60] Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In In Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks. Citeseer, 2010.

[61] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In Francis Bach and David Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 2152–2161, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/romera-paredes15.html.

[62] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In International conference on machine learning, pages 2152–2161. PMLR, 2015.

[63] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In Proceedings of the IEEE international conference on computer vision, pages 2830–2839, 2017.

[64] Masaki Saito, Shunta Saito, Masanori Koyama, and Sosuke Kobayashi. Train sparsely, generate densely: Memory-efficient unsupervised training of high-resolution temporal gan. International Journal of Computer Vision, 128(10):2586–2606, 2020.

[65] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. Advances in neural information processing systems, 29, 2016.

[66] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1568–1576, 2018. doi: 10.1109/WACV.2018.00181.

[67] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. Advances in Neural Information Processing Systems, 33:7462–7473, 2020.

[68] Ivan Skorokhodov, Savva Ignatyev, and Mohamed Elhoseiny. Adversarial generation of continuous images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10753–10764, 2021.

[69] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. Advances in neural information processing systems, 26, 2013.

[70] Young Chol Song, Iftekhar Naim, Abdullah Al Mamun, Kaustubh Kulkarni, Parag Singla, Jiebo Luo, Daniel Gildea, and Henry A Kautz. Unsupervised alignment of actions in video with text descriptions. In IJCAI, pages 2025–2031, 2016.

[71] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. CoRR, abs/1212.0402, 2012. URL http://dblp.uni-trier.de/db/journals/corr/corr1212.html#abs-1212-0402.

[72] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv:1212.0402, 2012.

[73] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. Advances in Neural Information Processing Systems, 33:7537–7547, 2020.

[74] Zineng Tang, Jie Lei, and Mohit Bansal. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In Proceedings of the

2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2415–2426, 2021.

[75] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In 2015 IEEE International Conference on Computer Vision (ICCV), pages 4489–4497, 2015. doi: 10.1109/ICCV.2015.510.

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

[77] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 6450–6459, 2018.

[78] Larissa T Triess, Andre Bühler, David Peter, Fabian B Flohr, and J Marius Zöllner. Point cloud generation with continuous conditioning. arXiv:2202.08526, 2022.

[79] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. MoCoGAN: Decomposing motion and content for video generation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1526–1535, 2018.

[80] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1526–1535, 2018.

[81] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin

Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. arXiv:1812.01717, 2018.

[82] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. Advances in neural information processing systems, 30, 2017.

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[84] Vinay Kumar Verma and Piyush Rai. A simple exponential family framework for zero-shot learning. In Joint European conference on machine learning and knowledge discovery in databases, pages 792–808. Springer, 2017.

[85] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 4281–4289, 2018.

[86] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2):1–37, 2019.

[87] David S Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, et al. Hmdb: the human metabolome database. Nucleic acids research, 35(suppl_1):D521–D526, 2007.

[88] Chenfei Wu, Lun Huang, Qianxi Zhang, Binyang Li, Lei Ji, Fan Yang, Guillermo Sapiro, and Nan Duan. Godiva: Generating open-domain videos from natural descriptions. arXiv:2104.14806, 2021.

[89] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. N\" uwa: Visual synthesis pre-training for neural visual world creation. arXiv:2111.12417, 2021.

[90] Jiaxin Wu, Sheng-Hua Zhong, and Yan Liu. Mvsgcn: A novel graph convolutional network for multi-video summarization. In Proceedings of the 27th ACM International Conference on Multimedia, pages 827–835, 2019.

[91] Jiaxin Wu, Sheng-hua Zhong, and Yan Liu. Dynamic graph convolutional network for multi-video summarization. Pattern Recognition, 107:107382, 2020.

[92] Zhiyong Wu, Lianhong Cai, and Helen Meng. Multi-level fusion of audio and visual features for speaker identification. In International Conference on Biometrics, pages 493–499. Springer, 2006.

[93] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly. CoRR, abs/1707.00600, 2017. URL http://arxiv.org/abs/1707.00600.

[94] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5542–5551, 2018.

[95] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 3485–3492, 2010.

[96] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. arXiv:2104.10157, 2021.

[97] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In International Conference on Learning Representations, 2022.

[98] Yitian Yuan, Tao Mei, and Wenwu Zhu. To find where you talk: Temporal sentence localization in video with attention based location regression. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 9159–9166, 2019.

[99] Shuangfei Zhai, Yu Cheng, Rogerio Feris, and Zhongfei Zhang. Generative adversarial networks as variational training of energy based models. arXiv:1611.01799, 2016.

[100] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2021–2030, 2017.