S3VAE: Self-Supervised Sequential VAE for Representation Disentanglement and Data Generation

Yizhe Zhu\textsuperscript{1,2}, Martin Renqiang Min\textsuperscript{1}, Asim Kadav\textsuperscript{1}, Hans Peter Graf\textsuperscript{1}
yizhe.zhu@rutgers.edu, \{renqiang, asim, hpg\}@nec-labs.com
\textsuperscript{1}NEC Labs America, \textsuperscript{2}Department of Computer Science, Rutgers University

Abstract

We propose a sequential variational autoencoder to learn disentangled representations of sequential data (e.g., videos and audios) under self-supervision. Specifically, we exploit the benefits of some readily accessible supervisory signals from input data itself or some off-the-shelf functional models and accordingly design auxiliary tasks for our model to utilize these signals. With the supervision of the signals, our model can easily disentangle the representation of an input sequence into static factors and dynamic factors (i.e., time-invariant and time-varying parts). Comprehensive experiments across videos and audios verify the effectiveness of our model on representation disentanglement and generation of sequential data, and demonstrate that, our model with self-supervision performs comparable to, if not better than, the fully-supervised model with ground truth labels, and outperforms state-of-the-art unsupervised models by a large margin.

1. Introduction

Representation learning is one of the essential research problems in machine learning and computer vision \cite{5}. Real-world sensory data such as videos, images, and audios are often in the form of high dimensions. Representation learning aims to map these data into a low-dimensional space to make it easier to extract semantically meaningful information for downstream tasks such as classification and detection. Recent years have witnessed rising interests in disentangled representation learning, which tries to separate the underlying factors of observed data variations such that each factor exclusively interprets one type of semantic attributes of sensory data. The representation of sequential data is expected to be disentangled into time-varying factors and time-invariant factors. For video data, the identity of a moving object in a video is regarded as time-invariant factors, and the motion in each frame is considered as time-varying ones \cite{36}. For speech data, the representations of the timbre of speakers and the linguistic contents are expected to be disentangled \cite{27}. There are several benefits of learning disentangled representations. First, the models that produce disentangled representations are more explainable. Second, disentangled representations make it easier and more efficient to manipulate data generation, which has potential applications in entertainment industry, training data synthesis \cite{53,54} and several downstream tasks \cite{32,19,18,43,55}.

Despite the vast amount of works \cite{24,33,6,15,16,7,30} on disentangled representations of static data (mainly image data), fewer works \cite{27,36,23,46} have explored representation disentanglement for sequential data generation. For unsupervised models, FHV AE \cite{27} and DSV AE \cite{36} elaborately designed model architectures and factorized latent variables into static and dynamic parts. These models may well handle simple data forms such as synthetic animation data but fail when dealing with realistic ones as we will show later. Besides, as pointed out in \cite{37}, unsupervised representation disentanglement is impossible without inductive biases. Without any supervision, the performance of disentanglement can hardly be guaranteed and greatly depends on the random seed and the dimensionality of latent vectors set in the models. On the other hand, several works \cite{23,46} resort to utilizing label information or attribute annotation as strong supervision for disentanglement. For instance, VideoVAE \cite{23} leveraged holistic attributes to constrain latent variables. Nevertheless, the
costly annotation of data is essential for these models and prevents them from being deployed to most real-world applications, in which a tremendous amount of unlabeled data is available.

To alleviate the drawbacks of both unsupervised and supervised models discussed above, this work tackles representation disentanglement for sequential data generation utilizing self-supervision. In self-supervised learning, various readily obtainable supervisory signals have been explored for representation learning of images and videos, employing auxiliary data such as the ambient sounds in videos [41, 3], the egomotion of cameras [1, 29], the geometry cue in 3D movies [20], and off-the-shelf functional models for visual tracking [51], and optical flow [42, 50]. However, how self-supervised learning benefits representation disentanglement of sequential data has barely been explored.

In this paper, we propose a sequential variational autoencoder (VAE), a recurrent version of VAE, for sequence generation. In the latent space, the representation is disentangled into time-invariant and time-varying factors. We address the representation disentanglement by exploring intrinsic supervision signals, which can be readily obtained from both data itself and off-the-shelf methods, and accordingly design a series of auxiliary tasks. Specifically, on one hand, to exclude dynamic information from time-invariant variable, we exploit the temporal order of the sequential data and expect the time-invariant variable of the temporally shuffled data to be close to if not the same as that of the original data. On the other hand, the time-varying variable is expected to contain dynamic information in different modalities. For video data, we allow it to predict the location of the largest motion in every frame, which can be readily inferred from optical flow. For audio data, the volume in each segment as an intrinsic label is leveraged as the supervisory signal. To further encourage the representation disentanglement, the mutual information between static and dynamic variables are minimized as an extra regularization.

To the best of our knowledge, this paper is the first work to explicitly use auxiliary supervision to improve the representation disentanglement for sequential data. Extensive experiments on representation disentanglement and sequence data generation demonstrate that, with these multiple freely accessible supervisions, our model dramatically outperforms unsupervised learning-based methods and even performs better than fully-supervised learning-based methods in several cases.

2. Related Work

Disentangled Sequential Data Generation With the success of deep generative models, recent works [24, 33, 6, 7, 30] resort to variational autoencoders (VAEs) [34] and generative adversarial networks (GANs) [22] to learn a dis-entangled representation. Regularizations are accordingly designed. $\beta$-VAE [24] imposed a heavier penalty on the KL divergence term for a better disentanglement learning. Follow-up researches [33, 6] derived a Total Correlation (TC) from the KL term, and highlights this TC term as the key factor in disentangled representation learning. In InfoGAN [7], the disentanglement of a latent code $c$ is achieved by maximizing a mutual information lower-bound between $c$ and the generated sample $\hat{x}$.

Several works involving disentangled representation have been proposed for video prediction. Villegas et al. [48] and Denton et al. [12] designed two networks to encode pose and content separately at each timestep. Unlike video prediction, video generation from priors, which we perform in this work, is a much harder task since no frame is available for appearance and motion modeling in the generation phase.

To handle video generation, VAEs are extended to a recurrent version [17, 4, 10]. However, these models do not explicitly consider static and dynamic representation disentanglement and fail to perform manipulable data generation. More recently, several works have proposed VAEs with factorized latent variables. FHVAE [27] presented a factorized hierarchical graphical model that imposes sequence-dependent priors and sequence-independent priors to different sets of latent variables in the context of speech data, but did not take advantage of the sequential prior. Combining the merits of recurrent VAE and FHVAE, DSVAE [36] is capable of disentangling latent factors by factorizing them into time-invariant and time-dependent parts and applies an LSTM sequential prior to keep a better sequential consistency for sequence generation. Although with elaborately designed complex architectures, these models may only perform decently on representation disentanglement of simple data, the disentanglement performance degrades rapidly when the complexity of data increases. In contrast, our work explores both model and regularization designs for representation disentanglement and sequential data generation. Our model fully factorizes the latent variables to time-invariant and time-varying parts, and both the posterior and the prior of the time-varying variable are modeled by LSTM for dynamic consistency. The auxiliary tasks with readily accessible supervisory signals are designed to regularize and encourage representation disentanglement.

Self-Supervised Learning The concept of self-supervised learning traces back to the autoencoder [25], which uses the input itself as supervision to learn the representation. Denoising autoencoder [49] makes the learned representations robust to noise and partial corruption of the input pattern by adding noise to the input. Recent years have witnessed the booming interest in self-supervised learning. The sources of supervisory signals can be roughly categorized into three classes. (a) Intrinsic labels: Doersch et al. [13] explored
the use of spatial context in images, and Noroozi et al. [40] trained a model to solve Jigsaw puzzles as a pretext task. Several works [52, 35] showed that colorizing a gray-scale photograph can be utilized as a powerful pretext task for visual understanding. Temporal information of video is another readily accessible supervisory signal. [38] trained a model to determine whether a sequence of frames from a video is in the correct temporal order and [31] made the model learn to arrange the permuted 3D spatiotemporal crops. (b) Auxiliary data: Agrawal et al. [1] and Jayaraman et al. [29] exploited the freely available knowledge of camera motion as a supervisory signal for feature learning. Ambient sounds in videos [41, 3] are used as a supervisory signal for learning visual models. The geometry cue in 3D movies [20] is utilized for visual representation learning. (c) Off-the-shelf tools: Wang et al. [51] leveraged the visual consistency of objects from a visual tracker in the video clips. [42] used segments obtained by motion-based segmentation based on optical flow as pseudo ground truth for the single-frame object segmentation. Instead of learning visual features as in aforementioned methods, this work aims to achieve static and dynamic representation disentanglement for sequential data such as video and speech. To this end, we leverage supervisory signals from intrinsic labels to regularize the static representation and off-the-shelf tools to regularize the dynamic representation.

3. Sequential VAE Model

We start by introducing some notations and the problem definition. \( \mathcal{D} = \{ \mathbf{X}^t \}_{t=1}^T \) is given as a dataset that consists of \( M \) i.i.d. sequences, where \( \mathbf{X} \equiv \mathbf{x}_{1:T} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) \) denotes a sequence of \( T \) observed variables, such as a video of \( T \) frames or an audio of \( T \) segments. We propose a sequential variational autoencoder model, where the sequence is assumed to be generated from latent variable \( z \) and \( z \) is factorized into two disentangled variables: the time-invariant (or static) variable \( z_f \) and the time-varying (or dynamic) variables \( z_{1:T} \).

**Priors** The prior of \( z_f \) is defined as a standard Gaussian distribution: \( z_f \sim N(0, 1) \). The time-varying latent variables \( z_{1:T} \) follow a sequential prior \( z_t \mid z_{<t} \sim N(\mu_t, \text{diag}(\sigma_t^2)) \), where \( \mu_t, \sigma_t \) are the parameters of the prior distribution conditioned on all previous time-varying latent variables \( z_{<t} \). The model \( \phi_{prior}^r \) can be parameterized as a recurrent network, such as LSTM [26] or GRU [9], where the hidden state is updated temporally. The prior of \( z \) can be factorized as:

\[
p(z) = p(z_f)p(z_{1:T}) = p(z_f) \prod_{t=1}^{T} p(z_t | z_{<t}).
\]

**Generation** The generating distribution of time step \( t \) is conditioned on \( z_f \) and \( z_t \): \( \mathbf{x}_t \mid z_f, z_t \sim N(\mu_{x,t}, \text{diag}(\sigma_{x,t}^2)) \), where \( \mu_{x,t}, \sigma_{x,t} \) can be a highly flexible function such as a deconvolutional neural network [39]. The complete generative model can be formalized by the factorization:

\[
p(x_{1:T}, z_{1:T}, z_f) = p(z_f) \prod_{t=1}^{T} p(x_t | z_f, z_t) p(z_t | z_{<t})
\]

**Inference** Our sequential VAE uses variational inference to approximate posterior distributions:

\[
z_f \sim N(\mu_f, \text{diag}(\sigma_f^2)), \quad z_t \sim N(\mu_t, \text{diag}(\sigma_t^2)),
\]

where \( \mu_f, \sigma_f = \psi_f^{Encoder}(\mathbf{x}_{1:T}) \) and \( \mu_t, \sigma_t = \psi_t^{Encoder}(\mathbf{x}_{<t}) \). The static variable \( \psi_f^{Encoder} \) is conditioned on the whole sequence while the dynamic variable is inferred by a recurrent encoder \( \psi_t^{Encoder} \) and only conditioned on the previous frames. Our inference model is factorized as:

\[
q(z_{1:T}, z_f \mid \mathbf{x}_{1:T}) = q(z_f \mid \mathbf{x}_{1:T}) \prod_{t=1}^{T} q(z_t \mid z_{<t})
\]
Learning  The objective function of sequential VAE is a timestep-wise negative variational lower bound:

\[
\mathcal{L}_{VAE} = \mathbb{E}_{q(z_{1:T}, z_f|x_{1:T})}[-\sum_{t=1}^{T} \log p(x_t|z_f, z_t)] + \\
\text{KL}(q(z_f|x_{1:T})||p(z_f)) + \sum_{t=1}^{T} \text{KL}(q(z_t|x_{<t})||p(z_t|z_{<t}))
\]

(5)

The schematic representation of our model is shown in Figure 2. Note that DSVAE also proposes a sequential VAE with disentangled representation, but it either independently infers \(z_t\) only based on the frame of each time-step without considering the continuity of dynamic variables and thus may generate inconsistent motion, or assumes the variational posterior of \(z_t\) depends on \(z_f\), implying that the variables are still implicitly entangled. In contrast, we model both the prior and the posterior of \(z_t\) by recurrent models independently, resulting in consistent dynamic information in synthetic sequences, and ensure full disentanglement of \(z_f\) and \(z_t\) by posterior factorization.

4. Self-Supervision and Regularization

Without any supervision, there is no guarantee that the time-invariant representation \(z_f\) and the time-varying representation \(z_t\) are disentangled. In this section, we introduce a series of auxiliary tasks on the different types of representations as the regularization of our sequential VAE to achieve the disentanglement, where readily accessible supervisory signals are leveraged.

4.1. Static Consistency Constraint

To encourage the time-invariant representation \(z_f\) to exclude any dynamic information, we expect that \(z_f\) changes little when varying dynamic information dramatically. To this end, we shuffle the temporal order of frames to form a shuffled sequence. Ideally, the static factors of the original sequence and shuffled sequence should be very close, if not equal, to one another. However, directly minimizing the distance of these two static factors will lead to very trivial solutions, e.g., the static factors of all sequences converge to the same value and do not contain any meaningful information. Thus, we randomly sample another sequence as the negative sample of the static factor. With a triple of static factors, we introduce a triplet loss as follows:

\[
\mathcal{L}_{SCC} = \max \left(D(z_f, z_f^{pos}) - D(z_f, z_f^{neg}) + m, 0 \right).
\]

(6)

where \(z_f\), \(z_f^{pos}\) and \(z_f^{neg}\) are the static factors of the anchor sequence, the shuffled sequence as the positive data, and another randomly sampled video as the negative data.

Figure 3. The pseudo label generation for video datasets. (a) The left image is the input frame and the right image is the corresponding optical flow map that is split by a grid. (b) Three distances are used as the dynamic signals for the face dataset.

\(D(\cdot, \cdot)\) denotes the Euclidean distance and \(m\) is the margin, set to 1. This regularization makes \(z_f\) preserve meaningful static information to a certain degree while excluding dynamic information.

4.2. Dynamic Factor Prediction

To encourage the dynamic representation \(z_t\) to carry adequate and correct time-dependent information of each time-step, we exploit dynamic information-related signals from off-the-shelf tools for different types of sequential data and accordingly design the auxiliary tasks as the regularization imposed on \(z_t\). We have the loss \(L_{DFP} = \mathcal{L}(\phi_{a}(z_t), y)\), where \(\mathcal{L}(\cdot, \cdot)\) can be either cross-entropy loss or mean squared error loss according to the designed auxiliary task, \(\phi_{a}(\cdot)\) is a network for dynamic factor prediction and \(y\) contains supervisory signals.

Video Data  The dynamic representation \(z_t\) can be learned by forcing it to predict the dynamic factors of videos. Motivated by this, we expect the location of the largest motion regions can be accurately predicted based on \(z_t\). To this end, the optical flow maps of video are first obtained by commonly used functional model FlowNet2 [28] and then split into patches by grid, as shown in Figure 3.a. We compute the average of motion magnitudes for every patch and use the indices of patches with the top-k largest values as the pseudo label for prediction. For this task, \(\phi_{a}(\cdot)\) is implemented with two fully-connected layers and a softmax layer.

Apart from the optical flow, some freely obtainable data-specific signals can be exploited. For a human face dataset, the landmark of each frame can be readily detected and considered as a supervision for dynamic factors. We obtain the landmark from an off-the-shelf landmark detector [14]. To keep our model concise and efficient, we only leverage the distance between the upper and lower eyelids as well as the distance between the upper and lower lips in each frame as the dynamic signal, as shown in Figure 3.b. Here, \(\phi_{a}(\cdot)\) consists of two fully-connected layers to regress the three distances. We observe that our model can easily capture dynamic motions under this simple supervision.
Audio Data For the audio dataset, we consider the volume as time-dependency factor and accordingly design an auxiliary task, where $z_t$ is forced to predict if the speech is silent or not in each segment. The pseudo label is readily obtained by setting a magnitude threshold on the volume of each speech segment. $\phi_a(\cdot)$ consists of two fully-connected layers and performs a binary classification.

4.3. Mutual Information Regularization

Forcing the time-varying variable $z_t$ to predict dynamic factors can guarantee that $z_t$ contains adequate dynamic information, but this fails to guarantee that $z_t$ excludes the static information. Therefore, we introduce the mutual information between static and dynamic variables as a regulator $\mathcal{L}_{MI}$. The mutual information is a measure of the mutual dependence between two variables. By minimizing $\mathcal{L}_{MI}$, we encourage the information in these two variables are mutually exclusive. The mutual information is formally defined as the KL divergence of the joint distribution to the product of marginal distribution of each variable. We have

$$\mathcal{L}_{MI}(z_f, z_{1:T}) = \sum_{t=1}^{T} \mathbb{KL}(q(z_f | z_t) \| q(z_f)q(z_t)) \tag{7}$$

where $H(\cdot) = -\mathbb{E}_{q(\cdot)}[\log q(\cdot)] = -\mathbb{E}_{q(z_f, z_t)}[\log q(\cdot)]$. The expectation can be estimated by the mini-batch weighted sampling estimator [6],

$$\mathbb{E}_{q(z)}[\log q(z)] \approx \frac{1}{M} \sum_{i=1}^{M} \left[ \log \sum_{j=1}^{M} q(z_a | x_i) - \log(\text{NM}) \right], \tag{8}$$

for $z_a = z_f, z_t$ or $(z_f, z_t)$, where $N$ and $M$ are the data size and the minibatch size, respectively.

4.4. Objective Function

Overall, our objective can be considered as the sequential VAE loss with a series of self-supervision and regularization:

$$\mathcal{L} = \mathcal{L}_{VAE} + \lambda_1 \mathcal{L}_{SCC} + \lambda_2 \mathcal{L}_{DFP} + \lambda_3 \mathcal{L}_{MI}, \tag{9}$$

where $\lambda_1, \lambda_2$ and $\lambda_3$ are balancing factors.

5. Experiments

To comprehensively validate the effectiveness of our proposed model with self-supervision, we conduct experiments on three video datasets and one audio dataset. With these four datasets, we cover different modalities from video to audio. In video domain, a large range of motions are covered from large character motions (e.g., walking, stretching) to subtle facial expressions (e.g., smiling, disgust).

5.1. Experiments on Video Data

We present an in-depth evaluation on two problems, tested on three different datasets and employing a large variety of metrics.

5.1.1 Datasets

Stochastic Moving MNIST is introduced by [11] and consists of sequences of 15 frames of size 64 × 64, where two digits from MNIST dataset move in random directions. We randomly generate 6000 sequences, 5000 of which are used for training and the rest are for testing. Sprite [36] contains sequences of animated cartoon characters with 9 action categories: walking, casting spells and slashing with three viewing angles. The appearance of characters are fully controlled by four attributes, i.e., the color of skin, tops, pants, and hair. Each of the attributes categories contains 6 possible variants, therefore it results in totally $6^9 = 1296$ unique characters, 1000 of which are used for training and the rest for testing. Each sequence contains 8 frames of size 64 × 64.

MUG Facial Expression [2] consists of 3528 videos with 52 actors performing 6 different facial expressions: anger, fear, disgust, happiness, sadness, and surprise. Each video comprises of 50 to 160 frames. As suggested in MocoGAN [47], we crop the face regions, resize video to 64×64, and randomly sample a clip of 15 frames in each video. The 75% of dataset is used for training and the rest for testing.

5.1.2 Representation Swapping

We first perform the representation swapping and compare our method with DSVAE, a disentangled VAE model, as well as MonkeyNet [45], a state-of-the-art deformable video generation model. Suppose two real videos are given for motion information and appearance information, denoted as $V_m$ and $V_a$. Our method and DSVAE perform video generation based on the $z_f$ from $V_a$ and $z_{1:T}$ from $V_m$. For MonkeyNet, the videos are generated by deforming the first frame of $V_a$ based on motion in $V_m$. The synthetic videos are expected to preserve the appearance in $V_a$ and the motion in $V_m$. The qualitative comparisons on three datasets are shown in Figure 4.

For SMMNIST, the generated videos of our model can preserve the identity of digits while consistently mimic the motion of the provided video. However, DSVAE can hardly preserve the identity. For instance, it mistakenly changes the digit “9” to “6”. We observe that MonkeyNet can hardly handle the case with multiple objects like SMMNIST, because the case does not meet its implicit assumption of only one object moving in the video.

For Sprite, DSVAE generates blurry videos when the characters in $V_a$ and $V_m$ have opposite directions, indicat-
expressions on different faces in generated videos. For MUG, we evaluate the ability of a model to preserve the appearance characteristic, indicating that the characters in the generated videos by swapping \( z_f \) and \( z_{1:T} \), from our model and two competing models: DSVAE and MonkeyNet. Each column is supposed to have the same motion.

For MUG, the generated video of DSVAE can hardly preserve both the appearance in \( V_a \) and the facial expression in \( V_m \). For example, the person in the right has a mixed appearance characteristic, indicating \( z_f \) and \( z_t \) are entangled. Due to the deformation scheme of generation, MonkeyNet fails to handle the case where the faces in two videos are not well aligned. For instance, forcing the man with a smile to be fear results in unnatural expression. On the contrary, our model disentangles \( z_f \) and \( z_t \), supported by the realistic expressions on different faces in generated videos.

\[ V \]

\[ \text{Video A} \]

\begin{tabular}{cccccccc}
Real & Real & Real & Real & Real & Real & Real & Real \\
Ours & Ours & Ours & Ours & Ours & Ours & Ours & Ours \\
DSVAE & DSVAE & DSVAE & DSVAE & DSVAE & DSVAE & DSVAE & DSVAE \\
Monkey net & Monkey net & Monkey net & Monkey net & Monkey net & Monkey net & Monkey net & Monkey net \\
\end{tabular}

\[ \text{Video B} \]

\[ \]

Figure 4. Representation swapping on SMMNIST, Sprite and MUG datasets. In each panel, we show two real videos as well as the generated videos by swapping \( z_f \) and \( z_{1:T} \). from our model and two competing models: DSVAE and MonkeyNet. Each column is supposed to have the same motion.

\[ \]

\[ \text{Figure 5. Unconditional video generation on MUG. The upper and lower panels show the qualitative results of our model and MoCoGAN, respectively.} \]

\[ \]

5.1.3 Video Generation

Quantitative Results We compute the quantitative metrics of our model with and without self-supervision and regularization, denoted as full model and baseline, as well as two competing methods: DSVAE and MoCoGAN. All these methods are comparable as no ground truth labels are used to benefit representation disentanglement. Besides, the results of our baseline with full supervision from human-annotation baseline-sv are also provided as a reference.

To demonstrate the ability of a model on the representation disentanglement, we use the classification accuracy \( Acc \) [36], which measures the ability of a model to preserve an attribute when generating a video given the corresponding representation or label. To measure how diverse and realistic videos a model can generate, three metrics are used: \( IS \) [44], Intra-Entropy \( H(y|x) \) [23] and Inter-Entropy \( H(y) \) [23]. All metrics utilize a pretrained classifier based on the real videos and ground truth attributes. See Appendix for the detailed definitions of the metrics. The results are shown in Table 1.

For representation disentanglement, we consider generating videos with a given \( z_f \) inferred from a real video and randomly sampled \( z_{1:T} \) from the prior \( p(z_{1:T}) \) for SMMNIST. We then check if the synthetic video contains the same digits as the real video by the pretrained classifier. For MUG, we evaluate the ability of a model to preserve the facial expression by fixing \( z_{1:T} \) and randomly sampled
Table 1. Quantitatively performance comparison on SMMNIST, Sprite and MUG datasets. High values are expected for Acc, $H(y)$ and $IS$, while for $H(y|x)$, the lower values are better. The results of our model with supervision of ground truth labels baseline-sv* are shown as a reference.

<table>
<thead>
<tr>
<th>Methods</th>
<th>SMMNIST</th>
<th></th>
<th></th>
<th>Sprite</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>MUG</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc</td>
<td>IS</td>
<td>$H(y</td>
<td>x)$</td>
<td>$H(y)$</td>
<td>Acc</td>
<td>IS</td>
<td>$H(y</td>
<td>x)$</td>
<td>$H(y)$</td>
<td>Acc</td>
</tr>
<tr>
<td>MoCoGAN</td>
<td>74.55%</td>
<td>4.078</td>
<td>0.194</td>
<td>0.191</td>
<td>92.89%</td>
<td>8.461</td>
<td>0.029</td>
<td>2.011</td>
<td>70.51%</td>
<td>5.136</td>
<td>0.135</td>
</tr>
<tr>
<td>DSVAE</td>
<td>88.19%</td>
<td>6.210</td>
<td>0.167</td>
<td>2.052</td>
<td>94.12%</td>
<td>8.312</td>
<td>0.071</td>
<td>2.190</td>
<td>90.12%</td>
<td>6.543</td>
<td>0.167</td>
</tr>
<tr>
<td>baseline</td>
<td>90.12%</td>
<td>6.543</td>
<td>0.167</td>
<td>2.052</td>
<td>94.12%</td>
<td>8.312</td>
<td>0.071</td>
<td>2.190</td>
<td>90.12%</td>
<td>6.543</td>
<td>0.167</td>
</tr>
<tr>
<td>full model</td>
<td>95.09%</td>
<td>7.072</td>
<td>0.150</td>
<td>2.106</td>
<td>99.49%</td>
<td>8.637</td>
<td>0.041</td>
<td>2.197</td>
<td>70.51%</td>
<td>5.136</td>
<td>0.135</td>
</tr>
<tr>
<td>baseline-sv*</td>
<td>92.18%</td>
<td>6.845</td>
<td>0.156</td>
<td>2.057</td>
<td>98.91%</td>
<td>8.741</td>
<td>0.028</td>
<td>2.196</td>
<td>72.32%</td>
<td>5.006</td>
<td>0.129</td>
</tr>
</tbody>
</table>

Figure 7. Controlled video generation. (a) Video generation controlled by fixing the static variable and randomly sampling dynamic variables from the prior $p(z_f | T)$. All sequences share the same identity but with different motions for each dataset. (b) Video generation with changed facial expressions. Expression is changed from smile to surprise and from surprise to disgust in two sequences, respectively. We control it by transferring the dynamic variables smoothly.

$z_f$ from the prior $p(z_f)$. For Sprite, since the ground truth of both actions and appearance attributes are available, we evaluate the ability of preserving both static and dynamic representations, and report the average scores. It’s evident that our full model consistently outperforms all competing methods. For SMNIST, we observe that MoCoGAN have poor ability to correctly generate the digits with given labels while our full model can generate correctly digits, reflected by the high Acc. Note that our full model achieves 99.49% Acc on Sprite, indicating the $z_f$ and $z_t$ are greatly disentangled. Besides, full model significantly boosts the performance of baseline, especially in MUG where more realistic data is contained, the performance gets giant boost from 53.83% to 70.51%, which illustrates the crucial role of our self-supervision and regularization. For video generation, full model consistently shows the superior performances on IS, $H(y|x)$, $H(y)$. Especially in MUG, full model outperforms the runner-up MoCoGAN by 18.6% on IS, demonstrating that high quality of videos generated by our model. Note that our baseline is also compared favorably to DSVAE, illustrating the superiority of the designed sequential VAE model.

It is worth noting that our model with self-supervision full model outperforms baseline-sv in SMNIST on the representation disentanglement. The possible reason is with the ground truth labels, baseline-sv only encourages $z_f$ to contain identity information but fails to exclude the information in $z_t$, resulting in confusion in digit recognition when using various dynamic variables. For MUG, baseline-sv performs better on preserving motion. We conjecture it’s because that the strong supervision of expression labels forces $z_t$ to encode the dynamic information, and $z_f$ does not favor encoding temporal dynamic information and thus varying $z_f$ does not affect the motion much.

Qualitative results We first demonstrate the ability of our model to manipulate video generation in Figure 7. By fixing $z_f$ and sampling $z_t$, our model can generate videos with the same object that performs various motions as shown in Figure 7.a. Even in one video, the facial expression can be transferred by controlling $z_t$, as shown in Figure 7.b.

We also evaluate the appearance diversity of generate objects from our model, Figure 6 shows the frames our model generates with sampled $z_f$. The objects with realistic and diverse appearances validate our model’s outstanding capability of high-quality video generation.

Besides, we compare our model with MoCoGAN on unconditional video generation on MUG dataset, as shown in Figure 5. The videos are generated with sampled $z_f$ and $z_t$. MoCoGAN generates videos with many artifacts, such as unrealistic eyes and inconsistent mouth in the third video. Conversely, our model generates more realistic human faces with consistent high-coherence expressions.
Table 2. Ablation study of disentanglement on MUG dataset.

| Methods     | Acc  | TS  | \(H(y|v)\) | \(H(y)\) |
|-------------|------|-----|-------------|----------|
| No \(\mathcal{L}_{SCC}\) | 61.45% | 4.850 | 0.201 | 1.734 |
| No \(\mathcal{L}_{DFP}\) | 58.32% | 4.423 | 0.284 | 1.721 |
| No \(\mathcal{L}_{MI}\) | 66.07% | 4.874 | 0.175 | 1.749 |
| Full model  | 70.51% | 5.136 | 0.135 | 1.760 |

5.1.4 Ablation Studies

In this section, we present an ablation study to empirically measure the impact of each regularization of our model on its performance. The variant without a certain regularization is denoted as \(\text{No} \ \mathcal{L}_X\). In Table 2, we report the quantitative evaluation. We note that \(\text{No} \ \mathcal{L}_{SSC}\) performs worse than the full model. This illustrates the significance of the static consistency constraint to make \(z_f\) to be disentangled from \(z_t\). \(\text{No} \ \mathcal{L}_{DFP}\) degrades the performance on \(\text{Acc}\) considerably, indicating that \(\mathcal{L}_{DFP}\) as regularization is crucial to preserve the action information in the dynamic vector. Besides, after removing the mutual information regularization, \(\text{No} \ \mathcal{L}_{MI}\) again shows an inferior performance to the full model. A possible explanation is that our \(\mathcal{L}_{MI}\) encourages that \(z_t\) excludes the appearance information; thus the appearance information is only from \(z_f\). The qualitative results shown in Figure 8 confirms this analysis. We generate a video with \(z_f\) of the woman and \(z_t\) of the man in the first row by different variants of our model. Without \(\mathcal{L}_{MI}\), some characteristics of the man are still preserved, such as the beard, confirming that the appearance information partially remains in \(z_t\). In the results of \(\text{No} \ \mathcal{L}_{SSC}\), the beard is more evident, indicating that the static and dynamic variable are still entangled. On the other hand, without \(\mathcal{L}_{DPF}\), the woman in the generated video cannot mimic the action of the man well, which indicates the dynamic variable does not encode the motion information properly. Finally, the person in the generated video of baseline neither preserves the appearance of the woman nor follows the expression of the man. It illustrates the representation disentanglement without any supervision remains a hard task.

5.2. Experiments on Audio Data

To demonstrate the general applicability of our model on sequential data, we conduct experiments on audio data, where the time-invariant and time-varying factors are the timbre of a speaker and the linguistic content of a speech, respectively. The dataset we use is TIMIT, which is a corpus of phonemically and lexically transcribed speech of American English speakers of different sexes and dialects \([21]\), and contains 63000 recordings of read speeches. We split the dataset to training and testing subsets with a ratio of 5:1. As in \([27]\), all the speech are presented as a sequence of 80 dimensional Mel-scale filter bank features.

We quantitatively compare our model with FHVAE and DSVAE on the speaker verification task based on either \(z_f\) or \(z_{1:T}\), measured by the Equal Error Rate (EER) \([8]\). Note that we expect the speaker can be correctly verified with \(z_f\) as it encodes the timbre of speakers, and randomly guess with \(z_{1:T}\) as it ideally only encodes the linguistic content. The results are shown in Table 3. Our model outperforms competing methods in both cases. Especially when based on \(z_{1:T}\), our model doubles the score of the baseline, indicating our model significantly eliminate the timbre information in \(z_{1:T}\).

Table 3. Performance comparison on speaker verification. Small errors are better for \(z_f\) and large errors are expected for \(z_{1:T}\).

<table>
<thead>
<tr>
<th>model</th>
<th>feature</th>
<th>dim</th>
<th>EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>FHVAE</td>
<td>(z_f)</td>
<td>16</td>
<td>5.06%</td>
</tr>
<tr>
<td>DSVAE</td>
<td>(z_f)</td>
<td>64</td>
<td>4.82%</td>
</tr>
<tr>
<td></td>
<td>(z_{1:T})</td>
<td>64</td>
<td>18.89%</td>
</tr>
<tr>
<td>Ours</td>
<td>(z_f)</td>
<td>64</td>
<td>4.80%</td>
</tr>
<tr>
<td></td>
<td>(z_{1:T})</td>
<td>64</td>
<td>40.12%</td>
</tr>
</tbody>
</table>

6. Conclusion

We propose a self-supervised sequential VAE, which learns disentangled time-invariant and time-varying representations for sequential data. We show that, with readily accessible supervisory signals from data itself and off-the-shelf tools, our model can achieve comparable performance to the fully supervised models that require costly human annotations. The disentangling ability of our model is qualitatively and quantitatively verified on four datasets across video and audio domains. The appealing results on a variety of tasks illustrate that, leveraging self-supervision is a promising direction for representation disentanglement and sequential data generation. In the future, we plan to extend our model to high-resolution video generation, video prediction and image-to-video generation.
References

[19] Deng-Ping Fan, Wenguang Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In IEEE CVPR, 2019. 1
[27] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In Advances in neural information processing systems, pages 1878–1889, 2017. 1, 2, 8


[37] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In ICML, 2019. 1


