IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received June 30, 2020, accepted July 29, 2020, date of publication August 13, 2020, date of current version August 25, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3016313

Generating In-Between Images Through Learned Latent Space Representation Using Variational Autoencoders

PAULINO CRISTOVAO[1,2], **(Graduate Student Member, IEEE), HIDEMOTO NAKADA**[1,2], **YUSUKE TANIMURA**[1,2], **AND HIDEKI ASOH**[2]

[1]Department of Compute Science, University of Tsukuba, Tsukuba 305-8577, Japan
[2]National Institute of Advanced Industrial Science and Technology, Tsukuba 305-8560, Japan

Corresponding author: Paulino Cristovao (cristovao-paulino@aist.go.jp)

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant 19K11994.

ABSTRACT Image interpolation is often implemented using one of two methods: optical flow or convolutional neural networks. These methods are typically pixel-based; they do not work well on objects between images far apart. Because they either rely on a simple frame average or pixel motion, they do not have the required knowledge of the semantic structure of the data. In this paper, we propose a method for image interpolation based on latent representations. We use a simple network structure based on a variational autoencoder and an adjustable hyperparameter that imposes the latent space distribution to generate accurate interpolation. To visualize the effects of the proposed approach, we evaluate a synthetic dataset. We demonstrate that our method outperforms both pixel-based methods and a conventional variational autoencoder, with particular improvements in nonsuccessive images.

INDEX TERMS Image interpolation, latent variables, representation learning, variational autoencoder.

I. INTRODUCTION

The process of generating in-between images from a sequence of images is known as image interpolation. Image interpolation reveals the dynamics of objects in a scene by relating spatial features (i.e., distinct viewpoints) to temporal changes (i.e., different timestamps) [1]. Image interpolation methods are used in a wide variety of computer vision applications, including the movie and animation industry. It aims to enhance the quality of images displayed in different scenarios. In the digital and movie industry, original videos often have a high frame rate. Because of the limitations on network bandwidth, the rate has to decrease before transmission. This reduction is often made by skipping some frames [2]. Here, image interpolation can help restore clarity to the image.

Some of the challenges in image interpolation occur when the variations in pixel values are significant, i.e., objects in the images vary considerably, overlapping objects, occlusions, missing objects, and noise. Optical flow [3], [4] and convolutional neural networks (CNNs) [5]–[8] are two common approaches for image interpolation based on pixel motion. The former considers the pixel motion of the objects and performs a simple pixel average, and the latter learns optical flow feature representations by convolving input images with spatially adaptive kernels that account for pixel motion [9]. In both of these approaches, a pixel-based method algorithm is used to generate image interpolation of arbitrary sequences. However, when objects in input images are far apart, this may cause problems, given that temporal dependence between objects may be lost. The resulting generated images may appear with holes, overlapping objects, and ghost artifacts.

In this paper, we propose a novel method for the problem of image interpolation based on latent variables. Our model learns to encode the spatial and temporal structure of the image based on latent representations (inherent action) and not image context (pixel motion). The model then generates the in-between image based on the learned representations. Because the model relies on stochastic latent representations of the data, it is not very straightforward to assess whether the generated structure is accurate. To address this limitation, we introduce a loss function that constrains the latent space information capacity.

We investigate image interpolation using a variational autoencoder (VAE) because it offers stability during training,

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

the ability to provide meaningful representations, and the latent space allows semantic operations with vector space arithmetic [10]. We found that by limiting the latent space, i.e., putting pressure on Kullback–Leibler's divergence term by adding an adjustable hyperparameter alpha (α), the model generates accurate semantics of the in-between image.

In summary, we make the following contributions. We design a simple model that relies on latent variables for image interpolation of nonconsecutive images. The model generalizes well to unseen objects (i.e., objects with occlusion or overlap). We reveal that constraining latent representations can lead to interpretable data representation. Furthermore, the beneficial effects of Kullback–Leibler's divergence are denoted.

The paper is organized as follows. In section II, we review the several approaches for image interpolation related to ours for image interpolation. In section III, we demonstrate our proposed model, particularly how it outperforms the conventional VAE. In section IV, we show the evaluation on different settings. In section V, we discuss the related work. Finally, section VI concludes the paper.

II. RELATED WORK

A. IMAGE INTERPOLATION METHODS

Research on image interpolation(motion flow, disparity, displacement) has a long history in computer vision. Two directions have been explored: optical flow and, more recently, convolutional neural networks (CNNs).

1) OPTICAL FLOW

Initial attempts at image interpolation were based on optical flow methods. Optical flow is used to describe the apparent shifting of pixel values in time-varying images, caused by illumination change, camera motion, or noise. Optical flow techniques compute the motion estimation vector for each pixel or group of a pixel in an image, and this involves having an initial image and at least one of its neighbors. A large part of the work on image interpolation is based on differential algorithms proposed by Lucas and Kanade [11] and Horn and Schunck [12]. These algorithms are based on several assumptions, such as brightness constancy and temporal consistency [13]. Lucas and Kanade assume that pixels surrounding a pixel being observed behave almost the same as the observed pixel (local variation), while Horn and Schunck consider the global variation in an image. This assumption means that the motion vectors of a pixel depend on the value of its neighbors. Since these algorithms are based on differential methods, to avoid aliasing caused by the significant differential between pixels, temporal smoothing between images is necessary.

To overcome the limitations of traditional methods two main directions have been explored, including motion-compensated frame interpolation (MCFI) techniques [3], [14]–[16] and phased-based methods [17], [18]. The former estimates the motion based on the previous image and current image and then generates the in-between by averaging the pixels in the images pointed by half of the obtained motion vectors. MCFI is based on assumptions that motion in images is smooth and continuous, which might work well on sequences with relatively small motions [14]; on large displacement, residual information of skipped frames is unavailable, and the generated image might include overlapped objects, holes, and blocking artifacts. In the second direction, phased-based methods assume that small motion can be encoded in the phase shift on an individual pixel's color. Meyer *et al.* [18] suggested extending flow-based methods to the path-based method; by using a path-based method, the motion accuracy was expanded, improving the range of the motion trajectory. Alternatively, Zhang *et al.* [19] extended the motion range by computing a disparity map, while Elgharib *et al.* [20] proposed combining a phased-based method with optical flow. These methods have largely improved the performance over differential algorithms but still cannot handle large displacement.

2) CONVOLUTIONAL NEURAL NETWORKS

Neural networks have achieved state-of-the-art performance in various applications. Recently, researchers have shown interest in applying CNNs for the task of image interpolation [21]–[26]. CNNs are well-known algorithms for extracting semantic knowledge from data. They learn the optical flow feature representations by convolving input images with spatially adaptive kernels that account for pixel motion. Dosovitskiy *et al.* [5] proposed two CNNs (FlowNetS and FlowNetC), which estimated the optical flow based on the U-Net denoising autoencoder [27]. The Dosovitskiy *et al.* model takes an input pair of images and outputs the flow field. The image interpolation results have significant errors in the backgrounds. Alternatively, Ilg *et al.* [6] suggested combining deep learning with domain knowledge; their model has a small network concentrating on small motion and others on large motion. Jiang *et al.* [8] extended a single image generation to multi-images. Shu *et al.* [28] trained their age progression model with paired images of the same person with different ages. Although the training dataset is similar to our approach, their goal is to train the age progression dictionary, while our interest is to have better latent representation. Interpolation tasks using neural networks have been extended to text [29] and video [8], [9], [26], [30].

Despite the excellent performance, pixel-based methods rely on pixel motion. They are limited to highly similar images. They do not perform well on objects in images that are far apart (large displacement between objects in input images). Because the input images that are far apart may lose temporal dependence between objects, they do not have the required knowledge of the semantic structure of the input images. Thus, the generated in-between image may appear with some errors, such as occlusion, overlapping, and ghost artifacts. CNN models alleviate the problems of pixel-based models to some extent. In this work, we propose a novel method for the problem of image interpolation based on latent variables.

B. VARIATIONAL AUTOENCODERS (VAE)

VAE [31], [32] has shown promising results in various tasks, including image classification [33], image segmentation [34], text generation [29], and artistic applications [35]. The model is composed of the encoder network and decoder network. The role of the encoder network is to map the input data into a latent space distribution, whereas the decoder network maps the latent space representation back to the input.

The VAE models modify the autoencoder architecture by replacing the deterministic function with a probabilistic function. The latent variable z is sampled from the mean μ and standard deviation σ from a continuous latent space to make VAEs more useful for generative modeling. The μ vector controls where the encoding of the input should be centered, while σ controls the area, i.e., how much the encoding can vary. The decoder learns the data distribution rather than a single point, and this exposes a wide range of encoding for the same input during training. VAE models enable random sampling and arithmetic operations on its latent space. Following the general formulation introduced in [31], [32], the VAE loss function (1) minimizes the lower bound on the marginal loglikelihood.

$$L(\theta, \phi) = \mathbb{E}_{z \sim q_\theta(z|x)}[logp_\phi(x|z)] - D_{KL}(q_\theta(z|x)||p(z)) \quad (1)$$

The first term represents the reconstruction error; it measures how well the latent variable describes the image, and a pixel-wise quadratic error if often chosen between the actual image and the reconstructed image. The second term represents Kullback–Leibler's divergence (D_{KL}) between the prior $p(z)$ and the approximate posterior distribution $q_\theta(z|x)$; it assesses the regularization of the latent space, and (θ, ϕ) parameterizes the distributions of the encoder and decoder. VAE aims to generate new samples that are not present in the training set.

C. IMAGE INTERPOLATION BASED ON VARIATIONAL AUTOENCODERS

To connect our work with existing approaches for learning latent representations, we provide practical analysis of conventional VAE [31], [32] and β-VAE [36]. We attempted to generate image interpolation based on latent representations. We found that the results were not very encouraging, and it did not perform well. The generated image did not resemble the structure of the in-between image. We empirically assume that the latent space does not have any constraint under its learning representations, and the generated latent variables have certain degrees of freedom. Another possible explanation is that the latent space does not have the necessary structure that enables interpolation. We then designed a network structure to enforce the latent space to have the appropriate structure. Later in this paper, we compare our model with these baseline models.

D. LATENT REPRESENTATIONS

The data that are often in high-dimensional space can be represented in a lower dimension, often referred to as latent representations. These latent representations hold relevant information of the initial data, which are highly dependent on downstream tasks [37]. However, these representations are often unstructured and hard to control or interpret [4]; without the pressure to regularize the latent space, they do not exhibit the desired structure [38]. To address this limitation, Higgins *et al.* β-VAE [36] proposed to constraining the latent space capacity, forcing the model to learn salient features of the data, which results in a more interpretable representation of the data. In this work, we demonstrate the benefits of using learned latent representations for the task of image interpolation.

III. PROPOSED MODEL

A. METHOD OVERVIEW

The success of image interpolation is restricted to pixel-based approaches. The pixel-based approach works well on consecutive homogeneous images. Because these images are highly similar, they often do not require good knowledge of the semantic structure of the objects. However, when the motion is complex, such as the case of large displacements between objects, pixel-based approaches do not perform well; to restore the in-between image, semantic information is necessary [39]. Based on this insight, we propose a new approach based on latent variables to the objects' problems in images that are far apart from each other. The proposed model benefits from the ability to constrain the freedom of the latent representation

In this section, we begin the discussion by explaining and describing the motivation of our proposed network structure. To improve the performance of the proposed model, we introduce an additional loss function that restricts the latent space to probable structures. We also provide detail of the related hyperparameter.

B. PROPOSED NETWORK STRUCTURE

1) DETAILS OF THE NETWORK STRUCTURE

Our network structure Fig.1 follows the conventional VAE structure [31], [32]. The key components are the Z', which averages the latent space of input images (first image and second image), and the α component, which weigh the importance given to the average inputs and actual in-between latent representation. The α term penalizes the network if the generated image has deviated from the actual in-between. If Z' is ignored $\alpha = 0$ (which corresponds to conventional VAE), the model is not strongly penalized in case the generated in-between does not reflect the actual in-between—giving the model the freedom to sample any possible latent point. This scenario is not ideal; we have to control the latent space if we aim to learn an interpretable representation of the data manifold for the task of inbetweening. The effects of α are further explained in this work.

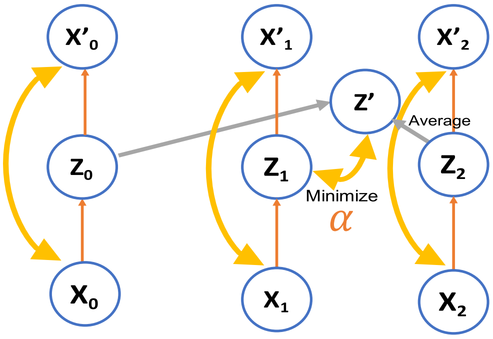


FIGURE 1. Network structure of our approach.

2) NETWORK IMPLEMENTATION

The network structure is based on three variational autoencoders, as illustrated in Fig.1. The network receives a pair of images (X_0, X_2) and actual in-between (X_1). Each network has an encoder X and decoder (X') network, and (z) corresponds to the latent space. To generate the in-between image, we average the latent representations of the adjacent networks(z_0, z_2) and the actual in-between (z_1). To reduce the model complexity, all the networks share the same weights. The weight-sharing technique is a method for building translation-invariant networks [40] and also used for multi-modal knowledge transfer [41], [42]. The encoders have 6 hierarchical layers, consisting of five convolutional layers and a fully connected layer. At each hierarchy, a pooling layer with stride two and 4 × 4 kernels, except the first layer, which has kernel size 3×3 and stride one. The decoders have 6 hierarchical layers, consisting of five deconvolutional layers and a fully connected layer. Having each stride one and kernel size 4 × 4, except for the last layer, which has kernel size five. We used AdamMax optimization with a learning rate of 0.0001, and the batch size was 100. Later, when we compared our results with FlowNet2.0 and SloMo, we increased the number of layers to 10 since we worked with images of 256 × 256 instead of 32 × 32. The learning rate was 0.005, and the batch was 30. The network was trained to capture salient features from the input data and to minimize the difference between (z') and (z_1).

C. PROPOSED LOSS FUNCTION

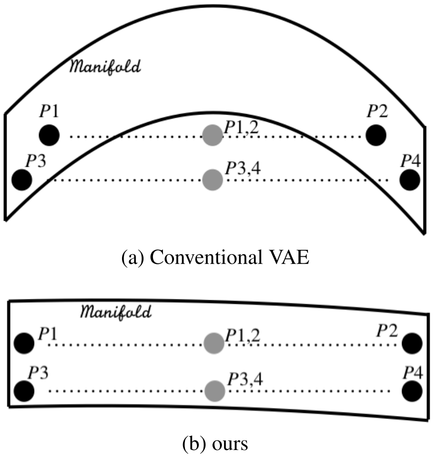
Initially, we attempted to generate image interpolation based on latent variables using conventional VAE. The generated structure of the in-between image did not resemble the actual in-between image. Because the latent representations are unstructured and lack easy understanding and controllability, the model is under no constraint to generate the structure, reflecting the actual in-between image. In addition, conventional VAE achieves limited application in tasks, such as discovering new factors of variation in the data.

In this work, we propose a loss function (2) that is a modification of the conventional VAE loss function. We also demonstrate the beneficial effects of the KL divergence term and its role in the generative process. Kim and Mnih [43] and Higgins *et al.* [36] demonstrated the beneficial effects of limiting the capacity of latent representations, this approach forces the model to learn salient features from the data. We limit the information capacity of latent space to generate the actual structure of the in-between image. We demonstrate that with the proposed loss function, the model generates the actual structure of the in-between image.

$$L(X_0, X_1, X_2) = L_{VAE}(X_0) + L_{VAE}(X_1) + L_{VAE}(X_2) + \alpha(D_{KL}(q_{(X_1)}||\frac{q_{(X_0)} + q_{(X_2)}}{2})) \qquad (2)$$

1) A LOSS FOR ENFORCING FLAT MANIFOLD

Often probabilistic models depend on the way we constraint the learning representations. In Fig.2, we show the task of interpolating between two points (P1 and P2; P3 and P4). The conventional VAE approach often generates a curved manifold, as shown in Fig.2 (top). The task becomes complex because the manifold is curved, and the generated point lies off of the data manifold (P1,2 and P3,4). Linear interpolation traverses the shortest path in terms of Euclidean distance between the two points. The generated in-between is more likely to be unrealistic. On the other hand, our loss function forces the manifold to be locally flat, as shown in the Fig.2 (bottom). Interpolation between two points on flat manifold lies on the manifold, and the generated samples from interpolated representation (such as P1,2 and P3,4) will be more plausible. Bengio *et al.* [44], Verma *et al.* [45], have explored the relationship between interpolation and flat data manifold in the context of representation learning.



(a) Conventional VAE

(b) ours

FIGURE 2. Conventional VAE is likely to interpolate in a curved manifold. Our model forces the manifold to be flat, resulting in smooth interpolation.

2) ADJUSTABLE HYPERPARAMETER ALPHA (α)

Conventional VAE ($\alpha = 0$) [31] latent information did not learn the structure of the in-between image due to a lack of constraint on the latent information bottleneck. There was no signal to the model to generate the structure of the in-between image. To learn the latent space that represents the structure of the in-between image, we hypothesize that it is relevant to tune ($\alpha > 0$). Alpha balances the relative importance given to the difference between ground truth loss and average loss. Alpha ($\alpha > 0$) places a stronger constraint on the latent bottleneck, unlike the conventional VAE. This ($\alpha > 0$) limits the capacity of latent space z, which, combined with the pressure to maximize the loglikelihood of the training data, and encourages the model to learn the most salient representations of the data [36]. Because the data are generated using some conditional independent ground truth and Kullback–Leibler's divergence term of the loss function, this encourages conditional independence, and higher values of α should promote learning.

While tuning α, two factors must be considered: the latent dimension and the complexity of the dataset.

IV. EXPERIMENTS

In this section, we present the datasets, the scenarios tested with individual results and evaluations. We also expose the effects of the hyperparameter and the gains of our proposed method.

A. DATASET AND DEGREES OF FREEDOM

For clear visualization of the intended image interpolation result, we relied on a collection of synthetic images, namely dots, face, teapot, and 2D shapes. These datasets allowed us to create and replicate various possible scenarios. Training samples were obtained, by randomly sampling 10000 triplets of nonconsecutive images (large displacement between objects in input images), with 10 to 40 degrees from one image to another, and testing random 5000 triplets with 30 to 60 degrees from one image to another. The range prevents the use of consecutive images that are visually very similar. Additionally, by randomly sampling a triplet, we hypothesize that the model does not memorize the training sequence. We do not control the angles between the first and second images. The initial samples consisted of 32×32 image size, except when comparing our approach to Super SloMo and FlowNet2.0. Here, we normalize to 256×256 image resolution. Primarily, we tested ''one degree of freedom'' where the object is rotated 360 degrees on the $x-axis$ and then on ''two degrees of freedom'' where the object rotates 360 degrees on the $x-$ and $y - axis$.

B. IN-BETWEEN IMAGE GENERATION

Our model was evaluated far apart images (large displacement between objects in input images). We initially tested image interpolation based on the latent space of conventional VAE ($\alpha = 0$). There was no constraint applied to the model learning representation. The results show that without limitation ($\alpha = 0$), the generated image interpolation did not preserve an accurate structure of the actual in-between image.

We then applied a constraint to the latent space representation by tuning an adjustable hyperparameter. If tuned ($\alpha > 0$), the model could generate an image that preserves the in-between image's structure. The results demonstrate that our proposed method outperformed conventional VAE on image interpolation. This is explained by the fact that constraining the latent space encourages the model to learn the more salient structure of the in-between image. Next, we show the interpolation results for different scenarios.

1) ONE DEGREE OF FREEDOM

We demonstrated two examples using one degree of freedom. This example represented a simple scenario, with a total of 360 possible angles. The goal was to test the structure of the in-between image (location, angle). As shown on the right side of Fig.3 and Fig.4, our proposed model preserved the structure of the in-between image in every scene illustrated in the images. This was opposed to conventional VAE, which failed to preserve the structure of the in-between image.

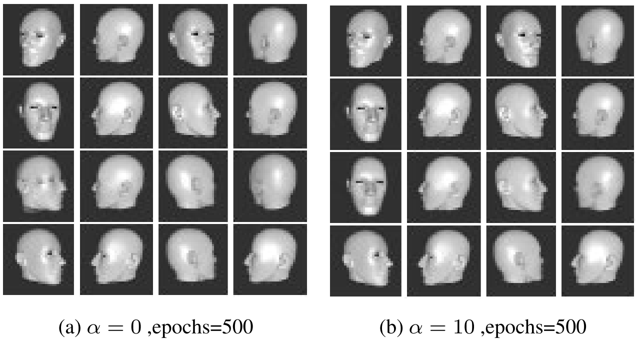

(a) $\alpha = 0$,epochs=500 (b) $\alpha = 10$,epochs=500

FIGURE 3. Face-testing: 1st row:first frame, 2nd row: ground truth, 3rd row: in-between image, 4th row: second frame. a) The conventional VAE model failed to preserve the structure of the in-between image. b) Our model generated an accurate structure of the in-between image.

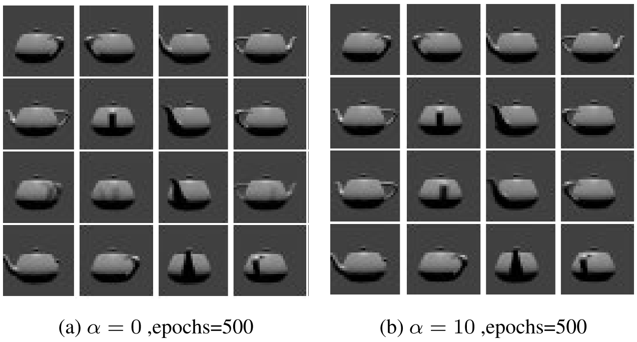

(a) $\alpha = 0$,epochs=500 (b) $\alpha = 10$,epochs=500

FIGURE 4. Teapot-testing: 1st row:first image, 2nd row: ground truth, 3rd row: in-between image, 4th row: second frame. a) The conventional VAE model failed to preserve the structure of in-between image. b) Our model generated an accurate structure of the in-between image.

2) TWO DEGREES OF FREEDOM

In the next phase of the experiment, we randomly rotated the object under the influence of two variables: ''two degrees of freedom.'' In the previous experiment (one degree of freedom), there were only 360 possible scenes, regardless of the number of samples. Working with two degrees squares the number of possible scenes. We randomly sampled the input images to ensure that the model did not see a scene twice. The results highlighted in Fig.5 demonstrate that our approach ($\alpha = 10$) preserved the structure of the in-between image, even in a complex scenario.

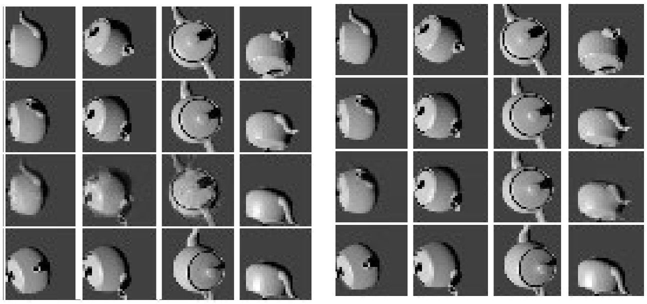

(a) $\alpha = 0$,epochs=500 (b) $\alpha = 10$,epochs=500

FIGURE 5. Teapot-testing: 1st row:first image, 2nd row: ground truth, 3rd row: in-between image, 4th row: second frame. a) Conventional VAE model failed to preserve the structure of the in-between image. b) Our model generated an accurate structure of the in-between image.

3) MOVING 2D SHAPES - MULTIPLE OBJECTS INTERPOLATION

To assess whether our model could generate interpolation in case of the presence of multiple objects in the image. We created new training data. Moving 2D shapes is a dataset containing three objects (moving randomly): a white square, a red triangle, and a blue circle. These data are similar to what we can expect in the real world, where different people and objects are moving in random directions. The model must capture the location, shape, and color of the objects. This example represented a more complex scenario since the model has to match similar shapes and colors during the interpolation. One particularity of these data is that small variation (motion) between the objects in the input image cannot be easily noticeable by human eyes. Fig.6 shows the results on both conventional VAE ($\alpha = 0$) and our proposed model ($\alpha = 100$). Conventional VAE failed to generate in-between objects. Additionally, when the objects were displayed, it did not preserve the structure of the in-between image. Despite the data complexity, our model preserves the accurate structure of the in-between image. Even when objects overlap, the model matches the shape, color, and location. We highlight the advantages of our model compared to conventional VAE, illustrated in Fig.7. Restricting the latent space information encourages the model to preserve the semantic structure of the in-between image.

C. EVALUATION

We have so far focused on demonstrating interpolation abilities; in this section, we evaluated our results.

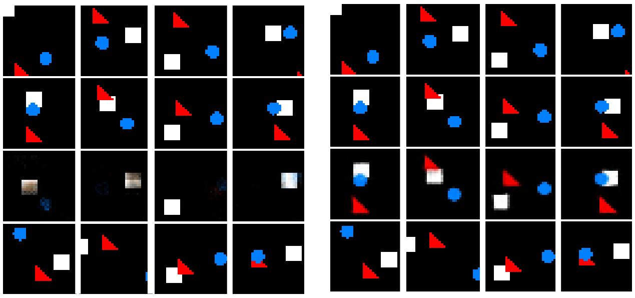

(a) $\alpha = 0$,epochs=2000 (b) $\alpha = 100$,epochs=2000

FIGURE 6. 1st row:first image, 2nd row: ground truth, 3rd row: in-between image, 4th row: second frame. a) Conventional VAE failed to preserve the spatial location of the objects. b) Our model preserved the structure of the in-between image.

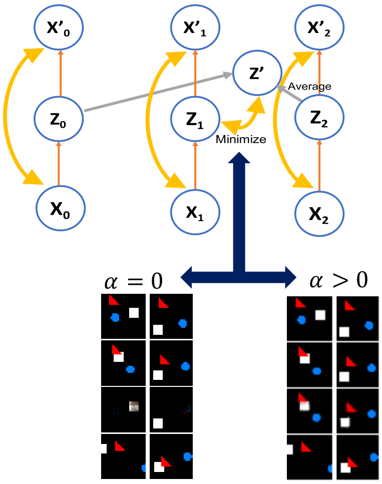
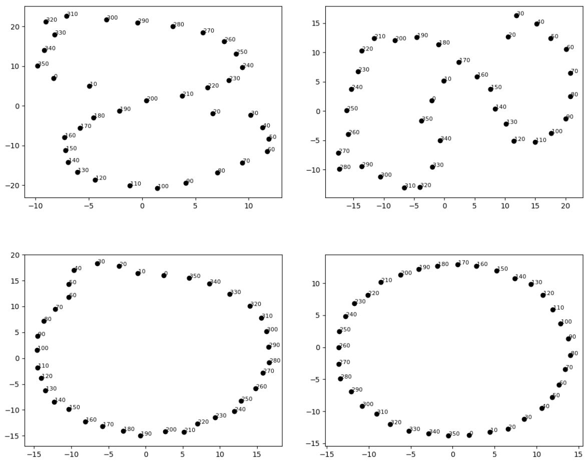


FIGURE 7. The effects α on the in-between image: For $\alpha = 0$: The generated in-between image lacked the structure (location of the moving object) of the in-between image. For $\alpha > 0$: The contextual structure of the in-between image was preserved.

1) QUALITATIVE EVALUATION OF LEARNED REPRESENTATION

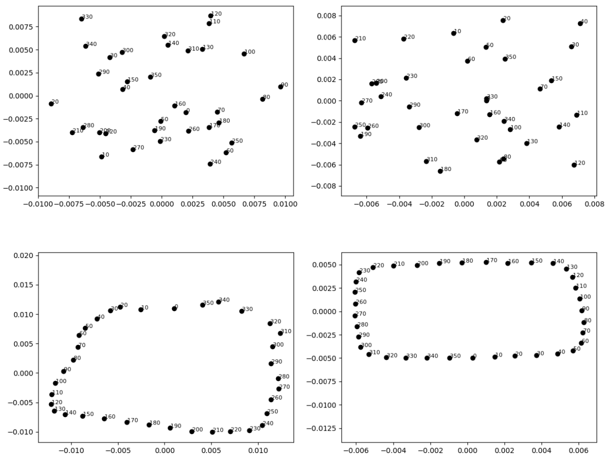
We evaluated the embedded structure of learned representations using two conventional approaches, principal component analysis (PCA) and T-SNE [46]. PCA is used to reduce the data dimensionality while preserving the variations [47]. T-SNE preserves the metric properties of the original high-dimensional data. It preserves the information indicating which points neighbor each other [48].

When projecting the latent representations z learned by the model using TSNE, we found that our model effectively showed a consistent loop, while latent representation produced by conventional VAE preserved the distance in the data but did not preserve the structure of the input images (Fig.8). While using PCA, we found that our model preserved the structure of the input data. Conventional VAE did not preserve the structure of the input dataset. From its definition, PCA preserves the variation in the data. Two neighboring points in the high dimension should be closer in the low dimen-



Epochs=50 Epochs=300

FIGURE 8. **Projection of the latent representation z after training using T-SNE. Each dot represents an angle with a corresponding value from the face dataset. First row: VAE ($\alpha = 0$), which shows a loop but does not have a consistent loop structure. Second row: Our model ($\alpha = 10$) shows a consistent loop since the dataset used for testing has angles ranging from 0 and 360 degrees.**

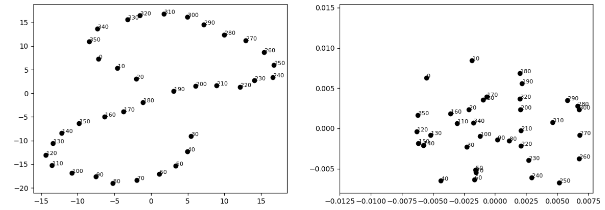


Epochs=50 Epochs=300

FIGURE 9. **Projection of the latent representation z after training using PCA. Each dot represents an angle with a corresponding value from the face dataset. First row: VAE ($\alpha = 0$) ignores the variance in the angle. Second row: Our model ($\alpha = 10$) shows a consistent loop since the dataset used for testing has angles ranging from 0 and 360 degrees.**

sion. Conventional VAE ignores the variance in the data, while our model keeps the fundamental structure of the input data (Fig.9).

β-VAE. We trained β-VAE [36] with different values of β; we found it to have the same behavior as conventional VAE. β-VAE does not have the necessary structure to generate the latent space that resembles the in-between image. We demonstrated the latent representation in Fig.10, and the results on TSNE suggest that conventional VAE and β-VAE might generate the structure of the in-between image if some form of penalty was imposed in the latent space or input signal is given to the model.



TSNE PCA

FIGURE 10. **β-VAE shows the same behavior as conventional VAE. The latent representation lacks the structure to generate the in-between image.**

2) COMPARISON WITH STATE-OF-THE-ART METHODS - LARGE DISPLACEMENT

a: QUANTITATIVE EVALUATION

This work lies between image interpolation and latent representations. Since existing works on latent representations focus on disentangled representations, we cannot compare them. The objective of this work is to generate an in-between image based on latent variables. In disentangled representation work, there is an assumption about the number of hidden variables presented in the data, and the data are often arranged to prove this assumption. We did not arrange the training data to disentangle the factors of variations present in the data.

We compared our approach with state-of-the-art approaches on image interpolation based on optical flow and neural networks, including Super SloMo [8], FlowNet [6] and a conventional VAE. To evaluate the error between the actual in-between and predicted image interpolation, we follow some baseline metrics presented in [49], including the peak signal-to-noise ratio (PSRN), structural similarity index (SSIM), L2, and L1 scores. In Tables 1,2 and 3, we demonstrate the performance of FlowNet and its versions, SloMo, conventional VAE and our approach. In Table2 and Table3, we used the face and dots datasets respectively. Our model achieves the best performance on all metrics. Despite good accuracy on all metrics, in Table1, for PSRN and L2, our model presents values slightly lower than FlowNet2.0 and FlowNet2S. The performances of our model indicate a plausible generalization capability for distinct datasets.

b: VISUAL EVALUATION

We compare our approach with two state-of-the-art works on image interpolation based on CNN and optical [8] and latent representation learning [31]. Our model achieved the best performance, particularly where the object is facing and produces fewer artifacts (Fig. 11). We highlight in a yellow box the errors presented by other models. Optical flow-based methods seem to have more problems with large displacement. It generates the in-between; however, the image resembles one of the input images, not the actual in-between,

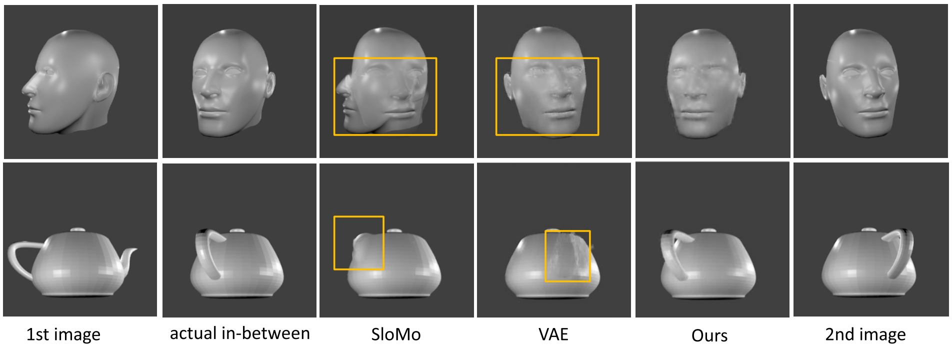


FIGURE 11. Illustration on teapot and face dataset. Our model produces plausible in-between and less artifacts around the object shown in yellow box.

TABLE 1. The results on "Teapot" dataset.

	PSRN	SSIM	L1	L2
FlowNet2CSS	64.10	0.0047	0.126	0.025
FlowNet2CS	62.37	0.0018	0.166	0.038
FlowNet2SD	54.09	0.0003	0.414	0.254
FlowNet2S	60.01	0.0010	0.180	**0.001**
FlowNet2C	60.39	0.0862	0.113	0.060
FlowNet2.0	**77.13**	0.3873	0.023	**0.001**
Super SloMo	69.97	0.9077	0.020	0.007
VAE	69.15	0.8442	0.023	0.008
Ours	73.19	**0.9078**	**0.015**	0.003

TABLE 2. The results on "Face" dataset.

	PSRN	SSIM	L1	L2
FlowNet2CSS	57.80	0.0015	0.292	0.108
FlowNet2CS	57.80	0.0005	0.374	0.175
FlowNet2SD	54.19	0.0004	0.406	0.248
FlowNet2S	60.44	0.0029	0.198	0.059
FlowNet2C	59.17	0.0005	0.196	0.079
FlowNet2.0	64.00	0.1538	0.068	0.026
Super SloMo	70.91	0.7653	0.032	0.005
VAE	74.04	0.8087	0.018	0.003
Ours	**75.46**	**0.8276**	**0.016**	**0.002**

TABLE 3. The results on "Dots" dataset.

	PSRN	SSIM	L1	L2
FlowNet2CSS	59.87	0.0003	0.219	0.067
FlowNet2CS	57.08	0.0002	0.127	0.127
FlowNet2SD	56.61	0.0007	0.241	0.142
FlowNet2S	60.64	0.0032	0.148	0.056
FlowNet2C	58.87	0.0030	0.224	0.084
FlowNet2.0	70.71	0.3338	0.033	0.006
Super SloMo	50.58	0.9396	0.130	0.570
VAE	64.853	0.863	0.024	0.021
Ours	**72.221**	**0.9437**	**0.010**	**0.004**

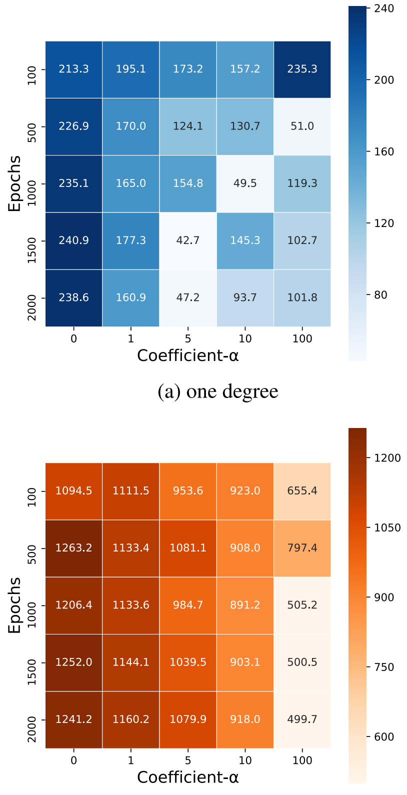
as illustrated in the figure. Conventional VAE does not capture the direction of the object and presents some artifacts in the generated image. One explanation is that learning from a pixel-based approach does not allow predicting large motion since it does not learn the embedding representations of the data.

c: IMPACT OF DEGREES OF FREEDOM - ADDITIONAL EVALUATION USING MSE

To learn more general data representations, we argue that it is essential to introduce diversity in the training samples. The model is assessed on different degrees of freedom using the mean squared error (MSE). The primary objective is to evaluate the complexity of the datasets, both on the degree of freedom and generalization. The same object is evaluated in two scenarios, one and two degree(s) of freedom: the same epochs, coefficient (α), and latent dimension (z). Fig.12 indicates that two degrees of freedom represent a more complex scenario. To generate a plausible in-between image in one degree of freedom, $\alpha = 5$ and $epoch = 1,500$ are required, whereas $\alpha = 100$ and $epoch = 2,000$ are required for generating a suitable in-between image in two degrees of freedom. These results are due to the differences in the number of possible scenarios between one degree (360) and two degrees (360 × 360).

d: IMPACT OF LATENT DIMENSION ON DIFFERENT DEGREES OF FREEDOM

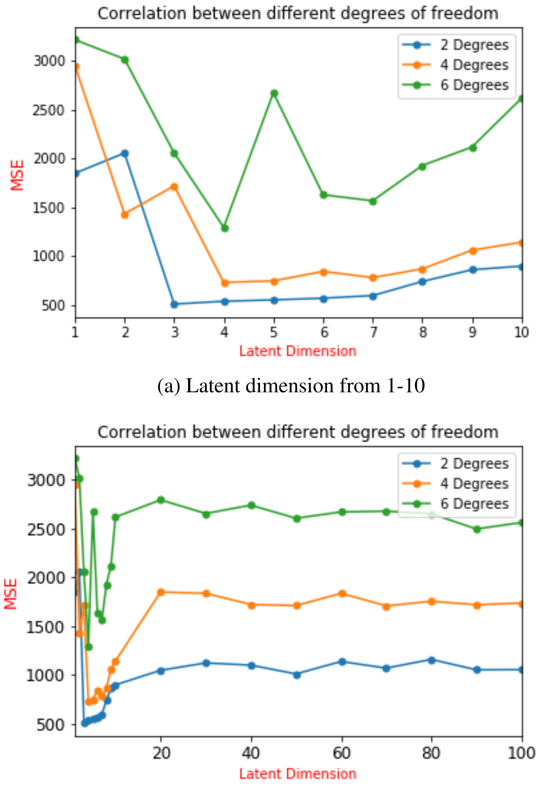
Latent variables are compressed representations (salient features of the data) of high-dimensional data. In VAE, the latent variables can be found in the bottleneck layer. Depending on the number of variables passed, the output quality might change. To date, the results have been assessed on a single latent dimension $(d_z) = 10$, except for "moving 2D shapes. As shown previously, the decoder can reconstruct the output, with only 10 variables passed to the bottleneck. We investigated the impact of the latent dimension on different degrees of freedom using "moving 2D shapes". The model was trained for 5,000 epochs with different latent dimensions (1 to 100). The model stabilized on latent dimension $z = 20$,



(a) one degree

(b) two degrees

FIGURE 12. a) MSE loss for one degree of freedom. b) MSE loss for two degrees of freedom. For one degree, the minimum loss for coefficient (α) = 5 and *epochs* = 1, 500. For two degrees, the minimum loss for coefficient (α) = 100 and *epoch* = 2, 000. This indicates a higher complexity of two degrees of freedom.



(a) Latent dimension from 1-10

(b) Latent dimension from 10-100

FIGURE 13. Effects of latent dimensions using MSE. There are two degrees, four degrees, and six degrees. The model begins stabilizing on the latent dimension z = 20. For this specific dataset, with 20 dimensions, the decoder could reconstruct the output.

as illustrated in Fig.13. For good generalization, passing 20 variables to the bottleneck could be sufficient. The decoder may be able to reconstruct the output.

D. LINEAR LATENT SPACE INTERPOLATION

Autoencoders can generate a semantically meaningful combination of features from two distinct data points. David *et al.* [50] have explored autoencoders in the context of regularization to improve linear interpolation. Ideally, latent variables of the data are close to each other but different. This characteristic enables smooth interpolation and stimulates creative design [51], [52]. Sampling latent variables through arithmetic operations can generate diverse outputs [44] suggests that models that preserve smooth interpolation between points might be relevant for disentangling explanatory factors of variation in data. Another critical application of continuous linear latent interpolation is to test if the model has not merely memorized the training data. By decoding the latent space of two data points, it is possible to visualize a smooth change from one image to the next, as illustrated in Fig.14.

V. DISCUSSION

There are two main lines of research relevant to our work. The first is similar to [5], [6], [8], and seeks to generate image interpolation based on a pixel-based approach. The second line is similar to [36], which revolves around seeking to learn controllable and interpretable latent representations of data. Of particular relevance to our work are approaches that explore latent space in the context of learning representations. Several works on (unsupervised) learning representations are based on VAE. Prior works [36], [38], [43], [53], enhance the quality of learned representation by modifying the conventional VAE objective function. These works often considered controlling the level of regularization of the latent space through KL divergence at the cost of reconstruction.

KL divergence allows the model to normalize and smoothly interpolate the latent space [54]. However, if not well-tuned, KL divergence can also induce the network model to a suboptimal [55]. The model does not exploit all the latent variables for generation, the so-called over-pruning/variable collapse discussed in [56]. Placing importance in the KL divergence term leads to a more controllable latent space, which may lead to a better quality of generated samples. A state-of-the-art study on unsupervised disentanglement

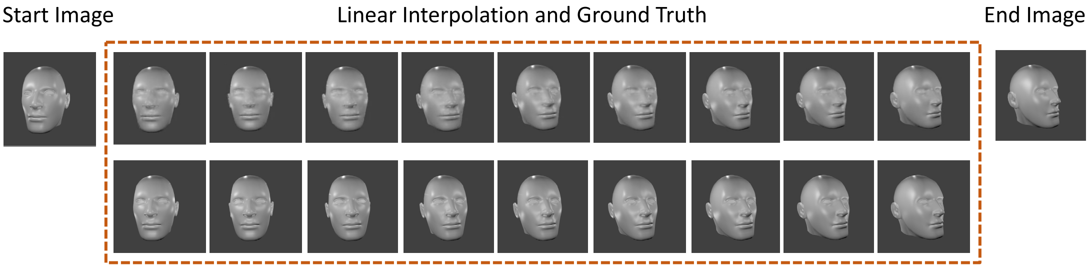


FIGURE 14. Continuous linear latent space interpolation.

representations β-VAE [36], gave relative importance to the KL divergence term by introducing a hyperparameter β to the VAE loss function. The authors argued that this modification encourages the model to learn interpretable representations of the data. In the same line of research, [57] enhanced β-VAE by modifying the training process. The authors claimed that increasing the information capacity of the latent codes during training enables the model to see more factors of variations continuously, thus resulting in better disentanglement. Our objective function is similar to β-VAE, but we do not aim to disentangle factors of variation in the data.

A different path to learning latent representations was taken by Chen *et al.* [58]. The authors proposed InfoGAN, a model based on a generative adversarial network (GAN). The model encourages disentanglement by penalizing the total correlation [59], i.e., the mutual information between the data and latent representation. Disentangled representation models have been shown to discover factors of variations in the data; the application is still restricted to a synthetic dataset. Locatello *et al.* [60] argued that disentangling a specific factor is nearly impossible without any forms of inductive bias on both the model and the data. Furthermore, the authors were not clear about the relevance of disentanglement for downstream tasks.

VI. CONCLUSION

This paper presented a simple approach to improving image interpolation. Our model produces good performance on all datasets. In addition, the model outperforms some baseline approaches on large displacements between images. The key to the success of this approach is dedicated to latent variables. Learning latent representations of the data and limiting the freedom of latent space has been demonstrated to have an impact on the generated in-between image structure. Previous works are pixel-based except conventional VAE; however, VAE does not have control over generated in-between. We propose a model that has the ability to control the latent space.

ACKNOWLEDGMENT

This work is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

REFERENCES

[1] H. Raveshiya and V. Borisagar, "Motion estimation using optical flow concepts," *Int. J. Comput. Technol. Appl.*, vol. 3, no. 2, pp. 1–5, 2012.

[2] H. Rezaee Kaviani, "Novel image interpolation schemes with applications to frame rate conversion and view synthesis," Ph.D. dissertation, McMaster Univ., Hamilton, ON, Canada, 2018.

[3] S. Rimac-Drlje and D. Vranjes, "Fast frame-rate up-conversion method for video enhancement," in *Proc. Int. Conf. Syst., Signals Image Process. (IWSSIP)*, May 2016, pp. 1–4.

[4] J. Klys, J. Snell, and R. Zemel, "Learning latent subspaces in variational autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 6444–6454.

[5] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.

[6] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2462–2470.

[7] Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4473–4481.

[8] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High quality estimation of multiple intermediate frames for video interpolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9000–9008.

[9] S. Niklaus, L. Mai, and F. Liu, "Video frame interpolation via adaptive separable convolution," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 261–270.

[10] T. White, "Sampling generative networks," 2016, *arXiv:1609.04468*. [Online]. Available: http://arxiv.org/abs/1609.04468

[11] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," Dept. Comput. Sci., Carnegie-Mellon Univ., Pittsburgh, PA, USA, Tech. Rep., 1981.

[12] B. Horn and B. Schunck, "Determining optical flow artificial intelligence," Artif. Intell. Lab., Massachusetts Inst. Technol., Cambrige, MA, USA, Tech. Rep., 1981, vol. 17.

[13] A. M. G. Pinto, A. P. Moreira, P. G. Costa, and M. V. Correia, "Revisiting Lucas-Kanade and horn-schunck," *J. Comput. Eng. Informat.*, vol. 1, no. 2, pp. 23–29, Apr. 2013.

[14] J. Zhai, K. Yu, J. Li, and S. Li, "A low complexity motion compensated frame interpolation method," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2005, pp. 4927–4930.

[15] J. Li and S. Li, "Low complexity motion compensated frame interpolation method," U.S. Patent 8 018 998, Sep. 13, 2011.

[16] D. Wang, A. Vincent, P. Blanchfield, and R. Klepko, "Motion-compensated frame rate up-conversion—Part II: New algorithms for frame interpolation," *IEEE Trans. Broadcast.*, vol. 56, no. 2, pp. 142–149, Jun. 2010.

[17] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, "Phase-based video motion processing," *ACM Trans. Graph.*, vol. 32, no. 4, pp. 1–10, 2013.

[18] S. Meyer, O. Wang, H. Zimmer, M. Grosse, and A. Sorkine-Hornung, "Phase-based frame interpolation for video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1410–1418.

[19] Z. Zhang, Y. Liu, and Q. Dai, "Light field from micro-baseline image pair," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3800–3809.

[20] M. A. Elgharib, M. Hefeeda, F. Durand, and W. T. Freeman, "Video magnification in presence of large motions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4119–4127.

[21] D. Gadot and L. Wolf, "PatchBatch: A batch augmented loss for optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4236–4245.

[22] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

[23] C. Bailer, K. Varanasi, and D. Stricker, "CNN-based patch matching for optical flow with thresholded hinge embedding loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3250–3259.

[24] J. Thewlis, S. Zheng, P. H. S. Torr, and A. Vedaldi, "Fully-trainable deep matching," 2016, *arXiv:1609.03532*. [Online]. Available: http://arxiv.org/abs/1609.03532

[25] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[26] S. Meyer, A. Djelouah, B. McWilliams, A. Sorkine-Hornung, M. Gross, and C. Schroers, "PhaseNet for video frame interpolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 498–507.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[28] X. Shu, J. Tang, Z. Li, H. Lai, L. Zhang, and S. Yan, "Personalized age progression with bi-level aging dictionary learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 905–917, Apr. 2018.

[29] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," 2015, *arXiv:1511.06349*. [Online]. Available: http://arxiv.org/abs/1511.06349

[30] Y. Li, D. Roblek, and M. Tagliasacchi, "From here to there: Video inbetweening using direct 3D convolutions," 2019, *arXiv:1905.10240*. [Online]. Available: http://arxiv.org/abs/1905.10240

[31] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: http://arxiv.org/abs/1312.6114

[32] D. Jimenez Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," 2014, *arXiv:1401.4082*. [Online]. Available: http://arxiv.org/abs/1401.4082

[33] T. Salimans, D. Kingma, and M. Welling, "Markov chain monte carlo and variational inference: Bridging the gap," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1218–1226.

[34] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 3483–3491.

[35] C. Chan, S. Ginosar, T. Zhou, and A. Efros, "Everybody dance now," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5933–5942.

[36] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, and A. Lerchner, "β-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–13.

[37] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[38] E. Mathieu, T. Rainforth, N. Siddharth, and Y. Whye Teh, "Disentangling disentanglement in variational autoencoders," 2018, *arXiv:1812.02833*. [Online]. Available: http://arxiv.org/abs/1812.02833

[39] V. Samsonov, "Deep frame interpolation," 2017, *arXiv:1706.01159*. [Online]. Available: http://arxiv.org/abs/1706.01159

[40] C. M. Bishop, *Pattern Recognition and Machine Learning* (Information Science and Statistics). Secaucus, NJ, USA: Springer-Verlag, 2006.

[41] J. Tang, X. Shu, Z. Li, G.-J. Qi, and J. Wang, "Generalized deep transfer networks for knowledge propagation in heterogeneous domains," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 12, no. 4s, pp. 1–22, Nov. 2016.

[42] X. Shu, G.-J. Qi, J. Tang, and J. Wang, "Weakly-shared deep transfer networks for heterogeneous-domain knowledge propagation," in *Proc. 23rd ACM Int. Conf. Multimedia (MM)*, 2015, pp. 35–44.

[43] H. Kim and A. Mnih, "Disentangling by factorising," 2018, *arXiv:1802.05983*. [Online]. Available: http://arxiv.org/abs/1802.05983

[44] Y. Bengio, G. Mesnil, Y. Dauphin, and S. Rifai, "Better mixing via deep representations," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 552–560.

[45] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, A. Courville, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," 2018, *arXiv:1806.05236*. [Online]. Available: http://arxiv.org/abs/1806.05236

[46] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[47] I. Jolliffe, *Principal Component Analysis*. Cham, Switzerland: Springer, 2011.

[48] A. C. Müller and S. Guido, *Introduction to Machine Learning With Python: A Guide for Data Scientists*. Newton, MA, USA: O'Reilly Media, 2016.

[49] S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *Int. J. Comput. Vis.*, vol. 92, no. 1, pp. 1–31, Mar. 2011.

[50] D. Berthelot, C. Raffel, A. Roy, and I. Goodfellow, "Understanding and improving interpolation in autoencoders via an adversarial regularizer," 2018, *arXiv:1807.07543*. [Online]. Available: http://arxiv.org/abs/1807.07543

[51] D. Ha and D. Eck, "A neural representation of sketch drawings," 2017, *arXiv:1704.03477*. [Online]. Available: http://arxiv.org/abs/1704.03477

[52] K. Hamada, K. Tachibana, T. Li, H. Honda, and Y. Uchida, "Full-body high-resolution anime generation with progressive structure-conditional generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 1–9.

[53] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," 2018, *arXiv:1802.04942*. [Online]. Available: http://arxiv.org/abs/1802.04942

[54] A. A. Alemi, B. Poole, I. Fischer, J. V. Dillon, R. A. Saurous, and K. Murphy, "Fixing a broken ELBO," 2017, *arXiv:1711.00464*. [Online]. Available: http://arxiv.org/abs/1711.00464

[55] A. Asperti and M. Trentin, "Balancing reconstruction error and Kullback-Leibler divergence in variational autoencoders," 2020, *arXiv:2002.07514*. [Online]. Available: http://arxiv.org/abs/2002.07514

[56] S. Yeung, A. Kannan, Y. Dauphin, and L. Fei-Fei, "Tackling over-pruning in variational autoencoders," 2017, *arXiv:1706.03643*. [Online]. Available: http://arxiv.org/abs/1706.03643

[57] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE," 2018, *arXiv:1804.03599*. [Online]. Available: http://arxiv.org/abs/1804.03599

[58] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2172–2180.

[59] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM J. Res. Develop.*, vol. 4, no. 1, pp. 66–82, Jan. 1960.

[60] F. Locatello, S. Bauer, M. Lucic, G. Rätsch, S. Gelly, B. Schölkopf, and O. Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," 2018, *arXiv:1811.12359*. [Online]. Available: http://arxiv.org/abs/1811.12359

PAULINO CRISTOVAO (Graduate Student Member, IEEE) received the M.S. degree in computer science from the University of Tsukuba, Japan, in 2019, where he is currently pursuing the Ph.D. degree. From 2016 to 2019, he stayed in the Advanced Industrial Institute of Science and Technology (AIST) to study machine learning. His research interest includes disentangling latent representations that are learned through variational autoencoders. He is a member of JSAI.

HIDEMOTO NAKADA received the Ph.D. (Eng.) degree from The University of Tokyo, in 1995. He joined the Electrotechnical Laboratory, in 1995. He is currently working with the National Institute of Advanced Industrial Science and Technology (AIST), in 2001. Since 2015, he has been an Adjunctive Professor with the Cooperative Graduate School, University of Tsukuba. His research interests include parallel and distributed computing, including grid, cloud, and HPC- and machine learning-related technologies. He is a member of ACM, AAAI, IPSJ, and JSAI.

YUSUKE TANIMURA received the Ph.D. degree in engineering from Doshisha University, in 2004. He is currently a Senior Research Scientist with the National Institute of Advanced Industrial Science and Technology (AIST), Japan. He is also an Associate Professor with the Cooperative Graduate School, University of Tsukuba. His research interests include distributed storage systems, big data analytics, cloud computing, and high-performance computing.

HIDEKI ASOH received the B.Eng. degree in mathematical engineering and the M.Eng. degree in information engineering from The University of Tokyo, in 1981 and 1983, respectively. In 1983, he joined in the Electrotechnical Laboratory as a Researcher. From 1993 to 1994, he worked at the German National Research Center for Information Technology as a Visiting Research Scientist. He is currently a Principal Research Manager of the Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST). His research interest includes constructing intelligent systems that can learn through interactions with the real world.