IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 26, 2018, accepted November 27, 2018, date of publication December 4, 2018,
date of current version January 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2885006

Using a Vertical-Stream Variational Auto-Encoder to Generate Segment-Based Images and Its Biological Plausibility for Modelling the Visual Pathways

XUE-SONG TANG[1], HUI WEI[2], AND KUANGRONG HAO[1]

[1]Engineering Research Center of Digitized Textile Apparel Technology, Ministry of Education, College of Information Science and Technology, Donghua University, Shanghai 201620, China
[2]Laboratory of Cognitive Model and Algorithm, School of Computer Science, Fudan University, Shanghai 201203, China

Corresponding author: Kuangrong Hao (krhao@dhu.edu.cn)

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2232017D-08, in part by the Shanghai Sailing Program under Grant 17YF1426100, in part by the National Natural Science Foundation of China under Grant 61771146 and Grant 61806051, and in part by the International Collaborative Project of the Shanghai Committee of Science and Technology under Grant 16510711100.

ABSTRACT Human beings have a strong capability to identify objects in different viewpoints. Unlike computer vision that requires sufficient training samples in various scales and rotations, biological visual systems can efficiently recognize objects in diverse spatial states. To achieve this objective, images are processed into a segment-based representation and then a vertical stream variational auto-encoder (VSVAE) is utilized to generate images based on the preprocessed segments in this study. The novel structure of the two vertical streams can be also considered as a computational model for the interaction between the ventral pathway and the dorsal pathway in the visual cortex. The reconstructive capability of the VSVAE is testified by using a series of geometric information sets to enhance the segment-based representation. By visualizing the learnt features in the hidden layers of VSVAE, the biological plausibility of the model is discussed. In addition, the proposed methodology is able to facilitate the classification accuracy, especially when the images are severely transformed.

INDEX TERMS Invariant recognition, representation argumentation, segment-based image representation, variational auto-encoder, visual cortex modeling.

I. INTRODUCTION

Cognitive computational simulation in brain-like intelligence is one of the core issues in artificial intelligence. Amongst various brain mechanisms, the most comprehensively-researched area is the visual system in the high-level mammals. Many years of researches on the visual cortex, which is the central processor of the visual system, have divided it into multiple functional areas with different functionalities and several visual pathways that are responsible for processing different types of information. Deep neural network is the major technical framework used to model the visual pathways. Deep learning technologies have achieved breakthrough experimental results in the fields of machine vision such as image classification, target detection and video analysis. These methods mainly use deep neural networks to learn the distributions of features such as texture, brightness, and color of images in different regions. At present, their performances depend heavily on the massive, high-quality training samples, complex structures and sophisticated hyper-parametric tuning. They still need improvements especially in generalization ability, interpret-ability, and training efficiency. Therefore, how to use the visual characteristics and biological evidences to imitate the visual cognitive mechanism and modularize the integration of different visual pathways becomes an important topic for the cross-discipline researches of brain science and brain-like artificial intelligence.

Utilizing operations such as convolution and pooling, mainstream models still can hardly learn features that are invariant to certain transformations of the objects.

Enhancing the capability of learning such features usually requires to build quite deep and complex networks. However, the brain's visual system does not require repeated training to recognize the same object in different spatial states. When the inputs are rotated, scaled, and shifted, the activation states of the entire convolutional neural network will change significantly. However, the neurobiologists scanned the human brain and found that the biological activation varied slightly in our brain [1]. The current mainstream models can learn the corresponding invariant features through augmented training samples under different levels of rotations and scales for the objects. In the biological recognition process, the brain can autonomously rotate and scale the object to be an easily-recognized state, and then judge the object's category. This illustrates that the biological vision system has a strong ability to perform spatial transformation. Therefore, the research of biological visual mechanisms aimed to achieve a brain-like intelligent method that can effectively recognize objects in different spatial states is one of the key scientific challenges in artificial intelligence. The solution of this problem is expected to establish a methodology that truly enables the 'Equivariance-like' human visual system [2], which is genuinely different from the invariant feature learning at different scales.

Utilizing biological mechanisms for modelling the visual cortex to solve practical problems in computer vision is a crucial issue in the last decades. The recent researches mainly focus on the individual modelling of the primary vortex V1 [3], V2 [4], V4 [5] and IT [6] in the ventral pathway. For the high-level visual areas, a research proposed a method based on the sparse auto-encoder to model the V4 ventral pathway and successfully monitored its shape selectivity [7]. In another research, the authors discovered that the higher-level cells can integrate the contour fragments detected from the lower areas in the visual cortex [8]. Most of these researches can effectively model individual region separately yet they barely explore the relationships and hierarchy amongst different visual functional areas. Riesenhuber *et al.* proposed a classical visual model based on the research of Hubel and Wiesel [9], which uses iterative simple and complex cell layers for modelling the increasingly complexity of the information processing along the ventral pathway [10]. They also designed a feed-forward visual information and effectively applied it in practical object recognition missions [11]. Using deep learning methods to design neural computational models is prevail and achieved remarkable successes. The current visual system can be clearly categorized into two pathways, which are the ventral pathway and the dorsal pathway. The ventral flow is considered to be the main mediator transforming the visual signals to memory, cognition and consciousness, while the dorsal flow is mainly related with the spatial information of the object and motional controls. Fig. 1 displays the biological structure of the two streams. By imitating the double-stream structure of the visual cortex, researchers proposed novel models for vision understanding and image classification [13]–[15], and found

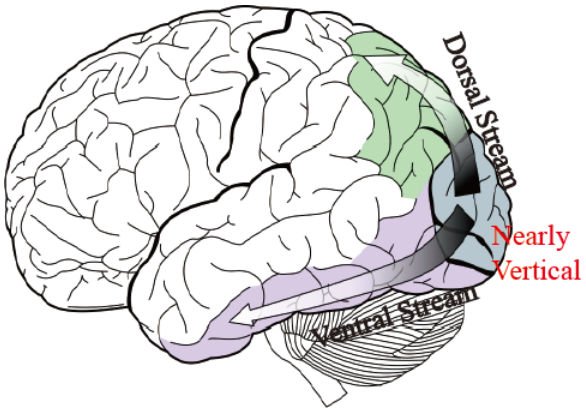

FIGURE 1. The biological structure of the dorsal/ventral streams. The biological directions of both streams are nearly vertical from each other.

that the structure can facilitate the feature learning process and the classification performances by fusing the results of different neural pathways. Another study proposed a system that can automatically design novel computational models by an artificial dorsal stream (ADS) for the task of head tracking [16].

Image has another important feature, which is the contour of the object. It has been widely used for shape construction [17], image segmentation [18] and object recognition [19], [20] in the past and achieved excellent experimental performances compared to the pixel-based methods. Segment-based representation is a powerful approximation to the object contour, which provides a compact, flexible and well-structured object description. There are many approaches for line segments detection. Line Segment Detection (LSD) [21] is a quite effective one, whose time complexity is proportional to the quantity of pixels to be processed in the image. Researchers also designed a segment detection method based on the visual cortex mechanism, which can achieve more accurate and intrinsic results [22]. In the computer vision fields, using line segments for visual missions has been studied for years and still plays an important role in many advanced applications. In particular, this paper uses a well proven Discrete Contour Evolution (DCE) [23] to approximate the segment distribution. Deep learning has also been well applied in some contour detection researches [24], [25]. Furthermore, with the explosive growth of touch screens, sketch-based images can be retrieved easily, which are also a form of contour-based representation. Hand-drawn shapes are well studied by researchers for sketch parsing [26], image retrieval [27], face sketch synthesis [28] and photo matching [29]. Those researches proved that contour-based images can achieve an effective reconstructive efficiency and a superior performance in discrimination.

Variational Auto-Encoder (VAE) [30] is one of the most important extensions to enhance the regular auto-encoders. With a high degree of data correlation, the regular auto-encoders can only compress data that is similar to the training

data, because the extracted features are generally highly relevant to the original training set. One famous application of the VAE is called Adversarial Auto-Encoder (AAE) [31], which is a probabilistic auto-encoder that uses the Generative Adversarial Networks (GAN) to perform a variational inference. Another extended version of VAE called PixelGAN auto-encoders is proposed, in which the generative path is a convolutional autoregressive neural network on pixels [32]. In another research [33], scientists proposed a series of latent representation models for improving the network anomaly detection by introducing new regularizers on a classical AE and a VAE. Another research presented a dual autoencoder network model based on the retinex theory to perform a low-light image enhancement, which analyzed and compared the stacked and convolutional autoencoders with the constraint terms of the variational retinex model [34]. Also, the VAE-based method can be developed to model images, as well as associated labels or captions [35]. In this paper, we propose a VAE-based framework to model the two pathways in the visual cortex. The early integration of both pathways is studied by using a novel vertical structure. The reason of using a latent variable model is that it offers an effective mechanism for the manifold learning, which can visualize the generated segments in a particular low-dimension space. The characteristic enriches the flexibility of the visualizations and the generative capability. Also, the utilization of VAE can effectively alleviate the dilemma between the expansion of representations and the computational efficiency. The GAN and its extended versions are not applied in this work due to their sophisticated implementations and difficulties for convergences in training.

In the traditional deep learning methods, the size of the input layer is decided by the resolution of the processed images, thus high-resolution images normally require a large network to process, which makes the training quite time-consuming. One effective way is to use convolutional operations to preserve spatial relation of an image and pooling operations to learn features at a more global scale and finally use a few full-connected layers for classification. These CNN-based models were proposed in 1990s and have been significantly developed recently with the great developments of the computational capacity and establishments of massive databases. However, the CNN-based technical frameworks heavily rely on the computational resources and the integrity of the databases, their computational efficiency, interpretability and train-ability are still problems urge to be solved. Particularly, the pixel-based models are vulnerable to transformations in scales and rotation. For a rotating object, when it is noticed, we understand that it is the same object whose orientation is changing but other properties remain the same. For a pixel-based feature learning system, if an image is rotated, it is clear that the activations of the neuron from the first few layers will be quite different compared with its original state, which means that the variances can be only achieved by learning deep features of the object. In another word, the invariances cannot be achieved without a large number of training samples. However, the biological mechanism can perform such invariant recognition easily, a person can recognize a rotating object in different spatial conditions without any training phases. In this paper, we use a segment-based representation formed by the geometric information, which can achieve higher invariances than the pixel-based representations. It uses a vertical-stream geometric structure, which complies with both visual pathways in the visual cortex and can be considered as an effective computational modelling for the integration of the two streams. The main contributions are presented below.

1) It verifies the feasibility of using segments-based representation to generate various samples in differentiated scales, rotations, translations and deformations.

2) The model is further developed to facilitate the image classification.

3) The visualizations can be compared with the biological evidences obtained by the neuroscientists.

The remainder of the paper is organized as follows. In Section II, we briefly introduce the preprocessing method for the original images. The overall method is described in Section III. Section IV analyses the characteristics of the proposed representation and evaluates the experimental results on various datasets. Section V discusses the possible developments of the proposed method in the future and concludes the paper.

II. A TWO-STAGE PREPROCESSING FOR LINE-SEGMENT NORMALIZATION

The major difficulty of training a neural network by using line segments is that the number of segments for each image can be arbitrary. To solve this problem, we propose a two-stage preprocessing method to normalize the images to be represented by an equal quantity of segments. The most intrinsic principle of normalizing the image is to use a reasonable number of segments to depict a major proportion of the object contour. However, due to the limited capacity of the segment detection algorithm, many clutters have been detected and the contours of the labelled target could be partially missing. To normalize the segment-based images, we need to select an appropriate number, then for those images with exceeding numbers of segments, we need to approximate them to a smaller number of segments. On the other hand, we can split segments to increase the total number. Notably, the splitting of the segments does not lead to the loss of contour information while the approximation does. Particularly, for a dataset with a large number of images, the line segment detection results can be varied from tens to hundreds, which makes the normalized number quite hard to be determined. Given a small normalized number, a large number of segments will be approximated, which causes significant geometric information loss for those images with exceeding numbers of segments and also makes the approximation quite time-consuming. On the contrary, a large normalized number will generate a massive neural network. Therefore, we initially use the salience of detected segments to filter a certain number of

segments in order to compress the quantity of the segments to a certain extent, which is the first step of the normalization. To determine which segments are less important in the image, we use the parameter of width in the LSD results, which indicates the detective level of each segment. Afterwards, we use the approximate operation to further adjust the number of segments so as to achieve the objective of normalization. For approximating the segments, the DCE algorithm is used, which can iteratively merge two segments into one and preserve as much geometric information as possible. The details of the DCE algorithm can be viewed in [23].

The algorithm of the two-stage preprocessing can be viewed in Algorithm 1 and Fig. 2 is a simple demonstration of the proposed two-stage approximation for the segments.

Algorithm 1 Preprocessing and Normalization

Input:
$S = \{s_0, s_1, \ldots s_{n-1}\}$;
Nor_num;
Output:
S_{nor};
while $n > Nor\_num$ **do**
  if $n > 1.3 * Nor\_num$ **then**
    Find s_m with the smallest width in S;
    Remove s_m from S;
  else if $n < 1.3 * Nor\_num$ **then**
    $S \leftarrow DCE(S)$;
  end if
  $n \leftarrow n-1$;
end while
while $n < Nor\_num$ **do**
  Find s_l with the longest length in S;
  Split s_l to s_l^1 and s_l^2 with equal lengths;
  Replace s_l with s_l^1 and s_l^2 in S;
  $S \leftarrow S = \{s_0, s, \ldots s_l^1, s_l^2 \ldots s_{n-1}\}$;
  $n \leftarrow n + 1$;
end while
Calculate the centroid P_cen of S;
Calculate the mean length L_p of S;
$RS \leftarrow [X_{P_cen} - L_p/2, Y_{P_cen}, X_{P_cen} + L_p/2, Y_{P_cen}]$;
Find s_k in S with the shortest distance to RS;
Set the nearer endpoint of s_k as the starting point of the directed segment and the further endpoint of s_k as the ending point of the directed segment;
$\overrightarrow{s_0} \leftarrow \overrightarrow{s_k}$;
Remove s_k from S and add $\overrightarrow{s_0}$ in S_{nor};
for $i = 1; i < Nor\_num - 1; i++$ **do**
  Find $\overrightarrow{s_k}$ as a directed segment with the shortest distance to $\overrightarrow{s_{i-1}}$ in S;
  Remove $\overrightarrow{s_k}$ from S;
  $\overrightarrow{s_i} \leftarrow \overrightarrow{s_k}$;
  Add $\overrightarrow{s_i}$ in S_{nor};
end for
$S_{nor} \leftarrow \{\overrightarrow{s_0}, \overrightarrow{s_1}, \ldots \overrightarrow{s_{Num_nor-1}}\}$;

III. METHOD

A. VERTICAL STREAM VARIATIONAL AUTO-ENCODER

The visual cortex of brain mainly processes visual information in two pathways, which are the ventral stream and dorsal stream respectively. The current typical convolutional neural networks that implement a hierarchical feature learning structure can be considered as a methodology inspired partially from the ventral pathway. One major problem with this technical framework is that they are vulnerable to transformations in scales and rotations. The main reason is that the CNNs merely use down-sampling and convolution to acquire global features and enhance the invariance. Therefore, the network cannot learn features that are genuinely invariant to the geometric transformations whereas the biological system does not require sophisticated training for achieving such functionality. Particularly, the biological visual systems have an effective interactive mechanism between the two pathways, thus it is conceivable to broaden the networks so as to implement a dorsal stream for better invariances. In this study, we do not attempt to change the intermediate processes of the traditional networks, such as the synaptic plasticity. Instead, we broaden the input layer to form a vertical stream that uses a series of geometric variations to achieve translation, scale, rotation and other forms of transformations. The presented representations in geometric information directly describe the spatial information of the objects, which are expected to model the dorsal stream and further explore the computational model of the interactions between the ventral and dorsal streams. Compared to the work in [14] that uses multiple streams to handle different geometric information, the VSVAE can integrate multiple sets of geometric features and fuse their invariances together by using a single VAE.

As shown in the top of Fig. 3, the inputs of the segments are geometrically transformed into a series of variations according to the original segments. The inputs are expanded to integrate different forms of transformations, such as translation, scale, rotation and deformation. The transformed geometric representations are appended after the original segments and a horizontal combination of different geometric representations can be formed. Because of the expansion of the inputs, the auto-encoder in the vertical direction changes accordingly but the dimension of the output is still equal to the dimension of the original segments. As a result, the network can be considered as a form of asymmetric auto-encoder. Compared to the existing multi-stream frameworks, both streams in the proposed model are vertical with each other, which can be better accordance with the biological mechanism of the visual cortex. The overall framework of the proposed model can be seen from Fig. 3.

B. GEOMETRIC TRANSFORMATIONS

For clarity, some notations and operators used in this paper are listed in the Table 1. The segments are preprocessed into a standard normal distribution with a mean 0 and a variance 1. All weights are initialized by using a random uniform distribution between -1 and 1.

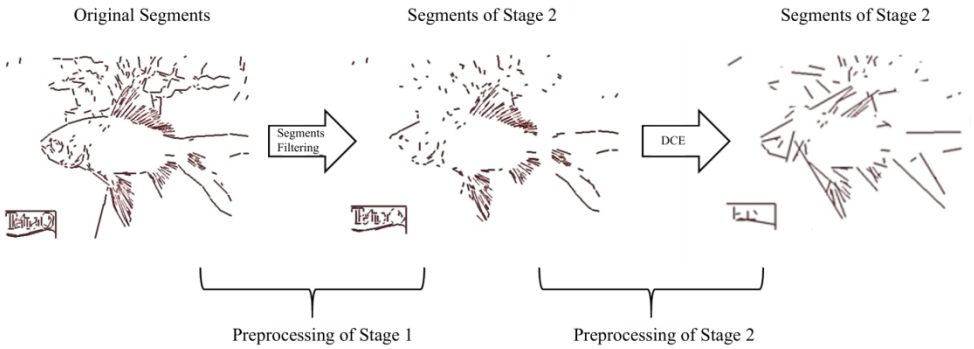

FIGURE 2. The two-stage preprocessing for approximating the segments. The segments with less importance according to their detected widths are removed in the first stage. Afterwards the remained segments are further approximated by using the DCE algorithm.

TABLE 1. Notations and definitions.

Notation	Definition
S_{nor}	preprocessed set of line segments
$\boldsymbol{\theta}$	weights and biases in the encoder
$\boldsymbol{\varphi}$	weights and biases in the decoder
$\mathbf{z}$	hidden representation
$\mathbb{E}(\cdot)$	reconstruction loss between the inputs and outputs
$\mathbb{D}_{KL}(\cdot)$	the information loss when using q to approximate p
$\mathbb{L}(\cdot)$	loss function
$p_{\boldsymbol{\theta}}(\cdot)$	encoder distributions
$q_{\boldsymbol{\varphi}}(\cdot)$	decoder distributions
$\Psi(\cdot)$	transformed geometric information set
$\phi(\cdot)$	reverse transformation from geometric information to segments
$\mathbf{W}_{L2}$	the L2 regularization by using the weight decay.
λ	parameter to control the importance of weight decay term
$\mathcal{N}$	A normal distribution

The geometric transformations used in this method are different from the affine transformations as they follow an encoder-decoder paradigm. The original distributions of the line segments are formed in a simple two-endpoint representation, as show in

$$s_i = \{x^b, y^b, x^t, y^t\}, \tag{1}$$

where x^b and y^b are the coordinates of the beginning point of s_i, and x^t and y^t are the coordinates of the terminal point of s_i. For a given image, it is preprocessed into a set of segments S_{nor} by using the Algorithm 1, which can be seen in

$$S_{nor} = \{s_0, s_1 \ldots\ldots s_{n-1}\}, \tag{2}$$

where S_{nor} is a $4{*}n$ matrix. In this study, we proposed a series of k geometric transformations in

$$\Psi = \{\psi_0, \psi_1 \ldots\ldots \psi_{k-1}\}. \tag{3}$$

For a given set of Ψ, an original set of S_n can be transformed into

$$\Psi(S^k) = \{\psi_0(S_n), \psi_1(S_n).....\psi_{k-1}(S_n)\}, \tag{4}$$

where $\Psi(S^k)$ is a $4{*}n{*}k$ matrix. And for any ψ_j, we have ϕ_j as the reverse transformation of ψ_j:

$$\phi_j(\psi_j(S_n)) = S_n. \tag{5}$$

ϕ_0 is a special normalized form of the original segment distribution, where all the starting points and ending points are normalized into a $\mathcal{N}(0, 1)$. It can be seen from Fig. 4 that the invariant information is distributed in different subsets of the geometric representations, where each of them is able to identify a particular type of transformation.

C. TRAINING

The training implementation in this paper is different from the traditional VAE, where the loss function not only contains the reconstruction loss, the Kullback-Leibler divergence loss, but also a regularization term of weight decay. In our condition, the loss function becomes:

$$\begin{aligned}\mathbb{L}(\boldsymbol{\theta}, \varphi; s_i) = {} & \mathbb{E}_{q_{\boldsymbol{\varphi}}}(\mathbf{z}|\Psi(s_i))(\log p_{\boldsymbol{\theta}}(\Psi(s_i)|\mathbf{z})) \\ & - \mathbb{D}_{KL}(q_{\boldsymbol{\varphi}}(\mathbf{z}|\psi_0(s_i)) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})) \\ & + \lambda * \mathbf{W}_{L2}(\boldsymbol{\theta}, \boldsymbol{\varphi}).\end{aligned} \tag{6}$$

Equation (6) is the reconstruction loss, or expected negative log-likelihood of the i_{th} segment. The expectation is taken with respect to the encoder's distribution over the representations. The first term is the cross entropy, which encourages the decoder to learn how to reconstruct the data. If the decoder's output does not reconstruct the data well, it will incur a large cost in this loss function.

The second term is an imposed regularization, which is the Kullback-Leibler divergence between the encoder's distribution and decoder's distribution. This divergence measures how much information is lost when using q to represent p. It is also a measure of how closed between q and p.

The third term is the newly added regularization term of the weight decay, which indicates the magnitude of the weights

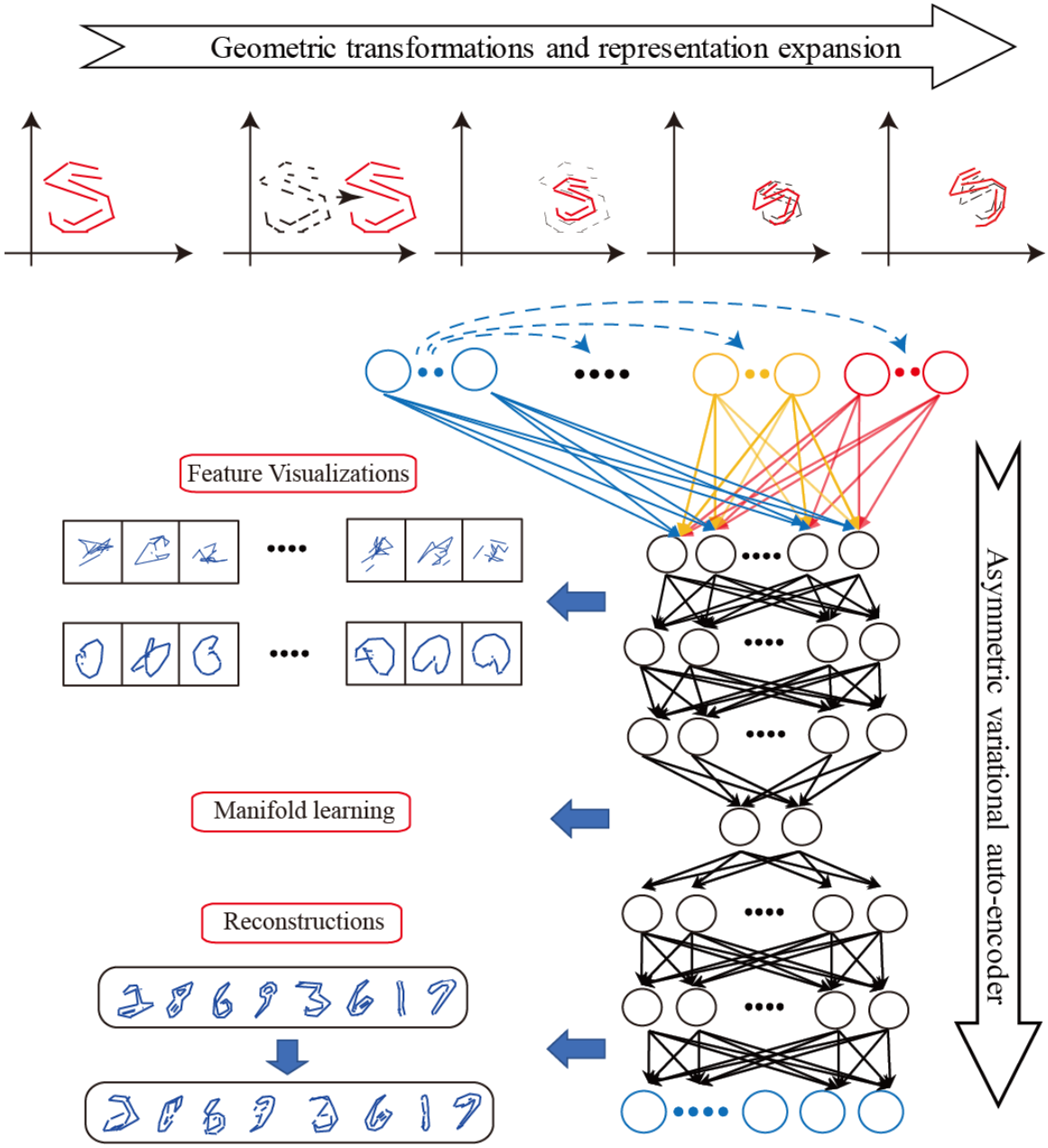

FIGURE 3. The overall framework of the proposed method. Top: the geometric transformations as an input expansion stream for the images, where the segment-based images are translated, scaled, rotated and deformed to a certain extent. Bottom: an asymmetric variational structure uses the expanded representation to form the original segments as a feature learning stream, where the two streams are vertical in this structure.

and helps preventing over-fitting. As a type of Bayesian regularization, the term tends to make the large weights suffer and the model is regularized to be simple and of better generalization power. λ is the control parameter to determine the relative importance amongst the terms, which is ranged from 0.0001 to 0.1 and tuned by a grid search in the experiments. The definition of weight decay is defined in

$$\mathbf{W}_{L2}(\boldsymbol{\varphi}, \boldsymbol{\theta}) = \sum_{\mathbf{w}_i \in \boldsymbol{\varphi}} \mathbf{w}_i^2 + \sum_{\mathbf{w}_j \in \boldsymbol{\theta}} \mathbf{w}_j^2. \tag{7}$$

IV. EXPERIMENTS

The experimental part consists of five subsections. First of all, the experimental configurations of the experiments are given and the datasets are introduced. Secondly, the capability of reconstructions of the proposed model is verified on different datasets. Thirdly, the manifold learning results are exhibited and the superiority of our method is presented by comparing with other pixel-based models. Afterwards, we discuss the biological plausibility by visualizing the hidden neurons in the VSVAE. Lastly, we verify the ability of the generated

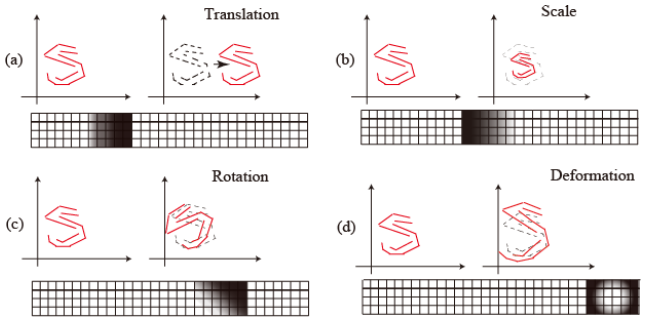

FIGURE 4. The different geometric transformations and their corresponding changing areas in the segment-based representation. When a segment-based image is translated (a), scaled (b), rotated (c) and deformed (d), only the corresponding regions in the representation are changed, while the rest regions remain the same.

samples to facilitate the performances for the image classification.

A. DATASET INTRODUCTIONS

The MNIST Handwritten Digit Dataset: a well-known dataset for handwritten numbers from 0 to 9, has a training set of 60,000 examples and a test set of 10,000 examples. The digits have been size-normalized and centered in a fixed-size image of 28∗28. The handwritten digits can be effectively represented by a dozen of segments. In this paper, the number of segments used to normalize the images in the MNIST dataset is set to 30.

Aaron Koblin Sheep Dataset: a set of crowd sourced drawings of around 8000 sheep. The drawings are all organized in a 3-stroke data structure, which indicates the drawing sequences. In this paper, the original representations are reorganized in the two-endpoint representation. Furthermore, the drawings are more complex than the handwritten digits, so the numbers of segments used to normalize the drawings are set to 50, 100 and 200, respectively.

The Street View House Numbers Dataset (SVHN): a dataset contains over 600,000 labelled digits cropped from street view images (a training set and an extra set) and 26032 test images. The goal of this task is to classify the digit in the center of each cropped 32∗32 color image. This is a difficult real-world problem. We preprocessed these samples in the same way as the MNIST dataset.

B. SEGMENT RECONSTRUCTION AND RECOVERY

In this subsection, the reconstructive capability of segment-based images from the VSVAE are qualitatively assessed. It can be seen that our model has a good train-ability in reconstruction, which means that the learnt geometric features can form up the original distributions of segments effectively. The variational approximation from Section III were used for the encoder and decoder respectively, where they have a same number of hidden units.

We compared the reconstructive results between the VSVAE and the original VAE in Fig. 5. It can be observed that the proposed model converges very quickly and gives birth

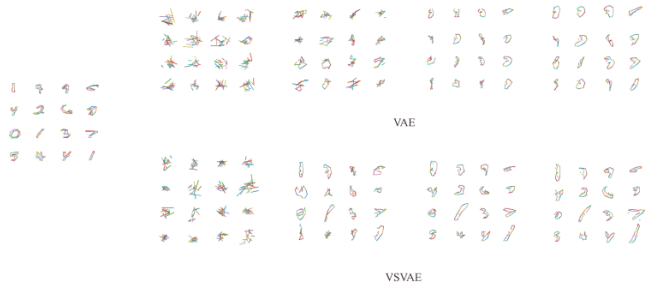

FIGURE 5. The comparison of the reconstruction results between the proposed VSVAE and the original VAE on the MNIST dataset.

to better reconstructions. It can be seen from Fig. 6 that the number of segments also affects the reconstructive capability, since the sheep drawings are of more complex topological structures than the handwritten digits. When the normalized number of segments are too small, the reconstructed shapes are heavily distorted.

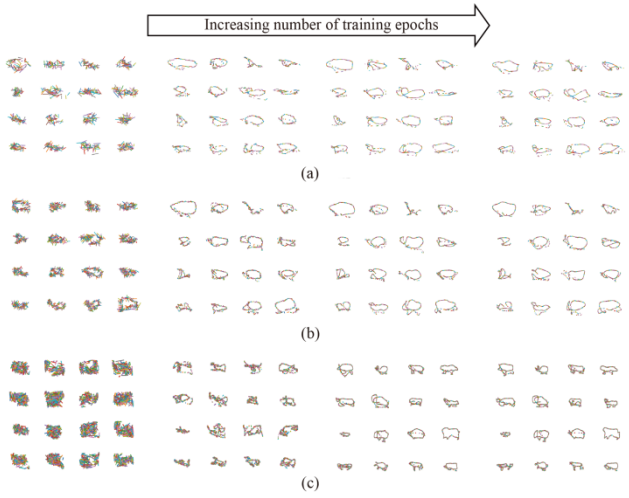

FIGURE 6. The reconstruction results of the Aaron Koblin Sheep Dataset by different quantities of segments. The sizes of the latent spaces are set to 2. It can be observed that the number of segments resulted from the preprocessing significantly affects the performance of reconstruction. (a) 50 segments. (b) 100 segments. (c) 200 segments.

Furthermore, we testify the reconstructive capability of the proposed method by evaluating its ability to recover segments when they are significantly denoised. For the proposed segment-based representations, the segments can be denoised by cutting off its individual length according to the midpoint of each segment. The ability to recover the partial segments reflects the reconstructive capability and general robustness. The segments can be considered to be shrunk to a certain level. Fig. 7 is some samples of different shrinking levels. It can be seen from the results that the reconstructive results remain satisfactory even when the geometric patterns are significantly weakened.

C. ANALYSIS OF THE MANIFOLD LEARNING RESULTS

In this subsection, we analyze the proposed method by visualizing the learnt latent spaces. For each $\mathbf{z}$, we plotted the

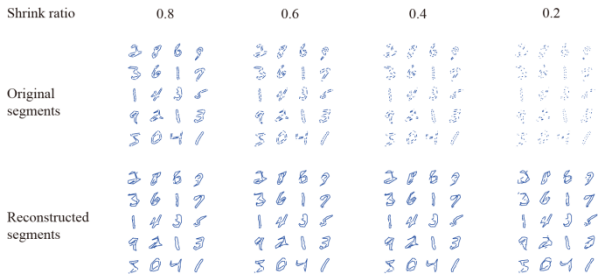

FIGURE 7. The comparison of the reconstruction results between the proposed VSVAE and the original VAE on the MNIST dataset.

corresponding $p_{\theta}(x|\mathbf{z})$ with the learned parameters $\boldsymbol{\theta}$. The capability of the proposed auto-encoder to impose a specified prior distribution $p(\mathbf{z})$ on the coding distribution is studied by comparing with the traditional VAE. The manifold learning results exhibit the variations amongst different classes. It can be seen from our results that the generated samples in the manifold are of obvious variations compared with the traditional ones. That is, the samples are generated in different scales and toward different orientations. It can be observed from the dimension reduction results that samples within a same class tend to be more centralized, which makes it superior than the original VAE from the perspective of separability. One essential issue in the proposed method is the selection of the geometric features. The visualization results of the manifold also significantly contribute to the optimization for a desired combination of geometric representations.

In the presented manifold in Fig. 8, the distributions of digits are considerately separated. Unlike the traditional VAE, where the digit 9s are entangled with digit 4s, the digit 9s are entangled with digit 6s in the proposed method, and digit 6s are entangled with a large proportion of 5s. The reason 9s and 6s are entangled is probably due to their similarity in geometry. Furthermore, digits within a same class are more centralized than the traditional pixel-based model, such as digit 1s and 7s, which are relatively simple in their geometric distributions. In Fig. 9, we presented the separated manifold results for the digit 0s, 1s and 2s. It can be seen from the results that, the segment-based VSVAE results have various geometric deformations in shapes. Besides, the scales and rotations are more obvious than the pixel-based results.

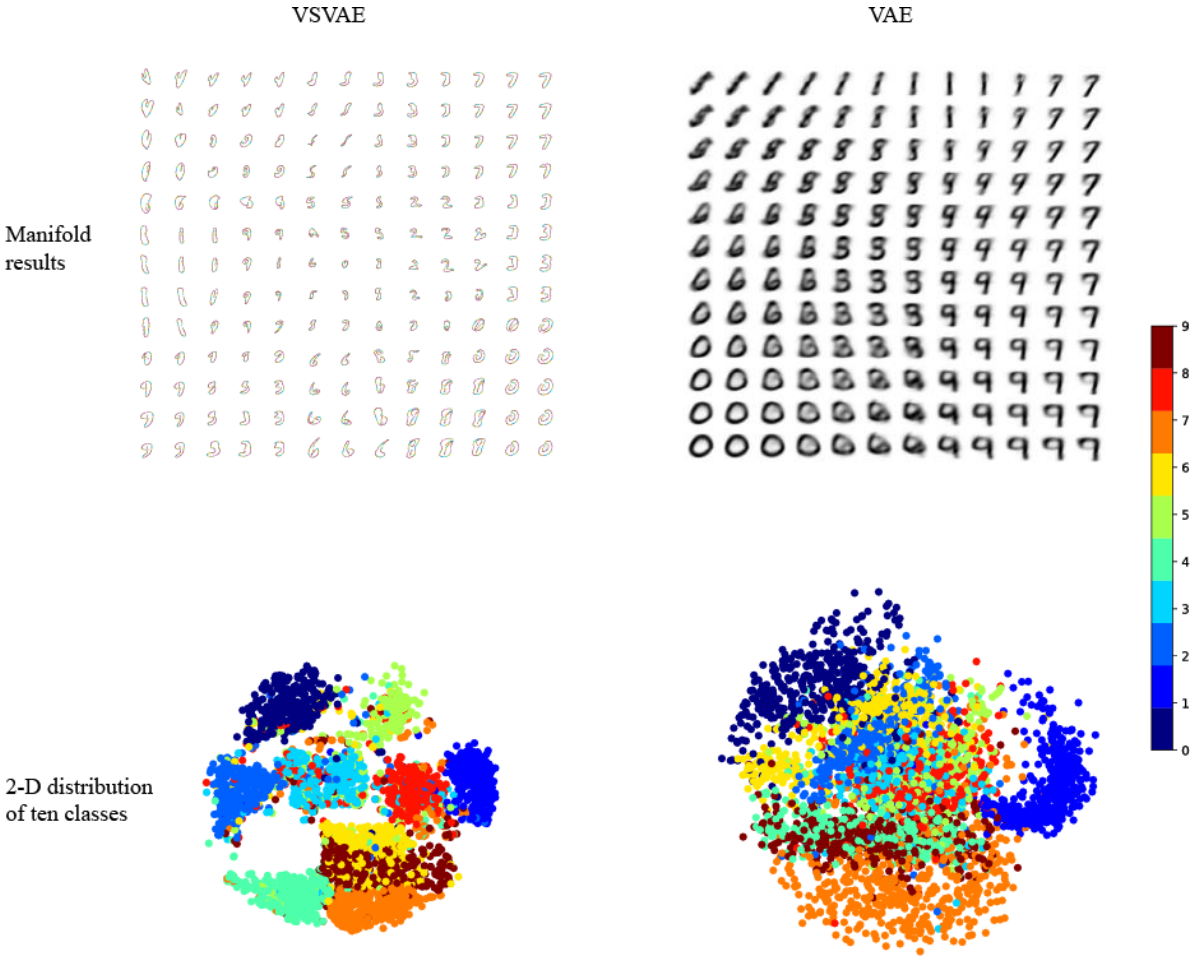

FIGURE 8. Comparison of the visualizations of the learned data manifolds for the generative models with a 2-dimensional latent space on the MNIST dataset.

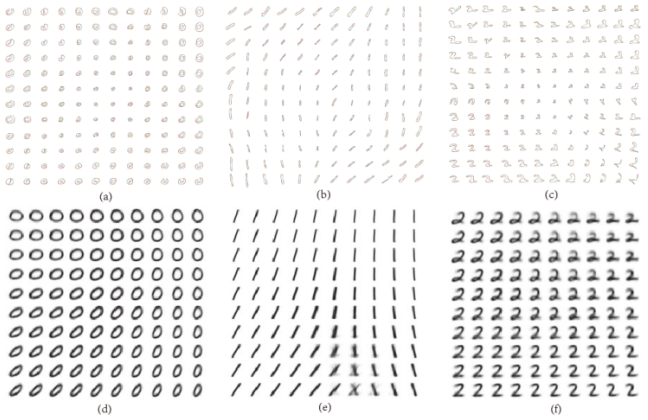

FIGURE 9. The comparisons of the manifold learning results. (a-c): digits of 0s,1s and 2s in 30 segments; (d-f): digits of 0s,1s and 2s in 784 pixels. It is obviously that our model can generate samples with better diversities, especially in scales and rotations.

D. BIOLOGICAL PLAUSIBILITY

Feature learning methods based on deep learning techniques are often regarded as black boxes. In most cases, those models are hard to be explained and not applicable for logical reasoning. Researchers have to rely on experiences and trail-and-errors to design a satisfactory network. Since the geometric features have special characteristics, their visualization results can be cross-validated with the neuro-biological findings. This provides a mutually beneficial research approach for the optimization of the deep learning models. By visualizing the neurons in the hidden layers, it is helpful to realize the simulation of the biological characteristics of the brain visual cortex and the approximation of biological data, and eventually to establish a parameter-tuning method with biological plausibility. Because the geometric features amongst line segments have explicit geometric meanings and topological relations, it is believed that deep hidden neurons learned through this representation can gain a better interpretability. There are many ways to visualize the neurons in the hidden layers. This paper uses the method of maximizing the activation value to find an input that can maximally activate the neuron, which can be regarded as a visual representation of the hidden neuron.

For the proposed geometric representation, effective visualizations of the hidden layers can enhance the interpretability of the model on some level, then can further provide guidance for the geometric feature selection and hyper-parameter configurations. Similar with the pixel-based learning paradigm, the learnt features in the shallow layers are simple and fundamental segments with different orientations. With the depth of the layers goes deeper, the hidden layer visualization exhibits increasingly complicated shapes, which have more concaves and bulges. The visualization results in Fig. 10 reveal the shape selectivity of the learnt features in deep layers, which is accordance with the dataset of shape stimulus used by neurobiology for verifying the shape selectivity of the V4 neurons in the ventral pathway.

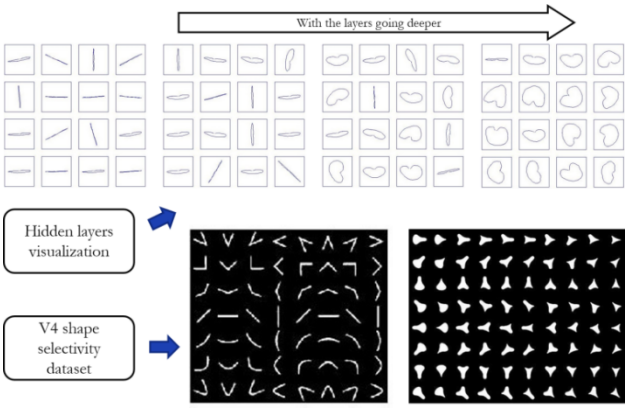

FIGURE 10. The visualizations of the neurons in the hidden layers.

It is found that the orientation columns located in the primary visual cortex form like a pinwheel structure [36], where multiple orientation columns converge. Orientation columns are organized radially around a center known as a singularity. In this study, it is found that the visualizations for the first few hidden layers are formed as a very similar pattern compared with the neurological discovery. The segment-based visualizations in the shallow networks also perform a pinwheel pattern when they are organized in a certain manner. The characteristics can be also performed by the traditional pixel-based method by using sparse coding. The comparison is shown in Fig. 11.

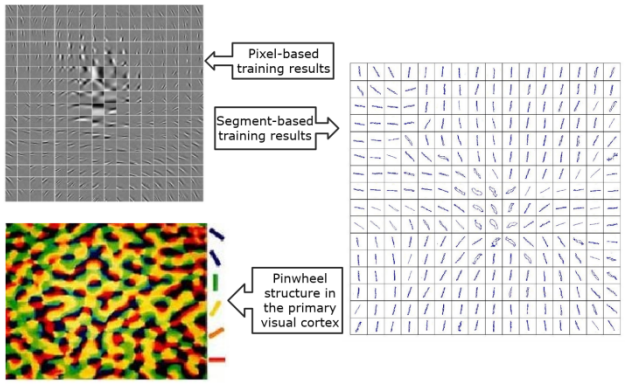

FIGURE 11. Pinwheel structure in the sparse coding. (a) training results based on pixels; (b) training results based on segments, where the orientations of the segment-based visualization are better converged; (c) pinwheel structure in the primary visual cortex.

Overall, the proposed method enables abundant feasibilities to interpret the learning system. Not only the reconstructive capability and manifold results, but also the consistency with the biological evidences can be utilized to select features and configure networks. By using those technical skills, the method is further testified on its classification performances.

E. PERFORMANCE EVALUATION

In this subsection, we firstly evaluate the overall performance of the proposed representation on the SHVN dataset, then the performance of the proposed method against to the

geometric transformations is particularly verified on several transformed versions of the MNIST dataset. The selected features are tuned by analyzing the reconstructions, visualizations and the class separability of the manifold. By utilizing these measures, the networks for classifications can be effectively constructed. The number of normalized segments still plays a vital role to the overall performance. In order to evaluate this important hyper-parameter, we preprocessed both datasets by using different normalization numbers. Hyper-parameters such as regularization parameters, learning rate and weight decay were fine-tuned on the validation sets, whose size is 5000 for both datasets. Stochastic gradient descent is used as the optimization method and the datasets are shuffled after each training iteration. In both tasks, we use a weight decay parameter of 0.0001, a sparsity parameter of 0.1 and the experiments are implemented on two NVIDIA GTX 1070 GPUs.

Unlike the MNIST dataset that has a clean background for each image, the SVHN dataset is more complicated as each digit has a real-world background. Even for human, the recognition rate is around 98%. For this reason, the LSD algorithm gives birth to quite poor results for a small proportion of the training samples in this dataset. In order to avoid negative effects of these samples, we only preserve the images with sufficient detected edges for training. The lowest boundary is set to 20. As a result, approximately 94% of images are remained for training, and all images are remained in the test set for a fair comparison. In this task, three normalization numbers of segments are used on the SVHN dataset, which are 30,45 and 60 respectively. The performance comparison can be viewed in Table 2. It can be seen from the results that our performances generally outperform the other methods. Notably, the number of segments affects the results significantly when the edges are not sufficiently provided.

TABLE 2. Comparison of performance on the SVHN dataset.

Model	No. of segments	Accuracy
VSVAE	30	92.4%
	45	95.8%
	60	95.6%
DBN	-	90.1%
K-means	-	90.6%
RDDL [37]	-	94.5%
CNN [38]	-	95.4%

In this paper, the proposed segment-based representation can acquire better invariances than the traditional pixel-based representations. In order to further evaluate the characteristics of the VSVAE, we implemented a series of experiments in this subsection. Initially, we introduce a set of augmented datasets of the MNIST dataset, where all the images are rotated, scaled and translated to a certain extent. We imposed random variations to the original dataset up to differentiated maximum extents from 0.3 to 0.9, which generate 7 variated datasets *VMNIST0.3* to *VMNIST0.9*. And the *VMNIST1.0* is the original dataset. For each image in the dataset *VMNISTX*, it is randomly rotated, scaled and translated to a certain level that is under the maximum value of X. Fig. 12 is an illustration about how we transform the images for the *VMNISTX* datasets.

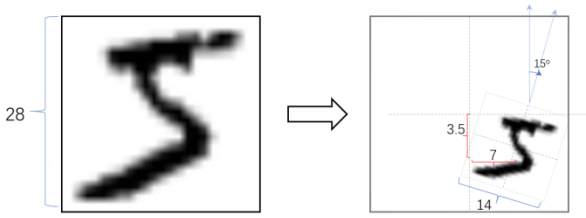

FIGURE 12. An illustration for the image transformations. The image is scaled to one quarter of the original. The rotation ratio is 15 degrees, which is 15/360=1/24. Besides, the scaled image is also translated to a certain extent.

In this task, the number of segments for the digits is normalized to 30. We implement a simple four-layer network for this classification mission. The numbers of neurons for the two hidden layers are set to 100. For each *VMNISTX* dataset, the experiment is run for 100 epochs. Our model is tested on each dataset with comparison to the AlexNet [39], the CapsuleNet [2] and a regular CNN that contains two convolution layers, two pooling layers and two fully-connected layers. Fig.13 shows that our model has an obvious advantage when the dataset is significantly transformed. It can be also noted that the overall performance is optimized when the dataset is slightly transformed.

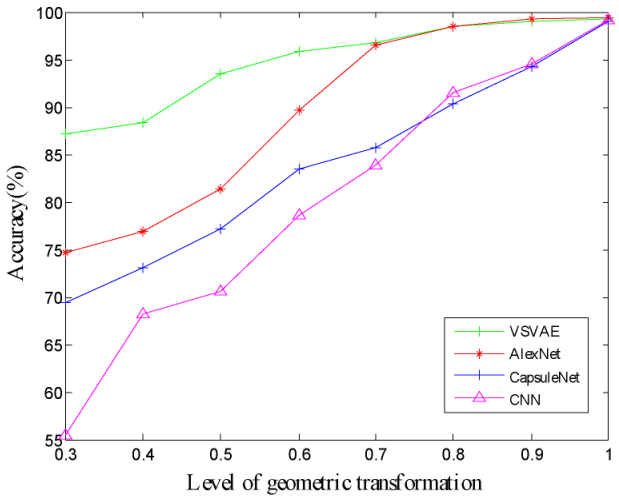

FIGURE 13. Comparison of the classification accuracies amongst the proposed model, CNN, the CapsuleNet [2] and the AlexNet [39].

Furthermore, we also evaluate the training efficiency of the proposed method for the *VMNISTX* datasets by using different batch sizes. Though it costs time to preprocessing images and generating geometric information, it still gets a better computational efficiency as it converges faster than the regular CNNs, which can be seen in Fig. 14. The results

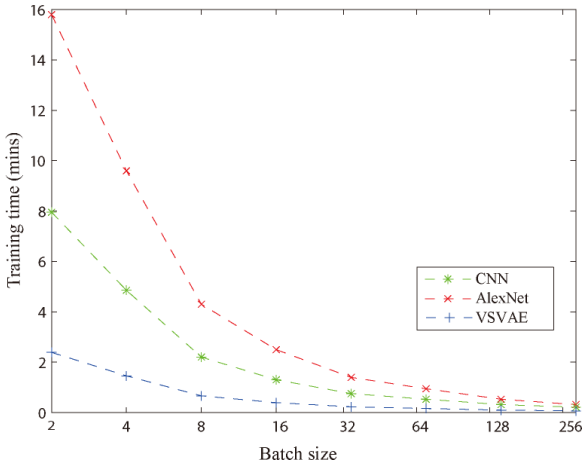

FIGURE 14. Comparison of computation efficiency amongst the proposed model, CNN and the AlexNet [39] by using different batch sizes.

of CapsuleNet is not presented as its training is much more time-consuming than the other methods.

V. CONCLUSIONS

This paper explores the possibility that uses a segment-based representation to build generative models for images. A vertical-stream generative model is proposed based on a variational encoder-decoder structure and a series of geometric transformations. It is discovered that the generated images are of better discrimination for geometric transformations than those pixel-based models. Furthermore, we visualize the hidden neurons as the learnt features to illustrate the biological plausibility of the proposed model. It can be seen from the results that the visualizations perform the same pattern as the neurons functionality in the superior visual functional areas of the visual cortex. The proposed method is not a simple form of data argumentation, but a comprehensive utilization of geometric characteristics for a better interpretability, train-ability and flexibility. It is found that the performances in both accuracy and training efficiency are excellent, especially when the images are significantly transformed.

The computational simulation of the visual cortex mechanism and its applications in machine vision have achieved certain research results. However, the research results of neurobiology and the machine vision models still have not been fully integrated. The study of the visual cortex double-stream mechanism has accumulated a large number of neurobiological achievements that can be utilized for modelling. It is imperative to further utilize those biological evidences to improve the cross-disciplinary research for the existing technical framework. Still, the theoretical proofs of the proposed method are not well studied in this paper. The features are manually selected by observing the visualizations and the characteristics of the explicit geometric representations. By combining the biological plausibility and effective mathematical measures, this work can be further extended to establish a self-adapted system.

REFERENCES

[1] D. H. Kelly, "Motion and vision. II. Stabilized spatio-temporal threshold surface," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 69, no. 10, pp. 1340–1349, Oct. 1979.
[2] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.
[3] H. Nakagama and S. Tanaka, "Self-organization model of cytochrome oxidase blobs and ocular dominance columns in the primary visual cortex," *Cerebral Cortex*, vol. 14, no. 4, pp. 376–386, 2004.
[4] A. Anzai, X. Peng, and D. C. Van Essen, "Neurons in monkey visual area V2 encode combinations of orientations," *Nature Neurosci.*, vol. 10, no. 10, pp. 1313–1321, 2007.
[5] A. Pasupathy and C. E. Connor, "Responses to contour features in macaque area V4," *J. Neurophysiol.*, vol. 82, no. 5, pp. 2490–2502, 1999.
[6] H. Goh, N. Thome, M. Cord, and J.-H. Lim, "Learning deep hierarchical visual feature coding," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 12, pp. 2212–2225, Dec. 2014.
[7] M. Szabo, M. Stetter, G. Deco, S. Fusi, P. Del Giudice, and M. Mattia, "Learning to attend: Modeling the shaping of selectivity in infero-temporal cortex in a categorization task," *Biol. Cybern.*, vol. 94, no. 5, pp. 351–365, 2006.
[8] S. L. Brincat and C. E. Connor, "Underlying principles of visual shape selectivity in posterior inferotemporal cortex," *Nature Neurosci.*, vol. 7, no. 8, pp. 880–886, 2004.
[9] D. H. Hubel and T. N. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *J. Physiol.*, vol. 148, pp. 574–591, 1959.
[10] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neurosci.*, vol. 2, no. 11, pp. 1019–1025, Nov. 1999.
[11] T. Serre, A. Oliva, and T. Poggio, "A feedforward architecture accounts for rapid categorization," *Proc. Nat. Acad. Sci. USA*, vol. 104, no. 15, pp. 6424–6429, 2007.
[12] H. Wei, Q. Li, and Z. Dong, "Learning and representing object shape through an array of orientation columns," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 7, pp. 1346–1358, Jul. 2014.
[13] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, and X. Xue, "Multi-stream multi-class fusion of deep networks for video classification," in *Proc. ACM Multimedia Conf.*, 2016, pp. 791–800.
[14] X.-S. Tang, K. Hao, H. Wei, and Y. Ding, "Using line segments to train multi-stream stacked autoencoders for image classification," *Pattern Recognit. Lett.*, vol. 94, pp. 55–61, Jul. 2017.
[15] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 1, 2014, pp. 568–576.
[16] G. Olague, D. E. Hernández, E. Clemente, and M. Chan-Ley, "Evolving head tracking routines with brain programming," *IEEE Access*, vol. 6, pp. 26254–26270, 2018.
[17] P. Parodi and G. Piccioli, "3D shape reconstruction by using vanishing points," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 2, pp. 211–217, Feb. 1996.
[18] S. Makrogiannis, G. Economou, S. Fotopoulos, and N. G. Bourbakis, "Segmentation of color images using multiscale clustering and graph theoretic region synthesis," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 35, no. 2, pp. 224–238, Mar. 2005.
[19] H. Wei and X.-S. Tang, "A genetic-algorithm-based explicit description of object contour and its ability to facilitate recognition," *IEEE Trans. Cybern.*, vol. 45, no. 11, pp. 2558–2571, Nov. 2015.
[20] X. S. Tang and H. Wei, "A segment-wise prediction based on genetic algorithm for object recognition," *Neural Comput. Appl.*, 2017, doi: 10.1007/s00521-017-3189-z.
[21] R. G. von Gioi, J. Jakubowicz, J.-M. Morel, and G. Randall, "LSD: A fast line segment detector with a false detection control," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 722–732, Apr. 2010.
[22] X. Liu, Z. Cao, N. Gu, S. Nahavandi, C. Zhou, and M. Tan, "Intelligent line segment perception with cortex-like mechanisms," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 45, no. 12, pp. 1522–1534, Dec. 2015.
[23] L. J. Latecki and R. Lakämper, "Shape similarity measure based on correspondence of visual parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1185–1190, Oct. 2000.
[24] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3982–3991.
[25] W. Ke, J. Chen, and Q. Ye, "Deep contour and symmetry scored object proposal," *Pattern Recognit. Lett.*, 2018, doi: 10.1016/j.patrec.2018.01.004.

[26] J. Wu, C. Wang, L. Zhang, and Y. Rui, "Offline sketch parsing via shapeness estimation," in *Proc. Int. Conf. Artif. Intell.*, 2015, pp. 1200–1206.
[27] C. Xiao, C. Wang, L. Zhang, and L. Zhang, "Sketch-based image retrieval via shape words," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2015, pp. 571–574.
[28] A. Akram, N. Wang, J. Li, and X. Gao, "A comparative study on face sketch synthesis," *IEEE Access*, vol. 6, pp. 37084–37093, 2018.
[29] S. Setumin and S. A. Suandi, "Difference of Gaussian oriented gradient histogram for face sketch to photo matching," *IEEE Access*, vol. 6, pp. 39344–39352, 2018.
[30] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–14.
[31] A. Makhzani *et al.*, "Adversarial autoencoders," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–16.
[32] A. Makhzani and B. J. Frey, "PixelGAN autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1975–1985.
[33] V. L. Cao, M. Nicolau, and J. McDermott, "Learning neural representations for network anomaly detection," *IEEE Trans. Cybern.*, 2018, doi: 10.1109/TCYB.2018.2838668.
[34] S. Park, S. Yu, M. Kim, K. Park, and J. Paik, "Dual autoencoder network for retinex-based low-light image enhancement," *IEEE Access*, vol. 6, pp. 22084–22093, 2018.
[35] Y. Pu *et al.*, "Variational autoencoder for deep learning of images, labels and captions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2352–2360.
[36] T. Bonhoeffer and A. Grinvald, "Iso-orientation domains in cat visual cortex are arranged in pinwheel-like patterns," *Nature*, vol. 353, pp. 429–431, Oct. 1991.
[37] V. Singhal, H. K. Aggarwal, S. Tariyal, and A. Majumdar, "Discriminative robust deep dictionary learning for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 9, pp. 5274–5283, Sep. 2017.
[38] P. Sermanet, S. Chintala, and Y. Lecun, "Convolutional neural networks applied to house numbers digit classification," in *Proc. Int. Conf. Pattern Recognit.*, Nov. 2013, pp. 3288–3291.
[39] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

Authors' photographs and biographies not available at the time of publication.