SSC-VAE: Structured Sparse Coding Based Variational Autoencoder for Detail Preserved Image Reconstruction

Hao Wang[*1], **Lu Wang**[*2], **Zhongyu Wang**[1],
Lixin Ma[1], **Ye Luo**[1†]

[1]School of Computer Science and Technology, Tongji University
[2]Institute for Infocomm Research (I2R), Agency for Science, Technology and Research (A*STAR)
2051475@tongji.edu.cn, wang_lu@i2r.a-star.edu.sg, 2152197@tongji.edu.cn, 2153085@tongji.edu.cn, yeluo@tongji.edu.cn

Abstract

Discrete latent representation techniques, such as Vector Quantization (VQ) and Sparse Coding (SC), have demonstrated superior image reconstruction and generation quality compared to continuous representation methods in Variational Autoencoders (VAEs). However, existing approaches often treat the latent representations of an image independently in their discrete representation space, neglecting both the inherent structural information within each representation and the correlations among them. This oversight leads to coarse representations and suboptimal generated results. In this paper, we address these limitations by introducing correlations among and within the latent representations of individual images in the latent discrete space of VAEs using sparse coding. We impose two-dimensional structural information through adaptive thresholding, enhancing local structure in image representations while suppressing noise via parsimonious representation with a learned dictionary. Empirical studies on three real benchmark datasets, including a clinical Ultrasound dataset, BSDS500, and mini-Imagenet, demonstrate that our proposed model preserves fine-grained details in image reconstruction and significantly outperforms baseline models of SC-VAE and VQ-VAE across objective and subjective image quality metrics. Particularly noteworthy are the substantial performance improvements observed on the ultrasound dataset, where structure information is crucial. Specifically, we observe significant performance improvements of 7.68 % and 17.03 % in SSIM, 3.25 dB and 6.58 dB in PSNR, 0.15 and 0.24 in LPIPS, 45.38 and 84.05 in FID over SC-VAE and VQ-VAE, respectively, indicating the superiority of our method in terms of image reconstruction quality and fidelity.

Code — https://github.com/rmEleven/SSC-VAE

1 Introduction

Learning rich representations without supervision is a fundamental and ongoing challenge in the field of deep learning. While supervised learning has achieved remarkable success by leveraging labeled data, the availability of large-scale labeled datasets is often limited, expensive, or impractical in many real-world scenarios. Unsupervised learning methods aim to address this limitation by extracting meaningful and useful representations directly from unlabeled data.

Variational autoencoders (Kingma and Welling 2014) offer an elegant solution for learning meaningful image representations by a continuous low dimensional latent vector under a predefined static prior. However, traditional VAEs often struggle to generate rich textures and tend to produce overly smooth images. To address these limitations, Variational Autoencoder with Vectored Quantization (VQ-VAE (van den Oord, Vinyals, and Kavukcuoglu 2017)) was introduced to generate a discrete representation that can be decoded to reconstruct the original data. The main advantage of VQ-VAE is its ability to separate relevant information from noise, making it suitable for tasks that require robust and compact representations. VQ-VAE loses local details when the code book is not large enough as it uses the same quantization index to embed similar image patches. Instead of using one quantization index for one image patch, Sparse Coding-based Variational Autoencoder (SC-VAE (Xiao, Qiu, and Sotiras 2023)) improves VQ-VAE by considering the latent representation as sparse linear combinations of atoms using sparse representation, thus retaining the subtle difference among similar patches. The activeness of multiple atoms enables to capture more complex dependencies and greatly reduces the code book size. However, the latent representations in both VQ-VAE and SC-VAE are treated independently when performing the coding and decoding, overlooking the inherent correlations present in the images.

We have observed significant correlations in the latent representations of individual images, both across image patches, i.e., spatial correlation, and across different atoms within each patch, i.e., internal dictionary atom correlation. As depicted in Figure 1, the latent representations in the second column demonstrate a notable correlation across various regions of both the natural and clinical ultrasound images, evidenced by their similar intensity levels. To illustrate the internal correlation across different dictionary atoms within single latent representation, we visualize its thresholds when discretized in sparse coding as a gray-level image map, considering all representations together. The thresholds, which regulate the sparsity of the coding of the latent feature, exhibit similar structures across atoms in the coding dictionary,

[*]These authors contributed equally.
[†]Corresponding author
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

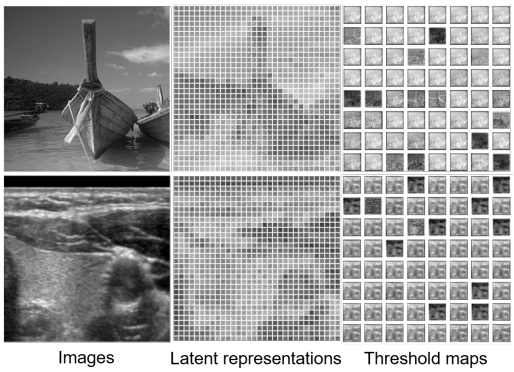

Figure 1: Visualization of the latent presentation correlations. The first column gives two sample images. The second column displays their latent representations across different image regions, which show high correlations in terms of similar intensity levels. The third column presents the threshold maps used in sparse coding discretization, revealing the high similarities among atoms and their particularly structural pattern.

as shown in the third column of Figure 1. These observations inspired us to explicitly model the two correlations in latent space thus to retain more local structures and details of the reconstructed images.

To incorporate two-dimensional structural information into the discretization of latent representations, we propose integrating an inference and refinement (two-step) controller for structured sparsity into the SC-VAE framework. This allows adaptive adjustment of thresholds in sparse coding, enhancing local structure in image representations while suppressing noise using a learned dictionary.

In the first step, we introduce a convolutional network responsible for inferring thresholds from the latent representations. This enables fine-grained control over representation sparsity, resulting in semantically related and information-rich thresholds that preserve intricate relationships and dependencies within the latent representations. In the second step, inspired by AKSVD (Liang et al. 2023), we refine sparsity using the Convolutional Block Attention Module (CBAM) (Woo et al. 2018), which integrates a refinement mechanism by applying channel-wise and spatial attention to the thresholds inferred in the first step. Channel-wise attention is aimed at capturing internal correlations within the representations, while spatial attention captures correlations among different representations.

Our main contributions can be summarized as follows:

- We introduce additional sparsity controller inference and refinement mechanism to SC-VAE, which we dubbed as Structured SC-VAE (SSC-VAE for short), where two-dimensional correlation are imposed through adaptive threshold in sparse coding to obtain image output with enhanced details.
- Spatial correlation among the latent representations helps to preserve weak but common structures across the different regions of one image.
- Internal correlation within a latent representation helps to enhance the dictionary learning by considering jointly the structure across the atoms and removing irrelevant information through structured sparsity.
- Experiments on different real-life datasets show state-of-the-art performance in image reconstruction and denoising, especially on the clinic ultrasound dataset.

2 Related Work

2.1 Continuous latent representation based VAE

The original Variational Autoencoder framework, introduced by Kingma and Welling (2013)(Kingma and Welling 2014), utilizes continuous variables to represent underlying data structures, laying foundation for subsequent development of various VAE variants. Research efforts have focused on refining the latent space in VAEs to enhance the model's understanding of semantic information in images. Kingma et al. (Kingma and Welling 2014) introduced β-VAE, incorporating a parameter β to regulate latent space smoothness, improving image reconstructions. Higgins et al. (Higgins et al. 2017) proposed β-TCVAE (Chen et al. 2018), which adds total correlation loss to foster latent representation independence, enhancing image disentanglement. Nouveau VAE (NVAE) (Vahdat and Kautz 2020) introduced a hierarchical architecture, improving generative model performance and image reconstruction quality. AEGAN (Wang et al. 2018) merges VAE and GAN, addressing blurry samples and mode collapse. The Multimodal and Dynamical Variational Autoencoder (MDVAE) (Leglaive 2022) organizes the latent space to differentiate between shared and unique modality factors in audiovisual speech representations, alongside a static variable encoding consistent information over time. Razavi et al. (Razavi et al. 2019) suggested posterior collapse arises from the discrepancy between the prior and approximated posterior distributions, and δ-VAE (Razavi et al. 2019) was introduced to mitigate this with a mean field posterior and correlated prior distribution.

In image reconstruction, Rezende et al. introduced VAE with PixelCNN Decoders (Rezende, Mohamed, and Wierstra 2014), enhancing image detail and realism. Maaløe et al. proposed ADGM (Maaløe et al. 2016), which models the underlying image probability distribution and refines image quality through layer-wise backpropagation. For image denoising, VAEs encode noisy images into latent variables and decode them back, effectively removing noise while preserving key features. NAVE (Cui et al. 2022) employs a Nouveau variational autoencoder with quantile regression loss for PET image denoising, enabling simultaneous uncertainty estimation.

2.2 Discrete latent representation based VAE

In contrast to continuous counterparts, Variational Autoencoders (VAEs) based on discrete latent representations utilize discrete latent variables, offering structured representations well-suited for tasks requiring explicit manipulation of categorical features, making them crucial for preserving semantic information. The Vector Quantized Variational Autoencoder (VQ-VAE), introduced by Razavi et al. (van den

Oord, Vinyals, and Kavukcuoglu 2017), encodes images into discrete latent representations using vector quantization, providing a compact and interpretable image representation. This allows for efficient storage and manipulation of image features, enhancing tasks like image generation and manipulation. Since its inception, various iterations (Esser, Rombach, and Ommer 2021)(Yu et al. 2022)(Zheng et al. 2022) have emerged to improve image reconstruction fidelity. A notable advancement is the Vector Quantized Generative Adversarial Network (VQ-GAN) by Esser et al. (Esser, Rombach, and Ommer 2021), which combines VQ-VAE with adversarial GAN training, achieving high-fidelity image generation with enhanced realism and diversity. Additionally, Ladder VAE, introduced by Kingma et al. (Kingma et al. 2014), presents a hierarchical VAE framework incorporating discrete representations at multiple levels, improving the model's ability to reconstruct images with rich features.

In image denoising tasks, discrete VAEs encode noisy images into discrete latent representations and then decoding them back, these models can effectively remove noise while preserving important categorical information. DVAE(Pu et al. 2016) attempts to train probabilistic models with discrete latent variables within the VAE framework, enabling backpropagation through the discrete latent variables and capturing both class information and pixel-level details from unsupervised data.

Our proposed SSC-VAE deals with both the tasks of image reconstruction and denoising by further considering the correlations among and within the discrete latent representations and achieve enhanced details and local structure preservation.

3 Structured Sparse-coding VAE

3.1 Preliminary Knowledge

Sparse coding is a technique used in signal processing and machine learning for representing data in terms of a sparse combination of basis elements or atoms. It aims to find a sparse representation that captures the essential features of the data. Specifically, for an input $X \in \mathbb{R}^n$, the optimal sparse code vector $Z \in \mathbb{R}^m$ need to be found to minimize the energy function that combines the square reconstruction error and the L_1 sparsity penalty:

$$E(X, Z) = 0.5\|X - DZ\|_2^2 + \alpha\|Z\|_1. \tag{1}$$

where $D \in \mathbb{R}^{n\times m}$ is a dictionary matrix whose columns are the atoms, α is a coefficient controlling the degree of sparsity penalty. The goal of sparse coding can be written as the following optimization problem:

$$Z^* = \arg\min_Z E(X, Z). \tag{2}$$

ISTA (Daubechies, Defrise, and De Mol 2004) is an iterative optimization algorithm widely used in sparse coding. It iterates the following recursive equation until it converges:

$$Z(t+1) = h_\theta(W_e X + SZ(t)), \quad Z(0) = 0. \tag{3}$$

And the related variables in Equation (3) can be defined consequently as:

$$\begin{aligned} \text{filter matrix:} \quad & W_e = \frac{1}{L}D^T \\ \text{mutual inhibition matrix:} \quad & S = I - \frac{1}{L}D^T D \\ \text{shrinkage function:} \quad & [h_\theta(V)]_i = \text{sign}(V_i)(|V_i| - \theta)_+ \end{aligned}$$

where L is a constant defined as an upper bound on the largest eigenvalue of $D^T D$. The function $h_\theta(V)$ is a component-wise vector shrinkage function with thresholds θ which are set to $\frac{\alpha}{L}$. The symbol i represents the different indices of the function's input. The sign function is used to perform the sign operation, while $|.|$ means to take the absolute value.

LISTA (Gregor and LeCun 2010) is a learned variant of ISTA that leverages deep learning techniques to improve sparse coding and signal reconstruction. It treats filter matrix, mutual inhibition matrix and θ in shrinkage function as learnable parameters. The architecture of LISTA can be viewed as a feed-forward network in which S is shared over layers and can be trained end-to-end using backpropagation to optimize the performance of the sparse coding task. LISTA has been successfully applied in various research domains, including image and signal processing. Its ability to learn an adaptive thresholding operation allows for improved sparse coding and reconstruction quality compared to traditional handcrafted methods like ISTA (Daubechies, Defrise, and De Mol 2004).

3.2 SSC-VAE Network

Overall Architecture Figure 2 describes the overall architecture of our method. $X \in \mathbb{R}^{C\times H\times W}$ is the input image, $E \in \mathbb{R}^{C'\times h\times w}$ represents latent representations, $Z \in \mathbb{R}^{K\times h\times w}$ is the sparse code for the representations. α is the threshold in the LISTA, α^+ is the refined threshold with the same dimension as Z. $\tilde{E}$ and $\tilde{X}$ are the reconstructed representations and image, respectively.

The encoder processes the input image to generate latent vector representations. Unlike SC-VAE, which directly applies LISTA to discretize the latent vector representations, our approach integrates inference and refinement, a two-step structured sparsity controller, to adaptively learn the thresholds of LISTA for sparse coding. We extend the simple scalar sparsity controller used in SC-VAE to a vector semantically related to each local features, enabling more fine-grained control and generating information-rich sparse codes. Once we obtain the sparse codes, the latent vector representations can be reconstructed by a linear combination of atoms in the dictionary. The reconstructed representations are then fed into the decoder, which processes them to generate the desired image.

Sparsity Controller Inference SC-VAE utilizes a scalar sparse coefficient to compute the sparse coding of a latent vector representation, as shown in the formula below:

$$z_k^* = \arg\min \|E_k - Dz_k\|_2^2 + \alpha_k\|z_k\|_1, \tag{4}$$

where E represents the latent representation; z denotes the sparse coding; z^* represents the optimal sparse code to

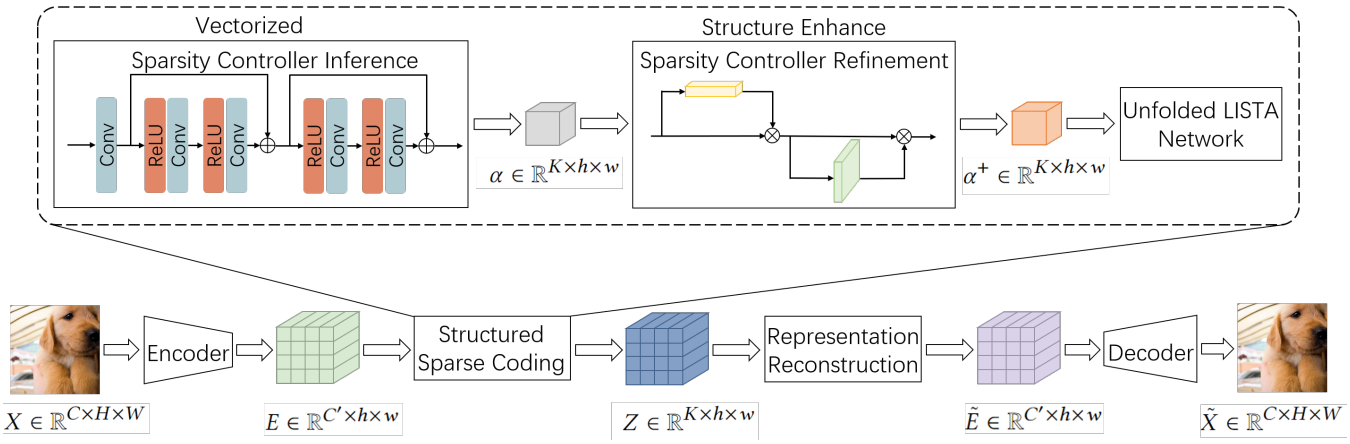

Figure 2: The overall architecture of the proposed Structured SC-VAE network. The Sparsity Controller Inference module extends scalar sparsity into vectorized sparsity, enabling diverse sparsity in the sparse representation and thereby capturing potential structures. The Sparsity Controller Refinement module introduces correlations both within and among image representations during sparse coding. Together, these modules mimic an adaptive sparse coding process, allowing the model to preserve more structural and detailed information.

estimate;z_k is the sparse code of the kd-th latent representations of the image; α is the sparsity threshold, which is a scalar for each vector z_k ; D is the dictionary.

We argue that it is insufficient to capture the structural characteristics of a latent representation using a scalar to control the sparsity of a vector. We propose to extend the scalar sparse controller $\boldsymbol{\alpha_k} = [\alpha_{k,1}, \alpha_{k,2}, ..., \alpha_{k,m}]^T$ of the same dimensionality as the sparse coding z_k, which allows diverse sparsity for each elements in z_k. This is equivalent to a modified objective with adaptive thresholding:

$$z_k^* = \arg\min\{\|E_k - Dz_k\|_2^2 + \sum\nolimits_i \alpha_{k,i}\|z_{k,i}\|_1\}, \quad (5)$$

where $\alpha_{k,i}$ represents the i-th elements in α_k.

The reconstruction loss can be define as:

$$L_{recon} = \|G(\tilde{E}(x)) - x\|_2^2, \quad (6)$$

where x is the input image, G is the decoder network. Each vector of reconstructed latent representation $\tilde{E}(x)$ can be calculated by the multiplication of sparse code z_k and dictionary D. And latent loss calculates the error in the sparse learning to reconstruct the latent representations:

$$L_{latent} = \frac{1}{n}\sum\nolimits_k \left(\|E_k(x) - Dz_k\|_2^2 + \sum\nolimits_i \alpha_{k,i}\|z_{k,i}\|_1\right), \quad (7)$$

where E is the encoder network, and n denotes the number of latent representations.

The final loss of our SSC-VAE simply combines the reconstruction loss and the latent loss:

$$Loss = L_{recon} + L_{latent}. \quad (8)$$

We propose using a convolutional network to infer the threshold vector from the latent representation. Leveraging the exceptional performance of the residual connection (He et al. 2016) in visual tasks, we apply it to the inference of the threshold vector. Specifically, a convolutional layer is used to adjust the dimensionality of the representation to match that of the sparse code. Then, through two residual blocks, each consisting of two convolutional layers with ReLU (Glorot, Bordes, and Bengio 2011) activation function, together with the input data itself, the initial inference is obtained as shown in the Sparsity Controller Inference module in Figure 2.

Sparsity Controller Refinement Till now, the discretization of each latent representation has been independently processed by LISTA, neglecting the correlation among different representations. Moreover, elements in the threshold vector are independently learned during the training process, failing to fully exploit the internal correlation within the representations. Therefore, we introduce Sparsity Controller Refinement to capture the overlooked two-dimensional structural information.

Our approach draws inspiration from the design of the CBAM (Woo et al. 2018) by imposing two-dimensional correlation mechanism within the LISTA network. Specifically, channel-wise and spatial attention are integrated on the initially inferred thresholds. The internal correlation within the representations is achieved through the application of channel-wise attention across the dictionary atoms. To capture the correlation information, we employ both average pooling and max pooling operations to reduce the dimensionality of the threshold vectors, resulting two descriptors with the same dimensions as the threshold vector. After the pooling operations, the descriptors are passed through a shared three-layer perceptron to calculate the channel-wise attention weights W_c:

$$W_c = \text{sig}(\text{MLP}(\text{AvgPool}(\alpha)) + \text{MLP}(\text{MaxPool}(\alpha))) \quad (9)$$

where *sig* is the sigmoid activation function, *MLP* is the multi-layer perceptron, α is the initially inferred threshold,

AvgPool and *MaxPool* are the average pooling and max pooling operation respectively.

The output of the channel-wise attention is further processed by the spatial attention, which captures the correlations among representations and reflects them in the thresholds. To achieve this, we utilize average pooling and max pooling operations to extract two quantities for each threshold vector, resulting in two feature maps. These two feature maps are concatenated and fed through a convolutional layer, followed by the sigmoid activation function to calculate the spatial attention weights W_s:

$$W_s = \text{sig}(f([\text{AvgPool}(\alpha); \text{MaxPool}(\alpha)])) \quad (10)$$

where f represents the convolution operation and $[;]$ represents the concatenation operation.

Therefore, the adapted threshold is achieved through:

$$\alpha^+ = W_s \otimes (W_c \otimes \alpha) \quad (11)$$

where α^+ denotes the refined threshold, $\otimes$ denotes element-wise product. The refined threshold will be fed into the unfolded LISTA network to calculate sparse coding while preserving the semantic information among and within the latent space.

4 Experiments and Analysis

In this section, image reconstruction and denoising tasks are used to showcase the superiority of the proposed model. The effectiveness of the proposed Inference and Refinement modules is further validated by successively introducing them to the baseline SC-VAE in the ablation studies.

4.1 Datasets

We evaluated our model on three public datasets: **Ultrasound**, **BSDS500** and **mini-Imagenet**.

Ultrasound is a clinic ultrasound dataset from the "Ultrasound Image Enhancement Challenge" organized as part of the 26th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). It includes 1,050 images of five different organs: breast, carotid, kidney, liver and thyroid[1].

BSDS500 short for the Berkeley Segmentation Dataset and Benchmark 500 (Mubashar et al. 2022), is a widely used benchmark dataset in computer vision. It comprises 500 images that cover a wide range of scenes and objects, including indoor and outdoor environments, people, animals, and natural landscapes.

Mini-Imagenet is a subset of the ImageNet dataset, which is a large-scale dataset commonly used in object recognition tasks with 60,000 images (Vinyals et al. 2016).

4.2 Implementation Details

The baseline model, SC-VAE adopts the encoder and decoder architecture of VQGAN (Esser, Rombach, and Ommer 2021). Our proposed SSC-VAE was built on the baseline with configuration reported the best performance in (Xiao, Qiu, and Sotiras 2023). In particular, the number of downsampling blocks of the encoder was set to 3, resulting in 32×32 latent vector representations for an input with a resolution of 256×256. The decoder contained 3 upsampling blocks to recover the dimensions of the input from latent vector representations. Additionally, for the purpose of comparison, the VQ-VAE and VQ-GAN also utilized the same encoder and decoder. The number of atoms in the dictionary for SSC-VAE and SC-VAE, and the size of the discrete latent space for VQ-VAE and VQ-GAN, were set to 512. The dimension of latent vector representations for all models was set to 256. For SSC-VAE and SC-VAE, the number of rollout steps in LISTA was set to 16. In the case of VQ-VAE and VQ-GAN, the commitment loss coefficient β was set to 0.25, which was also the choice in (van den Oord, Vinyals, and Kavukcuoglu 2017). The best models for testing are chosen with the lowest validation loss. We conducted our experiments in PyTorch framework on one RTX 4090 GPU with 24GB memory.

4.3 Evaluation Metrics

In order to evaluate the quality between reconstructed images and original images, we adopted Structural Similarity Index Measure (SSIM), Peak Signal to Noise Ratio (PSNR), Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), and Fréchet Inception Distance (FID) (Heusel et al. 2017) as evaluation metrics.

4.4 Image Reconstruction

We conducted image reconstruction experiments on the ultrasound dataset, BSDS500, and mini-Imagenet to investigate the superior ability of our proposed SSC-VAE model in preserving image details.

The ultrasound dataset contains 840 training images and 105 validation/test images each, all in 256×256 resolution. The BSDS500 dataset is split into 400 training and 50 validation/test images each, with 256×256 sub-images cropped via sliding-window processing. For mini-Imagenet, resolution filtering (>256px) yields 33,388 training, 8,960 validation, and 10,518 test images, using random cropping during training/validation and sliding-window cropping for testing.

Models of SSC-VAE, SC-VAE, VQ-VAE, VQ-GAN and VAE were trained on the ultrasound dataset for 400 epochs and on BSDS500 and mini-Imagenet for 200 epochs.

Quantitative Results: We evaluated the reconstructed images output by the model against the input images, in order to measure the model's performance on the image reconstruction task. Table 1 presents the results of our model and the baselines trained on three different datasets. These values are averages calculated after evaluating each image in the testing set. The best values for each metric are highlighted in bold.

It can be observed that our SSC-VAE model consistently outperforms the SC-VAE, VQ-VAE, VQ-GAN and VAE baselines across multiple datasets and evaluation metrics. Specifically, on ultrasound dataset, we observe significant performance improvements of 7.68 % and 17.03 % in SSIM, 3.25 dB and 6.58 dB in PSNR, 0.15 and 0.24 in LPIPS, 45.38

[1]https://grand-challenge.org/forums/forum/ultrasound-image-enhancement-challenge-2023-692/topic/dataset-1413/

Model	Dataset	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓
VAE	Ultra-sound	0.26	14.49	0.66	328.04
VQVAE		0.77	29.20	0.29	101.88
VQGAN		0.77	29.18	0.31	108.61
SCVAE		0.87	32.53	0.20	63.21
SSCVAE		**0.94**	**35.78**	**0.05**	**17.83**
VAE	BSDS-500	0.34	13.39	0.69	417.74
VQVAE		0.65	24.33	0.32	160.46
VQGAN		0.62	23.82	0.38	172.96
SCVAE		0.87	28.97	0.08	36.22
SSCVAE		**0.94**	**31.19**	**0.02**	**16.40**
VAE	mini-Imagenet	0.33	11.46	0.73	375.37
VQVAE		0.70	24.41	0.29	37.47
VQGAN		0.65	22.05	0.31	39.10
SCVAE		0.93	31.20	0.02	1.53
SSCVAE		**0.97**	**35.48**	**0.01**	**0.80**

Table 1: Quantitative results on three employed datasets for image reconstruction.

σ	Model	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓
10	VAE	0.26	14.47	0.66	374.26
	VQVAE	0.74	28.72	0.31	111.52
	VQGAN	0.75	28.87	0.31	113.39
	SCVAE	0.80	30.20	0.28	107.20
	SSCVAE	**0.88**	**33.16**	**0.14**	**44.55**
20	VAE	0.26	14.49	0.65	322.44
	VQVAE	0.74	28.71	0.32	114.38
	VQGAN	0.74	28.78	0.32	111.58
	SCVAE	0.79	30.00	0.28	103.28
	SSCVAE	**0.83**	**31.09**	**0.20**	**63.78**
30	VAE	0.26	14.48	0.67	339.63
	VQVAE	0.72	28.18	0.34	120.34
	VQGAN	0.71	28.29	0.34	114.21
	SCVAE	0.76	29.13	0.30	109.44
	SSCVAE	**0.78**	**29.72**	**0.25**	**77.71**

Table 2: Quantitative results on ultrasound dataset for image denoising with different noise levels σ.

and 84.05 in FID over SC-VAE and VQ-VAE, respectively, indicating the superiority of our method in terms of image reconstruction quality and fidelity.

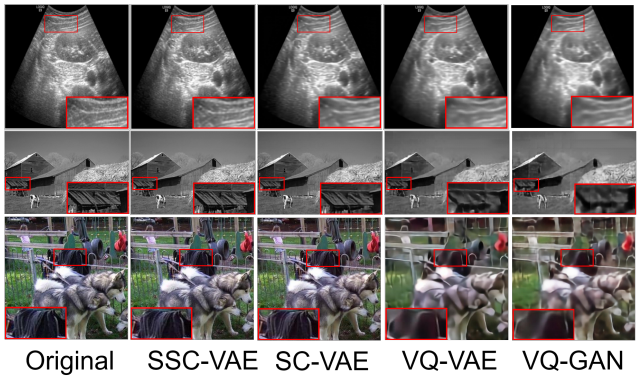

Figure 3: Image reconstruction performance by different models.

Qualitative Results: To more intuitively compare the reconstruction quality of different models, Figure 3 showcases the reconstructed images and real images for different models on 3 different datasets. The outputs of VAE are not displayed due to the poor reconstruction quality. More results can be found in the supplementary material. It can be observed that the VQ-VAE and VQ-GAN models fail to preserve rich detailed information compared to SSC-VAE and SC-VAE. Compared to the ground truth images, the reconstruction images generated by VQ-VAE and VQ-GAN appear more blurry, particularly in regions with complex image structures. The results of SC-VAE show a significant improvement compared to VQ-VAE and VQ-GAN, as it largely preserves the details present in the images. In contrast, due to the introduction of structure preservation layers, SSC-VAE excels in retaining fine-grained details in the images. The reconstruction results achieved by SSC-VAE exhibit higher visual quality and showcase a remarkable preservation of intricate features. Our proposed SSC-VAE model successfully captures and reproduces the intricate details of the input images, resulting in reconstructions that closely resemble the original images.

4.5 Image Denoising

To investigate the performance of the model on the image denoising task, we added Gaussian noise with standard deviations of 10, 20, and 30 to the ultrasound images. For each noise level, we trained our proposed SSC-VAE model as well as baseline models. Models received noisy input images and predicted outputs to approximate the the clean, ground-truth images. The dataset partitioning and preprocessing were the same as those used in the image reconstruction experiments. All models were trained for 400 epochs on the ultrasound dataset, 150 epochs on the BSDS500 dataset, and 75 epochs on the mini-Imagenet dataset.

Quantitative Results: Table 2 presents the denoising performance on the ultrasound dataset using SSC-VAE, SC-VAE, VQ-VAE, VQ-GAN and VAE under different Gaussian noise intensities of 10, 20, and 30, where averaged performances over all images in the testing set are reported. The best values for each metric are highlighted in bold. More comprehensive results on the other two datasets can be referred to the supplementary material. Regardless of the noise levels, SSC-VAE consistently achieves higher SSIM scores, affirming its ability to better preserve the structural similarity of denoised images. The PSNR values further support this trend, with SSC-VAE consistently outperforming SC-VAE, VQ-VAE, VQ-GAN and VAE, indicating superior fidelity and image quality in the denoised outputs. In addition, SSC-VAE also demonstrates significant advantages in terms of LPIPS and FID metrics compared to the other models.

Qualitative Results: Figure 4 displays the denoising outputs of each model on ultrasound dataset with different noise

Model	Dataset	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓
SCVAE	BSDS-500	0.87	28.97	0.084	36.22
SCVAE+In		0.90	29.55	0.055	28.37
SSCVAE		**0.94**	**31.19**	**0.016**	**16.40**
SCVAE	ultra-sound	0.87	32.53	0.202	63.21
SCVAE+In		0.88	32.93	0.179	66.95
SSCVAE		**0.94**	**35.78**	**0.051**	**17.83**

Table 3: Ablation study numerical results on reconstruction performance on BSDS500 and ultrasound dataset.

level. In the first row, the intensity of Gaussian noise σ is 10, and in the second and third rows, it is 20 and 30 respectively. The denoised images obtained by SSC-VAE exhibit a significantly higher level of clarity compared to the other models. In particular, the texture and details of the organs in the ultrasound images are remarkably close to those present in the ground truth images. In comparison, while SC-VAE yields better results compared to VQ-VAE and VQ-GAN, it still falls short of the denoising capabilities offered by our proposed SSC-VAE model. More results on the other two datasets can be referred to the supplementary material.

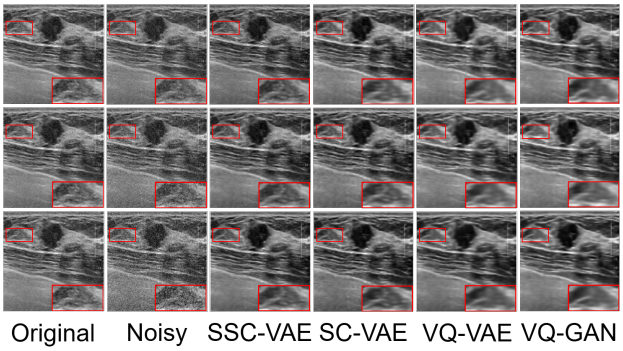

Figure 4: Image denoising on ultrasound dataset with different noise levels of $\sigma = 10, 20$ and 30 from top to bottom.

4.6 Ablation Study

Figure 5: Ablation study image results on reconstruction performance on BSDS500 dataset.

To validate the effectiveness of our proposed Inference and Refinement modules for the Sparsity Controller, we conducted a series of ablation experiments with the SC-VAE model as our baseline. We first constructed a variant "SCVAE+In" by adding only the Inference module to the baseline SC-VAE model to isolate the impact of the vectorized sparsity controller from the scalar sparsity controller. The Refinement module further enhances the structural information within the controller, preserving more accurate details. Our proposed model extends the baseline SC-VAE by introducing a vectorized sparsity controller and a structure-enhanced sparsity controller. This extension allows the model to capture potential structures through vectorization and refine those structures via our Refinement network. We focused on reconstruction experiments using these three models on the BSDS500 dataset, with their performances summarized in Table 3 and Figure 5. The results for the ultrasound dataset are included in the supplementary material.

Compared to the baseline SC-VAE with a scalar sparsity controller, the vectorized extension via the Inference module significantly improved the performance both in terms of metrics and visual results. Our model, which further incorporates the Refinement module, outperformed the simplified "SC-VAE+In" variant by enhancing potential structures within the image. These results validate, step by step, the necessity and impact of each module on overall model performance. It is emphasized that each component is crucial, serving its specific function in our proposed model.

Metrics	SSCVAE	SCVAE	VQVAE	VQGAN
FLOPs/GMac	160.42	156.52	152.16	150.57
Params/M	22.95	19.12	19.11	18.98

Table 4: Computational Complexity Analysis: FLOPs and model sizes.

4.7 Computational Complexity

The computational complexity of the SSC-VAE, SC-VAE, VQ-VAE, and VQ-GAN models are summarized in Table 4, taking into account their floating point operations (FLOPs) and model size. Slightly increase (2%) in complexity is observed for SSC-VAE due to the incorporation of vectorization and attention mechanisms. The increase is minimal comparing to the significant performance improvement.

5 Conclusion

In this paper, we propose a new variant of SC-VAE by introducing sparsity controller inference and refinement mechanism. By extending the scalar sparse coefficient for a latent representation, which is the threshold in LISTA, to a vector, our method enables to perform semantically fine-grained control over sparse code. Furthermore, two-dimensional correlation mechanism by CBAM is imposed on the initially inferred thresholds to utilize the correlation within and among representations, resulting more accurate and informative threshold estimation and sparse representation. Numerical results on the image reconstruction and denoising tasks demonstrate the effectiveness of the proposed method in preserving the image details and local structures.

Acknowledgements

This paper is supported by the General Program of the National Natural Science Foundation of China (NSFC) under Grant 62276189 and the Fundamental Research Funds for the Central Universities No. 22120220583.

References

Chen, R. T. Q.; Li, X.; Grosse, R.; and Duvenaud, D. 2018. Isolating sources of disentanglement in VAEs. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2615–2625.

Cui, J.; Xie, Y.; Joshi, A. A.; Gong, K.; Kim, K.; Son, Y.-D.; Kim, J.-H.; Leahy, R.; Liu, H.; and Li, Q. 2022. PET denoising and uncertainty estimation based on NVAE model using quantile regression loss. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 173–183.

Daubechies, I.; Defrise, M.; and De Mol, C. 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11): 1413–1457.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Glorot, X.; Bordes, A.; and Bengio, Y. 2011. Deep sparse rectifier neural networks. In *Proceedings of the 14-th international conference on artificial intelligence and statistics*, 315–323.

Gregor, K.; and LeCun, Y. 2010. Learning fast approximations of sparse coding. In *Proceedings of the 27th international conference on international conference on machine learning*, 399–406.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6629–6640.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C. P.; Glorot, X.; Botvinick, M. M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*, 1–13.

Kingma, D. P.; Rezende, D. J.; Mohamed, S.; and Welling, M. 2014. Semi-supervised learning with deep generative models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 3581–3589.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014*, 1–14.

Leglaive, S. 2022. A Multimodal Dynamical Variational Autoencoder for Audiovisual Speech Representation Learning. In *Proceedings of the 1st International Workshop on Methodologies for Multimedia*, M4MM '22, 3.

Liang, Y.; Wang, L.; Wang, J.; and Luo, Y. 2023. Attentive Deep K-SVD Network for Patch Correlated Image Denoising. In *2023 IEEE International Conference on Image Processing (ICIP)*, 1490–1494.

Maaløe, L.; Sønderby, C. K.; Sønderby, S. K.; and Winther, O. 2016. Auxiliary deep generative models. In *International conference on machine learning*, 1445–1453. PMLR.

Mubashar, M.; Khan, N.; Sajid, A. R.; Javed, M. H.; and Hassan, N. U. 2022. Have we solved edge detection? A Review of state-of-the-art datasets and DNN based techniques. *IEEE Access*, 10: 70541–70552.

Pu, Y.; Gan, Z.; Henao, R.; Yuan, X.; Li, C.; Stevens, A.; and Carin, L. 2016. Variational autoencoder for deep learning of images, labels and captions. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2360–2368.

Razavi, A.; van den Oord, A.; Poole, B.; and Vinyals, O. 2019. Preventing Posterior Collapse with delta-VAEs. In *International Conference on Learning Representations*.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, 1278–1286. PMLR.

Vahdat, A.; and Kautz, J. 2020. NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33: 19667–19679.

van den Oord, A.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6309–6318.

Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; and Wierstra, D. 2016. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 3637–3645.

Wang, J.; Zhou, W.; Tang, J.; Fu, Z.; Tian, Q.; and Li, H. 2018. Unregularized auto-encoder with generative adversarial networks for image generation. In *Proceedings of the 26th ACM international conference on Multimedia*, 709–717.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, 3–19.

Xiao, P.; Qiu, P.; and Sotiras, A. 2023. SC-VAE: Sparse Coding-based Variational Autoencoder. *arXiv preprint arXiv:2303.16666*.

Yu, J.; Li, X.; Koh, J. Y.; Zhang, H.; Pang, R.; Qin, J.; Ku, A.; Xu, Y.; Baldridge, J.; and Wu, Y. 2022. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as

a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zheng, C.; Vuong, T.-L.; Cai, J.; and Phung, D. 2022. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems*, 35: 23412–23425.