Improving Image Reconstruction using Incremental PCA-Embedded Convolutional Variational Auto-Encoder

Amir Azizi
CYENS CoE
Nicosia
CYPRUS
a.azizi@cyens.org.cy

Panayiotis Charalambous
CYENS CoE
Nicosia
CYPRUS
p.charalambous@cyens.org.cy

Yiorgos Chrysanthou
CYENS CoE
Nicosia
CYPRUS
y.chrysanthou@cyens.org.cy

ABSTRACT

Traditional image reconstruction methods often face challenges like noise, artifacts, and blurriness, requiring handcrafted algorithms for effective resolution. In contrast, deep learning techniques, notably Convolutional Neural Networks (CNNs) and Variational Autoencoders (VAEs), present more robust alternatives. This paper presents a novel and efficient approach for image reconstruction employing Convolutional Variational Autoencoders (CVAEs). We use Incremental Principal Component Analysis (IPCA) to enhance efficiency by discerning and capturing significant features within the latent space. This model is integrated into both the encoder and sampling stages of CVAEs, refining their capability to generate high-fidelity images. Our incremental strategy mitigates scalability issues associated with traditional PCA while preserving the model's aptitude for identifying crucial image features. Experimental validation utilizing the MNIST dataset showcases noteworthy reductions in processing time and enhancements in image quality, underscoring the efficacy and potential applicability of our model for large-scale image generation tasks.

Keywords

Image Processing, Image Reconstruction, Principal Component Analysis, Convolutional Variational Auto-Encoders

1 INTRODUCTION

Generative models like Variational Autoencoder (VAE) are a significant advancement in deep learning, using probabilistic approaches to generate new data from existing datasets.VAEs,consisting of an encoder and decoder pair, capture key properties [CGD$^+$20, BTLLW21].VAEs have a wide range of applications, from image [LSM$^+$21, YYSL16, WX21] ,video [YZAS21, WRO21, ZLS$^+$21, DLW$^+$22], text [LPL21, YDML21, SWL$^+$21, ZDYC21] and music generation [JY23, WT22, WY21] to anomaly detection [NYW20, LCB$^+$20], and they are easier to train than their competitors, such as Generative Adversarial Networks (GANs) [BTLLW21].They ensure training stability by offering a greater range of realistic and varied facts [DB21].CVAEs integrates the functionalities of VAEs with those of a CNNs, constituting a specialized form of deep generative model [YKK21].CVAEs possess an architecture that leverages CNNs as both encoders and decoders, harnessing the spatial relation-capturing capabilities of CNNs and the fidelity in data generation characteristic of VAEs [SDRM21].This method demonstrates exceptional proficiency in processing grid-based data, especially images, leveraging the spatial understanding capabilities inherent in CNNs [BLD22]. CVAEs stand as foundational components across a diverse range of applications [KSZ$^+$21, WML21, LPC22, CQWZ21], spanning image synthesis, reconstruction, and anomaly detection. Their robustness consistently delivers stable, realistic, and semantically coherent outcomes [JJ21].In Figure 1, the structure of CVAEs and the interaction between their components are illustrated.This paper is dedicated to enhancing both the efficiency and quality of image reconstruction through the utilization of CVAEs.
Our contributions are summarized as follows:

1. Introducing a novel method that combines IPCA with CVAEs, enhancing the sampling and encoding stages.

2. considerably reduces processing time, thus boosting the algorithm's overall efficiency.

3. Demonstrating substantial enhancements in image quality by implementing our proposed algorithm on MNIST datasets.

The paper is structured as follows: Section 2 delves into the related works; in Section 3, we outline our novel approach and methodology. Section 4 is our Experimental results. Finally Section 5 , discusses the research challenges, limitations, and potential future directions.

https://www.doi.org/10.24132/CSRN.3401.12

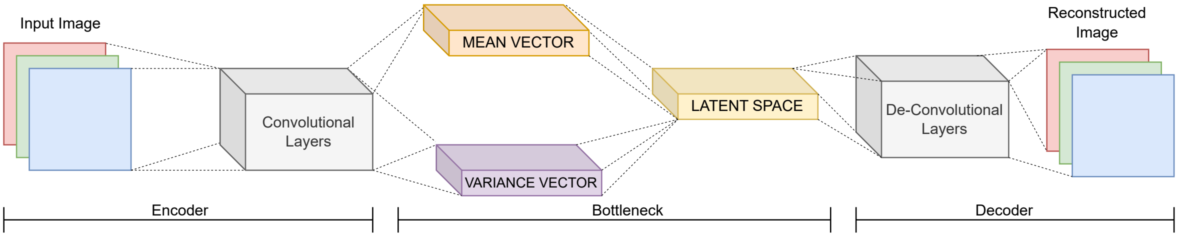

Figure 1: CVAEs utilize convolutional layers to encode input images, compress information into latent space, and decode samples, enabling efficient data representation and generation of new samples.

2 RELATED WORKS

The VAE, introduced by Kingma et al. [KW13].They applied variational Bayesian inference principles for image generation.It consists of two neural networks,Figure 2 shows an inference network for generating latent variable distributions, and a generation network for approximation.Deep learning and Artificial Intelligence have revolutionized image reconstruction and image generation from diverse data sources, with Variational Auto-encoders emerging as key methodologies for high-quality image generation [EEAMT22, LSC20, IB23, WCQ23]. The

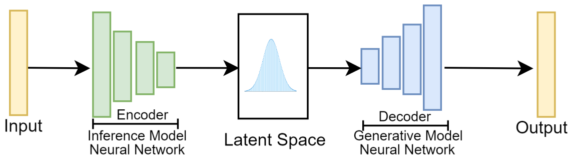

Figure 2: The model employs variational Bayesian inference principles for image generation, utilizing two neural networks for inference and generation to capture data structure and variability.

Conditional Variational Autoencoder (CVAE),which is shown in Figure 3, improves the unsupervised model by incorporating category information labels, transforming it into a semi-supervised mode within the CVAE framework [SLY15, HNW21].The Very Deep Variational Autoencoder (VDVAE) is an extension of the standard Variational Autoencoder (VAE) with increased depth in its neural network architecture [Chi20]. This depth allows for more intricate hierarchical representation of latent variables, enabling high-fidelity and nuanced reconstructions while navigating large-scale dataset complexities. VDVAE-SR [CHW$^+$23] uses transfer learning on pretrained VDVAEs to improve image super-resolution. Fusion-VAE [DVZN22] is a novel deep hierarchical variational autoencoder for generative image fusion, outperforming traditional methods. VAEL, a neuro-symbolic generative model, combines VAEs and probabilistic Logic Programming strengths, enabling the generation of new data points satisfying logical constraints while capturing complex relationships VAEL architecture illustrated in Figure 4 VDVAEs, while powerful for image generation, face computational complexity due to high-dimensional data, but ongoing advancements aim to improve efficiency and scalability [ASY$^+$22, VK20].

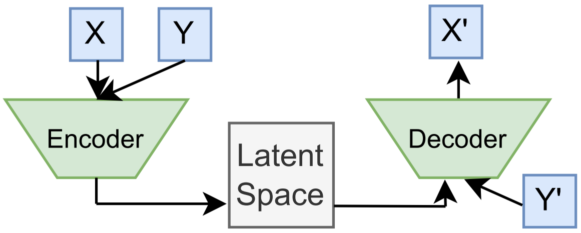

Figure 3: CVAE, similar to VAE,adds category information to input data, maximizing logarithmic marginal likelihood and lower bound function of variation, but not solving image blur or high synthetic data accuracy.

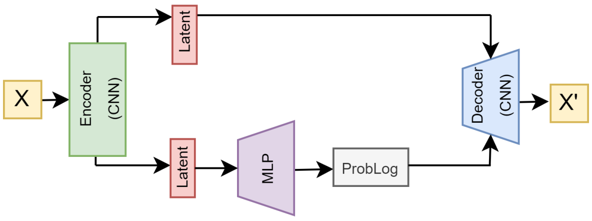

Figure 4: The VAEL model consists of three components: an encoder, a Prob-Log program, and a decoder, which compute latent variables, parameterize programs, and reconstruct images.

3 PROPOSED METHOD

The study uses IPCA and CVAEs to enhance image reconstruction. CVAEs extract hierarchical features from images using convolutional layers, preserving spatial information. They encode input images into a latent space, capturing meaningful representations of digits in a lower-dimensional space. Our novel approach lies in integrating IPCA within the training paradigm of CVAEs for image reconstruction. IPCA is used after encoding to reduce dimensionality in data obtained from the CVAE. It refines the latent space representation before decoding, enhancing efficiency and retaining meaningful information for the decoding step,

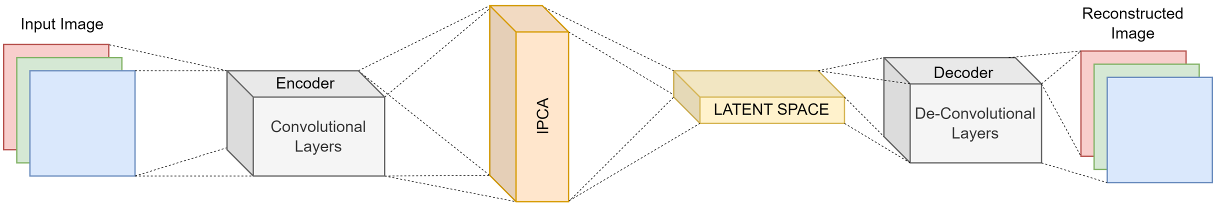

Figure 5: IPCA-CVAE: integrates incremental PCA and Convolutional Variational Auto-encoders to improve image reconstruction, enhancing efficiency and retaining meaningful information for decoding.

thereby optimizing the handling of latent representations The workflow of our proposed method , as illustrated in Figure 5 is structured as follows:

1. CVAEs are used to improve image reconstruction by extracting hierarchical features from images using convolutional layers. These layers preserve spatial information, ensuring image structural integrity. By encoding input images into a latent space, CVAEs capture meaningful representations of digits, allowing the model to learn essential features while reducing data dimensionality.

2. IPCA is a technique used in the training paradigm of CVAEs to optimize image reconstruction by refining the latent space representation obtained from the encoding step, thereby enhancing efficiency and retaining crucial information for the decoding step.

3. The decoder uses deconvolutional layers, also known as transposed convolutional layers or upsampling layers, to transform the refined latent representation from IPCA into a high-dimensional feature map. This process optimizes the handling of latent representations, enhancing the efficiency and accuracy of image reconstruction tasks. The integrated approach ensures meaningful information is retained and utilized throughout the decoding process.

The integration of IPCA enhances the CVAE model's efficiency by refining latent space representation before decoding, reducing data dimensionality while retaining crucial information for accurate image reconstruction. This approach enhances overall image reconstruction task performance and provides a novel solution for efficient dimensionality reduction and preservation of meaningful information. It is applicable across different domains and datasets, offering versatility and effectiveness in handling latent representations. Integrating IPCA into CVAEs training provides a robust and efficient solution for image reconstruction, contributing to image processing and deep learning techniques.

4 RESULTS

The study utilizes a PC with a Core i7-8700K CPU, 16GB RAM, and Google Colab for computational resources. The results of this work are presented in two parts:
i) Analyzing the impact of the IPCA on processing time.
ii) Exploring the relationship between re-constructed image quality and the proposed method.
We use two main datasets: MNIST [LC10] and Fashion-MNIST [XRV17].MNIST has 70,000 gray-scale images of ten handwritten digit classes. Fashion-MNIST is similar but covers ten categories of clothing, including items like t-shirts, trousers, and pullovers. To enhance compression, we utilized three models: CVAE, PCA-CVAE, and IPCA-CVAE. The architecture of the encoder and decoder involved a consistent structure. The encoder started with an input layer for gray-scale images, followed by two Conv2D layers implementing 32 and 64 filters. A flattening layer condensed the output, leading to a dense layer with 16 neurons activated by ReLU. Two dense layers generated parameters for shaping the variational distribution within the latent space. The decoder played a crucial role in the reconstruction process, starting with an input layer for 20-dimensional latent space representation. The output was rearranged into a (7, 7, 64) structure, enabling two Conv2DTranspose layers to progressively upscale the encoded latent space. The output layer, a Conv2DTranspose, reconstructed the original image dimensions, concluding the decoding phase.

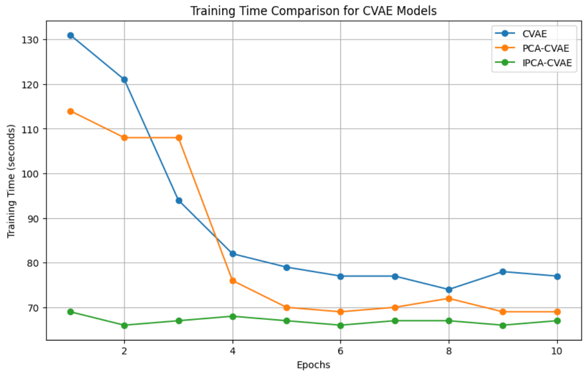

Figure 6: Training Time Comparison for CVAE Models on MNIST digits dataset.

In Figure 6, a comparative analysis of various methodologies is presented, focusing on the training duration. The visual representation of PCA-CVAE and CVAE

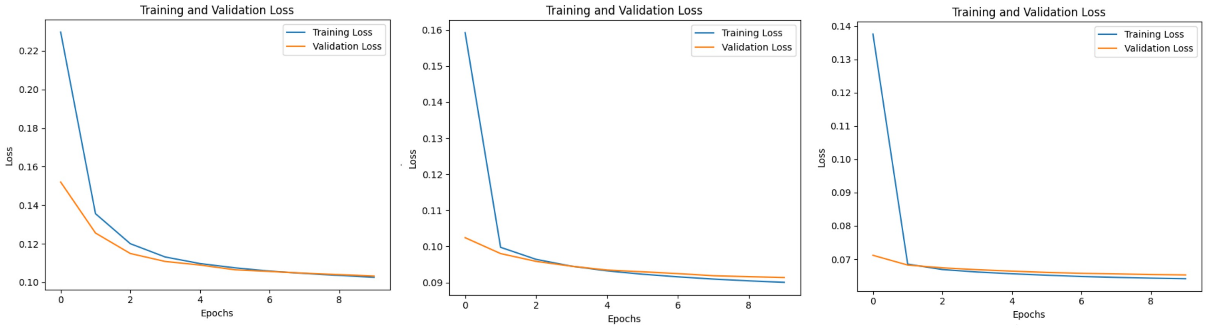

Figure 7: Training and Validation Loss for CVAE(left),PCA-CVAE(middle) ,IPCA-CVAE(right) on MNIST digits dataset

methodologies reveals significant differences. PCA-CVAE offers sustained functionality beyond the fourth epoch and reduces processing time, while IPCA-CVAE demonstrates accelerated training process completion, which is beneficial for large image databases. The proposed IPCA-CVAE method shows heightened efficiency, especially in scenarios with voluminous image collections, demonstrating its superior performance.

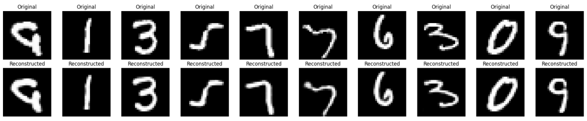

Figure 8: Reconstructed Images based on IPCA-CVAE on MNIST digits dataset.

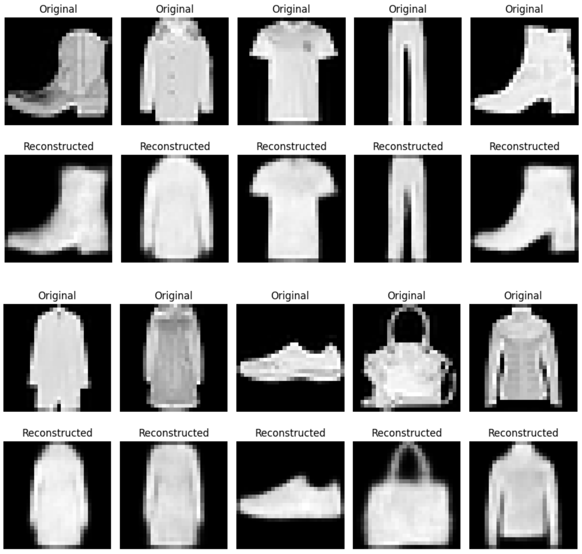

Figure 9: Reconstructed Images based on IPCA-CVAE on MNIST fashion dataset.

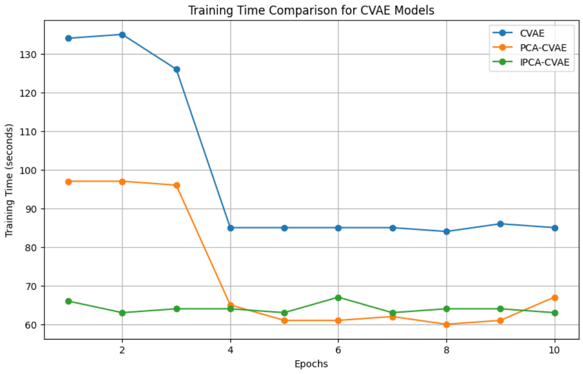

Figure 10: Training Time Comparison for CVAE Models on MNIST fashion dataset.

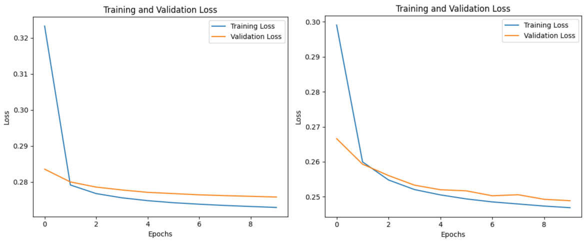

Figure 11: Training and Validation Loss for CVAE(left),IPCA- CVAE(right) on MNIST fashion dataset.

Figure 7 provides a comparison of Training and Validation Loss across various CVAE methods. Figure 8 and Figure 9 display reconstructed images generated using the IPCA-CVAE method applied to the MNIST datasets. Furthermore, Figure 10 highlights a noticeable discrepancy in training duration on the MNIST Fashion dataset, emphasizing the tangible advantages of employing the proposed algorithm. Significantly, the discernible reduction in processing time attests to the algorithm's efficacy in optimizing training efficiency. Additionally, Figure 11 presents a comparative analysis between CVAE and IPCA-CVAE methods. The left image corresponds to the CVAE method, the middle image represents PCA-CVAE, and the right image pertains to IPCA-CVAE. Notably, the Figure distinctly reveals a more pronounced reduction in loss during the training process in the proposed method compared to alternative approaches.

4.1 Metrics Evaluation

In this section, we quantitatively assess the performance of our proposed method, CVAE-IPCA, on the

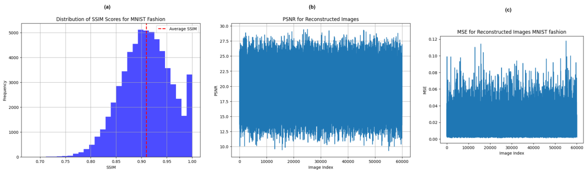

Figure 12: Metrics Evaluation after epoch 10 :(a) SSIM, (b) PSNR, and (c)MSE. The images demonstrate an average SSIM of 0.9097, PSNR of 21.0432, and an MSE of 0.0133. These values reflect the effectiveness of the proposed method.

MNIST fashion dataset for image reconstruction tasks. We employ three widely used image quality metrics: Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE), and Structural Similarity Index (SSIM). PSNR measures the ratio between the maximum possible power of a signal and the power of corrupting noise, defined as:

$$PSNR = 10 * log_{10}(peakval^2)/MSE$$

, where peakval is the maximum possible pixel value of the image.MSE calculates the average squared difference between the original and reconstructed images, which is defined for two images such as $\hat{g}(n,m)$ and $g(n,m)$as below :

$$MSE = 1/MN * (\sum_{n=0}^{M} \sum_{m=1}^{N})\,[\hat{g}(n,m) - g(n,m)]^2$$

Lastly, SSIM evaluates the similarity between two images based on luminance, contrast, and structure, and is defined as

$$SSIM(x,y) = [l(x,y)]^{\alpha} * [c(x,y)]^{\beta} * [s(x,y)]^{\gamma}$$

Here, L denotes luminance, C represents contrast, and S signifies structure. These parameters help gauge brightness, intensity range differences, and local pattern similarities between images, respectively.α ,β, and γ are constants used for computation stability. results of metrics evaluation illustrated in Figure 12.

4.2 Ablation Study

In our ablation study, we systematically varied three key parameters: learning rate, batch size, and latent space dimensionality, to understand their impact on the performance of our proposed method.

1. Learning Rate Variation: We tested learning rates of 0.001, 0.0001, and 0.01 to analyze their effect on model convergence and performance metrics such as PSNR, SSIM, and MSE. This experiment provided insights into how different learning rates influence training dynamics and model effectiveness.
2. Batch Size Variation We explored batch sizes of 64, 128, and 256 to assess their impact on training stability and computational efficiency. By observing training speed and model accuracy under different batch size settings, we gained an understanding of their trade-offs and implications.
3. Latent Space Dimensionality Variation : We varied latent space dimensions between 20, 50, and 100 to examine how they affect the model's ability to capture and represent input data features. Analyzing reconstruction quality across different latent space sizes provided insights into the dimensionality's role in feature representation.

The default parameter values for our method are learning rate = 0.001, batch size = 128, and latent space dimension = 20. These values serve as reference points for comparison against the variations tested in our ablation study.

4.2.1 Findings

1. Our investigation found that employing a learning rate of 0.0001 failed to yield any notable improvements. Although a rate of 0.01 showed some enhancements compared to 0.0001, the most consistent and optimal outcomes were achieved with a learning rate of 0.001.
2. Changing the batch size did not lead to any improvement in the results. However, the alterations in batch size yielded better outcomes compared to those from modifying the learning rate. Notably, the optimal results were observed with a batch size of 128. As anticipated, increasing the batch size to 64 resulted in longer processing times, while decreasing it to 256 reduced processing times.
3. Increasing the latent space dimensions to 100 yielded improved results. However, this enhancement came at the cost of significantly increased processing time. The results of the ablation study are shown in Table 1.

Figure 13 also presents the outcomes of PSNR.Figure 14 shows the reconstructed images with different parameters and finally, Figure 15 illustrates the results of SSIM, MSE, and training curves.

5 CONCLUSION AND FUTURE WORKS

This paper introduces a novel approach to image reconstruction using CVAEs.The efficiency of the model is optimized through IPCA, a technique used to identify and capture significant features in the latent space. This

Parameter	PSNR	SSIM	MSE	TT	Remarks
lr=0.01	20.439	0.8907	0.0167	1031	ld=20,bs=128
lr=0.001	21.0432	0.9097	0.0133	974	ld=20,bs=128
lr=0.0001	19.3043	0.8898	0.0193	1021	ld=20,bs=128
ld=20	21.0432	0.9097	0.01505	974	lr=0.001,bs=128
ld=50	20.8135	0.8992	0.0193	1255	lr=0.001,bs=128
ld=100	22.002	0.9101	0.0123	1245	lr=0.001,bs=128
bs=64	20.754	0.8998	0.0163	1182	lr=0.001,ld=20
bs=128	21.0432	0.9097	0.0133	1021	lr=0.001,ld=20
bs=256	20.9097	0.9001	0.0140	1084	lr=0.001,ld=20

Table 1: Summary of the Ablation Study(lr=Learning Rate , ld=Latent Dimension ,bs=Batch Size ,TT=Training Time(s))

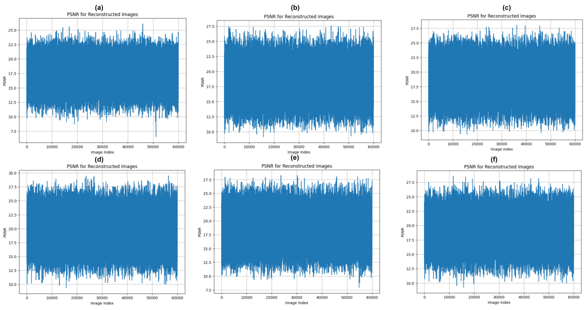

Figure 13: PSNR results (a)learning rate=0.0001, (b)learning rate=0.01 (c)=latent dimension=50 (d)latent dimension=100 (e)batch size=64 (f)batch size=256

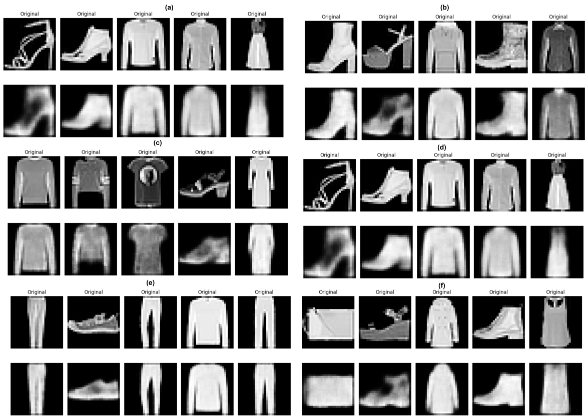

Figure 14: Reconstructed Images (a)learning rate=0.0001, (b)learning rate=0.01 (c)=latent dimension=50 (d)latent dimension=100 (e)batch size=64 (f)batch size=256

incremental strategy addresses scalability concerns associated with traditional PCA while preserving the model's proficiency in identifying essential image features. Experimental results on the MNIST dataset show :

1. Reduces processing time and improves image quality.

2. Applicable in large-scale image generation tasks.

3. CVAEs' sensitivity to input data distribution variations.

Our proposed method demonstrates significant potential across various image-related tasks, extending beyond the datasets examined thus far. In object detection, denoising, and other applications, such as image classification using extensive datasets like CIFAR-10 and ImageNet, integrating IPCA-CVAE could enhance feature extraction and dimensionality reduction. This enhancement may lead to improved classification accuracy and resilience against variations in image content. Moreover, in tasks like style transfer and image synthesis, where generating realistic and diverse images is crucial, IPCA-CVAE's refined latent representations could offer finer control over visual attributes, enabling the creation of more compelling images. Additionally, in medical imaging tasks like tumor detection and segmentation, where handling high-dimensional images is common, IPCA-CVAE integration could help capture meaningful anatomical structures while reducing computational complexity and memory usage. Furthermore, in satellite imagery analysis for environmental monitoring and disaster management, IPCA-CVAE could facilitate efficient feature extraction and anomaly detection, enabling timely identification of critical changes in land cover and environmental conditions.

Moving forward, future research directions should focus on adapting the model to diverse and more complex datasets, exploring real-time optimization strategies, and improving the robustness of the CVAE framework. Additionally, incorporating domain-specific knowledge or constraints into the model may further refine its applicability for specific image reconstruction tasks.

6 ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital

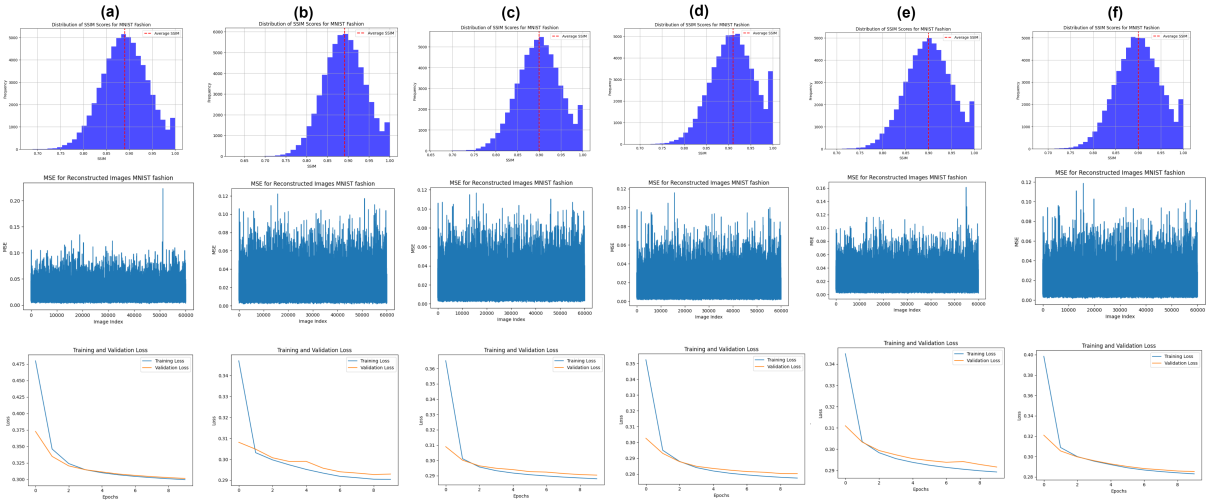

Figure 15: SSIM results(top) MSE results(middle) and Training curves(bottom) (a)learning rate=0.0001, (b)learning rate=0.01 (c)=latent dimension=50 (d)latent dimension=100 (e)batch size=64 (f)batch size=256

Policy. This work has received funding from the European Union's research and innovation program Horizon Europe under grant agreement No 10192889. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

7 REFERENCES

[ASY+22] Bilal Ahmad, Jun Sun, Qi You, Vasile Palade, and Zhongjie Mao. Brain tumor classification using a combination of variational autoencoders and generative adversarial networks. *Biomedicines*, 10(2):223, 2022.

[BLD22] Jichao Bao, Liangping Li, and Arden Davis. Variational autoencoder or generative adversarial networks? a comparison of two deep learning methods for flow and transport data assimilation. *Mathematical Geosciences*, 54(6):1017–1042, 2022.

[BTLLW21] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

[CGD+20] Taylan Cemgil, Sumedh Ghaisas, Krishnamurthy Dvijotham, Sven Gowal, and Pushmeet Kohli. The autoencoding variational autoencoder. *Advances in Neural Information Processing Systems*, 33:15077–15087, 2020.

[Chi20] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020.

[CHW+23] Darius Chira, Ilian Haralampiev, Ole Winther, Andrea Dittadi, and Valentin Liévin. Image super-resolution with deep variational autoencoders. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 395–411, Cham, 2023. Springer Nature Switzerland.

[CQWZ21] Dingliang Chen, Yi Qin, Yi Wang, and Jianghong Zhou. Health indicator construction by quadratic function-based deep convolutional auto-encoder and its application into bearing rul prediction. *ISA transactions*, 114:44–56, 2021.

[DB21] David Dehaene and Rémy Brossard. Reparameterizing vaes for stability. *arXiv preprint arXiv:2106.13739*, 2021.

[DLW+22] Yiping Duan, Mingzhe Li, Lijia Wen, Qianqian Yang, and Xiaoming Tao. From object-attribute-relation semantic representation to video generation: A multiple variational autoencoder approach. In *2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2022.

[DVZN22] Fabian Duffhauss, Ngo Anh Vien, Hanna Ziesche, and Gerhard Neumann. Fusionvae: A deep hierarchical variational autoencoder for rgb image fusion. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision*

– *ECCV 2022*, pages 674–691, Cham, 2022. Springer Nature Switzerland.

[EEAMT22] Mohamed Elasri, Omar Elharrouss, Somaya Al-Maadeed, and Hamid Tairi. Image generation: A review. *Neural Processing Letters*, 54(5):4609–4646, 2022.

[HNW21] William Harvey, Saeid Naderiparizi, and Frank Wood. Conditional image generation by conditioning variational auto-encoders. *arXiv preprint arXiv:2102.12037*, 2021.

[IB23] Ashhadul Islam and Samir Brahim Belhaouari. Fast and efficient image generation using variational autoencoders and k-nearest neighbor oversampling approach. *IEEE Access*, 11:28416–28426, 2023.

[JJ21] Anyue Jiang and Behnam Jafarpour. Deep convolutional autoencoders for robust flow model calibration under uncertainty in geologic continuity. *Water Resources Research*, 57(11):e2021WR029754, 2021.

[JY23] Shulei Ji and Xinyu Yang. Emomusictv: Emotion-conditioned symbolic music generation with hierarchical transformer vae. *IEEE Transactions on Multimedia*, 2023.

[KSZ$^+$21] Adam R Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Sona Mokrá, and Danilo Jimenez Rezende. Nerf-vae: A geometry aware 3d scene generative model. In *International Conference on Machine Learning*, pages 5742–5752. PMLR, 2021.

[KW13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[LC10] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[LCB$^+$20] Shuyu Lin, Ronald Clark, Robert Birke, Sandro Schönborn, Niki Trigoni, and Stephen Roberts. Anomaly detection for time series using vae-lstm hybrid model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4322–4326. Ieee, 2020.

[LPC22] Sang Min Lee, Sang-Youn Park, and Byoung-Ho Choi. Application of domain-adaptive convolutional variational autoencoder for stress-state prediction. *Knowledge-Based Systems*, 248:108827, 2022.

[LPL21] Ruizhe Li, Xutan Peng, and Chenghua Lin. On the latent holes of vaes for text generation. *arXiv preprint arXiv:2110.03318*, 2021.

[LSC20] Zhi-Song Liu, Wan-Chi Siu, and Yui-Lam Chan. Photo-realistic image super-resolution via variational autoencoders. *IEEE Transactions on Circuits and Systems for video Technology*, 31(4):1351–1365, 2020.

[LSM$^+$21] Kaikai Liu, Renjun Shuai, Li Ma, et al. Cells image generation method based on vae-sgan. *Procedia Computer Science*, 183:589–595, 2021.

[NYW20] Zijian Niu, Ke Yu, and Xiaofei Wu. Lstm-based vae-gan for time-series anomaly detection. *Sensors*, 20(13):3738, 2020.

[SDRM21] Kelathodi Kumaran Santhosh, Debi Prosad Dogra, Partha Pratim Roy, and Adway Mitra. Vehicular trajectory classification and traffic anomaly detection in videos using a hybrid cnn-vae architecture. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):11891–11902, 2021.

[SLY15] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[SWL$^+$21] Huajie Shao, Jun Wang, Haohong Lin, Xuezhou Zhang, Aston Zhang, Heng Ji, and Tarek Abdelzaher. Controllable and diverse text generation in e-commerce. In *Proceedings of the Web Conference 2021*, pages 2392–2401, 2021.

[VK20] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in neural information processing systems*, 33:19667–19679, 2020.

[WCQ23] Zhangkai Wu, Longbing Cao, and Lei Qi. evae: Evolutionary variational autoencoder. *arXiv preprint arXiv:2301.00011*, 2023.

[WML21] Hongzhuang Wu, Xiaoli Ma, and Songyong Liu. Designing multi-task convolutional variational autoencoder for radio tomographic imaging. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 69(1):219–223, 2021.

[WRO21] Jacob Walker, Ali Razavi, and Aäron van den Oord. Predicting video with vq-vae. *arXiv preprint arXiv:2103.01950*, 2021.

[WT22] Huiyao Wu and Maryam Tavakol. Musebar: Alleviating posterior collapse in recurrent vaes toward music generation. In *International Symposium on Intelligent Data Analysis*, pages 365–377. Springer, 2022.

[WX21] Yang Wu and Lihong Xu. Image generation of tomato leaf disease identification based on adversarial-vae. *Agriculture*, 11(10):981, 2021.

[WY21] Shih-Lun Wu and Yi-Hsuan Yang. Musemorphose: Full-song and fine-grained piano music style transfer with one transformer vae. *arXiv preprint arXiv:2105.04090*, 2021.

[XRV17] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[YDML21] Weijie Yuan, Linyi Ding, Kui Meng, and Gongshen Liu. Text generation with syntax-enhanced variational autoencoder. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[YKK21] Qien Yu, Muthu Subash Kavitha, and Takio Kurita. Mixture of experts with convolutional and variational autoencoders for anomaly detection. *Applied Intelligence*, 51:3241–3254, 2021.

[YYSL16] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 776–791. Springer, 2016.

[YZAS21] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[ZDYC21] Kun Zhao, Hongwei Ding, Kai Ye, and Xiaohui Cui. A transformer-based hierarchical variational autoencoder combined hidden markov model for long text generation. *Entropy*, 23(10):1277, 2021.

[ZLS$^+$21] Yizhou Zhou, Chong Luo, Xiaoyan Sun, Zheng-Jun Zha, and Wenjun Zeng. Vaeˆ 2: Preventing posterior collapse of variational video predictions in the wild. *arXiv preprint arXiv:2101.12050*, 2021.