

---

# Perceptual Loss in (Variational) Autoencoder

---

Nujitha Wickramasurendra - 7904790<sup>1</sup>

## Abstract

This work focuses on experimenting the task of image super-resolution using perceptual loss with Variational Autoencoders (VAEs). Traditional super-resolution techniques rely on pixel-wise loss functions, which may not capture high-level textural and contextual details that are crucial for visual perception. This study integrates perceptual loss, which is calculated using a pre-trained convolutional network like VGG, instead of solely relying on pixel-wise loss with the idea of trying to capture high-level features of images. By employing VAEs, we can take advantage of their probabilistic generative capabilities to model the distribution of high-resolution images. The preliminary results show how this method slowly attempts to preserve and achieve perceptual quality during both training and testing. Future work will explore the optimization of feature extractors and loss function compositions to further enhance the model's performance.

## 1. Introduction

### 1.1. Variational Autoencoders (VAEs)

VAEs are a type that is known to be a generative model which is popular in the space of image generation. The two major components of a VAE are encoder and decoder where the encoder takes input data, such as images, and then maps it into a lower-dimensional latent space representation, and the latent space can capture key features from the input data. Then the decoder transforms those samples and does the reconstruction work. These are usually applied and are capable of minimizing the difference between the input data and its reconstructed output while working on regularizing the latent space following a predefined probability distribution, usually a Gaussian distribution. This regularization encourages the model to learn a smooth and continuous latent

space representation, which makes it possible to interpolate meaningfully and generate new data points.

### 1.2. Importance of Perceptual Loss in Image Generation Tasks

Image generation tasks, such as generating realistic images from random noise or incomplete data, often face challenges in producing visually convincing results. Traditional loss functions like mean squared error (MSE) may not capture perceptual differences effectively, leading to blurry or distorted reconstructions. Perceptual loss addresses this limitation by incorporating high-level semantic information extracted from pretrained deep neural networks, such as convolutional neural networks (CNNs). By comparing feature representations of generated images with those of real images, perceptual loss measures the perceptual similarity between the two, resulting in more visually appealing and semantically meaningful reconstructions.

### 1.3. Incorporating Perceptual Loss in VAEs

VAEs are well known for capturing the statistical properties of the input data and generating convincing reconstructions. However, the limitation of generating quality outputs still may exist. As a solution perceptual loss offers the knowledge about deep features learned from deep neural networks to guide the generation process. By integrating the VAE with perceptual loss, the model can learn to generate images that not only match the input data statistically but also exhibit higher-level visual features, such as textures, shapes, and object semantics. This makes them more suitable for various applications, including image synthesis, in-painting, and style transfer.

## 2. Background

### 2.1. Image Generation

(Johnson et al., 2016) demonstrates in their study how perceptual loss can be used to enhance the performance of super-resolution applications. They implemented an image transformation network consisting of both pixel-wise and perceptual loss. They applied a similar approach in style transfer, where perceptual loss ensures that the stylized output remains close to the target image's content while show-

---

<sup>1</sup>Department of Computer Science, Brock University, St. Catherine's, ON, Canada. Correspondence to: Nujitha Wickramasurendra <u23qu@brocku.ca>.

ing artistic style inherited from a different image, resulting in a visually appealing synthesis. Another recent implement has been done by (Gatopoulos et al., 2020), who have developed a highly effective super-resolution VAE that comprises a unique DenseNet-based encoder, a DenseNet-based decoder, and a flow-based prior. This model has achieved the State-of-the-Art in terms of log-likelihood function among single-leveled VAEs. Additionally, they have introduced a new category of VAEs that feature a superresolution component for generating high-quality images. Interestingly few other studies have conducted the usability of perceptual loss on medical imaging (Yang et al., 2018) and another on earth image observations (Shi & Pun, 2019) demonstrating the usability of perceptual loss in variety of domains.

An excellent study done by (Pihlgren et al., 2023) systematically evaluates a host of benchmark studies including (Johnson et al., 2016) which are based on perceptual loss and pre-trained networks for different feature extraction points across four applications: perceptual similarity, super-resolution, image segmentation, and dimensionality reduction. Similarly several other studies like (Pihlgren et al., 2020) (Hou et al., 2024) and (Snell et al., 2017) have experimented on image generation utilizing perceptual loss.

## 2.2. Other Applications

Looking at the previous work done by industry and academia it can be seen that perceptual loss is not limited to computer vision and can be used in Natural Language Processing (NLP) and audio processing. In NLP, it can improve naturalness and semantic coherence in tasks such as text style transfer, similar to computer vision. Similarly, in audio processing, researchers have used perceptual loss to compare deep audio features to generate high-quality synthetic speech, showing its wider application across sensory modalities.

Perceptual loss is also being explored in reinforcement learning, where it could potentially improve how agents perceive and interact with their environments. (Mnih et al., 2015) demonstrated that deep learning in reinforcement learning can be extended by applying perceptual loss to help agents learn more meaningful environmental representations. An interesting study has been conducted by the MIT Media labs (MIT) introducing a loss function for audio neural networks that is motivated by perception. A research conducted by (Zhang et al., 2023) presents a way to enhance speech quality by using perceptual loss that is dependent on a person’s lip movement. Their study demonstrates that this approach is effective in improving speech enhancement performance.

These developments highlight the versatility and effectiveness of perceptual loss in various domains, making it a cornerstone in advancing machine learning techniques. By focusing on perceptual similarities instead of just pixel accu-

racy, researchers and practitioners can create more realistic and functionally effective models in diverse applications. Each of these areas benefits from a deeper understanding of how perceptual features can be utilized, leading to innovations that push the limits of what machines can perceive and accomplish.

## 2.3. Mathematical Foundation

**Mean Squared Error (MSE):** The Mean Squared Error (MSE) is a commonly used loss function in regression tasks, which calculates the average of the squared differences between the predicted and actual values. This function expresses the difference between the estimated values and actual values in terms of the square of the difference. The MSE formula is expressed mathematically as:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2$$

**Perceptual loss:** Perceptual loss is a type of loss function that compares the similarity of features extracted from intermediate layers of a deep learning model, usually a CNN, instead of comparing individual pixels. This comparison is made in the feature space and prioritizes perceptual similarity over pixel-by-pixel accuracy. Perceptual loss can be expressed as:

$$l_{feat}^{\phi,j}(\hat{y}, y) = \frac{1}{C_j H_j W_j} \|\phi(\hat{y}) - \phi(y)\|_2^2$$

Where:

- $\phi_l(y)$  and  $\phi_l(\hat{y})$  are the feature representations of the reference image  $y$  and the generated image  $\hat{y}$  at layer  $j$ .
- feature map of shape  $C_j H_j W_j$

## 2.4. Feature Space

The transformation of input data into a set of detectable features that represent various aspects of the data is known as the concept of "feature space". In the context of deep learning, specifically with CNNs, the feature space is built by the layers of the network that convert the raw input into a more compressed and abstract representation. Perceptual loss assesses the perceptual similarity between images within this feature space. It uses the deep feature maps derived from various layers of a pretrained CNN, like VGGNet (pyt), which is commonly employed for image-related tasks. By measuring the differences in these feature maps, perceptual loss can more effectively capture and emphasize discrepancies in texture, style, or structural elements of the images compared to traditional pixel-based loss functions. This

makes it particularly suitable for tasks where the output’s human-like perception quality is crucial, such as in style transfer or high-resolution image generation.

### 3. Methodology

We implemented a python code with PyTorch which utilizes a VAE enhanced with perceptual loss to tackle the task of image super-resolution. The architecture we referenced is similar to that of (Johnson et al., 2016), but we replace the image transformation network with a VAE that is similar to the specification by (Hou et al., 2024). This approach combines the probabilistic modelling capabilities of VAEs with the advanced perceptual evaluation afforded by perceptual loss. The goal is to generate high-resolution images from low-resolution inputs while ensuring both fidelity to the original images and high perceptual quality. The implemented code is available in the GitHub<sup>1</sup> repository.

#### 3.1. Data Preprocessing

The dataset used is the OxfordIIITPet (Parkhi et al., 2012) dataset, processed to generate pairs of low-resolution and high-resolution images. The dataset is freely available for academic and research purposes without licensing restrictions. This involves several transformation steps:

- High-resolution images are scaled to a fixed size and normalized using predefined mean and standard deviation values to match the input requirements of the VGG network.
- To reduce the resolution of input images, a Gaussian blur is applied with a sigma value of 1.0, followed by downsampling using bicubic interpolation with the given factor.

#### 3.2. Probabilistic Modeling in VAE

The VAE comprises two main components: an encoder and a decoder. The encoder transforms the input data into a latent space representation, characterized by a mean ( $\mu$ ) and a variance ( $\sigma^2$ ) that define a Gaussian distribution.

- The encoder network maps input image  $x$  to two parameters in a latent space,  $\mu(x)$  and  $\log \sigma^2(x)$ , defining the posterior distribution  $q(z|x) = \mathcal{N}(z; \mu, \sigma^2)$ .
- The latent variable  $z$  is then sampled using the reparameterization trick:

$$z = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I),$$

where  $\odot$  denotes element-wise multiplication.

<sup>1</sup><https://github.com/nujitha99/PerceptualLossVAE>

- The decoder then uses this latent variable  $z$  to reconstruct the input data, aiming to minimize the difference between the ground-truth image and its reconstruction.

The loss function of the VAE,  $L$ , is a combination of the reconstruction loss and the Kullback-Leibler divergence (KL divergence), balancing image fidelity and distributional alignment:

$$L_{\text{vae}} = \alpha L_{\text{reconstruction}} + D_{KL}(q(z|x)||p(z)),$$

where  $p(z)$  is the prior ( $\mathcal{N}(0, I)$ ), and  $\alpha$  is a hyperparameter managing the trade-off for which we chose value 0.001. After so much trial work it was noted that retaining the reconstruction loss alongside perceptual loss in a VAE allows for a more comprehensive and effective optimization framework and we noticed it tries to generate more semantically meaningful images.

#### 3.3. Perceptual Loss

By comparing the feature activations of the original and reconstructed images at intermediate layers (relu3\_3 in our work), the model focuses on perceptual similarity rather than pixel accuracy:

The perceptual loss,  $L_{\text{perceptual}}$ , can be formulated as the MSE between the feature representations of the reconstructed and target images:

$$L_{\text{perceptual}} = \text{MSE}(F(\hat{x}), F(x)),$$

where  $F$  denotes the feature extraction function provided by the pre-trained VGG,  $x$  is the target high-resolution image, and  $\hat{x}$  is its reconstruction from the VAE (image transformation network).

#### 3.4. Implementation Details

- The dataset operations include transformations to create low-resolution versions of high-resolution images. This involves downsampling, followed by blurring to simulate image degradation.
- During training, the model optimizes the combined loss:

$$L_{\text{total}} = \alpha L_{\text{vae}} + \gamma L_{\text{perceptual}},$$

where  $\alpha$  and  $\gamma$  are weights that balance the reconstruction and perceptual components of the loss which we chose 1 and 0.8 respectively.

- Regular evaluation during training helps monitor the progress and adjust parameters as necessary.

### 3.5. Experimental Setup and Training

The training of the model utilizes a blend of reconstruction and perceptual loss, with a small coefficient applied to the latter to ensure a balanced contribution against the primary reconstruction loss. The model processes the images to produce super-resolved outputs, which are evaluated both qualitatively and quantitatively using the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM).

## 4. Discussion

The integration of VAEs with the perceptual loss for the task of image super-resolution is a novel yet very challenging study since this method combines both the probabilistic generative capabilities of VAEs and the high-level feature sensitivity of perceptual loss, derived from pretrained convolutional network VGG. Here, we present the evaluation results, discuss the challenges faced during the implementation of this method for image super-resolution and suggest potential directions for future research.

### 4.1. Evaluation

We followed the same evaluation metrics as the previous study we referenced. Our evaluation was based on PSNR and SSIM for x2 super-resolution. The calculated scores were 1.32 and 0.02 respectively. In comparison, the study by (Johnson et al., 2016) reported an average PSNR score of 24.99 and SSIM score of 0.6731 for x4 super-resolution. The outputs generated from our solution did not produce the best results due to a number of limitations faced during training yet we can see how the model slowly trying to produce better results in Figure 1.

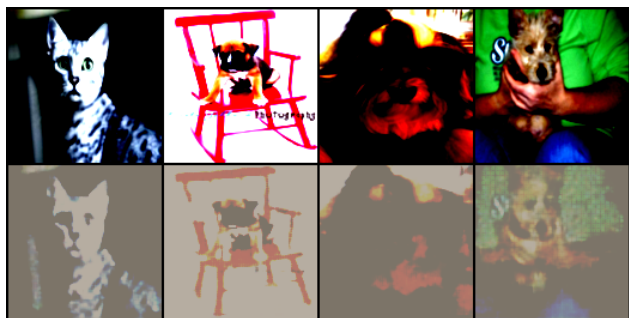


Figure 1. Low resolution images (top) and generated super-resolution images (bottom). We can observe the model’s attempt to enhance image quality by refining texture and perception. This result was produced after 200 epochs.

### 4.2. Challenges of using perceptual loss in VAEs

The utilization of perceptual loss to improve VAE performance has limitations and challenges that are critical and need to be handled carefully. One of these challenges is the computational overhead associated with perceptual loss, particularly when employing a separate deep neural network for feature extraction. To compute perceptual loss, pretrained networks such as VGG, require additional computational resources during both training and inference phases. With the current implementation and with average computing resource setup the model took around 6 hours to train.

We also experimented with the implemented system’s reconstruction loss alone without integrating the perceptual loss and apparently it produced more visually pleasing outputs. We suspect that the basic architecture of the VAE currently implemented or the layer of the VGGnet we used to calculate the perceptual loss may be the reason for the issue. It is worth noting that the two studies we based this experiment on had trained for around 200k iterations whereas we only experimented for just 200 iterations max. Therefore, as we already did, this work needs more in-depth level scrutiny to find the best parameters and weights for metrics like losses. This overhead may constrain the scalability and efficiency of VAEs, particularly when dealing with large-scale datasets or real-time applications. Therefore, developing perceptual loss metrics that capture the desired perceptual qualities across this task was a heavy challenge.

### 4.3. Strengths

In contrast to the challenges mentioned above and looking at previous work, the integration of perceptual loss in VAE training techniques seems a significant step forward in generative modelling. By combining traditional reconstruction loss with perceptual loss, the VAE is encouraged to produce reconstructions that not only match the input data at the pixel level but also capture higher-level features and structures found in real images. This results in more realistic textures, shapes, and object semantics in the generated images. Even though the implemented system did not achieve the best results after so much trial work, the previous works show integrating perceptual loss with different architectures and in different domains proves the versatility of perceptual loss producing more natural-looking quality outputs.

### 4.4. Future Research Directions

- **Exploring Different Feature Extractors:** While this implementation utilizes the ‘relu3\_3’ layer of the VGG network, future studies could experiment with other layers or different architectures altogether (such as ResNet or Inception) to determine the optimal feature extractor for perceptual loss in super-resolution.

- **Investigating Loss Function Compositions:** Further research into the optimal weighting and composition of loss functions, including the trade-offs between reconstruction loss, KL divergence, and perceptual loss, could lead to more refined models that better cater to specific application needs.

## 5. Conclusion


The use of VAEs with perceptual loss for image super-resolution offers promising results, particularly in terms of enhancing the perceptual quality of super-resolved images. Future explorations in this area could lead to substantial improvements not only in image quality but also in the adaptability and efficiency of super-resolution models across various imaging applications.

## References

- URL <https://www.media.mit.edu/projects/codec-perceptual-loss/overview/>.
- vgg16 2014; Torchvision main documentation — pytorch.org. <https://pytorch.org/vision/main/models/generated/torchvision.models.vgg16.html>. [Accessed 22-04-2024].
- Gatopoulos, I., Stol, M., and Tomczak, J. M. Super-resolution variational auto-encoders. *arXiv preprint arXiv:2006.05218*, 2020.
- Hou, X., Shen, L., Sun, K., and Qiu, G. Deep feature consistent variational autoencoder, 2024.
- Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part II 14*, pp. 694–711. Springer, 2016.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.
- Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.
- Pihlgren, G. G., Sandin, F., and Liwicki, M. Improving image autoencoder embeddings with perceptual loss, 2020.
- Pihlgren, G. G., Nikolaidou, K., Chhipa, P. C., Abid, N., Saini, R., Sandin, F., and Liwicki, M. A systematic performance analysis of deep perceptual loss networks: Breaking transfer learning conventions. *arXiv preprint arXiv:2302.04032*, 2023.
- Shi, C. and Pun, C.-M. Adaptive multi-scale deep neural networks with perceptual loss for panchromatic and multispectral images classification. *Information Sciences*, 490:1–17, 2019.
- Snell, J., Ridgeway, K., Liao, R., Roads, B. D., Mozer, M. C., and Zemel, R. S. Learning to generate images with perceptual similarity metrics, 2017.
- Yang, Q., Yan, P., Zhang, Y., Yu, H., Shi, Y., Mou, X., Kalra, M. K., Zhang, Y., Sun, L., and Wang, G. Low-dose ct image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE Transactions on Medical Imaging*, 37(6):1348–1357, June 2018. ISSN 1558-254X. doi: 10.1109/tmi.2018.2827462. URL <http://dx.doi.org/10.1109/TMI.2018.2827462>.
- Zhang, J., Feng, F., Hong, D., and Wang, X. Speech enhancement with lip perceptual loss. In *2023 IEEE 9th International Conference on Cloud Computing and Intelligent Systems (CCIS)*, pp. 187–192, 2023. doi: 10.1109/CCIS59572.2023.10263269.



## A. Course Feedback



Welcome Nujitha Wickramasurendra  
BrockU - SCES

English Sign Out

### My Home

Tasks

Q Search All Reset

Sort by Task Type

3 of 3 (filtered from 3 tasks)


	COSC 5F90 D5 S1 - MSC THESIS (PRO) Fri, Apr 26, 2024 11:59 PM	2023 Winter Open
	COSC 5P77 D3 S1 - Probabilistic Graphical Models and Neural Gen (LEC) Fri, Apr 19, 2024 11:59 PM	2023 Winter Completed

Figure 2.