Texture analysis and synthesis using neural networks

Lukáš Hudec*
Institute of Computer Engineering and Applied Informatics
Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava
Ilkovičova 2, 842 16 Bratislava, Slovakia
lukas.hudec@stuba.sk

Abstract

The structures and textures form the various patterns on the surface of objects. The analysis of these structures significantly contributes to the automatic processing of the scene by computer vision methods. The similarity of different textures and the structural variability in some materials (marble, wood, fur, ...) increase the difficulty of this problem. Modern texture analysis methods can be applied in digital pathology, where tissue scans structures resemble textures. According to the literature, the potential for improving the analysis and synthesis of textures are mostly the deep learning methods, neural networks. This paper presents a modern approach contributing to three areas of texture processing. The proper neural network architecture can determine the similarity of texture patches and represent extracted features in its own latent space. The similarity and dissimilarity classification of textures is applied in synthesis to recognize real and generated textures, and further exploited in analysis of histological data. We measure the success of our approaches on public datasets and compare them with state-of-the-art methods. Our approaches achieve the level of state-of-the-art methods or contribute to the interpretability of neural network decisions.

Our results contribute to the proof of the incredible power of deep learning and its potential in processing textures with different levels of heterogeneity. The proposed models of neural networks for support of cancer diagnostics can detect diagnostically interesting regions. They also point out that for histologists to understand the predictions of neural networks, it is necessary to process and present large tissue areas with context and detail, even larger than a simple network can evaluate, with further extension to the interpretability of network predictions and decisions.

*Recommended by thesis supervisor: Assoc prof. Vanda Benešová
Defended at Faculty of Informatics and Information Technologies, Slovak University of Technology in Bratislava on August 30, 2021.

Categories and Subject Descriptors

I.2.1 [**Applications and Expert Systems**]: Medicine and science; I.3.3 [**Picture/Image Generation**]: Texture generation; I.4.3 [**Enhancement**]: Miscellaneous; I.4.8 [**Scene Analysis**]: Object recognition; I.4.9 [**Applications**]: Miscellaneous; I.4.10 [**Image Representation**]: Multidimensional, Statistical

Keywords

texture similarity, synthesis, neural networks, histology

1. Introduction

Many machine vision, image processing and computer vision algorithms aim to decompose the scene into a set of simplified automatically generated descriptions, to help other applications to process data more easily. Traditional descriptors usually simplify the task by assuming the uniformity of objects' structure. This is is rarely a case in real world objects, where multiple structures form heterogeneous repeating patterns. The texture of object is one of visual characteristics utilized in the segmentation and recognition and significantly contribute to the understanding of a scene.

Texture analysis is a computer vision field of study that attempts to quantitatively describe intuitive qualitative visual characteristics of objects and regions in an image. The usual result of quantitative description methods is a vector of distinctive features extracted from the analyzed texture. The traditional approaches are *statistical*, *structural*, *model-based*, and *transform-based* methods [25]. These four categories are well-established, however, now considered out-of-date compared to the performance of more recent modern approaches. The *graph-based* and *entropy-based* approaches have become more popular since the 2000's. Additionally, the seventh category methods are based on *deep learning* and achieved high impact on computer vision and related fields. Their robustness and the self-learning approach assured their state-of-the-art results in all fields and superior performance over traditional methods. However, its application is limited to big datasets and low interpretability.

The computational texture analysis has achieved fine results in texture classification over the years [21]. The easy availability of large datasets with high-resolution photos allowed the application of neural networks to texture processing. The literature presents many papers analyzing texture similarity by trying to mimic human perception. Human visual perception ignores small per-pixel intensity variations and can recognize textures on global and local

level of detail. Nowadays the most promising methods resembling the perception of the human visual system are based on convolutional neural networks.

The synthesis techniques can be categorized into three major families: procedural, exemplar-based, and model-based texture synthesis methods [2]. Model-based methods and deep learning-based methods are able to create a descriptive model of a texture that can be used to reproduce the texture and synthesize the texture samples artificially. The goal of synthesis algorithms is to create a texture that is indistinguishable from the real-world textures. Deep learning-based methods learn to define the texture by optimized parameters and achieve superior quality [12, 14, 10]. However, most of the state-of-the-art approaches are robust enough to learn only one texture per model. Even though, the generation process with the optimized model is fast, the training and re-training of the model is very expensive.

A significant contribution and benefits of texture analysis application are achieved in biomedicine and medical imaging, because many medical images have characteristics similar to textures. Breast cancer is one of the most widespread causes of women's death worldwide. According to WHO, breast cancer has the second-highest incidence and is the fifth leading cause of death globally, estimated at 680 thousand deaths in 2020 [1]. During the diagnostics, if the series of non-invasive tests locate a anomaly, the biopsy extracts a tissue sample from the area for the further examination. The experts then examine the microscopy scan of the tissue. This process is time consuming and puts a lot of pressure on histologists. An automated support method can help speed up the process and relieve the doctors. However, such method requires a level of interpretability. Due to the black box nature and the lack of explanation of network decisions, deep learning techniques application in medical imaging is problematic.

2. Related work

2.1 Evaluating texture similarity

- **Distance measures:** The simplest option is to measure the spatial distance between 2D signals or extracted feature vectors. Kokare et al. [19] review 9 distance metrics on features extracted using Gabor wavelet to select the most suitable one for texture retrieval.

- **Spectrum, shape, statistics:** The statistics approaches assume the texture can be decomposed as a statistical repetition of structural patterns. Zujovic et al. [35] introduce a set of 5 structural texture similarity metrics (STSIM) using different texture statistics that characterize coefficients of a *multi-scale frequency decomposition* (steerable filters). The method compares the extracted feature vectors by *Mahalanobis distance* and achieves state-of-the-art results. Polec et al. [26] analyze arbitrary shaped areas (superpixels) and approximates texture by a set of orthogonal basis functions, which are particularly suitable for periodic or at minimum quasiperiodic textures.

- **Deep learning:** The approaches combined handcrafted and deep features than classify the texture by shallow machine learning models [23]. More complex solutions use well-known architectures to extract features from textures and compare them by another custom architecture [11]. Bruna and Mallat [6] reach the state-of-the-art results with imitating network using a sequence of Fourier decomposition layers and adding modulus average pooling to attain non-linearity.

2.2 Texture synthesis

The literature recognizes three main approaches of example-based synthesis:

- **Non-parametric:** The most straightforward approach is to copy parts of the input texture example and generate new ones by sticking these parts together by smoothing the borders. This approach is also referred as pixel/patch-based synthesis. Their main problem is a lack of the stochastic information of the original texture the visible repetitiveness.

- **Parametric:** More complex approaches learn the statistics of structure of texture to create computational descriptive model. The methods aim to model textures as a continuous 2D signal.

- **Deep learning generative models:** Many specialized architectures of the generative models are powerful in various generative tasks [22, 16]. They train on example texture to optimize their parameters and generate new samples from a random noise or by modifying the input example.

From the deep learning models, the Generative adversarial networks (GAN) achieved state-of-the-art quality for synthesized outputs. However, the Variational Auto-Encoders (VAE) are easier to train and to control the variations of the expected output, even though they produce textures of comparably lower quality.

Generative adversarial networks. Jetchev et al. [14] introduces GAN for arbitrary sized output to generate similar samples by interpolating on manifold of various samples. Zhou et al. [34] presents an example-based synthesis of non-stationary textures. The adversarial learning approach extracts the style of input texture to generate a new sample of the same style with increased dimensions of output sample. Fruhstuck et al. introduced TileGAN that achieved state-of-the-art quality in tiling non-stationary textures into large-scale image [10].

Variation Auto-Encoders. Chandra et al. [8] uses recurrent variational auto-encoder for synthesis of simple texture by iterative tiling of generated neighboring texture samples in each direction. However, each direction tiling must have individually trained model to ensure the correct tiling. The main drawback of VAEs is the blurriness of an output image frequently referenced in various studies and referred to as unsolved with several probable causes.

2.3 Medical imaging

Several proposed deep learning approaches have shown state-of-the-art results in different areas of computer vision and medical imaging, even for classification of breast cancer from histopathology images. In biopsy tissue image processing and analysis, it is important to correctly

identify structures deformed by cancer to specify its stage and form. The appearance of tumors of the same class is very different, the overall structure of tissue has many visual formations, complex texture, and high variance, which is still referred as a great problem even in recent studies [29, 13].

These problems are considerable challenges for automatic and precise classification of microscopy images for breast cancer. The approaches could be divided into two streams according to the data used: typical microscopy tissue images; more modern Whole Slide Images containing a very high-resolution scan of a magnified tissue sample in a pyramidal structure.

- **Microscopy tissue images:** Novel approaches try to overcome mentioned problems with additional heuristics, sometimes with architecture alternations [17, 3]. The usually used size of a cut-out sample is 128×128 and according to Araújo et al. [4] is small and may not contain the overall structure, so they recommend using samples of 512×512 pixels, depending on magnification. On the other hand, recently Celik et al. [7] experimented with advanced models to recognize Invasive Ductal Carcinoma. Their DenseNet and ResNet models were pre-trained on ImageNet and finetuned on a dataset with 227542 samples of 50×50 pixels, reaching F1 score to 94%.
- **Whole Slide Images:** Tripathi et al. [31] have used annotated WSIs and 400 cut-out samples. They state that if the prediction needs to be made for the whole tumor or gland, the network cannot lose spatial nor structural information.

3. Preliminaries

Neural networks are deep multi-layered robust models and the essence of deep learning. The layers are made up of computational units, the neurons without being interconnected in the same layer. To increase the computational power of neural network, the layers are connected one after another in various ways. The first layer receives input x, and the last layer exports the desired predictions $\hat{y}$ that should correspond to annotated ground truth y. The other layers are hidden and represent the network mapping function $y = f(x)$. Feed-forward operation propagates the input signal through the network. Backpropagation is a gradient descent algorithm widely used to train the networks. It propagates training error through all weights and computes their gradients as partial derivation. The optimizer then updates the weights according to gradient and learning rate. The training optimizes parameters to minimize the error - loss of the objective function approximation.

3.1 Siamese Neural Networks

The Siamese Networks are designed to compare two input patterns by computing the corresponding similarity measure. The networks extract feature vectors and compare them with a similarity metric, which can be cosine angle, cross entropy classification, and any distance metric. The standard architecture shares weights between two network branches, with other possible modifications.

Contrastive loss function. The most used loss function $\mathcal{L}$ is based on the principles of a spring system. According to the distance $D_{i\theta}$, the similar pairs are pulled together and dissimilar are pushed apart.

$$\mathcal{L}(\theta, y, x_1, x_2) = (1-y)\frac{1}{2}(D_\theta)^2 + y\frac{1}{2}\max(0, \mathbf{m} - D_\theta)^2 \quad (1)$$

The pairs with high dissimilarity would distort the total loss, so the contrastive loss contains a **margin** term $\mathbf{m}$ defining a maximal contribution radius of dissimilar pairs.

3.1.1 Distance metrics

L2 Euclidean: is a standard distance metric of Euclidean space $D_E(\vec{x_1}, \vec{x_2}) = \sqrt{\sum_{i=1}^{N}(x_{1i} - x_{2i})^2}$

Mahalanobis: computes a dissimilarity value for vectors of the **same distribution** from the **covariance matrix**. The absolute value assures the positive value of distance. $D_{Mah}(\vec{x_1}, \vec{x_2}) = \sqrt{|(\vec{x_1} - \vec{x_2})^T cov(x_1, x_2)(\vec{x_1} - \vec{x_2})|}$

Canberra: the numerator signifies the difference and denominator normalizes the difference. Thus, the intermediate divisions between vector values never exceeds 1, and is equal to 1 whenever either of the attributes is 0. $D_c(\vec{x_1}, \vec{x_2}) = \sum_{i=1}^{N} \frac{|x_{1i} - x_2|}{|x_{1i}| + |x_{2i}|}$

Bray-Curtis: is similar to the Canberra and directly related to the Sorensen similarity index. The difference to Canberra is that it can reach high negative values when two feature vectors are mutually opposite. $D_{BC}(x_1, x_2) = |\sum_{i=1}^{N} \frac{|x_{1i} - x_{2i}|}{x_{1i} + x_{2i}}|$

3.2 Variational Auto-Encoders

VAE is introduced as architecture that might be able to generate content [18]. Rezende et al. [28] compared them to denoising Auto-Encoders and stated they could be a realization of variational inference in latent variable models. VAEs encode the input x into constrained distribution of latent vector z in simple cases represented with isotropic multivariate Gaussian with diagonal covariance structure containing approximate inference values of mean μ and standard deviation σ. The vector z is sampled from the distribution $p_\theta(z)$ obtained by encoder/approximate inference network $q_\phi(z \mid x)$. The sample is then up-sampled by *decoder* network $p_\theta(\hat{x} \mid z)$. The general VAE loss represents the joint log-likelihood of the variables under the approximate posterior over the latent variables. The term $\mathcal{H}(q(z \mid x))$ represents an entropy of the approximate posterior. Maximizing this entropy increases the standard deviation of added noise to the predicted mean value, and also encourages the variational posterior to distribute z values that could generate x around a mean value rather than pushing them all to one point. The log-likelihood is usually a reconstruction error and VAE also use constraint, Kullback-Leibler divergence [15] pulling together the posterior distribution $q(z \mid x)$ and prior model distribution $p_{\texttt{model}}(z)$.

4. Research goals

According to the analyzed literature this paper focuses on developing and applying novel neural network based approach for texture analysis and synthesis. All of our research in the following topics have been published or is awaiting for the review.

- **Evaluation of similarity**. We apply a novel neural network-based approach for texture description. The approach uses the network as a feature extractor and texture descriptor computing the feature vectors of pre-defined size. Our research experiments with multiple distance metrics to compare

their influence on learning embeddings with contrastive loss function.

- **Texture synthesis**. The Variational Auto-Encoder is designed to generate new samples by modifying the latent layer parameters responsible for the reconstruction. The basic model suffers from blurry outputs with a decreased level of detail. Therefore, we modify the architecture and design a novel derivative with enhanced abilities to learn and add micro-texture details into reconstructed and synthesized texture images.
- **Breast cancer diagnostics**. Based on our research in texture analysis using deep neural networks, we utilize known approaches in the medical imaging. We develop a special support method for breast cancer diagnostics from microscopy tissue images. We propose a classification approach using a Siamese neural network regarding the implicit interpretation of steps leading to the final cancer class type prediction.

5. Datasets

5.1 Texture dataset

Unfortunately, most of the public datasets consists of a relatively small number of samples for each material, or the structures are too regular. One of only exceptions is Describable Textures Dataset. We introduce our dataset similar to the one used in [35] with a sufficient number of high-resolution textures that makes it applicable to training the neural network. The gathered dataset is available online[1].

Construction of the dataset. We have downloaded over 450 texture images for free from various online databases. We obtain images of different natural or artificial materials, and if more images contain identical material (grass, leather, fur, ...), the annotated similarity depends on the spatial and structural similarity of texture elements. This allows a simplified annotation of similar and dissimilar pairs. The resolution of downloaded images varies from 512×512 to 3200×2880 which is necessary for the sampling of the patches, as each texture class is represented by only one image. Finally, we filtered misleading images containing multiple texture classes (multi-label) with resolutions higher than a sampled patch.

5.2 Breast tissue microscopy dataset

We implemented and tested our method on the public dataset from 2018 **Grand Challenge on BreAst Cancer Histology images (BACH)** competition [5]. In this research, we use 400 cut-out slices annotated with one cancer class per slice (**normal, benign, In Situ** and **invasive carcinoma**). The tissue slices are stained by **H&E**. **Hematoxylin** binds to DNA and stains nuclei to dark violet. **Eosin** binds to proteins and stains other structures to pink. Other color variations are due to various concentration, mixing colors, and uneven light absorption. The training set consists of 400 classified images with uniform distribution of classes. The test set has 100 images without an explicit class label. The cut-outs are saved in .tiff format in resolution 2048×1536 with 1 pixel representing $0.42\mu m$ receptive area.

[1] https://bit.ly/texdat

6. Texture similarity evaluation

We have adopted the concept of Siamese Neural Network to measure the similarity between texture patches. The architecture of our twin network's branches is inspired by AlexNet [20]. We decrease the number of filters layer-wise, add two more convolutional layers, and substitute the last two dense layers with global mean pooling layer and a convolutional layer with 1×1 kernel. An activation function *LeakyReLU* follows each convolutional layer. To experiment with how the network learns the embedding, we set the size of last layer to 3 (3D coordinates) and to 128 like most of the traditional descriptors. The other training enhancement is batch normalization layer, that preserves the distribution of activation values of previous layer and provides an effective substitution for dropout layer. The batch normalization is added before every convolutional layer (except the first layer and after global mean pooling) (Table 1).

Table 1: Architecture of a single branch of used SiamNet - conv = convolutional layer, pool = max pooling, glob-p = global max pooling, conv/fc = convolution 1×1 substitution for original fully connected layer.

Layer	Kernel Size	Stride	Dims.	Output channels
conv1	3×3	1×1	150×150	32
pool1	2×2	2×2	75×75	
conv2	3×3	1×1	75×75	64
pool2	2×2	2×2	37×37	
conv3	3×3	1×1	37×37	64
conv4	3×3	1×1	37×37	64
pool3	2×2	2×2	18×18	
conv5	3×3	1×1	18×18	128
conv6	3×3	1×1	18×18	256
pool4	2×2	2×2	9×9	
conv7	3×3	1×1	9×9	256
flatten	-	-	81×1	256
glob-p	-	-	1×1	256
conv/fc	1×1	1×1	1×1	128 or 3

We experiment with *Canberra, Bray-Curtis*, and *Mahalanobis distance* and compare it to the performance of standard *Euclidean distance*.

The mini-batch size is constant for all experiments with a size of $m' = 64$ patch pairs. The distribution rate of similar vs. dissimilar pairs in the batch is approximately equal (50 : 50) and provided by `PyTorch DataLoader`. The pair's are uniformly sampled from classes so it is an equal probability for the class to be in a similar or dissimilar pair which does not adversely affect training. We sample new patches after each epoch, which adds a statistical variability of structures and computational stability. Each training can run a maximum of 200 epochs, and early stopping can stop it when the validation loss does not decrease in seven epochs (patience=7).

6.1 Results

To compare the effectiveness of training and achieved results, we keep the same architecture for all training runs. The table 2 below shows the number of epoch evaluated as best by early stopping, or as the last one with reasonable value, as the training with the selected network could fail, and the model would be unusable.

Table 2: Ending training with the epoch - optimal results or failed model.

Metric	3D	128D	Successful training
Euclidean	148	190	yes
Canberra	189	196	yes
Bray-Curtis	59	8	loss=nan from e70
Mahalanobis	11	-	no inverse matrix

Table 3: Evaluation and comparison of several similarity metrics introduced by Zujovic et al. in [35] and [36] respectively.

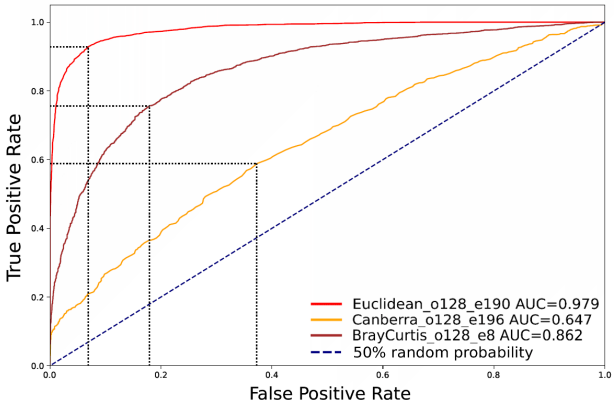
Algorithm	P@1	AUC	P@1	AUC
PSNR	0.04	0.753	0.14	0.50
(S-)SSIM	0.09	0.446	0.41	0.52
CW-SSIM	0.39	0.921	0.84	0.87
STSIM-1	0.74	0.967	-	-
STSIM-2	0.74	0.963	0.86	0.88
STSIM-M	0.96	0.985	0.84	0.80
Gabor f.	0.92	0.979	-	-
Wavelet f.	0.84	0.836	-	-
Do & Vetterli	-	-	0.79	0.79
Ojala	-	-	0.57	0.54

The order of steps for the evaluation:

1. Randomly sample 16000 pairs from validation dataset for evaluation - to compute the threshold for binary classification
2. Compute distances between pairs using trained neural network
3. Draw Receiver Operating Characteristics (ROC) curve and compute the optimal threshold
4. Randomly sample new 16000 pairs from test dataset for evaluation
5. Compute distances between pairs and by threshold classify binary similar/dissimilar samples
6. Compute classification metrics: accuracy, precision, recall, Area Under Curve (AUC)

We compared our results to the most successful STSIM-W metric and several others that Zujovic et al. [35] compared their results. The 128D network ROCs are visible in Figure 1 and accuracy comparisons are in Table 3. Our method outperform the results of Zujovic et al. The optimal threshold for binary classification is selected as maximal TPR with minimal FPR (black dotted lines) - maximal difference. Comparison of all distance metrics used by our network: Euclidean, Canberra, Bray-Curtis, and Mahalanobis. The bottom plot of ROC curves compares several of Zujovic et al. [36] proposed approaches to traditional methods Wavelet, Gabor, LBP features and to approaches of Do and Vetterli, and Ojala et al. (PSNR = peak signal-to-noise ratio, SSIM = structural similarity metric, STSIM = structural texture similarity metric, CW-SSIM = complex wavelet SSIM)

The results of our experiments show the state-of-the-art success of SiamNet at determining the similarity of texture patches. Even the small, memory-optimized architecture with a 3D feature vector can recognize the texture



Eu.: fpr:0.069, tpr:0.928, thr:0.98
Ca.: fpr: 0.372, tpr:0.588, thr:8.01
BC.: fpr:0.179, tpr:0.757, thr:29.51
Mh.: -

Zujovic et al. [36]

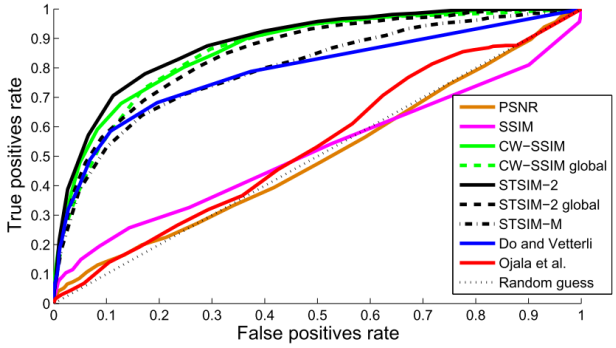


Figure 1: ROC curves (colored curves) with AUC, threshold and corresponding TPR/FPR. The bottom plot is from Zujovic et al. [36].

classes based on their structure. The comparison of distance metrics with different characteristics revealed that even though the metric can be good to compare Gabor wavelet extracted features [19], it does not necessarily be good to optimize the embeddings in the network's feature space.

6.2 Conclusion

We experimented with four distance metrics to examine their influence on training and embedding of feature space. We evaluated that each loss starts at a different value and preserves the converging oscillation. Like Mahalanobis or Bray-Curtis, some cases stopped very soon before the accuracy of a model could rise to successful values. In conclusion, the optimization of the neural network and the loss value depends on the selected distance function and may decrease rapidly or destroy the training.

To evaluate the correct understanding of which samples are similar, we examined the embeddings of extracted 3D and 128D feature vectors. On the contrary to the traditional feature extractors, in deep learning and especially in contrastive loss, the distance metric is not only comparing the extracted features but, more importantly, optimizing the embedding of feature space - unfolding the distribution manifold into space of a chosen distance metric.

We aimed to determine the similarity of textures with

high structural variations, which are more common in nature. We evaluated the performance of our method by several classification metrics and it reaches a considerable 92.7% accuracy, 92.8% precision, 92.7% recall, and integral of ROC curve covers 97.9% of the region. We compared our results to the related work of Zujovic et al., and we overperformed their approach, assuming the similarity of our datasets.

7. Texture reconstruction and synthesis

In contrast to VAE we designed our network as a Hierarchical VAE (HVAE) with multiple latent layers $\{z_1, ..., z_4\}$ learning features at a different levels of detail. HVAE can be represented as multiple variational auto-encoders joined in parallel to each other.

We are training both generative part $p(\hat{x}, z) = p(\hat{x} \mid z)p(z)$ along with inference part of the model $q(z \mid x)$ by optimizing the lower bound likelihood $p(\hat{x}) = \int p(\hat{x} \mid z)dz$. The optimization objective for hierarchical VAE after updating loss function of basic VAE with reconstruction error and distribution constraint of KL-divergence with special case when $z_0 = x$ and $z_{J+1} = 0$:

$$\mathcal{L}(\theta, \phi) = \sum_{j=0}^{J} E_{q_\phi, p_\theta}(z, x)[\log p(z_j|z_{>j})] - D_{\mathrm{KL}}(q_\phi(z|x)) \tag{2}$$

7.1 Architecture of HVAE

Our architecture uses pyramid approach for learning features from multiple-scales which also abstracts different levels of detail from texture and each latent level of the variational auto-encoder learns to encode features corresponding to the scale level. The architecture is built from the 2D convolutional layer blocks with 3×3 kernels, followed by a batch normalization and *LeakyReLu*. The down-sampling layers are replaced by convolution strides. We intend to generate textures with resolution of $(256 \times 256 \times 3)$. We included $1 \times 1 \times n_z$ convolutional layer before each latent layer for dimensionality reduction. Then, the output is flattened and fed in parallel to two fully connected layers with n_z units. They are optimized to represent the Gaussian distribution (μ and σ values) to sample the latent feature vectors z_i. The vector sampled from the distribution $z_i \sim \mathcal{N}(\mu_i, \sigma_i)$ is fed into another fully connected layer and reshaped into 2D output feature map with the exact same shape as before flattening. We experimentally choose $n_z = 40$ trainable parameters for each latent vector. It provides a good trade-off between the quality of reconstructed and generated samples, and overall compression of the input sample.

The standard Auto-Encoders and VAEs suffer from lossy compression in the latent space and generate blurry output lacking high frequency details. One of the main goals of our research is to **increase the quality** and **reduce the blurriness** of the reconstructed image. We separate the abstraction into multiple levels in order to synthesize new patches with increased precision in detail.

- The gradual pyramid addition of information from hierarchical latent layers contributes to improved reconstructions and decreases the blurriness.
- The separation forces the model to learn texture features of different scales at each level without affecting the resulting texture.

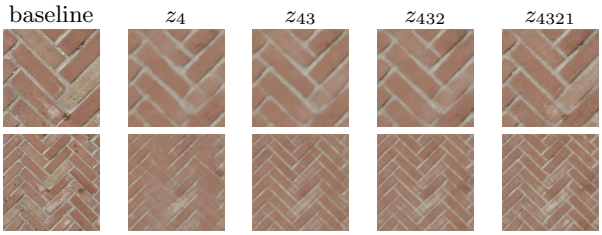


Figure 2: Visualization of reconstruction of selected baseline texture slices with gradual sampling from trained hierarchical layers. The columns present reconstructions from the combination of information on variational latent layers.

- The new textures are sampled from the latent sampling layers, which corresponds to the standard VAE behavior.

We modified the standard loss function composition for hierarchical VAE with the averaged KL divergence regularization (Equation 3) helping model to learn independent reconstruction information in parallel network branches. We used standard MSE to optimize the reconstruction capabilities of the network. Regularization of the distribution is held by averaged KL-divergence computed over all four hierarchical latent layers (number 4 in Equation 3).

$$L_{VAE} = L_{rec} + \frac{1}{4}\sum_{i=1}^{4}(L^j_{D_{\mathrm{KL}}}) \tag{3}$$

We prepare the input patches with dimensions $256 \times 256 \times 3$ from 4 different texture scales. The 4 scales are squares with size of the side 128, 256, 384, and 512 pixels respectively, and then reshaped to the required model input shape. This augments the train data with additional details at different frequencies for better model evaluation. The batch size is 16 images for all selected textures and the network is trained for 50,000 iterations for the reconstruction of the input image. The synthesis is a possible by-product of this architecture design.

7.2 Reconstruction

To evaluate the achieved quality, we perform the empirical qualitative comparisons of texture reconstructions from gradual sampling from the latent codes. We generate 4 sets of synthesized identical texture samples for each texture scale $512, 384, 256, 128$, on which we observe the gradual addition of details and independent changes in the reconstructed structures. Visual demonstration of the obtained texture samples is in Figure 2. The quantitative evaluation using Fréchet Inception distance on the reconstructed textures are in Table 4 and compared to Zhao et al. [33].

7.3 Synthesis

The main focus of our research and the introduction of our hierarchical architecture is the synthesis of texture samples. In order to synthesize new texture samples, the model must learn the continuous linear manifold on the encoded latent layers. The linear movement on the correctly optimized latent manifold should not introduce any artifacts into the generated image.

Our trained model is capable of synthesizing textures in 2 ways:

Table 4: Fréchet Inception Distance (FID) - Reconstruction task

Texture	multiscale		128px		256px		384px		512px	
	Zhao	Ours	Zhao	Ours	Zhao	Ours	Zhao	Ours	Zhao	Ours
bricks	282.1	**187.8**	320.3	**181.2**	349.8	**183.8**	276.5	**243.8**	243.3	**198.5**
camo	203.8	**135.4**	248.6	**144.4**	339.1	**114.7**	197.1	**159.3**	**200.4**	222.4
tiles	**136.9**	240.9	**111.8**	270.1	**131.7**	230.8	**173.7**	287	**201**	312.2
wall	**231**	236.5	**155.7**	278.2	**247.4**	319.6	288.9	**261.3**	291.6	**167.9**

1. *Random sampling* from all latent layers following the learned averaged Gaussian distributions $z_i \sim \mathcal{N}(\mu_i, \sigma_i)$
2. *Conditional sampling* and **modifying values** from encoded feature vectors of input texture example

Random sampling synthesis. It is very challenging and demanding on the quality of manifold optimization and its continuity. We rectified the random variables into the interval with the highest density of values $z \sim \mathcal{N}(\mu_\lambda, \sigma_\lambda)$ of the encoded latent vectors μ, σ based on several observations. Then we sample the latent vectors $z_1, \ldots, z_4$ from known distribution with $\mu_{i\lambda}, \sigma_{i\lambda}$ and use the latents as input to the decoder network, which generates new texture sample. Interpolating and sampling from the deepest latent layer z_4 provides high-intensity variations of the basic structure. The upper layers add details, noise, and illumination changes. The quantitative evaluation and comparison to at-the-time state-of-the-art is in Table 5.

Conditional synthesis by example. Generating new samples as variations of the example input is an easier approach. The encoder generates latent vectors μ, σ for each hierarchy level. The synthesis is equal to the random sampling with the condition to use encoded latent vectors with small variations or linearly interpolate between encoded vectors of two input examples. The decoder generates new variations of modified examples from sampled $z_1, \ldots, z_4$.

7.4 Conclusion

We designed an architecture for a generative neural network and a training strategy enforcing the hierarchical decomposition of real-world textures with details of various frequencies and to increase the reconstruction quality of VAE. We focused our designed network to synthesize texture because of its structural variability and challenging optimization. To confirm our claim, we generated several thousand samples and verified the quality of reconstruction and synthesis empirically and statistically by standard metrics used for measuring the quality of generated images and compared them to the state-of-the-art models. Our model reached comparable results and beat models designed for generating textures of lower resolution. However, the visual evaluation found the synthesized textures still lacking the photographic quality of real-world texture. In general GANs are more suitable for generating content thanks to the adversarial training that helps them optimize to learn various scales of details. On the other hand, using the hierarchical design allows the VAE model to decompose the texture into layers of structural details and noise.

8. Support for diagnosis of Breast cancer

Our first contribution to breast cancer diagnosis is a method with visually interpretable predictions to support the diagnosis. We use SiamNet to determine the similarity between pairs of tissue samples. We base our approach on retrieving the most similar atlas tissue samples to recommend the possible cancer type (class) according to the annotations. The visualized examples should draw the attention of the investigating histopathologist to the structure of a possible tumor, which occurs in both the predicted and in the evidence tissue slice.

We modify architecture of the Araujo et al. [4]. Our modification (Table 6) omits last 3 fully connected layers and instead used one with 512 neurons activated by ReLU activation function. The output is then used for computing the distance. The network input are 3 channeled color image pairs of resolution 512×512. We are using a single-step stride between convolutions and two-step stride in Max-Pooling to extract the most informative activations.

This research requires a large dataset of images with relevant information. According to the nature of histological image data, the rotation is irrelevant so we apply rotations by $3 \times 90^{\cdot}$ and mirroring in both directions. To unify the illumination and staining variations, we apply standard staining normalization by Macenko et al. [24]. The radius of a cell nuclei ranges from 3 to 11 pixels, therefore we sample patches of size 128×128 pixels to preserve the sufficient number of nuclei. After all augmentations we generate 280 new samples from each image making together 28000 samples per class. These are filtered and reduced of samples with diagnostically irrelevant regions - background, glass, fat, ...

Visual classes. The two cut-out samples even though classified as the same tumor type can be different due to tissue structures. This can cause issues during training as it would be difficult for network to map visually different images to one spatial cluster in latent space. Therefore we introduce Visual classes. To separate visually similar and different samples we experimented with RGB and HSV color models and Local Binary Patterns to extract a histogram as global descriptor of a sample. Then the samples with correlation greater than 0.75 are considered strongly similar and greater than 0.6 are weakly similar. The network then receives input mini-batches sampled with this constraints.

8.1 Classification strategies

After determining the similarity between analyzed and atlas samples it is required to state the probable tumor class. We experiment with 5 strategies based on retrieval techniques and atlas ordering.

1. **40 Sample Atlas predictions** The baseline approach is generating the 40 image atlas samples that

Table 5: Fréchet Inception Distance (FID) - Synthesis task

Texture	Zhao[33]	SGAN[14]	nSTGAN[34]	TexGAN[32]	Ours
bricks	315.5	286.17	**199.25**	460	270.59
camo	377.27	349.63	294.81	367.42	**278.13**
tiles	412.83	**187.39**	491.15	415.62	489.97
wall	**265.13**	363.7	279.37	402.13	403.43

Table 6: Architecture of branch networks used in Siamese network.

Layer	**Dimensions**	**Kernel size**	**Receptive area**(μm)
Input	$512 \times 512 \times 3$		0.4×0.4
Conv	$510 \times 510 \times 16$	3×3	1×1
MaxPool	$170 \times 170 \times 16$	3×3	2×2
Conv	$168 \times 168 \times 32$	3×3	4.6×4.6
MaxPool	$84 \times 84 \times 32$	2×2	5.9×5.9
Conv	$84 \times 84 \times 64$	3×3	11×11
MaxPool	$42 \times 42 \times 64$	2×2	13×13
Conv	$42 \times 42 \times 64$	3×3	24×24
MaxPool	$14 \times 14 \times 64$	3×3	34×34
Conv	$12 \times 12 \times 32$	3×3	63.8×63.8
MaxPool	$12 \times 12 \times 32$	3×3	94.1×94.1
FC+ReLU	512		215×215

all further approaches use for classification. The threshold (Equation 4) from ROC for binary classification determines the level of similarity. After the thresholding, the remaining atlas samples are ordered to show the most possible classes based on the similarity of patches.

$$\text{threshold} = \sqrt{(1 - \text{recall})^2 + (1 - \text{specificity})^2} \tag{4}$$

2. **3 closest samples** We can classify the sample according to 3 the most similar pairs. The class is determined by the threshold from ROC curve and according to the class of the remaining sample.

3. **3 closest samples with majority vote** Inspired by [4], the selection of K the most similar samples, but the class is determined by majority vote. If the voting is not successful, then the class is assigned based on priority queue Invasive carcinoma, In Situ carcinoma, benign tumor, normal tissue or by the class of the very strongly similar sample.

4. **Continuous classification** A gradual categorization based on a medical diagnosis perspective. The binary classification of the malignancy then specific subcategories. All predictions are sorted by most similar slices and the superclass with five samples in this series determines the tumor class.

5. **Mean distance** The similarity is a distance of feature vectors, and the lower distance the more similar are the samples. However, there may be outliers and some specific samples may be very similar even though the samples are from different tumor types. The class is determined according to the category of similar samples with the smallest mean and standard deviation of the distances.

8.2 Discussion

Table 7: The comparison to the methods for classification of Breast Cancer.

	Accuracy
AlexNet[9]	0.53
DenseNet-18[9]	0.71
DenseNet-121[9]	0.79
DenseNet-161[9]	**0.98**
Chennamsetty2018[9]	0.87
Spanhol2016[30]	0.86
Rakhlin2018[27]	v0.938
Araújo2017[4]	0.776
Alom2019[3]	**0.98**
Our Method	0.652

Our method reaches average results that can be caused by confusing pairs with insufficient visibly similar structures. The color histograms are global descriptors carrying too general information. Also, the LBP, even though a strong local descriptor, does not describe a complex and heterogeneous texture with large patterns sufficiently. However, the results of both methods are sufficient to prove the concept of explanation and interpretation of the reasoning behind the classification for histopathology processes and provide a robust visual tool recommending suspicious tissue samples for further analysis. After consultation with the histopathologist, they considered the slices too small, containing very little information about the potential tumor. But that is a general problem with datasets of cut-out samples.

We can conclude that the classification of malignancy is more successful and more stable than more specific classification into four cancer classes provided from dataset. The results demonstrate the positive impact of filtering the visual classes and training only the visually similar/dissimilar pairs without confusing the network. The main contribution of our research is an interpretable and self-explainable similarity approach using a Siamese convolutional neural network for classification. Our classification method is transparent in interpreting the reason for classifying the sample by visualizing similar samples with structures significant for the final classification and diagnosis and still utilizing the strengths of deep learning.

9. Conclusions

This paper focuses on processing texture as a basic visual surface characteristic of objects and related locations. We applied modern deep learning solutions to selected topics: texture similarity evaluation; texture synthesis; analysis of histology images.

We trained a Siamese Network to classify texture according to their similarity that was determined by the distance between their extracted representative feature vectors. Our experiments proved differences in optimizing networks by different distance metrics. The achieved results showed a significant downfall in classification accu-

racy with Mahalanobis and Bray-Curtis distances compared to the network trained with Euclidean distance.

We designed a hierarchical generative model for texture synthesis from random input and modifying the example inputs. We based the architecture on Variational Auto-Encoders and tried to increase the quality of generated output often suffering from blurriness. Our training approach forced the network to decompose the texture details to layers according to their frequency. This design contributed to the quality of reconstructed inputs and allowed the synthesis of new samples, even though the quality was far from real-world texture.

Furthermore, we applied the gained knowledge of texture analysis and determining similarity between patches into domain of histology and microscopy images. We developed an approach that compares the slices of tissue and determines the type of tissue and cell structures into benign and malignant tumors by comparing to atlas samples. The post-research consultations with medics marked the approach as interesting, however, using such small patches would not give them enough information for plausible diagnosis. The consultations with medics provided us precious information, that the classification of cancer is more reliable from larger regions and also that larger regions are required to find cell structures responsible for creation of tumors.

The future work may bring improvements with classification precision and quality of generated images. The approach for texture similarity may be used for discriminator in GAN networks. The classification of breast cancer benefits from larger regions with visible architectural structure, and the results can be improved with large dataset with equal distribution of all classes.

References

[1] Who, breast cancer fact sheets. *World Health Organization*, Mar 2021.

[2] A. Akl, C. Yaacoub, M. Donias, J.-P. Da Costa, and C. Germain. A survey of exemplar-based texture synthesis methods. *Computer Vision and Image Understanding*, 172:12 – 24, 2018.

[3] M. Z. Alom, C. Yakopcic, M. S. Nasrin, T. M. Taha, and V. K. Asari. Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network. *Journal of Digital Imaging*, 2019.

[4] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho. Classification of breast cancer histology images using Convolutional Neural Networks. *PLOS ONE*, 12(6):e0177544, 6 2017.

[5] G. Aresta, T. Araújo, S. Kwok, S. S. Chennamsetty, M. Safwan, V. Alex, B. Marami, M. Prastawa, M. Chan, M. Donovan, G. Fernandez, J. Zeineh, M. Kohl, C. Walz, F. Ludwig, S. Braunewell, M. Baust, Q. D. Vu, M. N. N. To, E. Kim, J. T. Kwak, S. Galal, V. Sanchez-Freire, N. Brancati, M. Frucci, D. Riccio, Y. Wang, L. Sun, K. Ma, J. Fang, I. Kone, L. Boulmane, A. Campilho, C. Eloy, A. Polónia, and P. Aguiar. BACH: Grand Challenge on Breast Cancer Histology Images. *Medical Image Analysis*, 56:122–139, 8 2019.

[6] J. Bruna and S. Mallat. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

[7] Y. Celik, M. Talo, O. Yildirim, M. Karabatak, and U. R. Acharya. Automated invasive ductal carcinoma detection based using deep transfer learning with whole-slide images. *Pattern Recognition Letters*, 133:232–239, 2020.

[8] R. Chandra, S. Grover, K. Lee, M. Meshry, and A. Taha. Texture synthesis with recurrent variational auto-encoder. *arXiv preprint arXiv:1712.08838*, 2017.

[9] S. S. Chennamsetty, M. Safwan, and V. Alex. Classification of Breast Cancer Histology Image using Ensemble of Pre-trained Neural Networks. pages 804–811. Springer, Cham, 6 2018.

[10] A. Frühstück, I. Alhashim, and P. Wonka. TileGAN: Synthesis of large-scale non-homogeneous textures. *ACM Transactions on Graphics*, 38(4), 2019.

[11] Y. Gao, Y. Gan, L. Qi, H. Zhou, X. Dong, and J. Dong. A perception-inspired deep learning framework for predicting perceptual texture similarity. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2019.

[12] L. Gatys, A. S. Ecker, and M. Bethge. Texture synthesis using convolutional neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 262–270. Curran Associates, Inc., 2015.

[13] Z. Han, B. Wei, Y. Zheng, Y. Yin, K. Li, and S. Li. Breast Cancer Multi-classification from Histopathological Images with Structured Deep Learning Model. *Scientific Reports*, 7(1):4172, 12 2017.

[14] N. Jetchev, U. Bergmann, R. Vollgraf, and Z. Research. Texture Synthesis with Spatial Generative Adversarial Networks. Technical report, 2016.

[15] J. M. Joyce. Kullback-leibler divergence. In *International encyclopedia of statistical science*, pages 720–722. Springer, 2011.

[16] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks, 2018.

[17] S. Khan, N. Islam, Z. Jan, I. Ud Din, and J. J. P. C. Rodrigues. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. *Pattern Recognition Letters*, 125:1–6, 6 2019.

[18] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. (Ml):1–14, 2013.

[19] M. Kokare, B. N. Chatterji, and P. K. Biswas. Comparison of similarity metrics for texture image retrieval. *TENCON 2003 Conference on Convergent Technologies for AsiaPacific Region*, 2:571–575, 2003.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, pages 1–9, 2012.

[21] L. Liu, J. Chen, P. Fieguth, G. Zhao, R. Chellappa, and M. Pietikäinen. From BoW to CNN: Two Decades of Texture Representation for Texture Classification. *International Journal of Computer Vision*, 127(1):74–109, 1 2019.

[22] T. P. M.-y. Liu, T.-c. Wang, J.-y. Zhu, and U. C. Berkeley. Semantic image synthesis with spatially-adaptive normalization. 2019.

[23] J. Lou, L. Qi, J. Dong, H. Yu, and G. Zhong. Learning perceptual texture similarity and relative attributes from computational features. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 2540–2546, 2016.

[24] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. IEEE, 2009.

[25] J. J. L. Manish H. Bharati and J. F. MacGregor. Image texture analysis: methods and comparisons. In *Chemometrics and Intelligent Laboratory Systems*, volume 72, pages 55–71, 6 2004.

[26] J. Polec, W. Benesova, R. Vargic, I. Ilčíková, and T. Csóka. Texture feature extraction using an orthogonal transform of arbitrarily shaped image regions. *Journal of Electronic Imaging*, 25(6):061413, 9 2016.

[27] A. Rakhlin, A. Shvets, V. Iglovikov, and A. A. Kalinin. Deep Convolutional Neural Networks for Breast Cancer Histology Image Analysis. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10882 LNCS, pages 737–744. Springer, Cham, 6 2018.

[28] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic Backpropagation and Approximate Inference. 2014.

[29] F. A. Spanhol, P. R. Cavalin, L. S. Oliveira, C. Petitjean, and L. Heutte. Deep features for breast cancer histopathological image classification. In *2017 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2017*, volume 2017-January, pages 1868–1873. IEEE, 10 2017.

[30] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte. Breast cancer histopathological image classification using Convolutional Neural Networks. *Proceedings of the International Joint Conference on Neural Networks*, 2016:2560–2567, 7 2016.

[31] S. Tripathi, S. K. Singh, and H. K. Lee. An end-to-end breast tumour classification model using context-based patch modelling - A BiLSTM approach for image classification. *Computerized Medical Imaging and Graphics*, 87:101838, jan 2021.

[32] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays. Texturegan: Controlling deep image synthesis with texture patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2018.

[33] S. Zhao, J. Song, and S. Ermon. Learning hierarchical features from deep generative models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4091–4099. JMLR. org, 2017.

[34] Y. Zhou, Z. Zhu, X. Bai, D. Lischinski, D. Cohen-Or, and H. Huang. Non-Stationary Texture Synthesis by Adversarial Expansion. 37(4), 2018.

[35] J. Zujovic, T. N. Pappas, and D. L. Neuhoff. Structural texture similarity metrics for image analysis and retrieval. *IEEE Transactions on Image Processing*, 22(7):2545–2558, 2013.

[36] J. Zujovic, T. N. Pappas, D. L. Neuhoff, R. Van Egmond, and H. De Ridder. Effective and efficient subjective testing of texture similarity metrics. *JOSA A*, 32(2):329–342, 2015.

Selected Papers by the Author

L. Hudec and W. Benesova. Texture similarity evaluation via siamese convolutional neural network. In *2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 1–5. IEEE, 2018.

M. Jakab, L. Hudec, and W. Benesova. Partial disentanglement of hierarchical variational auto-encoder for texture synthesis. In *IET Computer Vision*, 14, (8): 564–574, 2020.

D. Hradel, L. Hudec, and W. Benesova. Interpretable diagnosis of breast cancer from histological images using siamese neural networks. In *Twelfth International Conference on Machine Vision (ICMV 2019)*, volume 11433, page 1143321. International Society for Optics and Photonics, 2020.

A. Felšöová, T. Sloboda, L. Hudec, M. Koblížek, P. Pohunek, V. Martinů, Ž. Varényiová, S. Kadlecová, J. Uhlík. Quantitative Assessment of Ciliary Ultrastructure with the Use of Automatic Analysis: PCD Quant In *Diagnostics*, 11(8), 1363, 2021, doi: 10.3390/diagnostics11081363.

J. Mrva, Š. Neupauer, L. Hudec, J. Ševcech, P. Kapec Decision Support in Medical Data Using 3D Decision Tree Visualisation In *2019 E-Health and Bioengineering Conference (EHB)*, pages 1–4, IEEE, 2019.