β-VARIATIONAL AUTOENCODERS FOR LEARNING INVERTIBLE LOCAL IMAGE DESCRIPTORS

NINA ŽIŽAKIĆ AND ALEKSANDRA PIŽURICA

Group for Artificial Intelligence and Sparse Modelling (GAIM), TELIN, Ghent University, Ghent, Belgium
{nina.zizakic, aleksandra.pizurica}@ugent.be

Abstract. In this paper, we propose an efficient method for learning a local image descriptor and its inversion function using a modified version of a variational autoencoder (VAE) - a β-VAE. We examine different values of β in the loss function of the β-VAE to find the an optimal balance between incentivising the similarities between input patches to be preserved in latent space, and ensuring good reconstruction of the patches from their encodings in latent space. Our proposed descriptor demonstrates patch retrieval comparable to the reference autoencoder-based local image descriptor, and also shows improved reconstruction of patches from their encodings.

1 Introduction

Local image descriptors are an important component of many image processing tasks, such as object tracking, object recognition, image denoising, image stitching, and image retrieval.

Traditionally, local image descriptors have been designed using hand-crafted features, such as SIFT [13], HOG [7], GLOH [15], SURF [3], and BRIEF [4]. In recent years, the development of deep learning techniques has led to a new generation of learned local image descriptors [23, 17, 2, 9], showing excellent results [1].

Most of these learning approaches are supervised methods, which rely on labelled datasets. However, in many real-life applications, such datasets are not available. In contrast to supervised methods, unsupervised methods such as autoencoders and variational autoencoders, by definition, do not require labelled data. Autoencoders have already been used to learn local image descriptors [5, 20, 24, 21], showing promising results. However, the fundamental problem with autoencoders is that their latent space may not be continuous or may not allow for easy interpolation. These issues undermine the descriptors similarity preservation property. Variational autoencoders [12] have been created to tackle this problem in general, but have not been applied to the problem of learning local image descriptors.

Inverting local image descriptors has been an active area of research in the past decade, starting with the prominent work by Weinzaepfel et al. [22] on reconstructing an image from its SIFT descriptors. The authors used a database of descriptors and their corresponding patches to search for the nearest neighbour to the query descriptor,

and then take the patch connected to the retrieved nearest neighbour. Further works on inverting other descriptors followed, including the inversion of binary descriptors [8] and of HOG [19]. A more recent paper by Mahendran et al. [14] considers inverting descriptors back into patches using deep learning.

In this paper, we propose an unsupervised method that specialises in learning both a descriptor function that maps image patches to their encodings and an inverting function that decodes these encodings back into the original image patches. To the best of our knowledge, we are the first to present a descriptor that is optimised for inversion. Our method is using β-variational autoencoders, which we tweak to achieve an optimal balance between preserving similarities between patches and achieving good invertibility. We perform a thorough analysis of how the β value influences this trade-off. Due to their unsupervised nature, variational autoencoders do not require a labelled dataset. Furthermore, their generalisation, β-VAEs, are intrinsically well-suited for learning both the encoding function and its inversion, as we will show in Section 3. The existing autoencoder-based descriptors [5, 20, 24] do not present inverting results. Our experimental results clearly show a better inversion ability of the proposed method compared to the reference autoencoder-based approach [20]. To our knowledge, there are no other works using variational autoencoders to learn local image descriptors.

In the following section, we give a brief introduction to the classical and variational autoencoders. We describe our method in Section 3 and present the results of our experiments in Section 4 with discussion. Section 5 concludes this paper.

2 Preliminaries

Autoencoders are unsupervised neural networks used for learning efficient representations of data [11, 18, 6]. An autoencoder consists of two parts, an encoder and a decoder, and is trained by minimising the reconstruction error between the input and output, while imposing some constraints (usually dimensionality) on the middle layer.

The application of autoencoders to the problem of descriptor learning was first proposed by Chen et al. [5]. In our previous work [20, 24], we proposed autoencoder-based patch descriptors designed for applications with many patch comparisons within a single image. These approaches, however, have no way of enforcing the continuity of the latent space and thus, are unable to guarantee that the learned encodings are useful, i.e., that they posses the similarity preserving property – a key property for local image descriptors.

To tackle the problem of a lack of continuity in the latent space, Kingma et al. have proposed variational autoencoders (VAEs) [12]. Similar to classical autoencoders, VAEs consist of an encoder and a decoder, with a middle layer on which a dimensionality constraint is imposed. In contrast to classical autoencoders, however, variational autoencoders are probabilistic models that assume a prior distribution of the latent space, giving significant control over how we want to model the latent distribution. The data x has a likelihood $p(x|z)$ (the decoder distribution) that is conditioned on latent variables z. The posterior (typically Gaussian) is approximated with a family of distributions $q(z|x)$ (the encoder distribution). Apart from minimising the reconstruction loss, VAEs also minimise the Kullback–Leibler (KL) divergence between the true posterior $p(z)$ and its approximation $q(z|x)$. Given a dataset $X = \{x^{(1)}, x^{(2)}, ..., x^{(n)}\}$, the goal of a VAE is to minimise the negative log-likelihood lower bound:

$$\mathcal{L}(\theta, \phi; x^{(i)}) = \\ -\mathbb{E}_{q_\phi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)] + D_{KL}[q_\phi(z|x^{(i)})||p_\theta(z)], \tag{1}$$

where the encoder and decoder distributions are parametrised by ϕ and θ, respectively.

The first term promotes a good reconstruction of the input data samples, while the second term enforces that the distribution of the latent space is as close as possible to the multivariate Gaussian distribution.

Higgins et al. [10] have proposed a variant of a variational autoencoder named β-VAE. In a β-VAE, the loss function from Equation (1) is modified to add more weight on the second term, sacrificing the reconstruction capabilities of the VAE in order to make the latent space smoother and to allow for its better disentanglement:

$$\mathcal{L}(\theta, \phi; x^{(i)}) =$$

$$-\mathbb{E}_{q_\phi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)] + \beta D_{KL}[q_\phi(z|x^{(i)})||p_\theta(z)],$$

In the next section, we describe how we use β-VAEs to learn invertible local image descriptors.

3 β-VAEs for local image descriptors

We propose using a β-variational autoencoder for simultaneous learning of local image descriptors and their reconstruction back into image patches. Due to the nature of their architecture, both classical and variational autoencoders are ideal for the simultaneous learning of the descriptor function (the encoder part of the autoencoder) and the reconstruction function (the decoder part). However, unlike classical autoencoders, VAEs include additional regularisation that allows modelling the latent space to be continuous and to be easy to interpolate across, ensuring that similar input data samples (patches) get mapped to similar points in the latent space (encoding), and vice versa. This similarity-preserving property is a property of paramount importance for local image descriptors. We also hypothesise that the additional regularisation of VAEs will allow for learning sharper reconstructions in comparison to methods based on classic autoencoders, which we will show empirically in the next section.

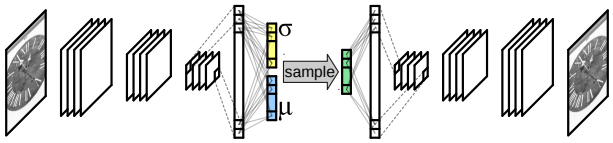

Fig. 1: Architecture of the variational autoencoder we used for learning local image descriptors.

In β-VAE, a generalisation of the loss function of VAEs is achieved by adding the β weight to the KL term. In this way, we can control the trade-off between learning to faithfully reconstruct the input patches and preserving patch similarities in the latent space. By setting the right value of β we can increase the influence of the reconstruction term to ensure good invertibility of the descriptor. In contrast to descriptors based on classical autoencoders, however, the KL term in the VAE loss function ensures the continuity of the latent space which could not be guaranteed when using the classical autoencoders.

Once we trained the β-VAE, the encoder part of it is our descriptor, and the decoder part is the inversion function that maps the patch encodings back to the original patches.

Figure 1 illustrates the architecture of the variational autoencoder used in this paper. The encoder consists of three convolutional layers followed by the fully-connected layers for the means and variances of Gaussian distributions. From these layers, we sample a vector that is the encoding of the input patch. We set the dimensionality of the latent space M (and therefore, the mean, variance, and the sampling layers) to be 128. The decoder architecture mirrors that of the encoder – at the beginning, there is one layer fully-connected to the sample (encoding), followed by three transposed convolutional layers. The dimensions of the output patch of our VAE are the same as the dimensions of the input.

Following the notation from [10], we use β_{norm} as the main hyperparameter that we vary. β_{norm} is defined as follows:

$$\beta_{norm} = \frac{\beta M}{N},$$

where M is the size of latent space and N is the input size. By normalising the β value, the analysis that we present in the following section can be applied to datasets of different patch size and different desired encoding sizes. We vary the β_{norm} values over several orders of magnitude – from 10^{-5} to 10^{2}. In the following section, we show how the β_{norm} value influences the patch retrieval of the descriptor and its invertibility.

We use rectified linear unit (ReLU) activation functions after all layers, except the last layer, where we use the sigmoid activation function instead. We use Adam optimiser to learn the weights of the VAE, which is trained on a dataset of 80k 56×56 patches that were extracted from the images from the ImageNet dataset using FAST (Features from Accelerated Segment Test) algorithm for feature detection [16]. The ratio between training, validation, and test set is $8 : 1 : 1$.

4 Experimental results

In this section, we show how the value of β_{norm} influences the proposed β-VAE–based local image descriptor and its performance with respect to patch retrieval (the main task for which local image descriptors are designed) and patch inversion from the patches' encodings.

We also evaluate both the retrieval and inversion capabilities of the proposed approach in comparison with a reference autoencoder-based descriptor. We compare our method only to this autoencoder-based descriptor, since non-autoencoder-based descriptors have no straightforward way of being inverted and thus give us no way of comparing their invertibility.

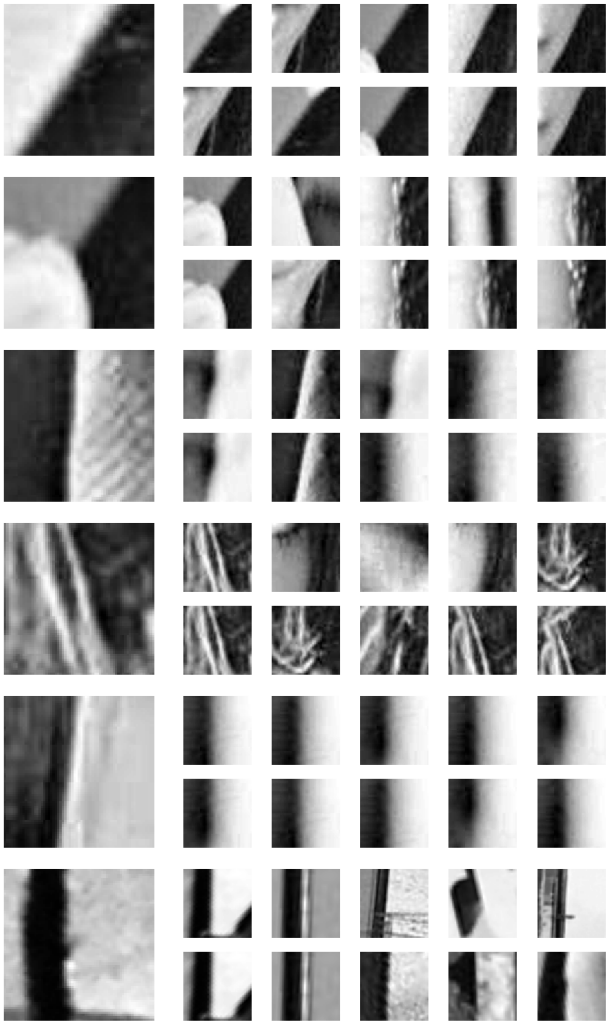

Fig. 2: Patch retrieval examples. Large patch is the query patch. Top rows: AE-based descriptor from [20]; bottom rows: proposed VAE-based descriptor.

4.1 Evaluation on patch retrieval

Patch retrieval evaluation is performed as follows. We select a set of query patches within a test dataset of patches. For each query patch, we retrieve the most similar patches by comparing their encodings as calculated by the descriptors. We show some examples of patches retrieved in such a way in Figure 2. The quality of patch retrieval is then evaluated based on two metrics (peak signal-to-noise

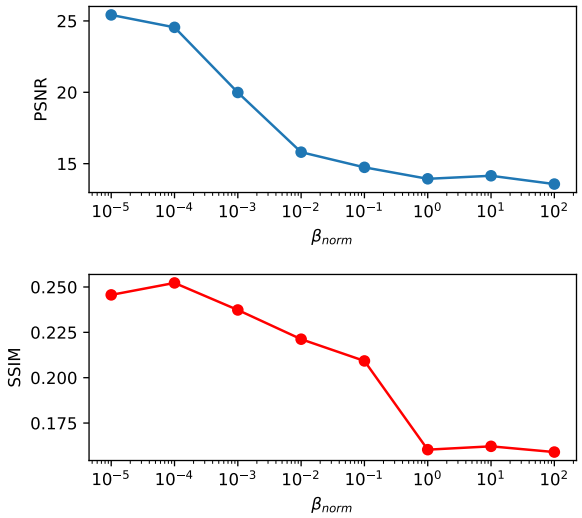

Fig. 3: Comparison of patch retrieval performance for different β_{norm} values.

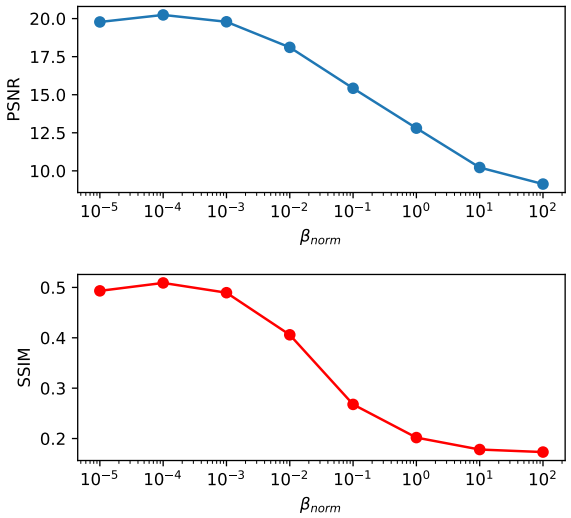

Fig. 4: Comparison of patch reconstruction performance for different β_{norm} values.

ratio (PSNR) and structural similarity index (SSIM)) between the query patches and patches deemed most similar to the queries based on the encodings computed with descriptors.

We first examine the patch retrieval capabilities of the proposed β-VAE–based descriptor for different β_{norm} values. In Figure 3, we see that, according to the PSNR metric, the patch retrieval seems to be the best when the β_{norm} value is the lowest, i.e., when the KL divergence term of the loss function is the closest to 0. However, when using a metric that better mimics human's perception of differences between images, SSIM, we see that adding a KL term is beneficial, as the patch retrieval in terms of SSIM shows a peak at β_{norm} value of 10^{-4}. In our case, this translates to β value of 0.0032.

Secondly, we compare the patch retrieval capabilities to an existing autoencoder-based descriptor [20]. We present our results in Table 1. We observe that the descriptor proposed in this paper is outperformed by the descriptor from [20] in terms of PSNR, however, in terms of SSIM, the proposed β-VAE–based descriptor shows slightly better performance. These results are consistent with the β_{norm} value analysis, since setting the β_{norm} to 0 would correspond to using a regular autoencoder.

Tab. 1: Patch retrieval performance comparison

	PSNR [dB]	SSIM
AE-based descriptor [20]	**26.0**	0.23
Proposed VAE-based descriptor	24.6	**0.25**

According to these experiments, we can claim that the proposed descriptor shows promising results in the main task for which descriptors are designed: retrieving patches.

4.2 Evaluation of invertibility

Now we evaluate the extent to which a descriptor can reconstruct the original patch from its encoding. For a test set of patches, we measure the difference between the original patch, and the patch reconstructed from the encoding via the descriptor.

We again first show the analysis for different β_{norm} values (Figure 4). Here we see the best performance (in terms of both PSNR and SSIM) for the β_{norm} value of 10^{-4}. We conclude that the KL divergence (albeit

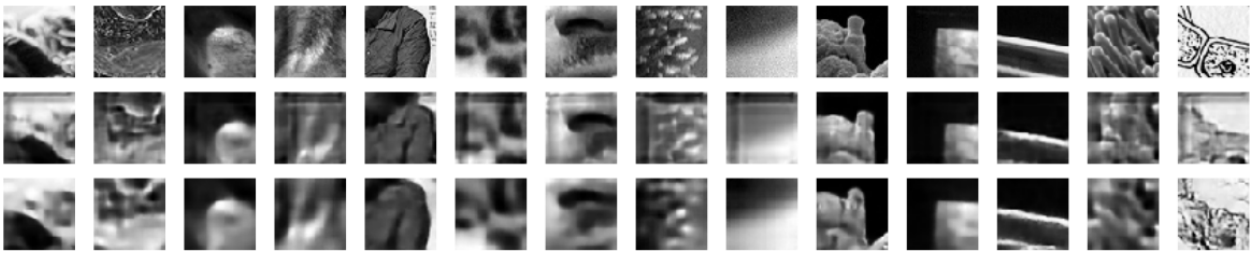

Fig. 5: Examples of patch reconstruction based on the descriptor's encoding. Top row: original patches; middle row: reconstructed patches using AE-based descriptor from [20]; bottom row: reconstructed patches using proposed VAE-based descriptor.

weighted very lightly) has a positive influence on the invertibility of the descriptor. Therefore, using β-VAE for invertible local image descriptor indeed makes sense – we can benefit from the regularisation by the KL divergence term and also adjust the extent to which it is taken into account.

We also compare our descriptor to the autoencoder-based descriptor from [20] (Table 2). The proposed descriptor shows better results than the descriptor from [20] across both metrics: PSNR and SSIM. In Figure 5, we show some examples of patches reconstructed with the proposed VAE-based descriptor. We can observe that the proposed descriptor outperforms the reference method and is able to reconstruct the patches with significant improvements in fidelity.

Tab. 2: Patch reconstruction performance comparison

	PSNR [dB]	SSIM
AE-based descriptor [20]	16.0	0.10
Proposed VAE-based descriptor	**20.2**	**0.51**

5 Conclusion

In this paper, we presented a novel approach based on β-variational autoencoders that combines the learning of a local image descriptor with the learning of its inversion. We showed that β-VAE is an excellent fit for this application, as it (being a VAE) introduces a KL divergence term to the loss function that acts as a regulariser and ensures smooth latent space, but at the same time, a β-VAE is more powerful than a general VAE since it allows slightly reducing this KL term and in this way optimising for the patch reconstruction from its encoding. We performed a thorough analysis of how the β_{norm} value influences the performance of the descriptor in patch retrieval and patch reconstruction from its encodings. We observed that setting β_{norm} to 10^{-4} gives the best performance on these tasks. Furthermore, we evaluated the proposed descriptor's patch retrieval abilities in comparison to a previous autoencoder-based method. Our VAE-based method showed improvements in terms of SSIM metric while appearing to perform slightly worse in terms of PSNR. Finally, we compared the invertibility of these two descriptors and showed that the proposed descriptor outperforms the reference descriptor from [20] in the two metrics that were assessed.

Acknowledgements

This research received funding from the Flemish Government (AI Research Program).

References

[1] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proceedings of the IEEE Conference on*

Computer Vision and Pattern Recognition (CVPR), July 2017.

[2] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 119.1–119.11, 2016.

[3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (SURF). *Computer vision and image understanding*, 110(3):346–359, 2008.

[4] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. BRIEF: Binary robust independent elementary features. In *European Conference on Computer Vision*, pages 778–792. Springer, 2010.

[5] Lin Chen, Franz Rottensteiner, and Christian Heipke. Feature descriptor by convolution and pooling autoencoders. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences-ISPRS Archives 40 (2015), Nr. 3W2*, 40(3W2):31–38, 2015.

[6] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011.

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, 2005.

[8] Emmanuel d'Angelo, Alexandre Alahi, and Pierre Vandergheynst. Beyond bits: Reconstructing images from local binary descriptors. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, pages 935–938. IEEE, 2012.

[9] Kun He, Yan Lu, and Stan Sclaroff. Local descriptors optimized for average precision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[11] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

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

[13] David G Lowe. Object recognition from local scale-invariant features. In *ICCV*, page 1150. IEEE, 1999.

[14] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5188–5196, June 2015.

[15] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, 27(10):1615–1630, 2005.

[16] Edward Rosten and Tom Drummond. Machine learning for high-speed corner detection. In Aleš

Leonardis, Horst Bischof, and Axel Pinz, editors, *Computer Vision – ECCV 2006*, pages 430–443. Springer Berlin Heidelberg, 2006.

[17] Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[18] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[19] Carl Vondrick, Aditya Khosla, Tomasz Malisiewicz, and Antonio Torralba. Hoggles: Visualizing object detection features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–8, 2013.

[20] Nina Žižakić, Izumi Ito, and Aleksandra Pižurica. Learning local image descriptors with autoencoders. In *Proc. IEICE Inform. and Commun. Technol. Forum ICTF 2019*, 2019.

[21] Nina Žižakić and Aleksandra Pižurica. Learned BRIEF – transferring the knowledge from hand-crafted to learning-based descriptors. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020.

[22] Philippe Weinzaepfel, Hervé Jégou, and Patrick Pérez. Reconstructing an image from its local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*, pages 337–344. IEEE, 2011.

[23] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[24] Nina Žižakić, Izumi Ito, Laurens Meeus, and Aleksandra Pižurica. Autoencoder-learned local image descriptor for image inpainting. In *BNAIC/BENELEARN 2019*, volume 2491, 2019.