# Invertible local image descriptors learned with variational autoencoders

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 local image descriptor and its inversion function using a variational autoencoder (VAE). We design a loss function of the VAE specifically for this purpose, which, on one hand, incentivises the similarities between input patches to be preserved in latent space, and on the other hand, ensures good reconstruction of the patches from their encodings in latent space. Our proposed descriptor demonstrates better patch retrieval compared to the reference autoencoder-based local image descriptor, and also shows improved reconstruction of patches from their encodings.**

***Index Terms*—local image descriptors, variational autoencoders, unsupervised deep learning**

## I. Introduction

Local image descriptors are a crucial 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 [1], HOG [2], GLOH [3], SURF [4], and BRIEF [5]. In recent years, the development of deep learning techniques has led to a new generation of learned local image descriptors [6]–[9], showing excellent results [10].

Most of these learning approaches are supervised methods, relying on relatively many annotated examples. Such specific labeled datasets are often not available. In contrast to supervised methods, unsupervised methods such as autoencoders and variational autoencoders, by definition, do not depend on labeled data. Autoencoders have already been used to learn local image descriptors [11]–[14], 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 [15] 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 research topic in the past decade, starting with the prominent work by Weinzaepfel et al. [16] on reconstructing an image from its SIFT descriptors. The authors used a database of descriptors and their corresponding patches to search for the nearest neighbor to the query descriptor, and then take the patch connected to the retrieved nearest neighbor. Further works on inverting other descriptors followed, including inverting binary descriptors [17] and inverting HOG [18]. A more recent paper by Mahendran et al. [19] 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 to the original image patches. To the best of our knowledge, we are the first to present a descriptor that is optimised for being inverted. Our method is based on variational autoencoders, in which we modify the loss function to achieve better invertibility. Due to their unsupervised nature, VAEs do not require a labeled dataset. Moreover, the nature of variational autoencoders makes them intrinsically well suited for learning both the encoding function and its inversion. The existing autoencoder-based descriptors [11]–[13] do not present inverting results. Our experimental results show clearly a better inversion ability of the proposed method compared to the reference autoencoder-based approach [12]. 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 III and present the results of our experiments in Section IV with discussion. Section V concludes this paper.

## II. Preliminaries

Autoencoders are unsupervised neural networks used for learning efficient representations of data [20]–[22]. 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. [11]. In our previous work [12], [13], 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,

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

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) [15]. 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:

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

where the encoder and decoder distributions are parametrised by $\phi$ and $\theta$, 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. [23] have previously proposed a modification to the loss function from Equation (1) that adds 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. In the next section, we describe our method where we add more weight to the reconstruction term to allow for a better reconstruction of the patches, while still keeping the second term as a form of regularisation that adds more smoothness to the latent space than the classical autoencoders.

## III. Proposed method

Our main contribution is a variational autoencoder used 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 Section IV.

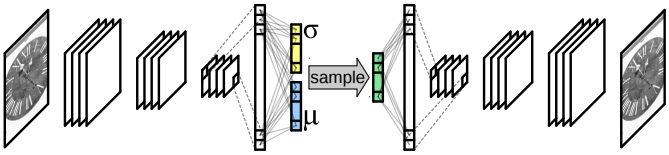


Fig. 1. Architecture of the variational autoencoder we used for learning local image descriptors.

We generalise the loss function for learning VAEs such to enable a trade-off between learning to faithfully reconstruct the input data samples, and preserving well patch similarities in the latent space. We therefore extend the expression in (1) with a weighing parameter $\gamma, \gamma > 1$ as follows:

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

thus increasing the influence of the reconstruction term. In contrast to descriptors based on classical autoencoders, however, the KL term in the VAE loss function ensures the continuity of the latent space.

Figure 1 illustrates the architecture of our variational autoencoder. 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 which is the encoding of the input patch. We set the dimensionality of the latent space (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 output patch of our VAE are the same as the dimensions of the input.

We use rectified linear unit (ReLU) activation functions after all layers, except the last layer, where we use sigmoid activation function instead. We use Adam optimiser to learn the weights of the VAE, which is trained on a dataset of 80k 5656 patches that were extracted from the images from the Imagenet dataset using FAST (Features from Accelerated Segment Test) algorithm for feature detection [24].

## IV. Experimental results

In this section, we evaluate both the retrieval and inversion capabilities of the proposed approach in comparison with a reference autoencoder-based descriptor.

### A. Evaluation on patch retrieval

We test the patch retrieval capabilities of our proposed descriptor by comparing it to an existing autoencoder-based descriptor [12]. The 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 3. The quality of patch retrieval is then evaluated based two metrics (peak signal-to-noise 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 can therefore claim that the proposed descriptor shows promising results at the task for which it was designed: retrieving patches.

In this paper, we compare our method only to an 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.

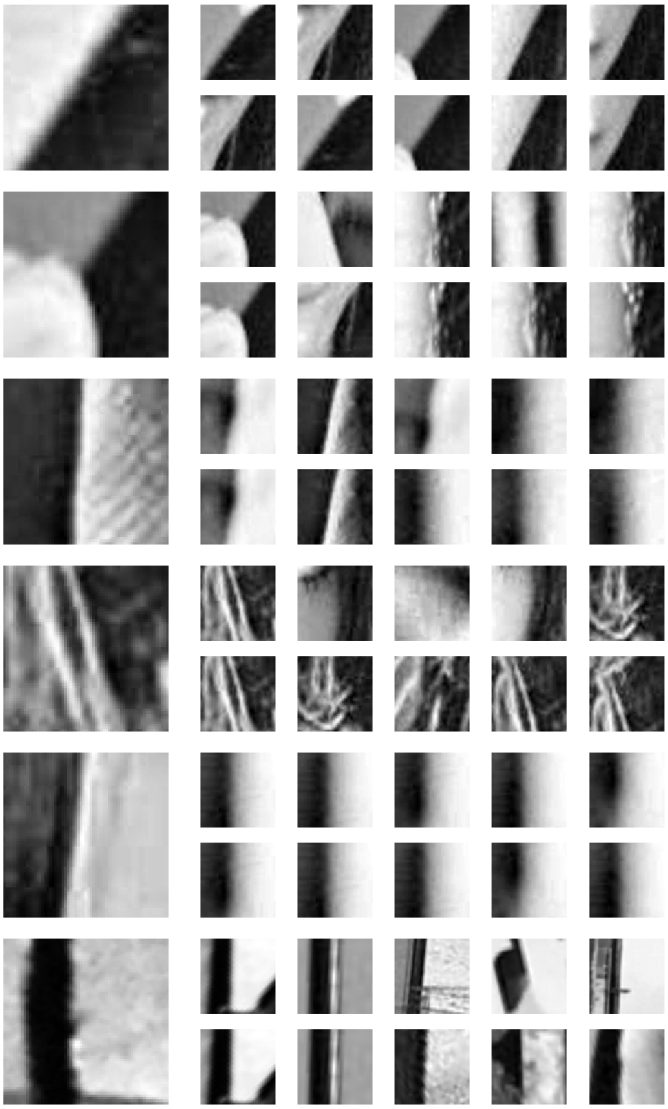

Fig. 2. Patch retrieval examples. Large patch is the query patch. Top rows: AE-based descriptor from [12]; bottom rows: proposed VAE-based descriptor.

We present our results in Table I. We observe that the descriptor proposed in this paper is outperformed by a small margin by the descriptor from [12] in terms of PSNR, however, when using a metric that better mimics a human's perception of differences between images, SSIM, the proposed descriptor shows slightly better performance.

TABLE I
PATCH RETRIEVAL PERFORMANCE COMPARISON

| | PSNR [dB] | SSIM |
|---|---|---|
| AE-based descriptor [12] | **29.2** | 0.32 |
| Proposed VAE-based descriptor | 29.1 | **0.33** |

### B. Evaluation of invertibility

Now we evaluate the extent to which the descriptor can reconstruct the original patch from its encoding. We again compare our descriptor to the autoencoder-based descriptor from [12]. For a test set of patches, we measure the difference between the original patch, and the patch reconstructed from the encoding via the descriptor. The proposed descriptor shows better results than the descriptor from [12] across both metrics: PSNR and SSIM. In Figure 3, 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.

TABLE II
PATCH RECONSTRUCTION PERFORMANCE COMPARISON

| | PSNR [dB] | SSIM |
|---|---|---|
| AE-based descriptor [12] | 16.0 | 0.10 |
| Proposed VAE-based descriptor | **24.5** | **0.50** |

## V. CONCLUSION

In this paper, we presented a novel method based on variational autoencoders that combines the learning of the local image descriptor with the learning of its inversion. We present a modification to the loss function of the VAE that results in learning better inversion, while keeping the KL term as a regularisation that also ensures the continuity of data point representations in the latent space. We have evaluated the proposed descriptor's patch retrieval abilities in comparison to a previous autoencoder-based method. Our VAE-based method shows improvements in terms of SSIM metric, while appearing to perform slightly worse in terms of PSNR. We also compare the invertibility of these two descriptors and show that the proposed descriptor outperforms the reference descriptor from [12] in both metrics that were assessed.

## REFERENCES


[1] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*. IEEE, 1999, p. 1150.

[2] 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)*, vol. 1, 2005, pp. 886–893 vol. 1.

[3] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 10, pp. 1615–1630, 2005.

Fig. 3. Examples of patch reconstruction based on the descriptor's encoding. Top row: original patches; middle row: reconstructed patches using AE-based descriptor from [12]; bottom row: reconstructed patches using proposed VAE-based descriptor.


[4] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[5] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary robust independent elementary features," in *European Conference on Computer Vision*. Springer, 2010, pp. 778–792.

[6] S. Zagoruyko and N. 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.

[7] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer, "Discriminative learning of deep convolutional feature point descriptors," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[8] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks." in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016, pp. 119.1–119.11. [Online]. Available: https://dx.doi.org/10.5244/C.30.119

[9] K. He, Y. Lu, and S. Sclaroff, "Local descriptors optimized for average precision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] V. Balntas, K. Lenc, A. Vedaldi, and K. 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.

[11] L. Chen, F. Rottensteiner, and C. 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*, vol. 40, no. 3W2, pp. 31–38, 2015.

[12] N. Žižakić, I. Ito, and A. Pižurica, "Learning local image descriptors with autoencoders," in *Proc. IEICE Inform. and Commun. Technol. Forum ICTF 2019*, 2019.

[13] N. Žižakić, I. Ito, L. Meeus, and A. Pižurica, "Autoencoder-learned local image descriptor for image inpainting," in *BNAIC/BENELEARN 2019*, vol. 2491, 2019.

[14] N. Žižakić and A. 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.

[15] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[16] P. Weinzaepfel, H. Jégou, and P. Pérez, "Reconstructing an image from its local descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011)*. IEEE, 2011, pp. 337–344.

[17] E. d'Angelo, A. Alahi, and P. Vandergheynst, "Beyond bits: Reconstructing images from local binary descriptors," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012, pp. 935–938.

[18] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, "Hoggles: Visualizing object detection features," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1–8.

[19] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 5188–5196.

[20] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[21] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[22] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 215–223.

[23] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, " -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.

[24] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds. Springer Berlin Heidelberg, 2006, pp. 430–443.