Variational Autoencoder with Gaussian Random Field prior: application to unsupervised animal detection in aerial images

Hugo Gangloff, Minh-Tan Pham, Luc Courtrai, Sébastien Lefèvre

Variational Autoencoder with Gaussian Random Field prior: application to unsupervised animal detection in aerial images

Hugo Gangloff, Minh-Tan Pham, Luc Courtrai, Sébastien Lefèvre
IRISA, Université Bretagne Sud, UMR 6074, 56000 Vannes, France

***Abstract*—In real world datasets of aerial images, the objects of interest are often missing, hard to annotate and of varying aspects. The framework of unsupervised Anomaly Detection (AD) is highly relevant in this context, and Variational Autoencoders (VAEs), a family of popular probabilistic models, are often used. We develop on the literature of VAEs for AD in order to take advantage of the particular textures that appear in natural aerial images. More precisely we propose a new VAE model with a Gaussian Random Field (GRF) prior (VAE-GRF), which generalize the classical VAE model, and we provide the necessary procedures and hypotheses required for the model to be tractable. We show that, under some assumptions, the VAE-GRF largely outperforms the traditional VAE and some other probabilistic models developed for AD. Our results suggest that the VAE-GRF could be used as a relevant VAE baseline in place of the traditional VAE with very limited additional computational cost. We provide competitive results on the MVTec dataset and two other datasets dedicated to the task of unsupervised animal detection in aerial images.**

***Index Terms*—Variational autoencoders, anomaly detection, Gaussian random fields, aerial images**

I. Introduction

This article introduces a new deep probabilistic model followed by its application to real world data. First, a new model of Variational Autoencoders (VAEs) with a Gaussian Random Field (GRF) prior is presented. This offers a relevant way to model images with strong spatial correlations. Second, the VAE-GRF is used in the context of Anomaly Detection (AD). More precisely, we address the real world application of unsupervised detection of animals in aerial images. Both of these topics are now presented.

A. Variational Autoencoders and some extensions

Our work focuses on VAEs. They are generative probabilistic models, widely used in the context of anomaly detection. They are popular for several reasons. First, they are derived from a sound probabilistic background and robust training procedures have been developed for such models. VAEs are also widely used in unsupervised settings such as ours. Finally, since a VAE is a generative model, samples can be generated from the model for any kind of purpose once the model has been trained.

In this article, we introduce the VAE-GRF model for images which makes use of a prior in the form of a stationary bi-dimensional GRF on the torus. Indeed, we construct a convolutional latent space in which we assume that the model learns a compressed representation of the input images. Therefore, such an approach offers a refinement of the prior distribution over the latent random variables as compared to the independent and identically distributed standardized Gaussian prior from the traditional VAE context. We study the advantages of the VAE-GRF in terms of modeling. We also demonstrate how the stationary and torus assumptions are used to develop a model with efficient computations despite the full covariance structure of the prior.

Over the last few years, several works have considered more complex priors for VAEs to be able to propose a more relevant modeling of the data. A recent review is available in [9]. The closest works to ours are VAEs with Gaussian Process priors, as explored in [6], [28] or in [17]. They propose a Gaussian Process (one-dimensional GRF) over the latent space. However the latent spaces they use are one dimensional which differs from the convolutional latent space we propose. Indeed, they aim at encoding inter-sample correlations between hidden random variables while we aim at encoding the spatial correlations in the latent space, for each sample. Therefore, as opposed to these approaches, we use a bi-dimensional GRF, which is more costly and requires different hypotheses to maintain a tractable model (stationarity and torus). In addition, since AD is not the focus of these approaches and since it remains unclear how Gaussian Process VAEs could be adapted for real-world image processing, we will discard them from comparisons.

B. Anomaly Detection in aerial images

In this article, we address the problem of unsupervised detection of animals in aerial images. Detecting, tracking and counting animals are real-world applications that are more and more studied in the literature, especially thanks to the unprecedented availability of Unmanned Aerial Vehicles (UAV) [1] [5] [24]. However, to the best of our knowledge, very few works consider the unsupervised context, which does not require the tedious and costly annotation step [3]. As the number of data keeps increasing, it becomes a necessity

This work was done as a part of the Game of Trawls project. We thank the European Maritime and Fisheries Fund (contract number 18/2216442) and France Filière Pêche (contract number 19/1000544) for funding. It was also supported by the SEMMACAPE project, which benefits from an ADEME (*Agence de la transition écologique*) grant under the "Sustainable Energies" call for research projects (2018–2019).

to develop reliable unsupervised approaches which can solve these tasks.

In the unsupervised context, we can not use any annotation to learn a representation of the animals. Moreover, in the context of animal detection with UAV, most of the captured images are empty, *i.e.*, they do not contain any animal. For these reasons, in this article, we resort to the principles of unsupervised AD in order to detect animals. AD is a vast research field and several reviews have already been published on the topic [31] [25]. An introduction to AD is proposed in Sec. III-A. In the context of this article, the AD approaches we will compare the VAE-GRF to are presented in Sec. III-C.

C. Outline of the article

The outline of the article is the following. We first present the new VAE-GRF model along with some theoretical backgrounds motivating the work. We then briefly review unsupervised AD and assess the new VAE-GRF on the AD task on a standard public dataset. Finally, we use the AD context to perform unsupervised animal detection in aerial images from two datasets. In all the experiments, the new model is compared with other state-of-the-art approaches.

Remark: The code of the VAE-GRF model to reproduce the experiments presented in this article is available at `https://github.com/HGangloff/vae_grf`.

II. Variational Autoencoders with Gaussian Random Field prior

A. Gaussian Random Fields

1) Definition: The probability density function of a GRF [30] with mean $\boldsymbol{\mu}$ and covariance matrix Σ (forming a set of parameters $\{\boldsymbol{\mu}, \Sigma\}$) is given by

$$p(\boldsymbol{x}) = (2\pi)^{-\frac{n}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\right), \quad (1)$$

where $|\cdot|$ refers to the determinant of a matrix and n is the dimension of $\boldsymbol{x}$. The GRF will be associated to the regular graph defined by an image: a node is associated to each pixel and nodes linked with each other in the graph are adjacent pixels (we consider the 8-nearest neighbors for each pixel). Then $\boldsymbol{x} = \{x\}_{s \in \mathcal{S}}$ where $\mathcal{S}$ is the set of nodes of the graph.

In the rest of the paper, we will consider GRFs with stationary mean and stationary covariance matrix, *i.e.*, the mean is the same for each x_s and the covariance between two x_s only depends on their distance (distance that will be defined later).

2) Spectral properties: In this article, along with the stationarity assumption, we will formulate the *torus assumption* on our images. In the torus assumption, the image borders are supposed wrapped like on a torus. One can observe that a stationary GRF defined on an image with torus assumption will have a covariance matrix which is block-circulant with circulant blocks [30]. Therefore, such a covariance matrix does not need to be fully stored, it is entirely defined by a smaller matrix called the *base* matrix. Indeed, for a matrix of dimension $l_x \times l_y$, its covariance matrix C has dimension $l_x l_y \times l_x l_y$. However, if both the stationarity and torus assumptions are made, the covariance matrix is block-circulant with circulant blocks and is entirely defined by its *base* matrix, $\text{base}(C)$, with dimension $l_x \times l_y$. All details about circulant and block-circulant matrices can be found in [30]. Matrix operations for block-circulant matrices with circulant blocks are efficiently computed with the Fourier transforms and are called the *spectral properties*[1]. The formulas we use in the article are given in App. A.

B. VAE-GRF: model definition

1) Generative model architecture: Recall that VAEs are generative probabilistic models which aim at modeling a distribution over the observations $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ with the help of latent random variables $\boldsymbol{z}$, such that, $p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$. However, in VAEs, computing the posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) / p_{\boldsymbol{\theta}}(\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{z}) / \int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) p_{\boldsymbol{\theta}}(\boldsymbol{z}) \mathrm{d}\boldsymbol{z}$ is intractable (the model likelihood, at the denominator, is, in general, an intractable integral). In VAEs, we thus introduce a variational distribution $q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})$ that aims at approximating $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ and whose parameters $\boldsymbol{\varphi}$ are learnt during the optimization process.

The VAE-GRF model we propose is first composed of a stochastic encoding network, with input $\boldsymbol{x}$, which maps to a convolutional latent space associated to the realizations of a random variable $\boldsymbol{z}$, following the ideas from [32] [11]. Thus, $\boldsymbol{z}$ has dimension $N = n_x \times n_y \times n_z$ (width×height×depth). The outputs of the encoder, $L = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$ and $\boldsymbol{m} \in \mathbb{R}^N$, parametrize a variational posterior distribution, $q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})$, chosen as independent Gaussian random variables; we then have $q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{m}, L)$. For a reason that will be clarified later, we factorize on the depth dimension such that $q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x}) = \prod_{k=1}^{n_z} \mathcal{N}(\boldsymbol{z}_k; \boldsymbol{m}_k, L_k)$, and then $L_k = \text{diag}((\sigma_k^2)_1, \dots, (\sigma_k^2)_{n_x n_y})$. The model is then composed of a stochastic decoder network whose output $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{l_x l_y})$ parametrizes a product of independent Continuous Bernoulli random variables [23], $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) = \prod_{k=1}^{l_x l_y} \mathcal{CB}(x_k, \lambda_k)$. A realization of this stochastic decoder corresponds to a reconstruction, denoted $\hat{\boldsymbol{x}}$, of the input image $\boldsymbol{x}$ by the model.

2) GRF prior: Let us first be more specific about the structure of the latent space with GRF prior. We consider the zero-mean stationary and toroidal GRF prior on the $n_x \times n_y$[2] dimension and we consider the components of $\boldsymbol{z}$ to remain independent on the depth dimension n_z. Hence $p_{\boldsymbol{\theta}}(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \Sigma) = \prod_{k=1}^{n_z} \mathcal{N}(\boldsymbol{z}_k; \boldsymbol{0}, \Sigma_k)$, where Σ_k are Symmetric Positive Definite (SPD) matrices, and the computations can still be done in a parallel manner on this dimension because the computations fall back to n_z parallel computations involving bi-dimensional GRFs which share parameters. We go one step further by sharing the parameters between these GRFs, thus, $p_{\boldsymbol{\theta}}(\boldsymbol{z}) = \prod_{k=1}^{n_z} \mathcal{N}(\boldsymbol{z}_k; \boldsymbol{0}, \Sigma)$. Note that such a parameter sharing is also proposed in the one dimensional case of the

[1] For example a one-dimensional circular convolution of a M-length signal has computational complexity $\mathcal{O}(M^2)$. This reduces to $\mathcal{O}(M \log M)$ thanks to the Fast Fourier Transform algorithm [10].

[2] In all the following we consider that $l_x = l_y$ and $n_x = n_y$: square input images and square latent space images.

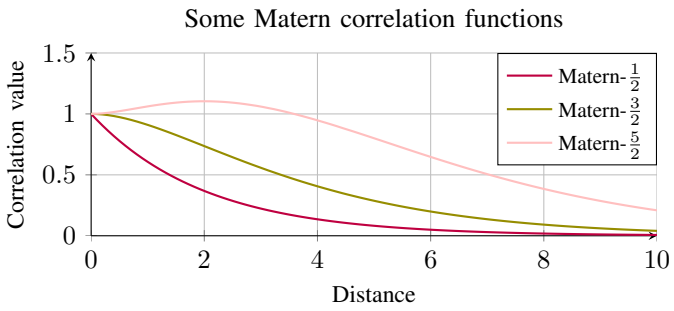

Fig. 1: Illustration of some specific instances of Matern correlation function. We have $r = 2$ for all curves.

factorized Gaussian Process VAEs of [17]. Note also that, since the diagonal covariance matrix used in the standardized Gaussian prior in the classical VAE model is comprised in the set of the SPD matrices yielding a GRF, the VAE-GRF model is a strict generalization of the classical VAE model.

We emphasize the fact that, even when it is not mentioned, we never fully compute the full covariance matrix Σ but only its base matrix. Then, note that the estimation of the covariance matrix Σ needs to yield a SPD matrix. In our work, to ensure this requirement, the covariance matrix Σ is assumed to be generated by Matern correlation functions [21] which form a class of correlation functions defined in $\mathbb{R}^2$ by the general equation: $\forall a \in \mathbb{R}^2, \forall b \in \mathbb{R}^2$,

$$\rho_\nu(a,b;r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{\|a-b\|_t}{r} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{\|a-b\|_t}{r} \right), \tag{2}$$

where Γ is the gamma function, K_ν is the modified Bessel function of the second kind, ν is a non-negative parameter, r is the non-negative range parameter and $\|.\|_t$ is the Euclidean distance on an image (of dimensions $l_x \times l_y$) with torus assumption. For $a = (a_1, a_2) \in \mathbb{R}^2, b = (b_1, b_2) \in \mathbb{R}^2$, the Euclidean distance on the torus is defined by

$$\begin{aligned} \|a-b\|_t = \Big(& \min(|a_1 - b_1|, l_x - |a_1 - b_1|)^2 \\ & + \min(|a_2 - b_2|, l_y - |a_2 - b_2|)^2 \Big)^{\frac{1}{2}}. \end{aligned} \tag{3}$$

In the rest of this article, we will work with two specific Matern correlation functions. First, the Matern-$\frac{1}{2}$ (Eq. (2) with $\nu = \frac{1}{2}$) correlation function on the torus which reads

$$\rho^{1/2}(a,b;r) = \exp\left(-\frac{\|a-b\|_t}{r} \right). \tag{4}$$

Second, the Matern-$\frac{3}{2}$ (Eq. (2) with $\nu = \frac{3}{2}$) correlation function on the torus which reads

$$\rho^{3/2}(a,b;r) = \left(\frac{\|a-b\|_t}{r} + 1 \right) \exp\left(-\frac{\|a-b\|_t}{r} \right). \tag{5}$$

These two instances of the Matern correlation function are commonly used in spatial statistics [12] [21]. They differ in the way they induce a decrease with the distance in the correlation between the random variables. Fig. 1 illustrates three different Matern correlation functions.

Note that to form the covariance matrix, it remains to multiply the correlation function with the GRF variance σ^2, leading to the following covariance between two random variables $(z_k)_s$ and $(z_k)_{s'}$, $\forall k \in \{1, \dots, n_z\}$:

$$Cov((z_k)_s, (z_k)_{s'}) = (\Sigma)_{s,s'} = \sigma^2 \rho(s, s'; r). \tag{6}$$

We also note that that the variance parameter σ^2 and the range r are constants for the whole GRF to respect a stationary covariance structure (and for all the n_z GRFs). Since we assume that the mean parameter $\boldsymbol{\mu} = \mathbf{0}$, the set of parameters defining the GRF prior is $\{r, \sigma^2\}$. In our approach, the prior parameters will be learnt as additional parameters for the network by log-likelihood maximization of the observed image $\boldsymbol{x}$ over which we also assume the same GRF structure as prior. This latter GRF has parameter $\bar{\boldsymbol{\theta}} = \{\bar{r}, \bar{\sigma}^2\}$. Using the estimation of $\bar{\boldsymbol{\theta}}$, we get the GRF over $\boldsymbol{z}$ parameter with the assumed relations:

$$\begin{cases} r_k = \lambda \bar{r}, \\ \sigma^2 = \bar{\sigma}^2, \end{cases} \tag{7}$$

where λ is set to the ratio of the latent image size over the input image size: $\lambda = \lfloor \frac{n_x}{l_x} \rfloor = \lfloor \frac{n_y}{l_y} \rfloor$. The details about these computations are given in Sec. II-C2.

Remark: Importantly, in the VAE-GRF, the encoder, the latent space and the decoder of the model have exactly the same number of parameters as the classical VAE and, overall, the VAE-GRF does not make use of any additional module. We only need to store two more scalar parameters in the VAE-GRF model: the range and the variance of the prior. Indeed, we propose a refinement in the prior modeling, yielding improved results, thus with a very limited additional computational cost.

C. Training the model

Classically, VAEs are trained by maximizing a lower-bound on the log-likelihood, called the Evidential Lower Bound (ELBO), denoted $\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x})$, which reads

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq \underbrace{\mathbb{E}_{q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{KL}(q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}))}_{\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x})}, \tag{8}$$

where KL denotes the Kullback-Leibler divergence between two probability distributions.

For more flexibility in the training, we use the β-ELBO [16] (with an additional stop gradient operator) which reads

$$\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi},\beta}(\boldsymbol{x}) = \mathbb{E}_{q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] - \beta \mathrm{KL}(q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})||\mathrm{sg}[p_{\boldsymbol{\theta}}(\boldsymbol{z})]), \tag{9}$$

where sg refers to the stop gradient operator. Then, the VAE-GRF is trained by maximizing the β-ELBO (with stop gradient) plus the log-likelihood prior over the observed image. The final loss is then defined by:

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x}) = \mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi},\beta}(\boldsymbol{x}) + \log p_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}). \tag{10}$$

The reason for this stop gradient operator is the following: we want the prior parameters σ and r to be updated only from the log-likelihood maximization term. Updating the prior parameters through the KL term would drag Σ towards L, *i.e.*, it would drag the prior towards a Gaussian with diagonal covariance matrix, which is the opposite of our purpose. The

reason for introducing the β scalar is that we can modulate how strong we want the posterior to fit to the prior, in other words, how strong the GRF assumption is.

1) The ELBO term: The ELBO is composed of a first term similar to a cross-entropy which favors reconstructions similar to the input, it is called the *reconstruction term* and its computation is the same as classically done in VAEs. The Kullback-Leibler divergence term[3] can be interpreted as a regularization term which pushes the posterior to match the prior during the training. We have by definition

$$\mathrm{KL}(q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z})) = \mathbb{E}_{q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})}{p_{\boldsymbol{\theta}}(\boldsymbol{z})}\right], \tag{11}$$

and we can show that

$$\begin{aligned}\mathrm{KL}(q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z})) &= \mathrm{KL}(\mathcal{N}(\boldsymbol{z};\boldsymbol{m},L)||\mathcal{N}(\boldsymbol{z};\boldsymbol{0},\Sigma)),\\ &= \frac{1}{2}\sum_{k=1}^{n_k}\Big(-\log|L_k| - n_x n_y + \log|\Sigma|\\ &\quad + \mathrm{tr}(\Sigma^{-1}\boldsymbol{m}_k\boldsymbol{m}_k^T) + \mathrm{tr}(\Sigma^{-1}L_k)\Big),\end{aligned} \tag{12}$$

where tr refers to the trace operator. In Sec. II-D, we explain how the computations of Eq. (12) can remain cheap in our VAE-GRF model, thanks to the stationary GRF prior on torus.

Remark: In the common case of a VAE with a standardized Gaussian as prior [19], Eq. (12) falls back to

$$\begin{aligned}\mathrm{KL}(q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z})) &= \mathrm{KL}(\mathcal{N}(\boldsymbol{z};\boldsymbol{m},L)||\mathcal{N}(\boldsymbol{z};\boldsymbol{0},I)),\\ &= \frac{1}{2}\sum_{i=1}^{N}\big(-\log\sigma_i^2 - 1 + m_i^2 + \sigma_i^2\big).\end{aligned} \tag{13}$$

2) The log-likelihood term: The second term of the loss function is the log-likelihood of the observed image upon which we assume a stationary GRF prior on the torus. With $\bar{\boldsymbol{\theta}}$ defined as before, we have

$$\log p_{\bar{\boldsymbol{\theta}}}(\boldsymbol{x}) = -\frac{l_x l_y}{2}\log 2\pi - \frac{1}{2}\log|\bar{\Sigma}| - \frac{1}{2}\boldsymbol{x}^T\bar{\Sigma}^{-1}\boldsymbol{x}, \tag{14}$$

where $\bar{\Sigma}$ is the covariance matrix generated from the selected correlation function and the parameter $\bar{\boldsymbol{\theta}}$. Again, the spectral properties are usable, and the details for the computation of Eq. (14) are given in Sec. II-D.

D. Computational efficiency

Recall that the stationarity of the GRF and the torus assumption make the properties detailed in Sec. II-A2 usable. The latter are a critical element in our approach. We now detail how the loss function (Eq. (10)) can be efficiently computed both in terms of time and memory complexity. We review each of the terms that are found in Eqs. (12) and (14). A naive computation of these terms would be impossible because they would lead to excessive memory allocations for standard GPUs. Thus, in the VAE-GRF:

- All determinants can be computed directly with Eq. (19).
- All matrix inversions can be computed with Eq. (17).
- The term $\mathrm{tr}(\Sigma^{-1}\boldsymbol{m}\boldsymbol{m}^T)$ can be computed by first computing $\Sigma^{-1}\boldsymbol{m}$. We then multiply this result element-by-element with $\boldsymbol{m}^T$ and sum the resulting vector.
- The term $\mathrm{tr}(\Sigma^{-1}L)$ is equivalent to multiplying each element of the diagonal matrix $L = \mathrm{diag}(\sigma_1^2,\dots,\sigma_n^2)$ by the element at position $(0,0)$ of matrix Σ^{-1} and then summing the resulting matrix.

Note that all the operations involve the base matrices of the full covariance matrices and thus, no matrix with greater dimension than $n_x \times n_y$ is stored or allocated. Moreover, an increase in time efficiency is also due to the use of Fourier transforms in the spectral properties. This way, all these operations can be implemented on GPU, differentiated and computed in a batched manner thanks to the Pytorch [27] library.

Example: Let us consider, for example, the naive computation of $\Sigma^{-1}\boldsymbol{m}_k$ (with a non stationary Σ^{-1}). In such case Σ^{-1} has dimension 1024×1024 and we would need to allocate batches of size $1024 \times 1024 \times 256$. Consider a batch size of 16 and 32-bit float precision, the latent space alone would require $16 \times 1024 \times 1024 \times 256 \times 32$bits $\approx$ 16GBytes without our specific hypothesis. This represents allocations for the computation of the KL term only and thus illustrates that the procedure we develop is crucial for the VAE-GRF to be implemented on standard GPUs.

III. Unsupervised anomaly detection

A. Principle

Anomaly Detection (AD) refers to the task of detecting observations that deviate from some underlying concept of normality [14]. It is an popular research topic with a vast literature [29] but recently, it has been revolutionized by deep learning approaches which have yielded new state-of-the-art results thanks to the unprecedented possibilities of capturing and modeling the normality [31]. The most popular family of approaches to AD is composed of the reconstruction-based methods, in which distances (ℓ-2 distance, SSIM distance [34], etc.) are computed between the inputs and the reconstructions in order to locate the anomalies. The summary works of [2] and [35] illustrate these approaches upon which we base our work.

The principle of unsupervised AD is the following. A representation of the normality is learnt thanks to a deep model which is trained on normal samples devoid of anomalies. Then, at testing time, metrics can be used to detect a change in behaviour of the model when the latter is presented an anomalous sample. VAEs and VQ-VAEs are popular models for this unsupervised task. The former models have appeared first [2] [35] [20] [8] while the latter models have been explored very recently [33] [11].

Remark: Strictly speaking, in the setting of this article, the approach should be called *weakly supervised* since some supervision is needed to make sure the training is done

[3] In this section, we ignore the stop gradient operator whose role has been clarified in the previous section.

only on images without anomalies. However, in this context, *unsupervised* and *weakly supervised* are found interchangeably in the literature.

B. VAE-GRF for Anomaly Detection

We now precisely describe VAE-GRF model for AD. The model is classically trained on a dataset of normal samples to learn a representation of the normality. Then, we propose to detect the anomalies at testing time from an original metric defined on the latent space coupled with a reconstruction based metric.

We show that the VAE-GRF is a robust approach, competitive with comparable approaches from the state-of-the-art provided that some assumptions hold on the datasets. Indeed, we will show that the new metric defined on the latent space requires, as expected, that the stationary and torus assumptions made for the GRF hold. This is the case for some textures images as we illustrate in Sec. III-E, and also for some aerial images in remote sensing as we present in Sec. IV.

Let us now introduce our new metric for AD that will be used in all the following experiments. Let MAD (from Mean Absolute Deviation) be the $n_x \times n_y$ anomaly map that we compute from the latent space. Each of the pixel of the anomaly map is computed as the mean absolute deviation from the mean of the same location of the output of the convolutional encoder $\boldsymbol{z}$, *i.e.*, $\forall x \in \{1, \dots, n_x\}, \forall y \in \{1, \dots, n_y\}$,

$$MAD_{x,y} = \frac{1}{n_z} \sum_{k=1}^{n_z} \left| z_{x,y,k} - \frac{1}{n_z} \sum_{k=1}^{n_z} z_{x,y,k} \right|. \qquad (15)$$

Then, the MAD anomaly map is upsampled to the dimension of the original image and AD can be performed.

Let us also define a reconstruction-based anomaly map, called SM, which uses the Structural Similarity Index Measure (SSIM) [34]. For each pixel i, we have:

$$SM(x_i) = \text{SSIM}(\boldsymbol{p}_i, \boldsymbol{q}_i) = \frac{(2\mu_{\boldsymbol{p}}\mu_{\boldsymbol{q}} + c_1)(2\sigma_{\boldsymbol{pq}} + c_2)}{(\mu_{\boldsymbol{p}}^2 + \mu_{\boldsymbol{q}}^2 + c_1)(\sigma_{\boldsymbol{p}}^2 + \sigma_{\boldsymbol{q}}^2 + c_2)}, \qquad (16)$$

where $\boldsymbol{p}_i$ (resp. $\boldsymbol{q}_i$) is a patch around pixel i of $\boldsymbol{x}$ (resp. $\hat{\boldsymbol{x}}$). μ_p, σ_p and σ_{pq} represent, respectively, the mean, the standard deviation and the covariance of the patches. The scalars are set to $c_1 = 0.01$ and $c_2 = 0.03$ [34].

From the two previous anomaly maps we form a new original anomaly map using an element-wise multiplication, which is denoted $MAD \odot SM$. The two anomaly maps MAD and $MAD \odot SM$ are at the core of our contribution for improved AD, along with the GRF prior. These anomaly maps will then be compared with state-of-the-art models presented in the next section.

For completeness, Fig. 2 graphically illustrates the model and MAD anomaly map. The encoder-decoder structure is, however, identical to that of a traditional VAE. This is remarkable since the improvements we propose are based only on an original refinement of the probabilistic model.

C. Related work

To date, most of the best performing methods are based on feature extraction [7] [22]. The latter always perform slightly better than VAE-based approaches. However, generative models offer other possibilities to work with aspects related to the probabilistic framework, sample generation or latent space interpretability [33] [17].

Such aspects are also crucial to our study, thus we will compare the VAE-GRF with generative models from the literature: comparable generative models are be based on a baseline VAE and also rely on a modelization refinement. Moreover, we favor related works with no additional modules attached (*e.g.* discriminator modules [32]). In such setting, we are able to fully evaluate the gain offered by the GRF prior and the new AD metrics.

We now list the most notable comparable approaches:

- We compare our MAD and $MAD \odot SM$ metrics to the already existing ℓ-2 and SM metrics introduced in [4].
- An interesting idea quite similar to us is the Visually Explained Variational Autoencoder (VEVAE) [20]. However the code is not available and we were not able to replicate their study. Hence we compare the VAE-GRF to the VEVAE on the MVTec experiment only in Sec. III-E.
- AD metrics similar to ours have also been introduced in [35]. The magnitude of the gradient of the loss $|\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x})|$ is used as a metric to localize anomalies and it can be multiplied by another anomaly map. Hence we also straightforwardly test the metric $|\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x})| \odot SM$.
- We also test the iterative procedure proposed by [8]. It consists in a refinement of the VAE reconstruction. We limit this approach to 15 iterations of projection because it is very slow and empirically performed best. The refinement is performed before computing the SM anomaly map: we refer to this approach as SM grad.

Thus, except for VEVAE, all the comparable approaches are from our reimplementations because the code were unavailable. All these reimplementations have the same classic VAE baseline described in Sec. III-D (except for the VEVAE).

D. Network architecture

All the inputs are resized to the size 256×256 ($l_x = l_y = 256$). Then:

- The encoder consists in the first three layers of the ResNet18 neural network [15], followed by a 1×1 convolutional layer setting the depth dimension to $2{*}256 = 512$ (kernel size 1, stride 1 and padding 0).
- The latent space image has width and height 32 ($n_x = n_y = 32$) and depth $n_z = 256$.
- The decoder is first composed of a deconvolutional layer with input dimension 256 and output 128 (kernel size 1, stride 1, and padding 0). It is then stacked with three deconvolutional layers (kernel size 4, stride 2 and padding 1), each followed by a ReLU activation and Batch Normalization.

This architecture originates from ideas found in [20] or [11].

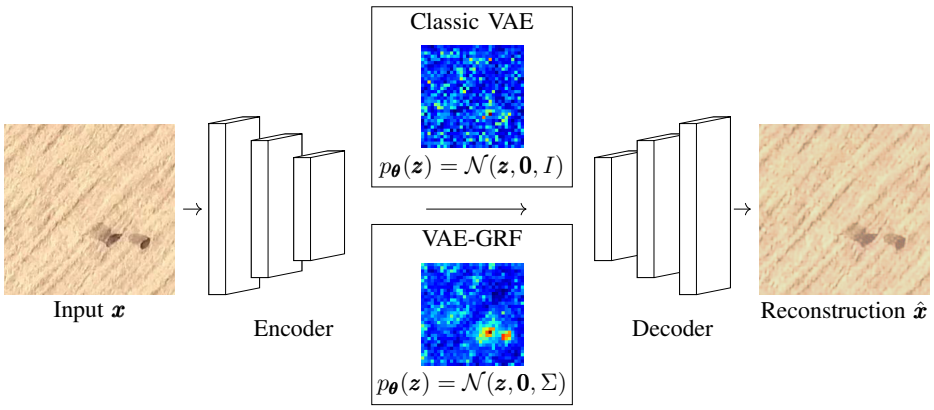

Fig. 2: The classic VAE and VAE-GRF architectures. We illustrate the pixel-wise MAD metric (Eq. 15) computed from the latent space in both cases. The interpretability and improved results of the metric can be seen in the case of the VAE-GRF as opposed to VAE with classical standardized Gaussian prior.

E. Effects of the GRF prior: experiments on MVTec textures

1) Presentation: In this section we experimentally investigate the role of the prior in the VAE-GRF model. This experiment is based on the textures images from the MVTec dataset [4] which is the standard dataset for AD. The dataset provides normal and abnormal (defective) RGB images of industrial goods from 15 different categories.

2) Results: Tab. I and II summarize the results in terms of pixel-wise area under the receiver operating characteristic curve (ROCAUC), computed with the final anomaly map and the ground truth. This is the classical score used for AD. Compared to the traditional VAE model (Tab. I), on the images which respect the most the stationary GRF assumption, *i.e.* texture images, the VAE-GRF performs better, especially when it comes to the new metrics we introduced (up to 8 points of improvement on the tile texture). On images where the hypothesis of a stationary GRF clearly does not hold, the MAD metric becomes useless and both the VAE and VAE-GRF perform similarly with the $MAD \odot SM$ metric. This result could be expected: indeed the VAE-GRF model is a strict generalization of the classical VAE model as stated before. We then can see the interest of the refinement in the VAE modeling and its limitations when the stationary GRF prior assumption is violated. That is why we only focus our analysis on MVTec texture images. Similar conclusions can be made when we compare to the other approaches in the literature (Tab. II). The VAE-GRF is always on par or better than the state-of-the-art results. The best gains coming from the texture images.

Fig. 3 illustrates the evolution of β-ELBO, $\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi},\beta}$, in function of the epochs. We conclude that despite the fact that the VAE-GRF results are always better or equivalent, the β-ELBO values for the VAE-GRF and VAE seem unpredictable. We suppose that this result can be linked with the known fact that a theoretically worse lower bound can lead to better final results when the modelization is more relevant (see, *e.g.*, [18], Sec. 5.2).

Note also that, globally, both our VAE and GRF-VAE are competitive against state-of-the-art approaches which also suggests that the proposed architecture and training procedures are highly relevant and optimized (see Sec. III-D). Finally, Fig. 4 illustrates the experiment on some images of the MVTec database.

Remark: The choice of the correlation function as well as the choice of the hyperparameter β is empirically made in our study. Further research on the correlation function type and its generalization is out of scope of this paper.

IV. Unsupervised animal detection in aerial images with VAEs-GRF

In this section we present how the VAE-GRF model can be used to solve the real world problem of detection in aerial images, where the landscape can be assumed to be well modeled by a stationary texture. We focus on the task of unsupervised animal detection on two different datasets. The context of AD described previously is used here. Indeed, many images are empty in our datasets and we can consider animals as anomalies.

In the following experiments, we show that the VAE-GRF model and the associated metrics are relevant and provide competitive results against state-of-the-art approaches.

A. Livestock dataset

1) Presentation: The open Livestock dataset [13] regroups aerial images of livestock over grassland in which animals have been annotated. We again wish to perform an unsupervised detection of the animals using an AD approach. It is possible since we have a total of 3430 empty images to perform training and 890 images containing at least one animal.

2) Experiments & Results: The network described in Sec. III-E is used here, as well as the pixel-wise AD metrics from Sec. III-B.

Table III gives the score in terms of pixel-wise ROCAUC, for all the models (we have discarded some of the worst performing approaches, ranking according to the MVTec results). We can see that the VAE-GRF performs best with both the MAD and the $MAD \odot SM$ metrics. Again, the

	Category	VAE				VAE-GRF			
		ℓ-2	SM	MAD	$MAD \odot SM$	ℓ-2	SM	MAD	$MAD \odot SM$
Stationary textures	Leather	0.78	0.92	0.80	0.91	0.77	0.93	0.95	**0.98**
	Tile	0.64	0.78	0.55	0.77	0.66	0.81	0.70	**0.85**
	Wood	0.70	0.75	0.67	0.77	0.71	0.80	0.74	**0.81**
Non stationary textures	Carpet	0.63	0.90	0.59	0.91	0.63	**0.92**	0.58	0.91
	Grid	0.72	**0.93**	0.51	**0.93**	0.71	0.92	0.54	**0.93**
	Hazelnut	0.88	**0.98**	0.64	0.97	0.95	**0.98**	0.62	**0.98**

TABLE I: ROCAUC scores for pixel-wise AD on the texture images from the MVTec dataset. The VAE-GRF approach does not exhibit an advantage over the VAE approach for *Carpet*, *Grid* and *Hazelnut* possibly because the stationary GRF assumption is clearly wrong for these images. For all the experiments with the VAE-GRF model, we have $\beta = 1$ except for *Wood* where we set $\beta = 0.1$. Similarly, we used the Matern-$\frac{3}{2}$ correlation everywhere except for *Tile* where we used Matern-$\frac{1}{2}$. Best scores appear in bold and second best are underlined. Scores from the main contribution of this article are in purple.

	Category	VAE-GRF	Liu et al. [20]	Zimmerer et al. [35]		Dehaene et al. [8]
		$MAD \odot SM$	VEVAE	$\nabla_{\boldsymbol{x}} \mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x})$	$\nabla_{\boldsymbol{x}} \mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x}) \odot SM$	SM grad
Stationary textures	Leather	**0.98**	0.95	0.55	0.94	0.95
	Tile	**0.85**	0.80	0.62	0.78	0.78
	Wood	**0.81**	0.77	0.54	0.76	0.74
Non stationary textures	Carpet	**0.91**	0.78	0.61	0.89	0.89
	Grid	**0.93**	0.73	0.54	0.92	0.90
	Hazelnut	**0.98**	**0.98**	0.80	**0.98**	0.92

TABLE II: ROCAUC scores for pixel-wise AD on the texture images from the MVTec dataset. Comparisons with other methods. Best scores appear in bold and second best are underlined. Scores from the main contribution of this article are in purple.

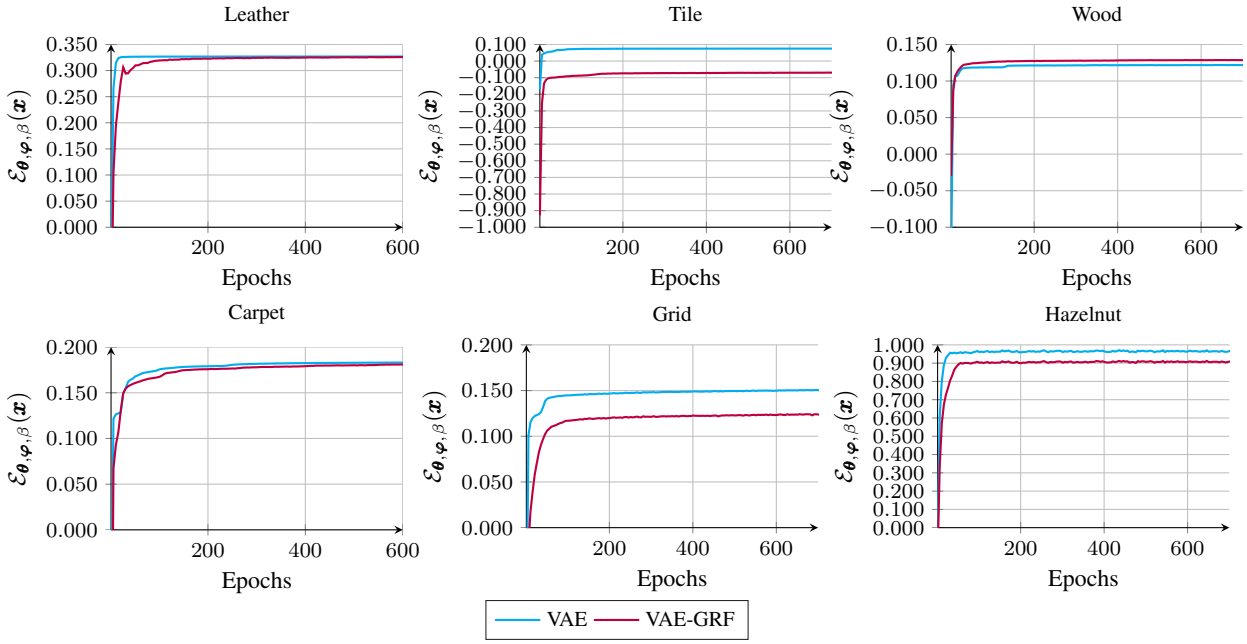

Fig. 3: Evolution of the β-ELBO ($\mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi},\beta}(\boldsymbol{x})$) in function of epochs for the three textures. As noted in Table I, $\beta = 1$ everywhere except for the VAE-GRF of the *Wood* experiment. While we observe convergence of the training procedure, it is not possible to extrapolate the relative performance of the model according to the relative ELBO values reached for each model.

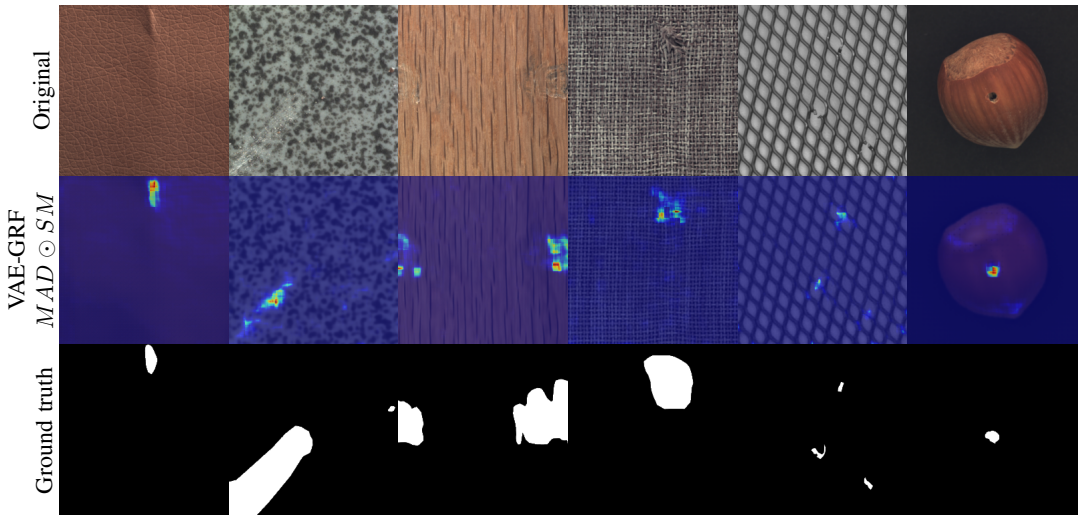

Fig. 4: Selected illustrations for the MVTec experiment. Anomaly maps and reconstructions from our proposed approach. *Top row:* original image. *Middle row:* the anomaly maps overlaying the VAE-GRF reconstruction. *Bottom row:* The segmented anomalies. From left to right: the *Leather*, *Tile*, *Wood*, *Carpet*, *Grid* and *Hazelnut* categories.

improvements brought by the GRF prior could be expected since the images seem to really fit the stationary GRF hypothesis. Fig. 5 provides a graphical illustration of the experiment over some images of the dataset.

Remark: The MAD metric is all the more interesting as it is a metric which is only computed from the latent space, *i.e.*, at testing time, we do not need to perform any computations with the decoder. This then represents an interesting computational gain (one can calculate that this saves more than 40.000 two dimensional convolutions when using the decoder described in Sec. III-D). Moreover, being independent of the reconstruction can be very interesting on complex dataset where reconstructions are unreliable as it will be shown in the next experiment.

B. Semmacape dataset

1) Presentation: The Semmacape dataset[4] comprises 165 aerial images collected in the Gironde estuary and Pertuis sea Marine Nature Park, France, in 2020. Birds and dolphins have been manually annotated, the images were then subdivided into patches giving rise to 345 images with an animal and $138,544$ empty images. Such a large number of empty images enables us to learn the normality (surface of the sea) and then to detect animals as anomalies. We will perform pixel-wise AD assessment with the previously introduced models.

2) Experiments & Results: The network described in Sec. III-E is used here; we are in the same experimental setting as before.

Table IV gives the scores for the models in terms of pixel-wise ROCAUC. First of all, probably because of the complexity of the dataset (many images are corrupted by the sun glare), the SSIM anomaly map seems particularly unreliable. We argue that this is the reason for the decrease in performance of the $MAD \odot SM$ metric. However, among all comparable and simpler models, the VAE-GRF with MAD metric seems to be the best performing. We also provide as a comparison to the best results achieved so far on this dataset in [3] with a much more complex model combining the PaDiM approach [7] and Normalizing Flows (NF) [26]. We can conclude that our VAE-GRF is competitive with the PaDIM + NF (which is not a generative approach), in particular because the stationary GRF assumption is relevant on such images of the sea surface. We also notice that, when brought to this complex and real dataset, the performances of

[4] https://semmacape.irisa.fr/

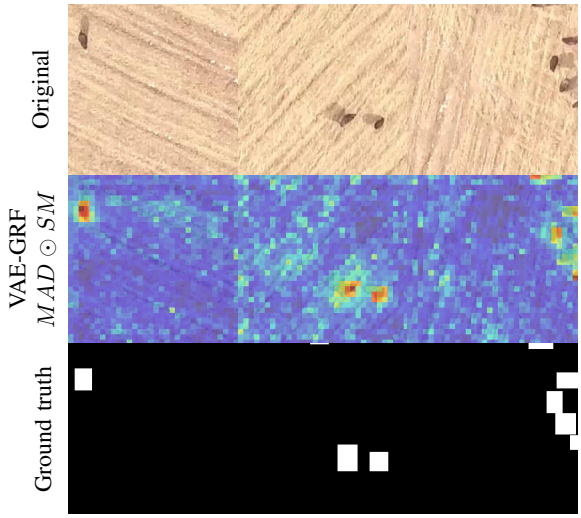

Fig. 5: Selected illustrations of the detection of our model in the Livestock dataset experiment using the VAE-GRF model with Matern-$\frac{3}{2}$ correlation function. *Top row:* original image with ground truth. *Middle row:* VAE-GRF reconstruction with the anomaly map overlaid. *Bottom row:* ground truth segmentation.

	VAE	VAE-GRF			Zimmerer et al. [35]		Dehaene et al. [8]
	ℓ-2	SM	MAD	$MAD \odot SM$	$\lvert\nabla_{\boldsymbol{x}} \mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x})\rvert$	$\lvert\nabla_{\boldsymbol{x}} \mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x})\rvert \odot SM$	SM grad
ROCAUC	0.67	0.76	**0.84**	<u>0.83</u>	0.77	0.52	0.68

TABLE III: ROCAUC scores for pixel-wise AD on the texture images from the Livestock dataset. For all the experiments with the VAE-GRF model, we have $\beta = 0.1$ and we used the Matern-$\frac{3}{2}$ correlation function. Best scores appear in bold and second best are underlined. Scores from the main contribution of this article are in purple.

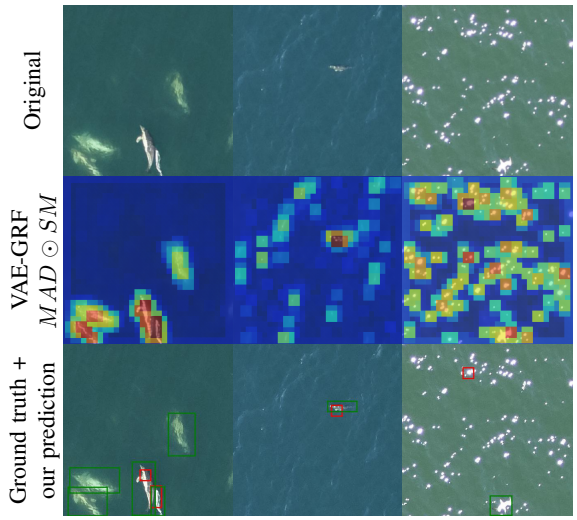

Fig. 6: Selected illustrations of the detection of our model in the Semmacape dataset experiment using the VAE-GRF model with Matern-$\frac{3}{2}$ correlation function. *Top row:* original image. *Bottom row:* VAE-GRF reconstruction with the anomaly map overlaid. *Bottom row:* with ground truth bounding box (green) and prediction (red). The right column describes a typical very complex images, with a lot of sun glares which are moreover confused with the white birds.

the approaches from [35] and [8] collapse. Fig. 6 graphically illustrates the experiment.

V. Conclusion

We have introduced a stationary GRF prior in the VAE model and shown how such a more complex prior (with full covariance matrix) can be embedded in a VAE model while preserving efficient computations. We have also demonstrated, in the context of AD, that such refinement in the modeling is relevant for specific images, notably, for texture images. To do so, we have introduced two new metrics yielding competitive results against comparable state-of-the-art VAE models for AD tasks on several datasets. Our results suggest that the VAE-GRF might replace the VAE baseline for many tasks, as we show how the stationary assumption does not introduce any additional computational cost. Indeed, VAE-GRF might offers an efficient and relevant prior for many practical applications, especially for the processing of images that exhibit textures.

Since the stationary assumption for the GRF prior remains a strong assumption, future work might consider relaxing this assumption and we might study ways to introduce non-stationary GRF prior while preserving the tractability of the model.

References

[1] J. G. A. Barbedo, L. V. Koenigkan, T. T. Santos, and P. M. Santos. A study on the detection of cattle in UAV images using deep learning. *Sensors*, 19(24):5436, 2019.

[2] C. Baur, S. Denner, B. Wiestler, N. Navab, and S. Albarqouni. Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. *Medical Image Analysis*, page 101952, 2021.

[3] P. Berg, D. Santana Maia, M.-T. Pham, and S. Lefèvre. Weakly supervised detection of marine animals in high resolution aerial images. *Remote Sensing*, 2021.

[4] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9592–9600, 2019.

[5] L. B. Boudaoud, F. Maussang, R. Garello, and A. Chevallier. Marine bird detection based on deep learning using high-resolution aerial images. In *OCEANS 2019-Marseille*, pages 1–7, 2019.

[6] F. P. Casale, A. V. Dalca, L. Saglietti, J. Listgarten, and N. Fusi. Gaussian process prior variational autoencoders. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10390–10401, 2018.

[7] T. Defard, A. Setkov, A. Loesch, and R. Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489, 2021.

[8] D. Dehaene, O. Frigo, S. Combrexelle, and P. Eline. Iterative energy-based projection on a normal data manifold for anomaly localization. In *International Conference on Learning Representations*, 2020.

[9] V. Fortuin. Priors in Bayesian deep learning: A review. *International Statistical Review*, 2022.

[10] C. Fox and R. A. Norton. Fast sampling in a linear-gaussian inverse problem. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1191–1218, 2016.

[11] H. Gangloff, M.-T. Pham, L. Courtrai, and S. Lefèvre. Leveraging vector-quantized variational autoencoder inner metrics for anomaly detection. 2022. URL `https://hal.archives-ouvertes.fr/hal-03541964`.

[12] P. Guttorp and T. Gneiting. Studies in the history of probability and statistics XLIX on the Matérn correlation family. *Biometrika*, 93(4):989–995, 2006.

[13] L. Han, P. Tao, and R. R. Martin. Livestock detection in aerial images using a fully convolutional network. *Computational Visual Media*, 5(2):221–228, 2019.

	VAE	VAE-GRF			Zimmerer et al. [35]		Dehaene et al. [8]	Berg et al. [3]
	ℓ-2	SM	MAD	$MAD \odot SM$	$\lvert\nabla_{\boldsymbol{x}} \mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x})\rvert$	$\lvert\nabla_{\boldsymbol{x}} \mathcal{E}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x})\rvert \odot SM$	SM grad	PaDiM + NF
F1	0.296	0.247	0.464	0.317	0.281	0.138	0.099	**0.530**
Recall	0.298	0.239	0.473	0.288	0.349	0.134	0.130	**0.757**
Precision	0.295	0.255	**0.455**	0.352	0.236	0.143	0.080	0.408

TABLE IV: ROCAUC scores on image-wise AD on the Semmacape dataset. Scores for OrthoAD and PadiM + NF are from [3]. For all the experiments with the VAE-GRF model, we have $\beta = 1$ and we used the Matern-$\frac{3}{2}$ correlation function. Best scores appear in bold and second best are underlined. Scores from the main contribution of this article are in purple.

[14] D. M. Hawkins. *Identification of Outliers*. Monographs on applied probability and statistics. Chapman and Hall, 1980.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[17] M. Jazbec, M. Pearce, and V. Fortuin. Factorized Gaussian process variational autoencoders. *arXiv preprint arXiv:2011.07255*, 2020.

[18] D. Kingma. *Variational Inference and Deep Learning: A New Synthesis*. Ridderprint, 2017.

[19] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

[20] W. Liu, R. Li, M. Zheng, S. Karanam, Z. Wu, B. Bhanu, R. J. Radke, and O. Camps. Towards visually explaining variational autoencoders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8642–8651, 2020.

[21] Y. Liu, J. Li, S. Sun, and B. Yu. Advances in Gaussian random field generation: a review. *Computational Geosciences*, 23(5):1011–1047, 2019.

[22] P. Liznerski, L. Ruff, R. A. Vandermeulen, B. J. Franks, M. Kloft, and K.-R. Müller. Explainable deep one-class classification. In *International Conference on Learning Representations*, 2021.

[23] G. Loaiza-Ganem and J. P. Cunningham. The continuous Bernoulli: fixing a pervasive error in variational autoencoders. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 13266–13276, 2019.

[24] C. Padubidri, A. Kamilaris, S. Karatsiolis, and J. Kamminga. Counting sea lions and elephants from aerial photography using deep learning with density maps. *Animal Biotelemetry*, 9(1):1–10, 2021.

[25] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2021.

[26] G. Papamakarios, I. Murray, and T. Pavlakou. Masked autoregressive flow for density estimation. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2338–2347, 2017.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035, 2019.

[28] M. Pearce. The Gaussian process prior VAE for interpretable latent dynamics from pixels. In *Symposium on Advances in Approximate Bayesian Inference*, pages 1–12, 2020.

[29] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

[30] H. Rue and L. Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

[31] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K. Müller. A unifying review of deep and shallow anomaly detection. *Proc. IEEE*, 109(5):756–795, 2021.

[32] S. Venkataramanan, K.-C. Peng, R. V. Singh, and A. Mahalanobis. Attention guided anomaly localization in images. In *European Conference on Computer Vision*, pages 485–503. Springer, 2020.

[33] L. Wang, D. Zhang, J. Guo, and Y. Han. Image anomaly detection using normal data only by latent space resampling. *Applied Sciences*, 10(23):8660, 2020.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[35] D. Zimmerer, F. Isensee, J. Petersen, S. Kohl, and K. Maier-Hein. Unsupervised anomaly localization using variational auto-encoders. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 289–297, 2019.

APPENDIX

A. Spectral computations

We now detail the formulas mentioned in Sec. II-A2. More details on the elements of this section can be found in [30].

Very useful properties, linked with the Fourier transform, are available for circulant and block-circulant matrices. We only recall the formulas for block-circulant matrices that are used in this study. In the following properties, DFT2 (resp. IDFT2) is the 2 dimensional (resp. inverse) Fourier transform, $\odot$ is the element-wise matrix multiplication and $\bullet$ is the element-wise exponentiation of the elements of the matrix. Note that in the next equations, we consider orthonormal Fourier transforms (normalized by $\frac{1}{\sqrt{l_x l_y}}$, where $l_x \times l_y$ is the matrix dimensionality).

1) Inverse of a block-circulant matrix: Let C be a block-circulant matrix, then we have

$$\mathrm{base}(C^{-1}) = \frac{1}{l_x l_y}\mathrm{IDFT2}(\mathrm{DFT2}(\mathrm{base}(C)) \bullet (-1)). \quad (17)$$

2) Product of block-circulant matrices: Let C and D be two block-circulant matrices, then we have

$$\mathrm{base}(CD)=\sqrt{l_x l_y}\mathrm{IDFT2}(\mathrm{DFT2}(\mathrm{base}(C))\odot\mathrm{DFT2}(\mathrm{base}(D))). \quad (18)$$

3) Eigenvalues of a block-circulant matrix: Let C be a block-circulant matrix, then the matrix filled with the eigenvalues of C is

$$\Lambda = \sqrt{l_x l_y}\mathrm{DFT2}(\mathrm{base}(C)). \quad (19)$$

4) Product of block-circulant matrix with vector : Let $\boldsymbol{v}$ be the column major vector obtained from $n \times N$ matrix V, then $\boldsymbol{u} = C\boldsymbol{v}$ is obtained from

$$U = \sqrt{l_x l_y}\mathrm{DFT2}(\mathrm{DFT2}(\mathrm{base}(C)) \odot \mathrm{IDFT2}(U)). \quad (20)$$